WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 4

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 4

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

OPTICAL SOURCES

JENS BUUS
Gayton Photonics
Gayton, Northants
United Kingdom

1. INTRODUCTION

Since the early 1980s an increasing fraction of the worlds telecommunication traffic has been carried on optical fiber, and since the late 1980s virtually all new trunk lines have been based on fiberoptics. Optical fibers are now the dominating medium for a variety of communication systems, ranging from short-distance data links to transoceanic telecommunication systems.

The dramatic increase in traffic brought about by the Internet has made optical fibers even more attractive because of the several terahertz of transmission bandwidth they offer. The wide bandwidth can be exploited by multiplexing a number of optical sources operating on different optical frequencies. This technology is known as *wavelength-division multiplexing* (WDM).

Different fiberoptic communication systems place different requirements on the optical sources used in the systems. The basic properties and characteristics of these sources are reviewed in this article.

2. HISTORICAL DEVELOPMENT

Since the late 1960s there has been an interesting interaction between the development of optical fibers and optical sources. Semiconductor lasers were first demonstrated in 1962, but it was only after the demonstration of room-temperature continuous-wave (CW) operation in 1970 that the practical use of these lasers became a reality. We note in passing that this improved laser performance was due to the introduction of the *heterostructure*, Alferov and Kroemer shared half the 2000 Nobel Prize in Physics for their work on this topic.

The use of optical fibers for long-distance communication was proposed in 1966, and the fiber loss was reduced to 20 dB/km in 1970. In the early fibers the minimum loss occurred at relatively short wavelengths, well suited to the emission wavelength of GaAs based lasers, which is in the 800–900-nm range. The fibers at that time were multimoded, and reduced fiber losses meant that transmission distances were limited mainly by modal dispersion (i.e., different fiber modes having different group velocity, thus leading to pulse distortion).

Further improvements resulted in fiber losses below 0.2 dB/km by 1980. As the fiber losses were reduced, the spectral range where the losses were lowest moved toward longer wavelengths. A wavelength region of particular interest was around 1300 nm, where the fiber dispersion was minimized. This new wavelength range was exploited by introducing lasers and LEDs based on InGaAsP compounds using InP substrates, this development started around 1980. These devices can cover the wavelength range from about 1100 nm to about 1700 nm. One of the first examples of a commercial long-distance transmission system at these "long wavelengths" was the London–Birmingham link in the early 1980s, based on LEDs operating around 1300 nm.

From the early 1980s single-mode fibers were being introduced, thus eliminating modal dispersion. The lowest loss for these fibers occurs at wavelengths around 1550 nm, which is within the range accessible by InGaAsP/InP lasers. However, the InP-based lasers have a tendency to operate simultaneously in several longitudinal modes. From the laser cavity length (typically about 300 μm), it follows that the longitudinal mode spacing (in frequency) is about 120 GHz, corresponding to a spectral spacing of about 1 nm (in wavelength). At a wavelength of 1550 nm a standard single-mode fiber has a chromatic dispersion of about 17 ps/(km $\cdot$ nm). Consequently, multimode laser operation will give rise to dispersion problems for high-speed systems operated over a long fiber length, and the development of single-frequency lasers (i.e., lasers operating in a single longitudinal mode) then became a priority. Single-frequency operation can be achieved by incorporating a wavelength selective element in the laser, typically a grating as in the DFB (distributed feedback) laser.

The development of the fixed wavelength DFB laser in turn made the efficient use of wavelength-division multiplexing (WDM) possible. In these systems the signals from a number of lasers, operating at different optical frequencies, are multiplexed together and transmitted over a single fiber, thus increasing the transmission capacity of the fiber significantly. As an example, a spectral range of 30 nm (around a wavelength of 1550 nm) corresponds to a frequency range of about 3800 GHz; using lasers spaced in frequency by 50 GHz will allow 76 separate channels, each of which can carry data at a rate of, for example, 10 Gbps (gigabits per second), thus giving an aggregate capacity of 760 Gbps. This capacity can be increased even more by the use of a wider spectral range, and/or a higher spectral efficiency (ratio of data rate to channel spacing).

The development of wavelength selective lasers and tunable lasers also opens new possibilities. Not only can these lasers be used as flexible spares or as "uncommitted" wavelength sources; they also allow the use of wavelength routing, where the path of a signal through a network is entirely determined by the wavelength of the signal. Ultimately the use of lasers that can switch fast between wavelengths opens the possibility for packet switching on the optical level.

Other interesting developments include lasers and LEDs specifically designed for (short-distance) data links, and pump lasers for optical amplifiers.

The reader is referred to Refs. 1 and 2 for more details on the history of the development of semiconductor lasers.

3. LIGHT-EMITTING DIODES

Red *light-emitting diodes* (LEDs) are well known from their use in displays and as indicator lights. LEDs based on GaAs or InP emit in the near infrared and are used for communication purposes. The basis for the operation

of an LED is that carriers (electrons and holes) are injected into a forward-biased *p-n* junction and recombine spontaneously, thereby generating photons. The photon energy (and hence the wavelength of the generated light) is determined mainly by the bandgap of the region where the recombination takes place. However, the carriers have a spread in energy, leading to a spectral width of the emitted light of a few times kT, where k is Bolzmann's constant and T is the operating temperature. Consequently the spectral width for an LED operating in the 1300-nm spectral region will be of the order 100 nm.

Not all the electrical power supplied to the LED is converted to light: (1) some power is lost because of electrical losses, (2) the radiative (spontaneous) recombination competes with various nonradiative recombination processes, and (3) only a finite fraction of the generated light is able to escape from the structure. This last effect is due to the process of total internal reflection; since the refractive index of the (semiconductor) LED structure is high compared to that of the surrounding medium (air), only light propagating at an angle nearly perpendicular to the surface will be able to escape from the structure. Additional optical losses occur when the light is coupled into a fiber. For coupling into a standard multimode fiber with a core diameter of 50 μm the overall efficiency (optical power in the fiber compared to electrical power supplied to the LED) is typically of the order of 1%, corresponding to coupled power levels of the order of 100 μW.

The light output from an LED can be modulated directly by varying the current passed through the device. In the absence of parasitics, the maximum possible modulation speed is approximately given by the inverse of the recombination time. With recombination times in the nanosecond range, it follows that LEDs can typically be used at data rates of up to a few hundred megabits per second (Mbps).

It should be noted that LEDs can be optimized for power levels of up to several milliwatts, and higher coupled power levels can be achieved by using large core fibers. Obviously, the wide spectral width of LEDs leads to chromatic dispersion, and the use of multimode (in particular large-core) fibers leads to modal dispersion. However, since LEDs are used at moderate data rates, they are an attractive simple and low-cost solution for links of a modest length (i.e., up to a few kilometers).

Finally, it should be mentioned that near infrared LEDs are also widely used for very short range free space communication between computers and peripheral equipment.

4. LASERS

4.1. Laser Basics

In order to understand the workings of a laser, we consider a system where the constituents (electrons, atoms, ions, or molecules) have two possible energy states. Transitions between these two states are accompanied by the absorption or the emission of photons, where the photon energy is equal to the difference in energy between the two states. In 1917 Einstein explained the relation between the energy distribution of a gas of molecules and that of the radiation field (Planck's law) by assuming that the following three processes occur:

1. *Spontaneous Emission*. Transition from the higher to the lower state accompanied by the emission of a photon.
2. *Absorption*. Transition from the lower to the higher state brought about by the absorption of a photon.
3. *Stimulated Emission*. In this process the transition from the higher to the lower state is triggered by incoming photons with energy equal to the transition energy; the additional photon emitted in the transition is in phase with the incoming photons. The transition probability is proportional to the number of incoming photons.

In thermal equilibrium the higher-energy state is less densely populated than the lower one, and it follows that an incoming stream of photons will be attenuated. However, if a situation is created where the higher-energy state is more densely populated than would be possible at the lower amplification level, this would be known as *population inversion*. Such a system is shown schematically in Fig. 1.

The basis for the acronym *LASER* (light amplification by stimulated emission of radiation) becomes clear from this description. The normal use of the word *laser*, however, refers to light generation (rather than amplification), and in order to construct an oscillator working at the lasing frequency, it is also necessary to provide feedback. A laser is usually constructed by placing material with an inverted population between a pair of partly reflecting mirrors. Light moving back and forth between the mirrors is amplified as a result of the stimulated emission process.

The combination of amplification and feedback from the mirrors forms an oscillator, and oscillation takes place if the amplification balances the loss caused by light escaping through the mirrors. With a gain factor g, a cavity length

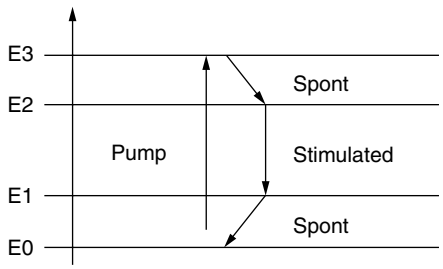


Figure 1. Schematic diagram of a four-level laser system. The upper energy level 3 is short-lived, giving a fast decay to the upper level 2 involved in the lasing transition, this level is long-lived. From the lower laser level 1 there is a fast decay to the ground level 0. "Pumping" from the ground level to the highest level is done by flashlamp, electric discharge, or the use of another laser, and population inversion can be achieved between levels 2 and 1.

L, and two mirrors both with a power reflectivity of R, this condition can be written as

$$\exp(gL)R\exp(gL)R = 1 \tag{1}$$

Since photons created by the stimulated emission process are emitted in phase with the incoming photons, the light emitted from the laser cavity is coherent.

A large number of laser types exist (gas lasers such as HeNe and CO_2, solid-state lasers such as Nd:YAG, etc.), and are being used in numerous fields (material processing, medical applications, etc.). The laser type of interest for optical fiber communication is the semiconductor laser. This laser type is also used in CD and DVD players, as well as in scanners and pointers.

4.2. Semiconductor Lasers

Nearly all semiconductor lasers are based on the double *heterostructure*. In this structure, a material with a relatively narrow bandgap — the *active layer* — (normally undoped) is sandwiched between a pair of n-type and p-type materials with wider bandgaps — the *confinement layers*. When this structure is under forward-biased *quasi-Fermi levels* are formed, and electrons and holes are injected into the active layer from the n-type and p-type materials, respectively. The Fermi level(s) determine the energy distribution of electrons in the conduction band and holes in the valence band. Population inversion, and thereby gain, is achieved when the quasi-Fermi level separation exceeds the bandgap of the active layer (see Fig. 2).

If the bandgap difference is sufficiently large, carriers injected into the active layer cannot escape over the heterobarrier, and carrier recombination can take place only in the active layer. Light generated in the active layer is not absorbed in the confinement layers since semiconductors are transparent to light with a photon energy lower than the bandgap. The photon energy is given by the product of Planck's constant h and the optical

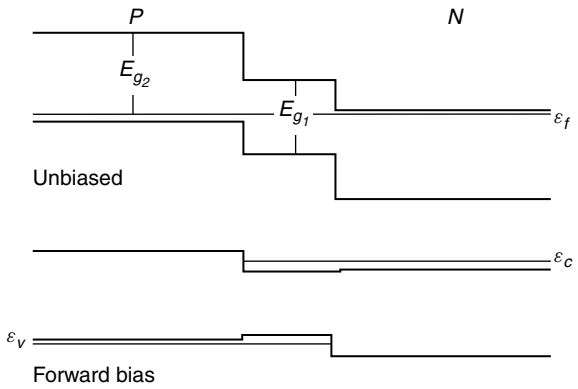


Figure 2. Double heterostructure formed with a material with a narrow bandgap E_{g1} placed between n-type and p-type materials with a wider bandgap E_{g2}. Carriers cannot have energies corresponding to states within the bandgap. Without bias the Fermi level ε_f is continuous. Under forward bias quasi-Fermi levels ε_c and ε_v are formed in the conduction and valence bands, respectively.

frequency ν, which is in turn equal to the speed of light in vacuum, c, divided by the wavelength λ, hence

$$h\nu = \frac{hc}{\lambda} \tag{2}$$

At a given optical frequency, a narrow bandgap semiconductor generally has a higher refractive index than a semiconductor with a wider bandgap. Consequently the structure shown in Fig. 2 also forms a planar dielectric waveguide with a high-index core between a pair of low-index cladding layers, analogous to an optical fiber. It is characteristic for a dielectric waveguide that only a part of the optical power is present in the core since the power distribution extends well into the cladding layers. The power fraction in the core is known as the *confinement factor*, and denoted by the symbol Γ.

The optical field distribution supported by the waveguide is known as a *mode*, and the structure is usually designed in such a way that only a single mode exists. The optical field propagates at a speed given by c/n_{eff}, where the *effective index* n_{eff} is higher than that of the cladding layers, but lower than that of the core.

The width of the optical power distribution is characterized by the *spotsize*, which is on the order of, or even less than, 1 μm. Since this is small compared to the wavelength, the output beam from a semiconductor laser is usually quite divergent. The optical mode in a fiber, on the other hand, has a spotsize in the 5–10 μm range. As the laser and fiber spotsizes are not compatible, lenses are required in order to ensure a reasonably efficient coupling of light from a semiconductor laser into a fiber.

The laser cavity forms a resonator, and the cavity length L and the effective index n_{eff} are related to the lasing wavelength λ by

$$n_{\text{eff}}L = \frac{M\lambda}{2} \tag{3}$$

which states that the optical length of the cavity is an integer number of half-wavelengths, where $M \gg 1$ is known as the (longitudinal) *modenumber*. The separation, *modespacing*, between two wavelengths (*longitudinal modes*) satisfying this condition (corresponding to modenumbers M and $M+1$) is

$$\Delta\lambda = \frac{\lambda^2}{2n_{\text{eff}}L} \tag{4}$$

A typical value for the modespacing (for a cavity length of about 300 μm) is about 1 nm. For a laser operating at a wavelength of about 1550 nm, this corresponds to a spacing between the optical frequencies of about 120 GHz.

As is indicated in Fig. 2, the lasing transition in a semiconductor laser is between *energy bands*, rather than between discrete energy levels. An important consequence of this is that the gain curve is quite wide, much wider than the modespacing given by Eq. (4). This wide gain can lead to simultaneous lasing in several longitudinal modes, thereby giving an effective spectral width of several nanometers. Such a wide spectral width will lead to

dispersion problems for systems operating at high data rates over a long length of dispersive fiber.

Another characteristic feature is that the gain levels can be very high. This means that the lasing condition, as expressed in Eq. (1), can be satisfied even for low values of the reflectivity R. Since semiconductors typically have refractive index values around 3.5, sufficient reflectivity (about 30%) can be obtained from a cleaved facet, and there is no need for special high reflectivity external mirrors as is the case for other laser types.

Two material systems are of particular importance for semiconductor lasers. The first is $Ga_{1-x}Al_xAs/GaAs$. Substituting a part of the group III element Ga by Al gives a material with a wider bandgap, but with nearly the same lattice constant. This means that structures containing varying amounts of Al can be grown on GaAs substrates without lattice mismatching. A band structure as shown in Fig. 2 is obtained by having a lower Al fraction in the active layer than in the confinement layers. Lasers based on this system usually operate in the 800–900-nm spectral region (the exact wavelength depends on the Al content in the active layer), and are widely used in CD players.

The material system of particular importance for fiberoptics is $In_{1-x}Ga_xAs_{1-y}P_y/InP$. By using two group III elements and two group V elements, there are 2 degrees of freedom in the composition. The first can be used to ensure lattice matching to an InP substrate, and the second can be used to adjust the bandgap. These materials are used in lasers for the important 1300-nm (minimum dispersion) and 1550-nm (minimum loss) fiber communication wavelength regions. Figure 3 shows a schematic of an InP-based communication laser.

The layers in a laser structure are normally grown by the *MOVPE* (metal-organic vapor-phase epitaxy) process, which allows the deposition of thin layers that are uniform in both thickness and material deposition. The active stripe is formed by a combination of photolithographic and etching processes followed by an overgrowth, and metallic contacts are formed to the n-type and p-type sides of the laser. Pumping of the upper laser level is performed simply by passing a current through the structure.

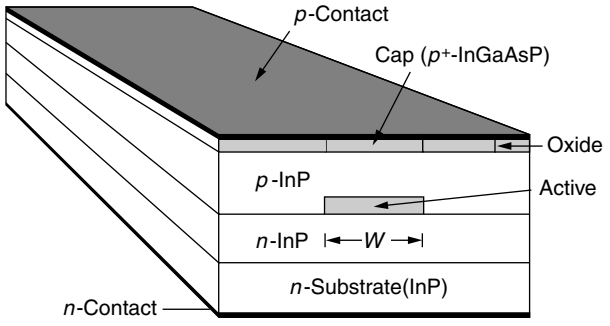


Figure 3. Diagram of communication laser. Typical dimensions are as follows: cavity length 300 μm, total width 100 μm, active region width (W) a few micrometers, substrate thickness 100 μm, active layer thickness a few tenths of a micrometer.

A semiconductor laser acts as a threshold device. For low values of the current, there is insufficient gain to satisfy the lasing condition, Eq. (1), and no laser light is emitted. Lasing starts as soon as the current is sufficient to give enough gain (this is known as the *threshold current,* I_{th}), and above this current level the output laser power increases in proportion with $(I - I_{\text{th}})$. Typically several milliwatts of optical power is emitted for a current in the range 10–100 mA. See Refs. 3–5 for more details on laser structures, and Ref. 6 for more advanced devices.

4.3. Laser Dynamics

As carriers (electrons and holes) are injected into the active region, they can recombine either spontaneously or by stimulated recombination brought about by the photon density in the active region. The photon density, on the other hand, is subject to both gain, due to the stimulated recombination of the carriers, and to losses, either internal losses or losses due to photons being emitted from the end facets of the laser. The interactions between the carrier density, N, and the photon density, S, is described by the so-called *rate equations* for the laser. In their simplest form these equations can be written as

$$\frac{dN}{dt} = \frac{I}{eV} - \frac{N}{\tau_s} - GS \tag{5}$$

$$\frac{dS}{dt} = GS - v_g\,(\alpha_{\text{int}} + \alpha_{\text{end}})\,S + \beta\frac{N}{\tau_s} \tag{6}$$

Equation (5) gives the time dependence of the carrier density. The first term on the right-hand side (RHS) is the pump term, where I is the current supplied to the laser, e is the unit charge, and V is the active volume of the laser. The second term accounts for spontaneous recombination, where τ_s is the spontaneous lifetime. Finally, the last term accounts for stimulated recombination, where G is the gain (per unit time). The second rate equation, Eq. (6), describes the time dependence of the photon density. The first term on the RHS is recognized as the stimulated recombination term. The second term accounts for losses, where α_{int} is the internal loss coefficient and α_{end} describes facet losses, both loss coefficients are losses per unit length, and v_g is the group velocity of the light in the laser. The final term occurs because a fraction β of the spontaneous emission events add a photon to the lasing mode.

The gain factor G is related to the gain per unit length in the active region, g_{act}, by

$$G = v_g \Gamma g_{act} \tag{7}$$

where the confinement factor Γ accounts for the fact that the laser active layer forms an optical waveguide with some of the power propagating outside the active region. The gain in the active region in turn is an increasing function of the carrier density N.

The facet loss α_{end} is caused by light being emitted from the ends of the laser. From Eq. (1) α_{end} can be found from

the gain required to balance the loss of photons through the facets

$$\alpha_{\text{end}} = \frac{1}{L} \ln\left(\frac{1}{R}\right) \tag{8}$$

It is a unique feature of a semiconductor laser that it can be modulated directly by varying the current [first term on the RHS of Eq. (5)], and the response of the laser can be found from the rate equations. Since these equations are nonlinear, due to the dependence of the gain on the carrier density, the rate equations cannot in general be solved analytically; however, a number of important results can still be derived from them. In the case of weak modulation, where the current I consists of a bias current plus a superimposed small-signal modulation current, the rate equations can be linearized. The result of this analysis shows that for low frequencies the optical output power will be modulated in proportion to the modulation current. For higher modulation frequencies the laser response has a resonance, with the resonance frequency increasing roughly in proportion to the square root of $(I - I_{\text{th}})$. The resonance frequency is typically in the gigahertz range. For modulation frequencies above the resonance frequency the laser response drops off rapidly. The resonance frequency provides a reasonable estimate on how fast the laser can be modulated directly, assuming that the laser response is not deteriorated by parasitic elements (such as the laser series resistance and parallel capacitance).

Other results that can be derived from the rate equations include harmonic distortion [7], and approximate expressions for the (large signal) turnon and turnoff times [8]. As spontaneous emission does not occur at a constant rate, but is a statistical process, it is possible to derive results on the laser intensity noise and its spectral distribution. Readers should consult Ref. 9 for more details on modulation and noise properties of semiconductor lasers.

4.4. Single-Frequency Lasers

In order to reduce the dispersion in optical fibers, it is necessary to restrict lasing to a single longitudinal mode. The conventional way to achieve this is by incorporating a periodic structure (*grating*) in the laser, as shown schematically in Fig. 4. In this structure the grating provides internal reflections at a wavelength determined by the grating period. This type of laser is known as a *distributed-feedback laser* (DFB).

The wavelength selected by the grating is given by

$$\lambda_{\text{DFB}} = 2n_{\text{eff}}\Lambda \tag{9}$$

where n_{eff} is the effective index and Λ is the grating period. Since the grating only provides efficient internal feedback for wavelengths very close to λ_{DFB}, any wavelength different from λ_{DFB} will have a higher rate of loss through the end facets, and as a result lasing will occur at λ_{DFB}. The discrete (and nonselective) reflections from the end facets will interfere with the distributed reflection from the grating, and usually the reflection from the front facet is suppressed by applying an *antireflection* (AR) coating.

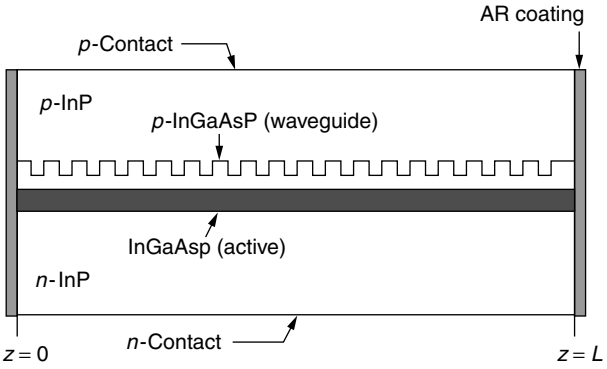


Figure 4. Outline of a DFB laser. The waveguide layer has a bandgap (and consequently a refractive index) between that of the active layer and InP.

The detailed theory for DFB lasers is rather involved, and readers are referred to Refs. 10 and 11 for more information on this topic.

Gratings can be fabricated by covering the waveguide layer with a photoresist, which is then exposed to an optical interference pattern, and the developed resist is used as an etch mask. After grating etching the remaining layers in the laser structure are grown.

The refractive index of a laser structure is quite sensitive to temperature; according to Eq. (9), this will lead to a temperature dependence of the lasing wavelength. A typical value is about 0.1 nm per degree (corresponding to a change in the optical frequency of about 10 GHz per degree). Whereas temperature tuning can be used to trim the wavelength to a given value, the temperature dependence also means that in order to ensure that the lasing frequency is within 10 GHz of a given value, the laser temperature has to be stabilized to within 1 degree.

In WDM systems signals from several lasers are multiplexed before transmission, and in order to ensure interoperability of equipment from different manufacturers, ITU has set a standard for optical transmission frequencies. This standard is based on a frequency grid with a 100-GHz spacing. Consequently the range from 192.1 THz (=1560.61 nm) to 195.9 THz (=1530.33 nm) consists of 39 channels.

4.5. Wavelength-Selectable Lasers

In order to achieve single-frequency lasing at a number of different wavelengths, arrays of DFB lasers can be formed. If these lasers are integrated with a combiner, several optical signals can be coupled into the same fiber. However, simultaneous operation of several closely spaced lasers will lead to crosstalk problems, and it may be more advantageous to consider *selectable* structures, where only one laser is operated at any time. An example of such a structure is shown in Fig. 5. This approach allows redundancy by having more than one laser per wavelength. The exact optical frequency is achieved by temperature tuning.

A different type of array is constructed by *cascading* of DFB lasers (several DFB lasers on a common optical

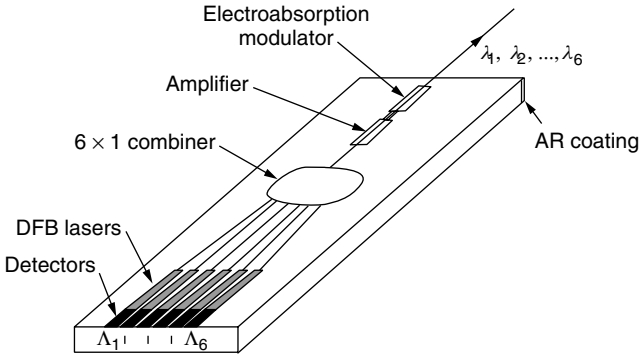


Figure 5. Selectable array with six DFB lasers and a combiner [5]. This *optoelectronic integrated circuit* also contains an amplifier to compensate for the combiner loss, a modulator for data encoding and monitor detectors for the lasers [12].

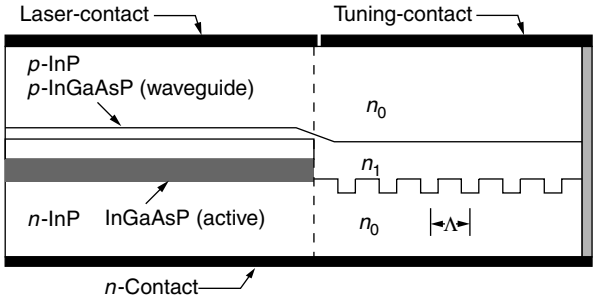


Figure 6. Two-section DBR laser. The output power is controlled by the laser current supplied to the active region. The tuning current supplied to the Bragg reflector region controls the carrier density in that region, and hence its refractive index. According to Eq. (9), this in turn tunes the wavelength at which the grating gives efficient reflection. Tuning ranges can be up to 10–15 nm [e.g., 14].

waveguide). By operating a single laser element above threshold and the other lasers close to threshold, the lasing wavelength is determined by the grating in the element operated above threshold. Several branches, each with several lasers, can be combined, and by using a high degree of temperature tuning, a wavelength range of 30 nm has been covered [13].

The wavelength range that can be covered by an array is limited by the number of array elements and by the degree of temperature tuning. For applications where many optical frequencies are required, or where a high degree of temperature tuning is undesirable, arrays may not be the best solution.

4.6. Tunable Lasers

The wide optical gain curve in a semiconductor laser makes it possible to achieve tuning of the lasing wavelength. As already mentioned, tuning is possible by changing the operating temperature. However, unless a large temperature variation is allowed, the tuning range will be limited to a few nanometers in wavelength (a few hundred gigahertz in frequency), and thermal tuning is comparatively slow (microsecond–millisecond range).

The fact that the refractive index depends on the carrier density can be applied for tuning. However, in a simple structure (such as Fig. 3 or 4), the carrier density is clamped to the value which is required to give sufficient gain to satisfy the lasing condition, and tuning by carrier density changes is not possible. This limitation can be overcome by using structures with two (or more) separate regions. One example is the *distributed Bragg reflector* (DBR) laser shown in Fig. 6. The tuning speed will be limited by the carrier lifetime in the tuning region (nanosecond range).

It should be noted that the tuning of a DBR laser is not continuous, but shows jumps between the wavelengths that satisfy the resonance condition given by Eq. (3).

Whereas the tuning range of a two-section DBR laser is limited by the extent to which the refractive index of the tuning section can be changed, wider tuning ranges can be achieved using somewhat more complicated structures.

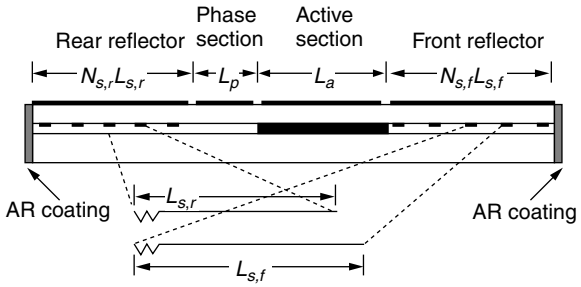


Figure 7. Sampled-grating DBR. Instead of a continuous grating the two reflectors have sampled gratings. These gratings give reflection spectra that have a comb of reflection peaks with a spacing determined by the sampling period. By using two different sampling periods, two reflection combs with different periodicities are obtained.

An example of such a structure is the *sampled-grating DBR* (SGDBR), which is shown in Fig. 7.

A small change of the refractive index of one of the tuning sections gives a large change in wavelength since a new pair of reflection peaks will coincide. This behavior is recognized as the *Vernier effect*; see Ref. 15 for more details. This principle leads to greatly enhanced tuning ranges; up to about 100 nm has been reported. Similar tuning ranges have also been achieved by combining a tunable codirectional coupler with a sampled grating [16].

A particular problem with the tunable laser structures described above is that several control currents are required for a specific combination of power and wavelength. Tunable lasers must therefore be characterized in sufficient detail to identify the current combinations required for various wavelengths, and the laser driver electronics must contain this information in such a form that tuning can be achieved in response to simple external instructions.

Other semiconductor laser structures capable of very wide tuning include external cavity lasers and vertical cavity lasers (see Section 4.7). Ultimately, the tuning range is of course limited by the width of the gain curve.

More details on various tunable laser types can be found in Ref. 17.

4.7. Vertical Cavity Surface-Emitting Lasers

In a *vertical cavity surface-emitting laser* (VCSEL), the direction of lasing is perpendicular to the active layer. Since this means that the active cavity is very short, it follows from Eq. (8) that very high end-reflectivities are required. Such high reflectivities can be achieved by having a stack consisting of a large number of layers with alternating high and low refractive index.

It is a considerable advantage of the short cavity that only one longitudinal mode exists because of the resulting wide modespacing [cf. Eqs. (3) and (4)]. This means that a VCSEL by its nature is a single frequency laser. Other major advantages include: the possibility of matching the laser spot size to that of a fiber, making coupling simpler and more efficient, and the use of on-wafer testing in the fabrication process. See Ref. 18 for more details and a review.

The various advantages of GaAs VCSELs make them highly suitable as relatively low-cost transmitters in short-distance systems operating at relatively short wavelengths, such as data links. The technology for VCSELs operating at the "telecoms" wavelength of 1300 and 1550 nm has proved to be considerably more difficult. One possible way of overcoming some of the problems is the use of 980-nm lasers for optical pumping as an alternative to electrical pumping.

Tunable VCSELs have been fabricated by incorporating an electrostatically deformable reflecting membrane at one end of the laser. A tuning range of up to 50 nm is then achieved by a simple voltage control of the cavity length [19].

4.8. Related Optical Components

A number of optical components are related to semiconductor lasers, because they are either of a similar structure or used together with semiconductor lasers:

Pump Lasers. Fiber amplifiers, used in long-haul linkage, require high-power optical pumping at specific wavelengths, usually 980 or 1480 nm. The pump power is supplied by specially designed high-power semiconductor lasers.

Semiconductor Optical Amplifiers (*SOAs*). These amplifiers are an alternative to fiber amplifiers and are very similar to lasers in structures. However, lasing is suppressed because SOA facet reflectivity is very low. Whereas SOAs can be integrated with other semiconductor components (see Fig. 5), the low coupling efficiency to fibers makes them less attractive for use as inline amplifiers in transmission systems.

Modulators. At data rates of 2.5 Gbps, directly modulated semiconductor lasers can be used, but direct modulation becomes increasingly problematic as the data rate increases, thus making dedicated modulators attractive in high data rate systems. Modulators are made from $LiNbO_3$ or semiconductors. Semiconductor-based electroabsorption modulators may be integrated with other elements, including lasers (see Fig. 5).

4.9. Packaging and Modules

In order to provide a fixed and robust coupling from a semiconductor laser to an optical fiber, the laser must be supplied in a suitably designed package. In addition to the *coupling optics*, a laser package may contain several of the following additional elements:

An *optical isolator* to prevent instabilities in the laser operation due to external reflection

A *thermoelectric element* to keep the laser temperature constant and prevent wavelength drift caused by variations in the ambient temperature

Drive electronics to provide bias current and modulation, in the case of external modulation a separate modulator may also be included

A *monitor detector* for control of the optical power from the laser

An example of a laser module is shown in Fig. 8.

BIOGRAPHY

Jens Buus was born in 1952 in Copenhagen, Denmark. He is an electrical engineer (MSc in electrophysics), graduated from the Technical University of Denmark (DTU) in August 1976. He also holds Lic. techn. (Ph.D.) and Dr. techn. (DSc) degrees from DTU. From 1979 to 1983 he was a postdoctoral fellow at DTU; from 1983 to 1992 he was with Marconi Caswell (formerly Plessey Research Caswell). Since January 1993 he has been a consultant at Gayton Photonics Ltd., United Kingdom. He

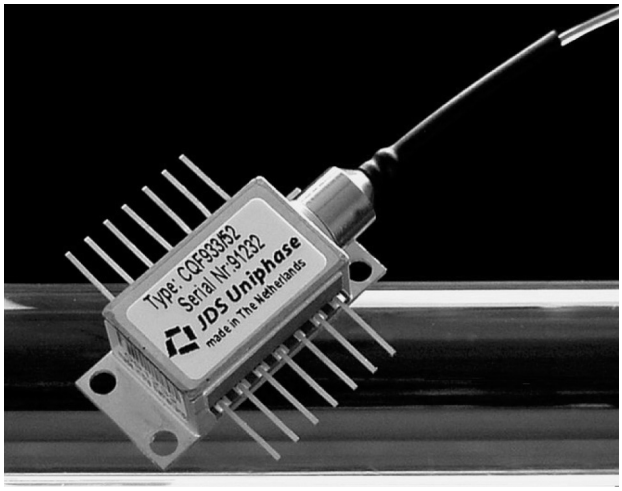


Figure 8. Packaged laser. This unit contains a fixed-frequency DFB laser (one of the standard ITU frequencies) and is designed for operation at a data rate of 2.5 Gbps. The package length is 30 mm, and the pins provide access to temperature control, monitor diode, as well as DC and AC input to the laser. (Courtesy of JDS Uniphase.)

has been project manager in the European RACE and ACTS programs and is currently project manager for a project under the IST program. Dr. Buus has served on several conference committees and given invited talks, tutorials, and short courses at several conferences; he has authored or coauthored about 60 papers, over 60 conference papers, and 2 books. During the academic years 1998–2000 he was a LEOS Distinguished Lecturer. He is a fellow of the IEEE, and a member of the Optical Society of America, the Institute of Electrical Engineers, and of the Danish Physical Society. His research has included contributions to the understanding of the properties of semiconductor lasers and optical waveguides, as well as contributions to work on gratings, integrated optics, and coherent optical communication.

BIBLIOGRAPHY

1. H. C. Casey and Jr., M. B. Panish, *Heterostructure lasers, Part A: Fundamental Principles*, Academic Press, Orlando, FL, 1978.
2. Special Issue on Semiconductor Lasers, *IEEE J. Quant. Electron.* **QE-23** (June 1987).
3. S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley, Chichester, UK, 1995.
4. L. A. Coldren and S. W. Corzine, *Diode Lasers and Photonic Integrated Circuits*, Wiley, Chichester, UK, 1995.
5. G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Van Nostrand-Reinhold, New York, 1993.
6. P. S. Zory, ed., *Quantum Well Lasers*, Academic Press, Boston, 1993.
7. T. E. Darcie, R. S. Tucker, and G. J. Sullivan, Intermodulation and harmonic distortion in InGaAsP lasers, *Electron. Lett.* **21**: 665–666 (1985).
8. R. S. Tucker, Large-signal switching transients in index-guided semiconductor lasers, *Electron. Lett.* **20**: 802–803 (1984).
9. K. Petermann, *Laser Diode Modulation and Noise*, Kluwer, Dordrecht, The Netherlands, 1988.
10. G. Morthier and P. Wankwikelberge, *Handbook of Distributed Feedback Laser Diodes*, Artech House, Norwood, MA, 1997.
11. J. E. Carroll, J. E. A. Whiteaway, and R. G. S. Plumb, *Distributed Feedback Semiconductor Lasers*, IEE, Stevenage, UK, 1998.
12. M. G. Young et al., Six wavelength laser array with integrated amplifier and modulator, *Electron. Lett.* **31**: 1835–1836 (1995).
13. J. Hong et al., Matrix-grating strongly gain-coupled (MG-SGC) DFB lasers with 34 nm continuous wavelength tuning range, *IEEE Photon. Technol. Lett.* **11**: 515–517 (1999).
14. F. Delorme, S. Grosmaire, A. Gloukhian, and A. Ougazzaden, High power operation of widely tunable 1.55 μm distributed Bragg reflector laser, *Electron. Lett.* **33**: 210–211 (1997).
15. V. Jayaraman, Z.-M. Chuang, and L. A. Coldren, Theory, design, and performance of extended tuning range semiconductor lasers with sampled gratings, *IEEE J. Quant. Electron.* **29**: 1824–1834 (1993).
16. P.-J. Rigole et al., 114 nm wavelength tuning range of a vertical grating assisted codirectional coupler laser with a super structure grating distributed Bragg reflector, *IEEE Photon. Technol. Lett.* **7**: 697–699 (1995).
17. M.-C. Amann and J. Buus, *Tunable Laser Diodes*, Artech House, Norwood, MA, 1998.
18. K. Iga, Surface-emitting laser — its birth and generation of new optoelectronics field, *IEEE J. Select. Top. Quant. Electron.* **6**: 1201–1215 (2000).
19. D. Vakhshoori et al., 2 mW CW singlemode operation of a tunable 1550 nm vertical cavity surface emitting laser with 50 nm tuning range, *Electron. Lett.* **35**: 900–901 (1999).

OPTICAL SWITCHES

K. L. EDDIE LAW
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

Optical transport networks have been deployed around the world for many years. As an article [1] in *Nature* indicated, the theoretical maximum bandwidth of a typical optical fiber in access networks is estimated to be about 150 Tbps (terabits per second). Even though it may be hard to estimate if this "glass ceiling" of 150 Tbps will actually hold, the speed of commercial transport systems has already been reaching 40 Gbps (i.e., OC[1]-768) in synchronous optical networks (SONETs). On the other hand, dense wavelength-division multiplexing (DWDM) systems can deliver information in a number of wavelengths in one optical fiber. NEC[2] was successful in transmitting 10.92 Tbps in a 117-km-long fiber with 273 wavelength channels at 40 Gbps per channel data rate. With 40 Gbps channels, we need 3750 channels to reach this fiber bandwidth limit. In the case that if we will be able to build a futuristic 160-Gbps channel, 900 channels will suffice to reach this limit. The technology of the optical transport system has been evolving rapidly. Obviously, the switching nodes are the bottlenecks in today's optical networks. Without optical logic technology, the switching nodes need to undergo signal conversions from photons to electrons in order to switch packets to their respective outgoing ports. Thereafter, the packets will be converted and delivered in the form of photons. As of today, all SONET optical switching systems carry out optical–electrical–optical (OEO) conversions.

[1] OC-N stands for optical carrier digital signal rate of $N \times$ 53 Mbps in SONET.

[2] NEC announced the DWDM transmission capacity world record on March 22, 2001. Exactly one year later, Lucent announced the transmission distance world record of 4000 km with 64 channels running 40 Gbps on March 22, 2002.

Active research on optical switches has been carried out with the goal of constructing all-optical networks that do not require any OEO conversions. A switching core may need to switch information from L incoming fibers to L outgoing fibers. Each fiber carries multiple, W, wavelength channels of information with the DWDM technology. Therefore, the designs of the optical cross-connects (OXCs) are getting complicated with the rapidly increasing number of wavelength channels per fiber. The architectural designs of OXCs are in two dimensions that involve the space and wavelength-switching domains [2,4–6]. That is, an ultimate OXC design should switch information from one particular wavelength channel at an input port to a specific output port with a selected outgoing wavelength channel. Therefore, given an $N \times N$ optical cross connect design, then $N \geq L \cdot W$ is required to provide an internally nonblocking switching matrix for any wavelength channel to any fiber. Figure 1a shows a system design of an OXC

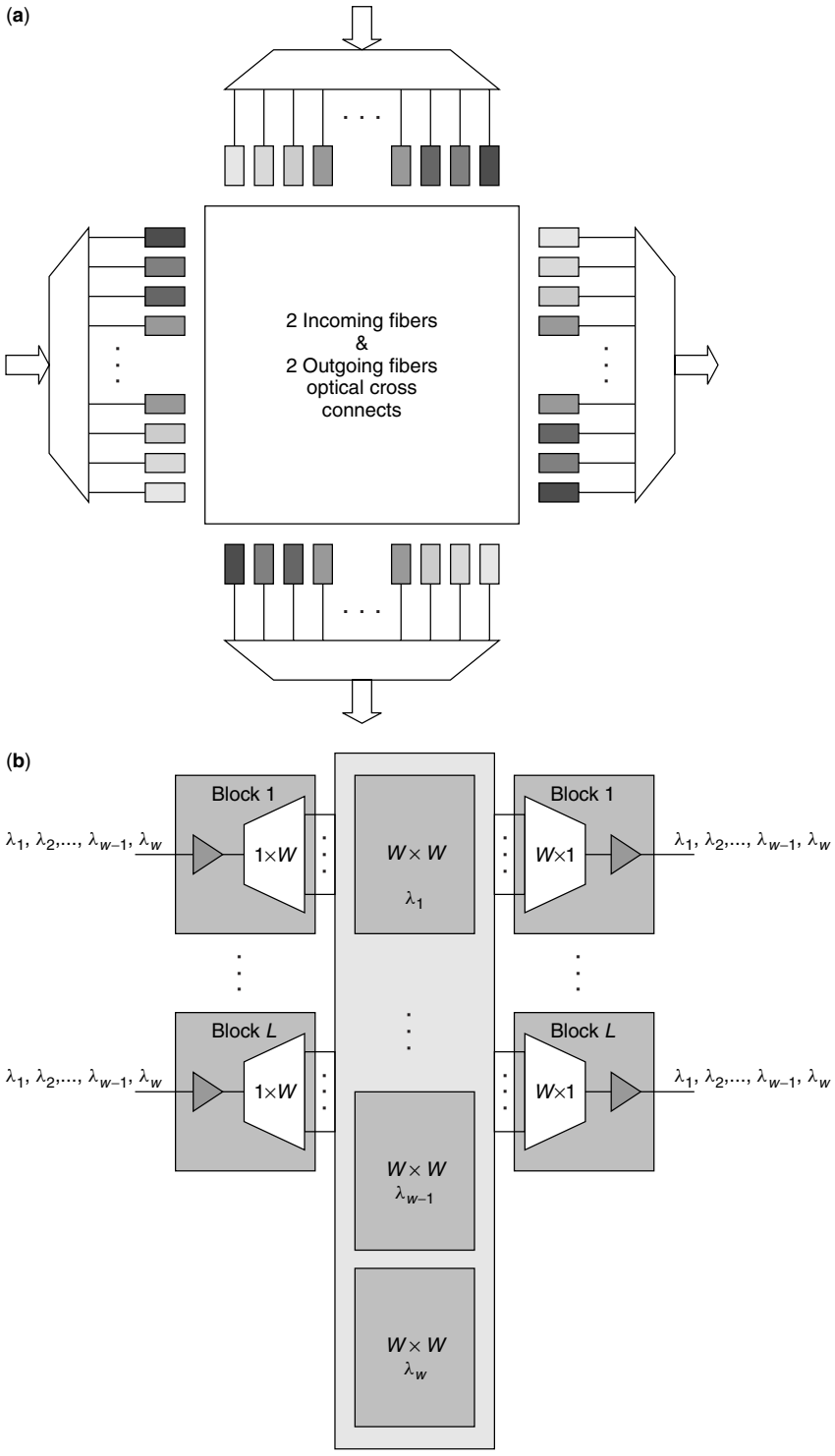


Figure 1. (**a**) Nonblocking OXC; (**b**) blocking 3-stage OXC.

with two incoming fibers and two outgoing fibers. We need to install wavelength filters and converters to produce a flexible design. Unfortunately, the cost of those converters is expensive. Traditionally, the Clos networks [3] are nonblocking multistage circuit-switched networks in the electronic domain. The alternative multistage design, as shown in Fig. 1b, has lower cost without any wavelength converters; however, it is not a nonblocking architecture. In this article, we discuss different architectural designs of the OXCs from a system perspective with consideration for device constraints.

2. DESIGN CONSTRAINTS

The performance of the next-generation all-optical networks relates to the optical properties and the functionalities of optical devices, and hence the throughput performance of the resulting OXC architectures. With different characteristics of optical devices, the selection criteria of the device components will definitely affect both the design architecture and the performance of the resulting optical cross-connects. Currently, the research on OXCs is still at an early phase. There are several technologies that are generally recognized to have the potential to construct the next-generation optical switches. They include the electrooptic, thermooptic, acoustooptic, thermobubble, liquid crystals, optomechanical, and beam-steering technologies. Among the initial investigations, some device level designs have already demonstrated that they are excellent candidates to be the basic building blocks for constructing the next-generation OXCs.

Before we describe the designs of optical switches, we consider several limiting factors of optical devices that may affect the architectural designs. With the wide bandwidth for transporting optical signals in fibers, the switching rate in OXC will be a limiting factor on the signal transfer rate. The switching time is determined through the rate of changing states on forwarding or detouring optical signals in devices. It may fall in the range from nanoseconds to milliseconds. The smaller the switching time, the better it is. It determines the information transfer rate in terms of bits, packets, and bursts. Moreover, it also indicates if the resulting optical networks can operate with circuit-switching, packet-switching, or burst-switching technology. Apart from switching time, port count, reliability, size, and cost are important design criteria with the space and wavelength switching in optical networks.

The overall loss budget is an important criterion that determines the power consumption and the placements of the optical amplifiers. There are different loss factors that include insertion loss, crosstalk, chromatic dispersion, polarization-dependent loss (PDL), and polarization-mode dispersion (PMD). Moreover, some architectures may be wavelength-dependent, and some may have a wide variation of losses between ports. In general, the maximum loss budget for an OXC should be around 25–30 dB. Therefore, it is seldom justifiable to create multistage networks if the per stage module has a high loss factor; otherwise, we need to provide signal amplification between stages. In the following, we focus on reviewing some important parameters: the insertion loss, crosstalk and switching time.

2.1. Insertion Loss

Whenever a photonic device is introduced in a lightpath, it introduces an insertion loss due to the mismatch at the interface. Ideally, it should be as small as possible in order to minimize the total loss budget, especially if the optical signal needs to travel through multiple OXCs. It is also important to determine a switching module that has uniform loss distribution with different interconnection patterns. If the insertion loss changes with different interconnection patterns, then a variable equalizer is required between switching stages. The resulting design is undesirable for it complicates the system control and increases the cost of the switch.

Free-space optical systems have the lowest insertion loss. For example, the microelectromechanical systems (MEMSs) belong to the optomechanical design class. The insertion loss of a MEMS OXC can reach as low as 1 dB. It is usually in the range from 1 to 6 dB depending on the size of the MEMS switch. Liquid crystal electrooptic is another switching technology for constructing OXCs. Its insertion loss is comparable to the MEMS switch as it can also reach 1 dB loss [23]. However, polarization loss may occur in the liquid crystal module. It relates to the Fresnel reflection on the glass–air interface and it can be as high as 3 dB. The Fresnel reflection occurs at a planar junction of two materials that have different refractive indices and is not a function of the angle of incidence. There are currently no reports on building large-scale liquid crystal switches.

For the other device technologies, for example, the insertion loss of thermooptic switches is usually in the range of 6.6–9 dB;[3] the lithium niobate ($LiNbO_3$) electrooptic switch[4] has an insertion loss of <9 dB. There has been some steady progress on improving both of these technologies. We can find moderately sized OXCs with these technologies in the market. However, the insertion loss is still considered to be comparatively high for next-generation large-scale OXCs. As a result, the $LiNbO_3$ switch is usually used in the external modulation rather than in the lightpath routing. This is because the external modulated information is usually amplified before it is transmitted.

2.2. Crosstalk

Crosstalk may be caused by either interference from signals on different wavelengths, the interband crosstalk, or interference from signals on the same wavelength on another source, the intraband crosstalk. Interband usually determines the channel spacing. Intraband crosstalk usually occurs in switching nodes where multiple signals

[3] The insertion loss of the 8×8 switch using a thermooptic Mach–Zehnder interferometer is <8 dB from NTT Electronics at *http://www.nel-dwdm.com/profile/profile.html*. Its switching speed is < 3 ms.

[4] The insertion loss of the 8×8 crossbar switch using $LiNbO_3$ planar lightwave circuit is <9 dB from Lynx Photonic Networks at *http://www.lynxpn.com/*. Its switching speed is <5 ns.

on the same wavelength are being switched from different inputs to different outputs. The degree of intraband crosstalk depends on the switch architectures.

In an optical device, crosstalk happens when a portion of the input signal "leaks" into another signal as they copropagate through the switch fabric. The ratio of the power at the unselected output port over the total input power in a switch element is referred to as the *crosstalk ratio* of the switch, since crosstalk is the noise usually introduced from the nearby connections. Therefore, crosstalk is usually more serious if the switch architecture design is complicated, especially if it has a large number of ports and connections. Since crosstalk measures the power of the loss signal to the input signal power [4–6], it is desirable if the value of crosstalk is as negative as possible in decibels.

With free-space optomechanical designs, MEMS optical switches provide the best crosstalk performance among all switching fabric technologies. Its crosstalk is in the range of −55 to −60 dB [29–33]. Besides, liquid crystal provides excellent insertion loss performance, and the crosstalk can reach −48 dB in general. There was a report on constructing an 8×8 crossbar liquid crystal switch with 1×8 switch arrays and the crosstalk could reach −59.5 dB [18,22,23]. This result is comparable to the MEMS switches. Unfortunately, there are still difficulties in building large-scale liquid crystal crossbar switches with the tradeoff between loss uniformity and the crosstalk level. Nevertheless, the liquid crystal switching technique is expected to improve with time, and is considered as a good candidate for building optical switching modules.

For the silica-based thermooptic switch using double-Mach–Zehnder interferometer (MZI) waveguide units, the crosstalk can reach −43 dB through a sophisticated hardware architecture design. There is a tradeoff between hardware complexity and the crosstalk level. To achieve this excellent crosstalk performance, we need to increase the hardware complexity by interconnecting 256 double-MZI units for a 16×16 silica thermooptic switch [12]. The resulting switch had an insertion loss and extinction ratio[5] of 17.5 and 32.9 dB, respectively. It will not be a cost-effective approach to construct an OXC with a large port count with thermooptic waveguide designs.

Some $LiNbO_3$ electrooptic switches are used to construct directional couplers by altering the refractive index of the waveguide with electric energy. These directional couplers were initially considered to have the potential to construct multi-stage switching networks. However, they cannot be used for OXCs because they have poor crosstalk isolation and a large insertion loss. On the other hand, there are $LiNbO_3$ acoustooptic tunable switches (AOTS). Acoustooptic switching technology uses surface acoustic waves to generate birefringence grating and alter the polarization of a lightbeam. Switching occurs at high speed, and it can reach as low as 3 μs.[6] An AOTS introduces about 5–6 dB insertion loss; however, its crosstalk ratio is about −20 dB [11–13]. The crosstalk of a single-channel acoustooptic 1×300 demultiplexer can reach about −35 dB. AOTS suffers both interband and intraband crosstalk. A double-stage devices or weighted coupling schemes may be required to reduce intraband crosstalk. Since interchannel interference may create intrinsic modulation of the transmitted signal, it affects the bit error rate (BIT) performance, and hence the device may not be working for long-haul optical systems. Both of these electrooptic and acoustooptic $LiNbO_3$ devices have yet to demonstrate that they can be used to build large-scale OXCs. Therefore, only small-scale OXCs [13] can be found in the market with these technologies. All in all, the optomechanical switch provides the best performance in crosstalk level compared to the other technologies.

2.3. Switching Time

The switching time describes the time it takes for a switch to establish an interconnection pattern. The desirable value must be as small as possible. As the data rate exceeds 10 Gbps per wavelength, a submicrosecond switching time is necessary to provide dynamic path provisioning, grooming, and path restoration on failure. This is an important parameter that determines the performance of future optical networks. Today's optical core networks are configured statically. When the optical device is able to switch states actively, future optical core networks are expected to provide dynamic path routing capability.

Among all the switch fabric technologies, the electrooptic switches have the fastest switching time compared to the other two technologies. The $LiNbO_3$ and semiconductor optical amplifier (SOA) switches are able to switch in the range of nanosecond response time. As discussed before, the $LiNbO_3$ device is suitable only for providing external modulations.

On the other hand, the thermooptic switch offers a switch time within the range of 1 ms. This response time is acceptable in optical path-switching applications. Unfortunately, the thermooptic switch's insertion loss is also considered to be too high when designing a large-scale switch. On the other hand, the mechanical switches, for example, fiber bundle switches [7], can achieve a good crosstalk level as well as low insertion loss. However, these fiber bundle mechanical switches usually have slow switching times. Fortunately, with the introduction of MEMS optical switching systems, apart from having excellent crosstalk ratio and low insertion loss, good mechanical design can also lead to good switching time, such as 700 μs [36]. At the moment, MEMS becomes the most appealing switch fabric technology for designing large-scale OXCs for future optical networks. In Table 1, we outline the characteristics of different available optical device technologies for building OXCs.

On concluding this part, we would like to outline the basic requirements for designing the OXCs as follows. The design should have (1) low insertion loss (typically <1 dB), (2) low crosstalk (typically <-50 dB), (3) low polarization-dependent loss (PDL), (4) switching time faster than or, at least, equal to millisecond range, (5) low power consumption, (6) long-term reliability, (7) small size, (8) low cost, (9) scalability to large port count, and

[5] The extinction ratio is defined as the ratio of the optical power transmitted for a bit "0" to the power transmitted for a bit "1."

[6] This is the 1×300 demultiplexer reported by the Light Management Group at *http://www.lmgr.net/*.

Table 1. Comparisons Between Different Optical Device Technologies

	Free-Space		Guided-Wave Integrated Optics		Guided-Wave Active Component
	MEMS	Liquid Crystal	Thermooptic, Bubble	Electrooptic	Semiconductor Optical Amplifier
Switching time	1–10 ms	2–5 ms[a]	1–10 ms	nsec	nsec
Insertion loss	Very good	Moderate	Moderate	Moderate	Acceptable
Crosstalk	Very good	Average	Average	Acceptable	Acceptable
Polarization-dependent loss	Good	Good	Good	Acceptable	Acceptable
Wavelength dependence	Good	Good	Average	Average	Acceptable
Bit rate transparency	Good	Good	Good	Good	Good
Power consumption	Good	Good	Bad	Good	Bad
Expandability/size	Large	Moderate	Small	Small	Small

[a]There was a report that could have a switching time of ~35 μs. However, the commercial products are usually in microseconds.

(10) self-holding or latching mechanism design. Since the total loss should be less than 30 dB, it then also limits the number of cascaded modules in the architectural design.

3. PROMISING SWITCHING FABRIC TECHNOLOGIES

Optical cross-connects can be classified into two broad classes: active and passive. Without optical logic devices, today's optical network can only offer a high-speed and large capacity transport system. Hence the optical routing paths are comparatively static and mostly preconfigured through the optical add/drop multiplexers (OADMs). In the near future, we expect higher deployments of the passive OXCs. These passive OXCs are usually designed with traditional doped waveguide technology. The switching characteristics are predefined and fixed; that is, an output signal pattern depends on the architectural design of a passive OXC and a specific arriving input signal pattern. In contrast, the switching points in active OXCs should be set according to the destination ports of the incoming signals. Most of the available technologies for constructing active OXCs are listed in Table 1. In Section 3.1, we will discuss the design of passive OXCs using arrayed waveguide gratings (AWGs). The technologies for constructing active OXCs is different from those for passive OXCs, and the designs can be found in Section 3.2.

3.1. Passive Optical Cross-Connects

Bulk optic or all-fiber filters and devices are used in a number of WDM applications. With the improvement of technology, the trend is to move toward monolithic integration of devices and components. One of those generic devices is the *arrayed waveguide grating* (AWG) multiplexer, also known as *waveguide grating router* (WGR). DWDM networks permit large capacity optical signal transfer. AWG can be used to split and combine optical signals of different wavelengths in the systems. The silica AWG allows the fusion splice of fiber to chip; however, it has low-contrast waveguide structure and its size is relatively large [8]. The AWG device can also be fabricated on indium phosphide (InP) [8] that provides high-index contrast and it is suitable for large-scale system integration. As shown in Fig. 2, an AWG [4–6,9] consists of two free-space couplers connected by a grating array. The first coupler has N inputs and N' outputs (where $N \ll N'$), while the second one has N' inputs and N outputs. For the first coupler, there is a regular angular distance between any adjacent input ports. Similarly, there is also another regular angular separation between any adjacent output arrayed waveguides. The setup of the second coupler is simply a mirror image of the first one. The grating array between couplers consists of N' waveguides, with lengths $l_0, l_1, \ldots, l_{N'-1}$, where $l_0 < l_1 < \cdots < l_{N'-1}$. The length difference between any two adjacent waveguides is constant. The constant difference in the lengths of the waveguides creates a phase difference in adjacent waveguides. This phase shift depends on the propagation constant in the waveguide, the effective refractive index of the waveguide, and the wavelength of the light. At the input of the second star coupler, the phase difference in the signal will be such that the signal will constructively recombine only at a single output port.

With this design, two signals of the same wavelength coming from two different input ports will not interfere with each other in the grating because of an additional phase difference created by the distance between any two input ports. The two signals will be combined in the grating but will be separated again in the second coupler and directed to different outputs. AWGs have been successfully demonstrated for a number of WDM enabling devices that include multiplexers, demultiplexers, channel dropping filters/equalizers, and tunable lasers.

Several important fundamental properties of AWGs enable the construction of passive OXCs [10]:

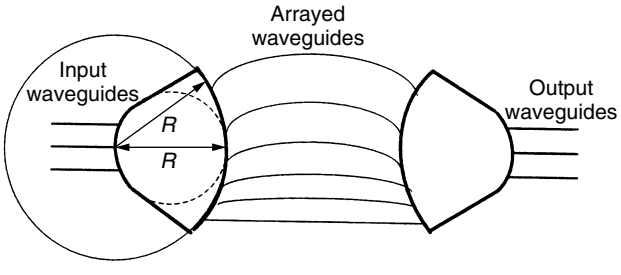


Figure 2. Arrayed waveguide grating (AWG).

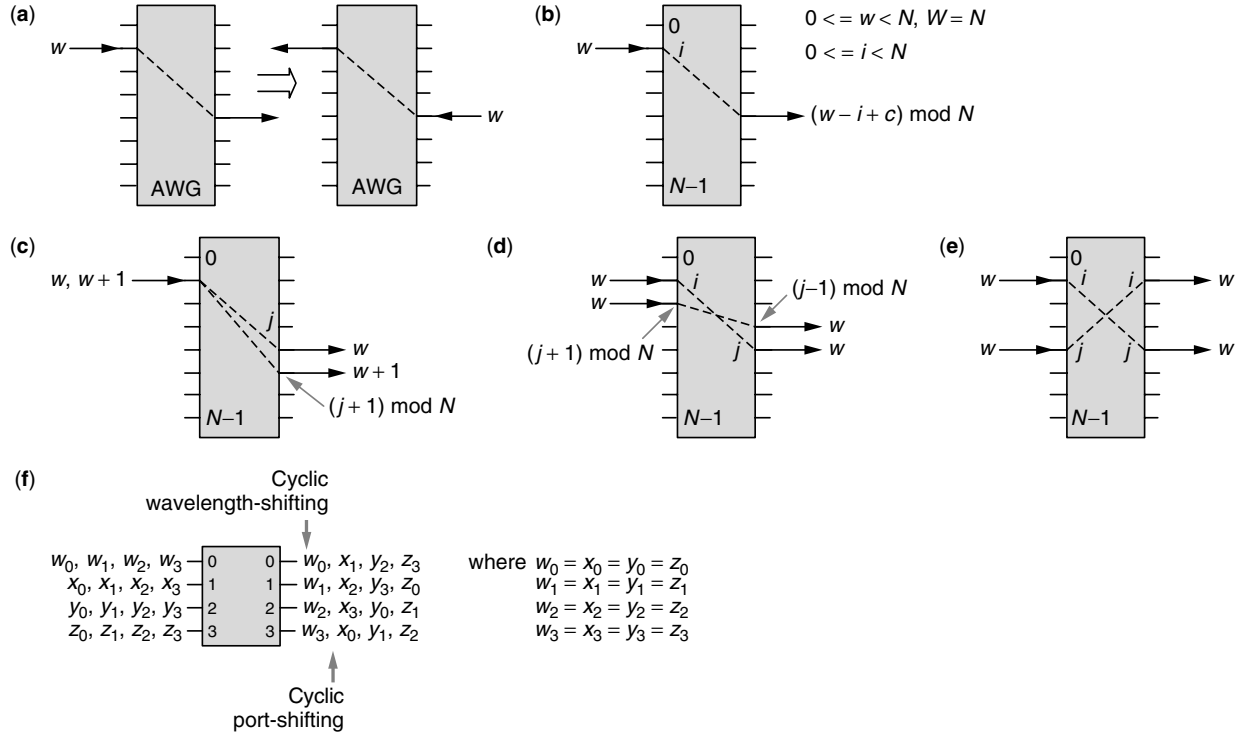


Figure 3. Routing properties of AWG.

1. *Reciprocity* (Fig. 3a). If a signal of a wavelength propagates from one input port to an output port, then any signal of the same wavelength injected at that output port will propagate backward to the input port in exactly the same way.
2. *Periodicity in Frequency* (Fig. 3b). A given frequency bandwidth may contain a number of wavelength channels. If all these wavelength channels in that frequency range follow the same transfer function in a device, then this frequency period is known as *free spectral range* (FSR). An $N \times N$ AWG has N input and output ports, and supports W wavelength channels, denoted by the sets $\mathcal{N} = \{0, 1, \ldots, N-1\}$, and $\mathcal{W} = \{0, 1, \ldots, W-1\}$ in an FSR, respectively. In general, $\mathcal{W} = \mathcal{N}$. Within an FSR, the wavelength channels have constant-frequency spacing instead of constant-wavelength spacing. For the periodicity property, a wavelength signal $w \in \mathcal{W}$ enters an input port $i \in \mathcal{N}$ is delivered to an output port $[(w - i + c) \bmod N]$. This c is an integer known as the FSR constant that depends on the selection of the FSR.
3. *Cyclic Wavelength Shifting* (Fig. 3c). If an input signal of wavelength w leaves AWG from port j, then any input signal of wavelength $w + 1$ entering the same port leaves the AWG from port $[(j + 1) \bmod N]$.
4. *Cyclic Port Shifting* (Fig. 3d). If an input signal enters port i and leaves AWG from port j, then any input signal of the same wavelength entering port $[(i + 1) \bmod N]$ leaves the AWG from port $[(j - 1) \bmod N]$.
5. *Symmetry* (Fig. 3e). If an input signal enters port i and leaves port j, then any input signal of the same wavelength entering from port j will leave the AWG from the port i.

As an example, the wavelength routing assignments of an AWG are shown in Fig. 3f. It is a 4×4 AWG with four wavelength channels in an FSR, specifically, $N = W = 4$. The four incoming ports are identified with w, x, y, and z from top to bottom. Assuming that c is zero and observing the top input port, then for an incoming wavelength numbered zero, w_0, the outgoing port is $[(w - i + c) \bmod N] = 0$, the top output port, with the periodicity property. With the cyclic wavelength-shifting property, we can arrange all four w wavelength channels sequentially at the outgoing ports as shown. Then observing the second top input port, the incoming wavelength numbered zero, x_0, goes to the outgoing port $[(-1) \bmod 4] = 3$ from the cyclic port-shifting property. The other wavelength channels can then be arranged with both the cyclic wavelength-shifting and cyclic port-shifting properties. The resulting wavelength assignment is shown in Fig. 3f. Moreover, it also satisfies the symmetry property.

There is a channel spacing concept in AWGs that allows more flexible wavelength assignments with the AWGs. In the following, a system with channel spacing of k is considered. From architectural point of view, two successive wavelengths that enter the same input port will be routed to two output ports x and $x + k$ with the wavelength-shifting property. From the setup of wavelength channels in an FSR, the two output ports are $[x = (w - i + c) \bmod N]$ and $[x + k = (w + k - i +$

c) mod N] with the periodicity property. This implies a spacing k between two successive wavelength channels at the input. Therefore, a wavelength set $\mathcal{S}$ is said to have a channel spacing of k if $\mathcal{S} = \{(s + ik) \bmod N : 0 \le i < |S|\} \subseteq \mathcal{W}$, for an $s \in \mathcal{S}$.

3.1.1. Compatible Ports. In order to have an AWG to perform as an OXC, we need to explore different routing properties that enable an AWG to do switching. Since one fiber can carry multiple wavelength channels, only some subsets of the N ports should be used for inputs and outputs. Given an $N \times N$ AWG with L incoming fibers, and if each fiber carries a set of wavelength channels $\mathcal{W}$, where $|\mathcal{W}| = W$, then we have $N \ge L \cdot W$ for the AWG to do proper switching.

A set of compatible ports $\mathcal{P}$ is defined with respect to $\mathcal{W}$ if for any pair of wavelengths in $\mathcal{W}$ that enter any two ports in $\mathcal{P}$ will be routed to two different output ports. With this compatible port concept, all incoming signals can be routed to some predetermined outgoing ports, and that will be useful for designing a passive OXC. In the following, two different cases for compatible ports will be examined. Given $\mathcal{N} = \{0, \ldots, N-1\}$, $|\mathcal{W}| = \mathcal{W}$ and $N \ge L \cdot W$ for both of them, $\mathcal{P} = \{p_i \in \mathcal{N} : 0 \le i \le L-1\}$ is defined as a set of L ports for L incoming fibers.

First, considering a channel spacing of 1, if $N = 8$, $W = 4$, and $L = 2$, then we have $\mathcal{P} = \{p_0, p_1 \in \mathcal{N}\}$. One possible arrangement of both the $\{p_0, p_1\}$, where $0 \le p_0 < p_1 < N$, can be found in Fig. 4a if $c = 0$. Mathematically, $\mathcal{P}$ is compatible with respect to $\mathcal{W}$ if and only if $p_i - p_{i-1} \ge W$, for all $1 \le i \le L-1$, and $p_{L-1} - p_0 \le N - W$. In particular, if $N = LW$, then $\mathcal{P}$ is compatible with respect to $\mathcal{W}$ if and only if $\mathcal{P}$ has an equal spacing of W. Second, if the channel spacing is L and N is a multiple of L, then the set of compatible port is $\mathcal{P} = \{p_i, p_j \in \mathcal{N} : p_i \bmod L \ne p_j \bmod L, \forall\, 0 \le i < j \le L-1\}$. The corresponding example can be found in Fig. 4b.

3.1.2. Self-Blocking Ports. With a selected set of compatible ports, only certain numbers of input and output ports are used. If some idle ports can be used for intermediate processing, then the number of AWGs may be possibly reduced to construct an OXC. For this purpose, a set of input ports $\mathcal{P}$ is defined as self-blocking with respect to $\mathcal{W}$ if they are mutually exclusive with their outgoing ports: $\mathcal{P} \cap \{(w - i + c) \bmod N : i \in \mathcal{P}, w \in \mathcal{W}\} = \emptyset$. This implies that the same set of self-blocking input ports of an AWG switch can be used as the output set and should not be used for any intermediate processing. An example shown in Fig. 4c has parameters $c = 1$, $W = 4$, and $N = 5$, and the top link is the self-blocking port.

It is desirable to construct a passive OXC with one AWG having such a set of input ports that is both compatible and self-blocking. For L incoming fibers, a necessary condition for the existence of L self-blocking and compatible ports with respect to $\mathcal{W}$ is $N \ge L \cdot (W + 1)$.

3.1.3. Compatible and Self-Blocking Ports. In order to design a preconfigured architecture with one AWG, it will be desirable to find a set of ports $\mathcal{P}$ that is both compatible and self-blocking with respect to $\mathcal{W}$:

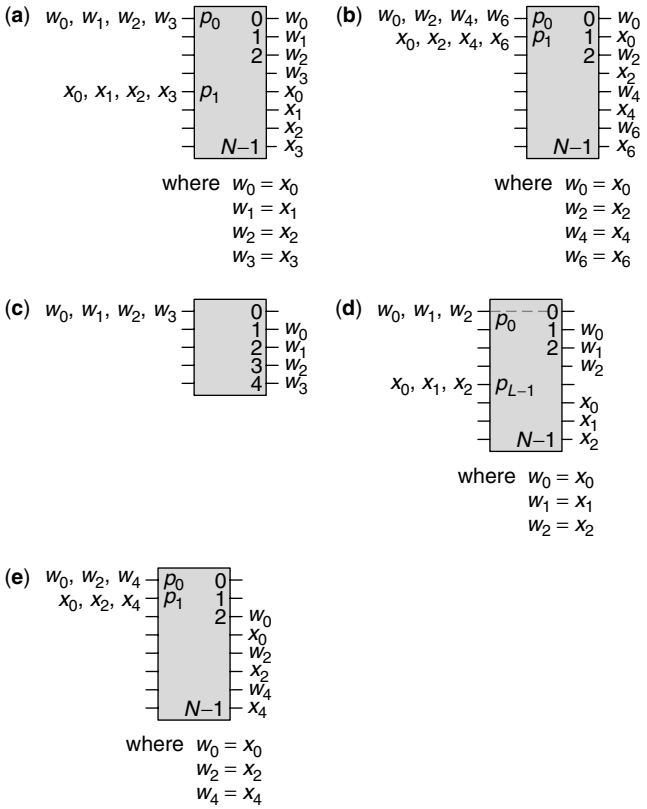


Figure 4. (**a**) Compatible ports with $k = 1$; (**b**) compatible ports with $k = 2$; (**c**) self-blocking port; (**d**) compatible and self-blocking ports with $k = 1$; (**e**) compatible and self-blocking ports with $k = 2$.

1. $\mathcal{P}$ is invariable with respect to some wavelength $w \notin \mathcal{W}$
2. $\mathcal{P}$ is compatible with respect to $\mathcal{W} \cup \{w\}$.

The invariable principle states that both the input and output ports are identical with respect to a specific wavelength w that is not in $\mathcal{W}$. The final design of the OXC includes the wavelength channels that are delivered to the set of compatible ports excluding the invariable ports with respect to w. These invariable ports are then the self-blocking ports with respect to $\mathcal{W}$.

As shown in Fig. 4d, a derived set of compatible and self-blocking ports is shown if the channel spacing is 1. This can be easily modified from the set of compatible ports shown in Fig. 4a. Mathematically, if N is a multiple of $(W+1)$, then the set of input ports can be described as $\mathcal{P} = \{i : 0 \le i < N, i \bmod (W+1) = q\}$ for some $0 \le q \le W$. If

$$[w - (2q+1) + c] \bmod (W+1) = 0$$

then $\mathcal{P}$ is compatible and self-blocking with respect to $\mathcal{W}$. On the other hand, when the channel spacing is L, then $\mathcal{W} = \{(w + Lx) \bmod N : 0 \le x < W\}$ for some integer w and $0 \le w \le N-1$. A possible configuration can be found in Fig. 4e, which can be easily obtained from Fig. 4d. Mathematically, if N is a multiple of L, then $\mathcal{P} = \{(q+i) \bmod N : 0 \le i \le L-1\}$, for some $0 \le q \le N$, is

compatible and self-blocking with respect to $\mathcal{W}$ if

$$[w - (2q + 2L - 1) + c] \bmod N = 0$$

or

$$[w + L(W - 1) - (2q - 1) + c] \bmod N = 0.$$

Detailed proofs can be found in Wan's treatise [10].

3.1.4. Implementations of Passive OXCs with AWGs. In this subsection, designs of different OXCs with AWGs will be discussed. It is assumed that there are L input fibers and each of them carries W wavelength channels. A design shown in Fig. 5a is equivalent to the one in Fig. 1b with $L = W$, a number of multiplexers, demultiplexers, and space switches. From Fig. 5b, only two AWGs are required to build a 2-line 4-wavelength OXC if the channel spacing is 1. The drawbacks are that some complicated crossover links must be set up between AWGs and the 2×2 space-switching modules. The designs of these 2×2 space-switching modules can be found in Section 3.2.

On careful investigation, there is a simple way to improve that design by selecting a wavelength set $\mathcal{W}$ with channel spacing L, where N is a multiple of L and L is even. As shown in Fig. 4b, a set of compatible ports, $\mathcal{P} = \{(q + i) \bmod N : 0 \le i < L\}$ for some $0 \le q < N$, is selected. Another desired requirement is to make sure that no waveguides are required to connect the top and bottom portions of an AWG to a space switch. This is achievable if any wavelength w from port $[(q + L - 1) \bmod N]$ is routed to a port that is a multiple of L, specifically, $\{[w - (q + L - 1) \bmod N + c] \bmod N\} \bmod L = 0$. An OXC can be built as long as q is selected such that $(w + c + 1 - q) \bmod L = 0$ is satisfied. If $w = c = 0$ and $q = 3$, then a cascaded configuration of a 2-line 4-wavelength OXC with a channel spacing of 2 is as shown in Fig. 5c.

So far, two AWGs are needed to build the designs shown in Fig. 5b,c. There are unused ports on the input sides in these designs. If a set of compatible and self-blocking ports with respect to $\mathcal{W}$ can be formulated as shown in Fig. 4d, then only one AWG with loopback links can be constructed as an OXC. Given $N \ge L\ (W + 1)$, the AWG operates as a multiplexer and demultiplexer in an OXC with the symmetric property. Furthermore, it may even reduce the insertion loss[7] in a single-AWG design. The AWG initially demultiplexes the L input fibers into $L \cdot W$ wavelength components. For those L identical wavelength channels, there is an $L \times L$ space switch. By taking advantage of the symmetric property, the signals can be rerouted to the input channels of the AWG. These

[7] The insertion loss of a 1-fiber 40-channel AWG from NTT Electronics is at worst 6 dB.

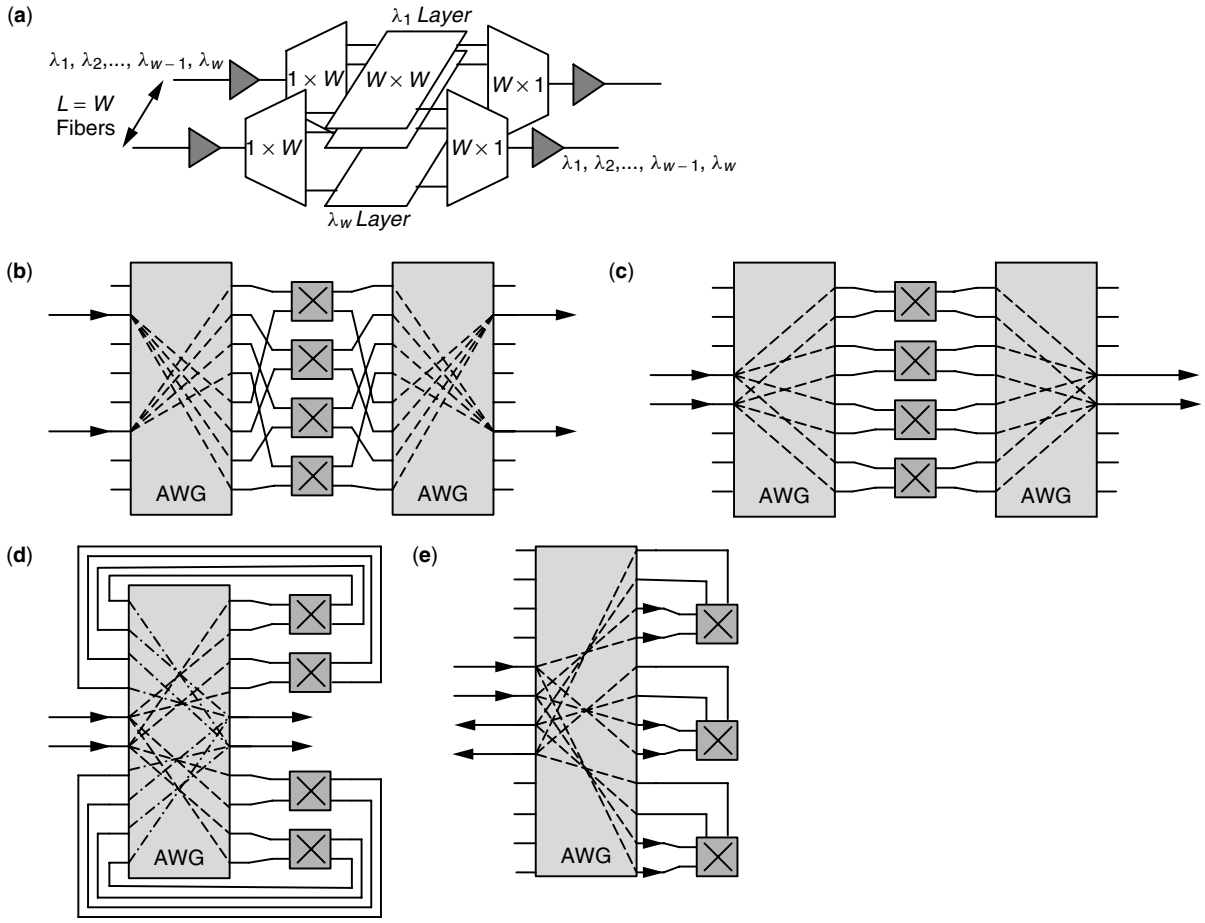


Figure 5. Optical cross-connects with AWGs.

multiple wavelength channels can then be remultiplexed at the L output ports. An example can be found in Fig. 5d with only one AWG. In this design, N is equal to $L \cdot (W+1)$ to minimize the cost, and the channel spacing is 1. To characterize the architecture mathematically, there is a set of compatible and self-blocking ports, $\mathcal{P} = \{0 \le i < N : i \bmod (W+1) = q\}$ with respect to the wavelength channels, $\mathcal{W} = \{(w+x) \bmod N : 0 \le x < W\}$. Provided $[w - (2q+1) + c] \bmod (W+1) = 0$ for some $0 \le w < N$ and $0 \le q \le W$, we obtain a passive loopback 2-line 4-wavelength OXC as shown in Fig. 5d with $w = 0, c = 1, q = 4$.

The design of a loopback architecture is effective on port utilization of an AWG. However, it may not be desirable to construct those recirculating fibers from switches to the input side of the AWG. Therefore, there is a foldback design as shown in Fig. 5e. In this design with $N \ge 2LW$, a set of double-sized compatible ports with respect to $\mathcal{W}$ is required. Suppose that $\mathcal{P}_1$ and $\mathcal{P}_2$ are the sets of L input and L output ports, respectively. It is desirable for $\mathcal{P}$ to be the set of compatible ports if $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Similar to the last design of OXC, the AWG initially demultiplexes incoming signals into $L \cdot W$ wavelength components. Subsequently, an $L \times L$ space switch does the proper signal switching for each wavelength channel. In this design, a switch accepts the wavelength signals from a set of L input ports, $\{(w - i + c) \bmod N : i \in \mathcal{P}_1\}$, and delivers the signals to another set of L output ports, $\{(w - i + c) \bmod N : i \in \mathcal{P}_2\}$. The design makes use of both the reciprocal and symmetry properties by folding back wavelength signals onto the L output lines. The final foldback design without any crossovers is found in Fig. 5e with $w = 0$, $c = 1$, and $q = 4$ for a 2-line 3-wavelength OXC. However, there are certain input ports that are not used to construct an OXC. This is the only drawback in this design.

3.2. Active Optical Cross-Connects

Space-switching architectural designs have been created since the invention of the telephone. Currently, space-switching architectures are quite mature. Unfortunately, these space-switching designs cannot be applicable to the optical domain because of unavailable optical logic and storage devices. These architectures are still useful mostly in constructing large-scale electronic switches. They probably can provide only a limited number of stages on switch expansions because of the loss issues in the optical domain. In the following, we outline the designs of several optical cross-connects that are based on three of the latest popular optical device technologies. The designs of these OXCs are still closely related to the device technology. They are built with the liquid crystal, thermobubble/thermocapillary and MEMS technologies.

3.2.1. Liquid Crystal Switches. Despite the name given, liquid crystals [18–23] are not truly liquid. They exist in a state between liquid and solid called *mesophasic*. A liquid crystal molecule has an elongated shape, and is often represented as a rod. Under the proper conditions, the orientation of these molecules can be changed so that they face in a certain direction. The orientation can affect the optical properties of the liquid crystal, which in turn affects the polarization of the light passing through it.

There are several types of liquid crystals, including nematic, discotic, cholesteric, and various kinds of smectic phases, which can be characterized by different arrangements of the molecules. Utilizing a magnetic or electric field can often change the optical properties of a liquid crystal. Liquid crystal switches have low insertion loss and excellent performance at the same time. A major advantage with the liquid crystal is its ability to add and drop different colors of light without having to demultiplex all wavelengths of the incoming signal.

In constructing basic liquid crystal switching modules, an early result was reported [20]. The 1×2 splitter is shown in Fig. 6a. The polarization beamsplitter (PBS) is used for lowering crosstalk. It divides the input light into two linearly polarized lightbeams. The twisted nematic liquid crystal (TN-LC) provides polarization switching. It provides 90° polarization rotation without an applied voltage, but it keeps the original polarization with an applied voltage. The function of the birefringent crystal block (BRB) is for extraordinary wave walkoff. The switching operations from one input to one of the two output ports are shown in Fig. 6b,c.

In the following, an architectural design of multichannel liquid crystal switches is described. There are three different basic liquid crystal components that can be used to construct a multistage interconnection of optical switches. These components are a 2×2 polarization switch, an optical beam router, and a beam shifter.

The 2×2 polarization switch is shown in Fig. 7a,b. It is constructed with the transmission-type twisted nematic liquid crystal spatial light modulator (LC-SLM) arrays. A thin nematic liquid crystal layer at the center is surrounded by two glass plates that are bonded to transparent electrodes, such as indium tin oxide. When it is in OFF state, it rotates the light with a 90° polarization angle. When it is ON, then no polarization state will be changed. Therefore, the 2×2 polarization switch operates like a switch between two orthogonal polarizations. For the second component, the optical beam router, its operation model is shown is Fig. 7c. Its goal is to exchange one of the polarization components with a lightbeam that propagates along an adjacent path. The PBS operates differently on the two polarization orientations, the polarizations that are parallel and perpendicular to the surface, the P component and the S component. As shown in Fig. 7d, a basic beam router is composed of five PBSs. Each of these PBSs reflects only the S component of a projecting lightbeam. With the arrangement of these PBSs, the two adjacent S components will be exchanged on leaving the beam router. However, this architecture requires high accuracy in assembly in order to stabilize the coupling loss. The third component is a beamshifter. Its functional model and hardware construction are shown in Fig. 7e,f, respectively. A basic beamshifter consists of three PBSs. Similar to the optical beam router, each PBS reflects only the S component of the lightbeam. It uniformly displaces one of the polarization components, and takes in another S component from its adjacent beam path. After designing all these basic components, we can move forward to construct

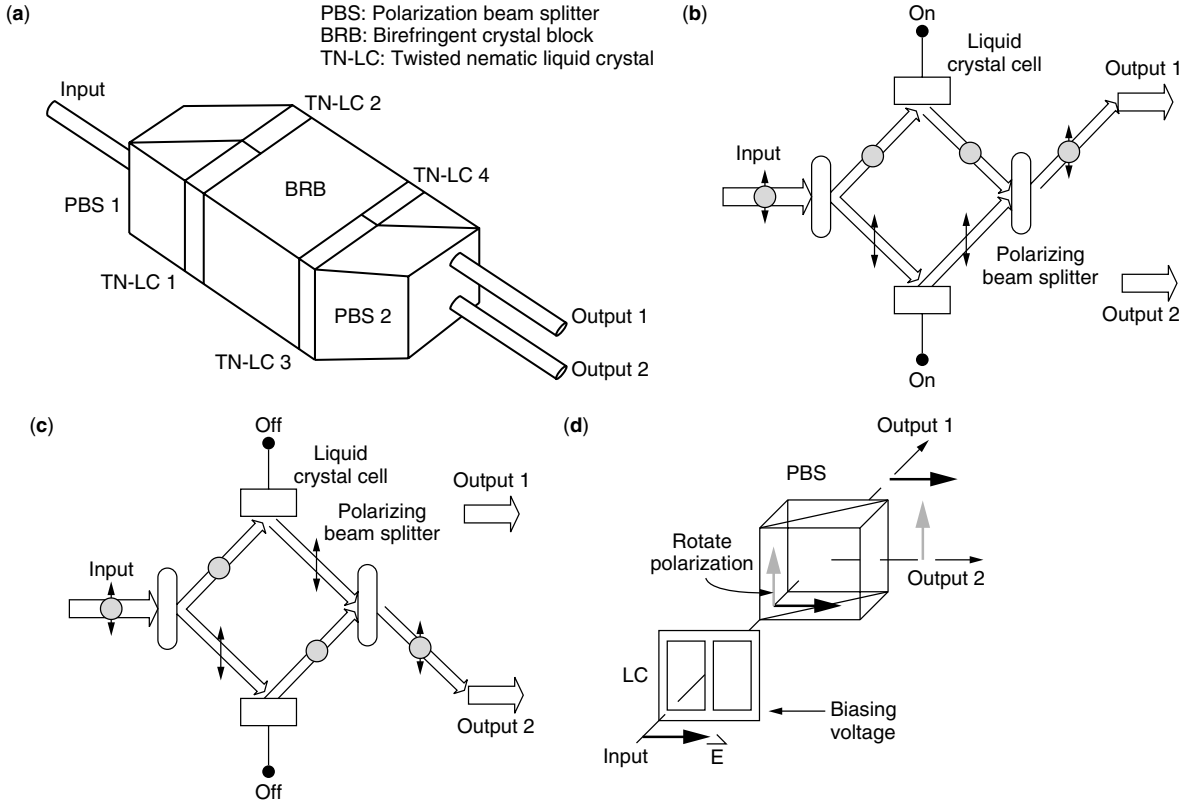


Figure 6. Liquid crystal switch and its operations.

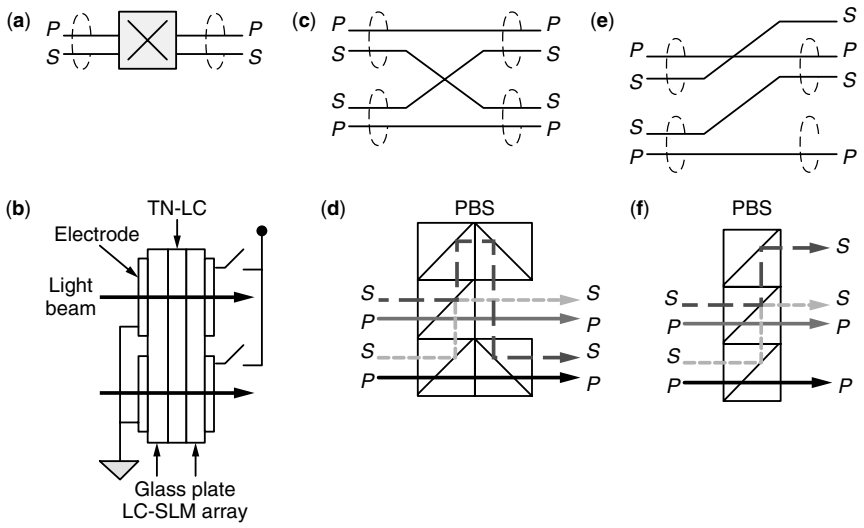


Figure 7. (**a**,**b**) 2×2 Polarization switch; (**c**,**d**) optical beam router; (**e**,**f**) beamshifter.

optical switch systems. Two optical switch architectures are as shown in Fig. 8. They demonstrate the feasibility of constructing multistage optical interconnection networks with the liquid crystal technology. Currently, the size of an optical liquid crystal switch will be limited because of the physical construction, alignment, and the signal loss issues.

In a relatively earlier design, the switching mechanism with the liquid crystal switch was based on the total internal reflection of the liquid crystal. A 2×2 switch consists of two glass prisms of equal refractive index and base angle [18]. The inside face of both prisms have been coated with a transparent electrode and thin polyamide layer. The prisms are bonded together using an epoxy edge seal loaded with spacers of the desired diameter in the range of few micrometers. Liquid crystal with OFF-state alignment is introduced between the prisms by vacuum filling, and the cell is sealed. The OFF-state alignment of the liquid crystal reflects light from the input port 1 back to the output port 1. The ON state is switched by applying a voltage to the electrode. It changes the alignment of the liquid crystal normal to the face of the prisms so that

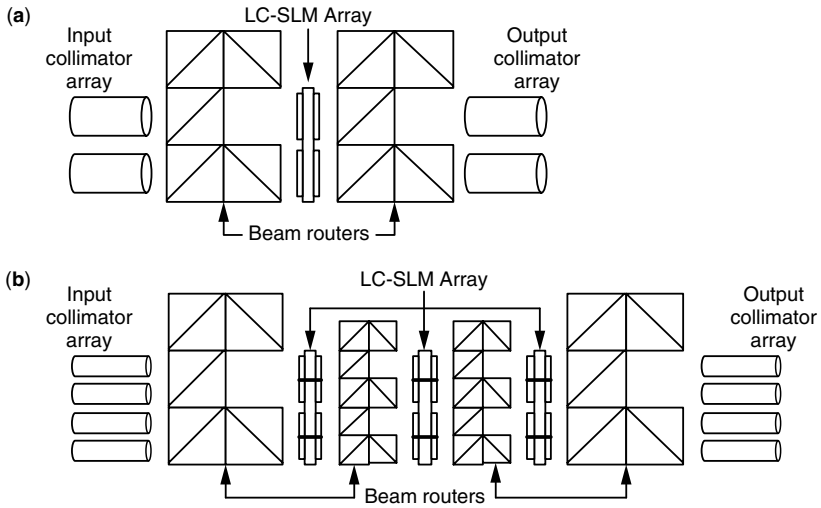


Figure 8. (**a**) 4×4 and (**b**) 8×8 optical switches.

light will be able to pass from input port 1 to output port 2 through the liquid crystal. However, this architecture is not easily extensible to construct large-sized optical switches. A design of 8×8 channels crossbar switch has been proposed [19]. The switch uses an array of eight 1×8 liquid crystal switches for signal divergence and eight 8×1 liquid crystal switches for signal convergence.

The switch performance of the 2×2 liquid crystal switch [18] has an average transmit state crosstalk of 33 dB, and the switching speed is <5 ms. The insertion loss obtained for the liquid crystal itself is 0.5 dB, excluding the Fresnel reflection at the glass–air interface and the 3 dB polarization loss [18]. For the 4×4 switch, it experiences an average crosstalk level of -22 dB and the insertion loss is 0.8 dB. The switching time should be longer because it is required to set up more electrodes along the switching path. The 8×8 liquid crystal crossbar switch described by Noguchi [19] has an average crosstalk level of -59.5 dB and an insertion loss of 3.44 dB. The improvement of the crosstalk level is due to the isolation of all the optical signal paths within the switch by a set of 1×8 and 8×1 liquid crystal array. The increase in insertion loss is due to the additional stage required within the switch body.

3.2.2. Bubble and Thermocapillary Switches. A bubble switch developed by Agilent is based on low-cost thermal inkjet bubble technology. The earliest report on bubble switching was provided by Jackel et al. [24]. The design concept of the basic bubble-switching component is simple. There are two core paths for light transmission. These two paths are crossing each other, and at the crosspoint, there is a trench holding refractive index-matching fluid. There is also a heater that makes the liquid boil, and forms a bubble. In the switch, this bubble is critical to reflecting the light onto a new path. When it forms, the bubble displaces the fluid from the trench, which makes the space more like air. It creates an interface between the glass and the bubble that shifts the light with the total internal reflection principle [25]. Even though there are no moving parts in the systems, the switching time for the device with the software control is around 10 ms as reported by Fouquet. The commercialized 32×32 two-dimensional crossbar switch from Agilent has a specification of -50 dB crosstalk.

Bubble switches are usually made on glass waveguides. The glass is etched to produce capillary channels that contain index-matching fluid and air bubbles. Usually, rough capillary walls are formed by reactive-ion etching or acid etching of glass, and they produce high scattering loss when the air bubble is located at the waveguide intersection. Typical losses of 2.2 dB have been reported [25] for glass waveguide devices. Switching in a bubble switch is shown in Fig. 9.

Bubble switch operations rely on the heating processes that create air bubbles in the index fluid. In order to keep the bubbles at their positions, continuous power has to be applied. On the other hand, there is a thermocapillary process being developed [26–28]. As reported by Sato et al., the structure of the thermocapillary switch is as shown in Fig. 10. A deep trench is formed at each cross

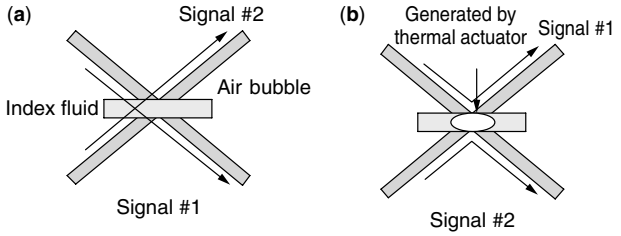


Figure 9. Optical bubble switch operations: (**a**) bar state; (**b**) cross state.

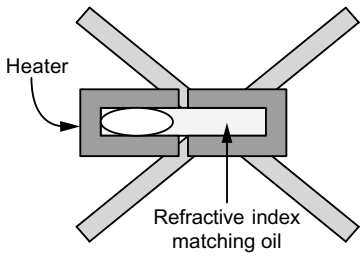


Figure 10. Thermocapillary switching module.

point of the waveguide cores. Refractive-index-matching oil is injected into the trench, which is sealed by a glass lid. The trench is half-filled with the oil; the other half, with a bubble. On top of the trench, there is a pair of microheaters to produce a thermal gradient along the trench. When the oil is heated, the surface tension of the bubble on the heater side is decreased, and then the bubble moves to the side of the actuating heater with a capillary force. This actuation mechanism with the thermal gradient is called thermocapillary. When the bubble is located at the cross point of the waveguides, the light is switched into the crossing waveguide because of the total internal reflection on the glass–air interface. This operation is identical to the operating principle in the bubble switch. However, the mechanism of the bubble action is completely different. For example, the bubbles in Agilent's switch are created through heating, while in the thermocapillary switch, the bubbles always exist and their motions are activated through the capillary force. On the performance side, this thermocapillary switch was reported to have an insertion loss of ~4 dB [27]. However, the switching time takes 50 ms to move the bubble for this specific design. The advantage of this design is that the bubble will latch on the wall and no extra power is required to keep it stationary at that position.

Further improvement can be made on both thermobubble and thermocapillary switches. In order to improve the performance, there are polymer-based thermocapillary switches. In this design, the fluid and air capillary is formed by precision laser ablation yielding a much smoother capillary wall. Optical surfaces achieved via laser ablation result in a polymer waveguide. This design has excellent insertion loss. The loss in the air bubble is less than 0.2 dB. When index fluid fills the capillary at the waveguide intersection, a loss of 0.1 dB is typical for polymer waveguides and glass waveguides.

As of today, there are several firms working on producing optical switches based on bubble technology. To date, only two-dimensional crossbar switches have been constructed with bubble technology.

3.2.3. Microelectromechanical Systems. We have already discussed the constructions of optical switches using the liquid crystal and bubble-switching technologies. Both of them show promising research results and they have the potential to be deployed for commercial use in the near future. However, as of today, the only OXC designs available on the market are made with the microelectromechanical systems (MEMS) technology [29–32], such as Lucent's LambdaRouter [33]. MEMS optical switches are different from the conventional mechanical switches, which are based on macroscopic bulk optics and utilize the advantages of free-space optics. These conventional mechanical switches suffer from large size and mass with slow switching time.

With the introduction of MEMS technology, MEMS optical switches not only retain their conventional advantages of free-space optics such as low losses and low crosstalk but also include additional advantages such as small size, small mass, and submillisecond switching times. Furthermore, MEMS fabrication techniques allow integration of microoptics, microactuators, complex micromechanical structures, and possibly microelectronics on the same substrate to realize integrated microsystems.

For the MEMS devices, their operating units are the micromirrors. The size of these mirrors may be as small as several hundreds of micrometers, and they are made with standard IC fabrication technology [32–38]. Since existing fabrication technology is quite mature, the cost is low for constructing MEMS devices. At the moment, there is a common fabrication process that is well accepted and is known as MUMPs (multiuser MEMS processes) from Cronos[8] for MEMS. The process is composed of both bulk and surface micromachining. Bulk micromachining, such as deep-silicon reactive-ion etching (DRIE), helps set up the overall outlook of a silicon system structure. The surface micromachining and LIGA[9] processes create the details of the final operating structure of a device. These MEMS optical switches consist of many moving mirrors. Therefore, there should be microactuators to drive or oscillate these moving parts in the MEMS device. For example, there may be moving micromirrors to switch laser beams from one fiber to another. There are currently a variety of methods to achieve these microactuation functions, for example, electrostatic, electromagnetic, piezoelectric, magnetostrictive, and thermal expansion. Currently, the electrostatic mechanism is the most common and best-developed method.

For the electrostatic actuators, several types of designs are suitable for constructing OXCs: parallel-plate capacitors, comb drives, and torsional bars. We roughly review the designs of the parallel-plate and comb drive mechanisms. The structure of a parallel-plate MEMS device is shown in Fig. 11a. In general, for all parallel-plate structures, the device stores some capacitance energy, that is, $W = CV^2/2$. When the plates move toward each other, the work done by the attractive force between them can be computed as a change in W with a displacement, x. Therefore, the force can be computed as $F = V^2(\partial C/\partial x)/2$. In the parallel-plate capacitor architecture, only attractive forces can be generated. The design will be more attractive if a larger force can be produced to carry out the heavier workload. It is desirable to offer a larger change in capacitance with respect to distance. This leads to the development of the electrostatic comb drives as shown in Fig. 11b. The comb drives consist of many interleaving fingers. When a voltage is applied, an attractive force is developed between fingers and they move toward each other. With this structural design, there is an increase in capacitance that is proportional to the number of fingers. Therefore, the larger number of fingers can generate larger forces. A potential problem is to carefully control the lateral gaps between fingers. A finger may swing and stick if the gaps are not identical on both sides.

As many reports [29–36] indicate, MEMS optical switches are able to demonstrate their superiority in

[8] Cronos is a division in JDS Uniphase.

[9] LIGA is a German acronym for lithography, electroplating, and molding.

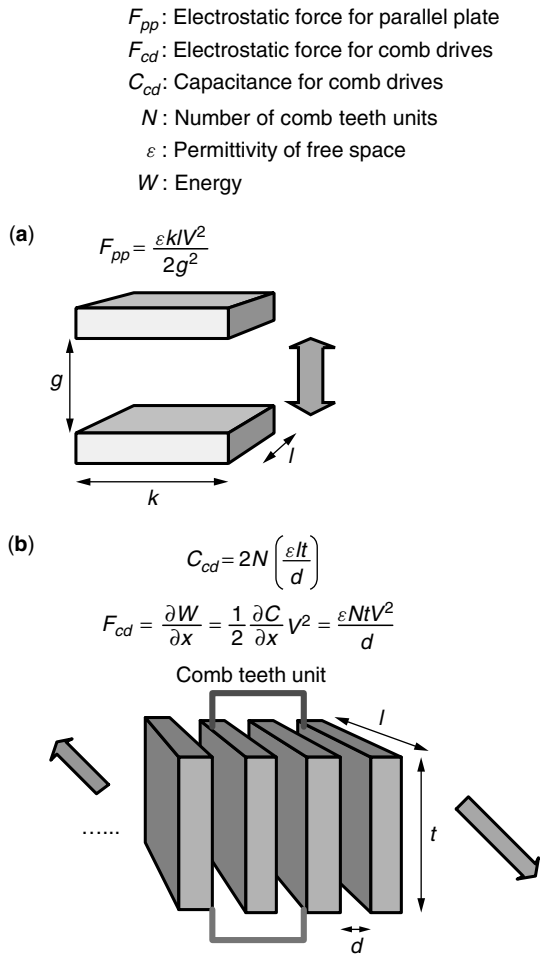


Figure 11. Electrostatic designs: (**a**) parallel-plate capacitor, (**b**) comb drives.

the areas of scalability, insertion loss, polarization-dependent loss (PDL), wavelength dependency, small size, low cost, crosstalk, switching speed, manufacturability, serviceability, and long-term reliability. Among these reports on system performance, the switching time is ~4 ms [33] with ~3 dB insertion loss. In general, we expect that the switching time may fall within the range 1–10 ms, and the insertion loss is between 1 and 6 dB. With steady advancement of the latest fabrication technology, the switching time will be performing even better in the future. For example, the switching time is less than 1 ms when scratch drive actuators are used [35].

In the following, we describe how MEMS technology helps us to construct optical switches. There are currently two broad approaches to implement MEMS optical switches: 2D and 3D MEMS optical switches. Even though both 2D and 3D MEMS optical switches operate on micromirrors in crossbarlike architectures, there are striking differences in terms of how the mirrors are controlled and their ability to redirect lightbeams. However, both of them have shown promise in finding their niche in telecommunication networks. There are already several large 2D MEMS optical switches in the market. Lucent/Agere's WaveStar LambdaRouter is the most sophisticated and the largest 3D MEMS optical switch available in the marketplace today. The delivery of the 1024 × 1024 WaveStar LambdaRouter is expected by the year 2002.

3.2.3.1. 2D MEMS Optical Switches. With the given MEMS technology, many two-dimensional crossbar optical switches were made. In this planar architecture, mirrors are always arranged in a crossbar configuration as shown in Fig. 12a. Each mirror has only two positions and is placed at the intersections of lightpaths between the input and output ports. They can be in either in the ON position to reflect light or in the OFF position to let light pass uninterrupted. The binary nature of the mirror positions greatly simplifies the control scheme. Typically, the control circuitry consists of simple transistor–transistor logic (TTL) gates and appropriate amplifiers to provide adequate voltage levels to actuate the mirrors. For an $N \times N$ switch, a total of N^2 mirrors are required to implement a strictly nonblocking optical crossbar switch fabric. For example, a 16×16-port switch will require 256 mirrors. Moreover, the capability of signal resynchronization within an optical switch is not possible,

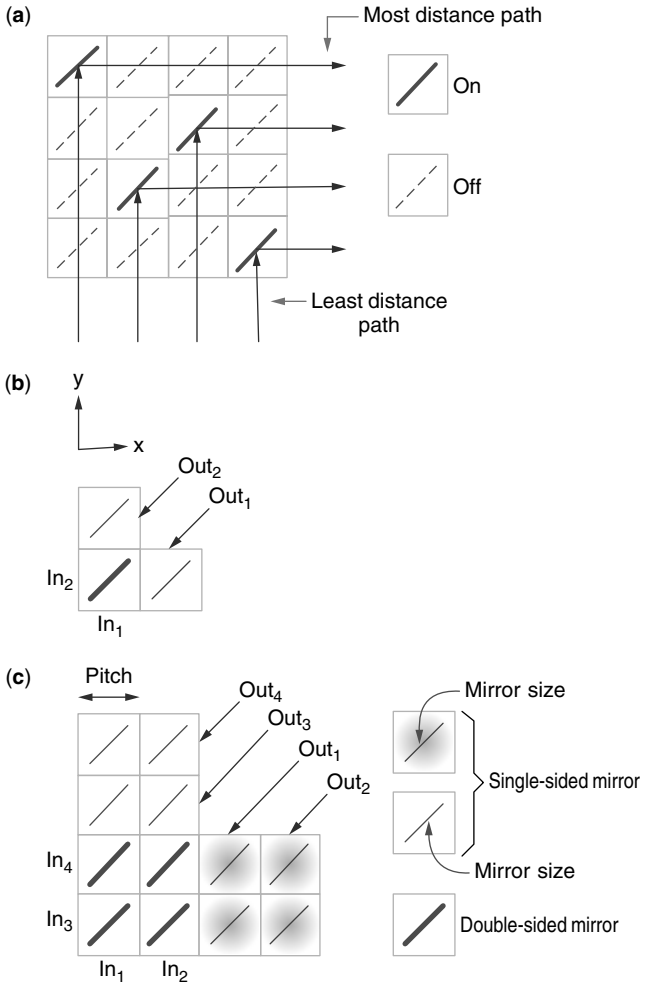


Figure 12. (**a**) Crossbar switch; (**b**) 2×2 and (**c**) 4×4 L-switching matrices.

and the free-space beam propagation distances among port-to-port switching are not constant. As a result, the insertion loss due to Guassian beam propagation is not uniform for all ports. Consequently, there are variations in losses among ports. The minimum and maximum insertion losses of AT&T's 8×8 switch with scratch drive actuators [36] are 3.1 and 3.5 dB, respectively. A simple way to improve the system performance of a 2D crossbar MEMS switch is to decrease the pitch size per mirror unit. It can then reduce the pathlength differences and signal loss as well as increase the port count in the 2D MEMS switch. Certainly, this also leads to smaller mirror size, which can cause signal loss due to the spreading of the Guassian beam. Therefore, more sophisticated and accurate fabrication may be needed to fabricate these systems.

An alternative approach to increasing port count is to interconnect smaller 2D MEMS switching modules to form multistage networks, for example, the three-stage Clos networks. However, this cascaded architecture typically requires up to thousands of complex interconnects between stages, thus decreasing serviceability of the overall switching system. Up to the current stage, extensive research is being performed on the device level. Yeow et al. [39] investigated double-sided mirror design to see if it can provide benefits in existing designs. A planar L-switching matrix design [39] was proposed. Figure 12b,c shows 2×2 and 4×4 L-switching matrices, respectively. The longest-distance path, l_{ldp}, in L-switching matrix is always 25% shorter than that of the regular two-dimensional crossbar switch; whereas the shortest-distance path in the L-switching matrix, l_{sdp}, is always about one-third of the l_{ldp}. As a result, when it is compared to the regular crossbar switch, the maximum path difference in the L-switching matrix grows slowly with the number of input or output ports. It helps slow down the impact of the loss nonuniformity issue [37,38] in 2D MEMS switches due to the pathlength difference problem. The current achievable port count of a 2D MEMS crossbar switch is 32×32, whereas we expect that the L-switching matrix should be able to scale to 64×64 without installing collimators with varying focal lengths for the system. Moreover, these L-switching modules can be used to construct larger-sized Clos networks. Comparison of the construction of three-stage Clos networks with that of the regular 2D crossbar MEMS switches reveals that the one with the L-switching matrix modules has substantially reduced accumulated insertion loss by almost 57% with the pathlength difference issue when only the pathlengths within the switches are counted [39]. However, there are shortcomings in the L-switching matrix; for instance, it may not be possible to establish a new connection without modifying existing connection configurations. Fortunately, the number of paths between an input and an output may have multiple possible paths. If the number of ports is N, the number of possible paths can be $N/2$ for some cases. As an example, you can find two setups for one set of connection requests in Fig. 13.

3.2.3.2. 3D MEMS Optical Switches. All lightbeams in a 2D MEMS switch reside on the same plane.

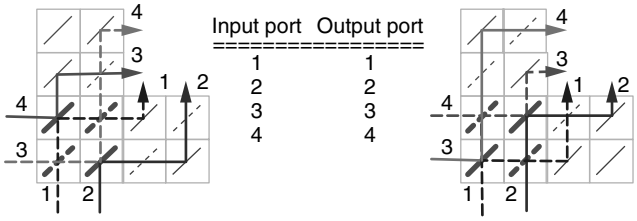


Figure 13. Two path configurations for one set of connection requests in a 4×4 L-switching module.

This arrangement usually results in unacceptable high and uneven loss for large port counts. The 3D MEMS switch [29,33] makes use of the three-dimensional space as an interconnection region that allows scaling far beyond 32 ports with acceptable optical losses. These analog or 3D MEMS switches have mirrors that can rotate freely on two axes as shown in Fig. 14, and light can be redirected precisely in space to multiple angles. The port count would be limited only by insertion loss that results from finite acceptance angle of fibers or lens. Another advantage is that the differences in free-space propagation distances among port-to-port switching are much less dependent on the scaling of the port count. Typically, the optical pathlength scales only as $\sqrt{N}$ instead of N, so port counts of several thousands are achievable with high uniformity in losses ($<$10 dB). Inevitably, much more complex switch design and continuous analog control are needed to improve stability and repeatability of the mirror angles.

To design 3D MEMS optical switches, N or $2N$ mirrors may be required. For example, Nortel Networks' 3D switching architecture [32,40] utilizes two sets of N mirrors. The first plane of N mirrors redirect light from N input fibers to the second plane of N mirrors. All mirrors on the second plane are addressable by each mirror on the first plane making nonblocking connections. In turn, mirrors on the second plane can each be actively and precisely controlled to redirect light into desired output fibers with minimum insertion loss. On the other hand, Lucent's WaveStar MEMS switches, shown in Fig. 14,

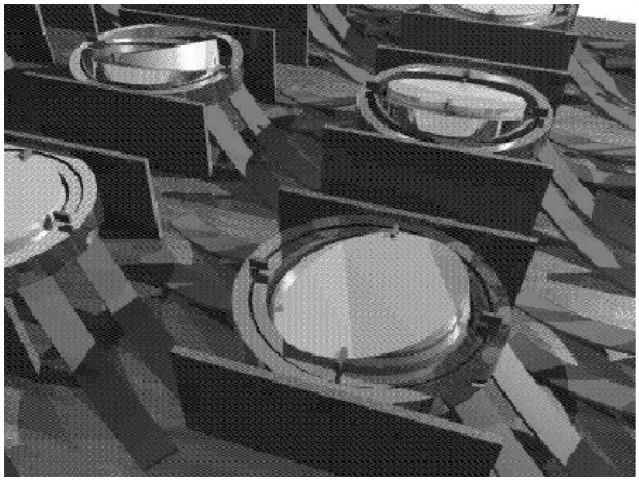


Figure 14. 3D MEMS mirrors.

use only N mirrors and a comparatively large and fixed reflective mirror tilted toward the mirrors. Light from an incoming fiber arrives at one 3D MEMS mirror and is reflected to the large reflective mirror. Then the light will return to another MEMS mirror on the same MEMS plane and be sent to an outgoing fiber. Typically, the mirror can rotate on two axes and is continuously controllable to tilt by at most $\pm 10^{\circ}$ [30]. Moreover, the reported switching module has a maximum insertion loss of 6 dB and a switching time of <10 ms. This 3D optical architecture clearly presents real hope for developing a scalable large-port-count OXC. A WaveStar LambdaRouter with more than one thousand ports is expected to be available in 2002.

4. CONCLUDING REMARKS

Optical switches are the most important components in the future all-optical networks. Numerous companies are producing the next-generation optical switches. Since the late 1990s, the properties of AWGs that allow predefined path setup in the optical domain have been more clearly understood. However, AWGs can provide a platform only for constructing passive OXCs. Therefore, the latest exciting research is on designing active OXCs. At the moment, multiple technologies have been showing promising results. Among them, 3D MEMS is the first one to show exciting and promising results for designing large-scale OXCs. Numerous issues still need to be investigated to further improve the performance of 3D MEMS switches, including the fabrication process, packaging, deposition uniformity on small mirror surfaces, and analog tilting open-loop and closed-loop control [30]. Until now, the research has focused on device-level technology. Novel architecture design on optical switches can also be investigated from the system level with the learnt hardware properties, for example, the L-switching matrix design [39] by integrating different component structures in one system design.

BIOGRAPHY

K. L. Eddie Law received the B.Sc.(Eng.) degree in electrical and electronic engineering from the University of Hong Kong, the M.S. degree in electrical engineering from Polytechnic University, Brooklyn, New York (USA), and the Ph.D. degree in electrical and computer engineering from the University of Toronto in Canada in 1995. From 1995 to 1999, he joined Nortel Networks in Ottawa, Canada, and worked in three different groups: Passport Research, Next Generation ATM Systems, and Computing Technology Lab. Since September 1999, he has been an Assistant Professor in the Communications Group in the Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto. His current research interests are on the active networks, policy-based management on the Internet, TCP/IP protocol design and development, reconfigurable network design, and photonic switch design.

BIBLIOGRAPHY

1. P. P. Mitra and J. B. Stark, Nonlinear limits to the information capacity of optical fibre communications, *Nature* **411**: 1027–1030 (June 28, 2001).
2. N. A. Jackson, S. H. Patel, B. P. Mikkelsen, and S. K. Korotky, Optical cross connects for optical networking, *Bell Labs Tech. J.* 262–281 (Jan.–March 1999).
3. C. Clos, A study of non-blocking switching network, *Bell Syst. Tech. J.* **32**: 406–424 (March 1953).
4. R. Ramaswami and K. N. Sivarajan, *Optical Networks: A Practical Perspective*, Morgan Kaufmann, San Francisco, 1998.
5. T. E. Stern and K. Bala, *Multiwavelength Optical Networks: A Layered Approach*, Addison-Wesley, Reading, MA, 1999.
6. A. Rogers, *Understanding Optical Fiber Communications*, Artech House, Boston, 2001.
7. J. E. Ford, D. J. DiGiovanni, and D. J. Reiley, $1 \times N$ Fiber Bundle, *Proc. Optical Fiber Commun. Conf.'98*, Feb. 1998, pp. 143–144.
8. M. K. Smit and C. van Dam, PHASAR-based WDM-devices: Principles, design and applications, *IEEE J. Select. Top. Quant. Electron.* **2**(2): 236–250 (1996).
9. Y. P. Li and C. H. Henry, Silica-based optical integrated circuits, *IEE Proc. Optoelectron.* **143**(5): 263–280 (Oct. 1996).
10. P.-J. Wan, *Multichannel Optical Networks*, Kluwer, 2002.
11. S. Morasca, D. Scarano, and S. Schmid, Application of $LiNbO_3$ acousto optic tunable switches and filters in WDM transmission at high bit rates, in G. Prati, ed., *Photonic Networks*, Springer-Verlag London Ltd., 1997, pp. 458–472.
12. A. Himeno, T. Kominato, M. Kawachi, and K. Okamoto, System applications of large-scale optical switch matrices using silica-based planar lightwave circuits, in G. Prati, ed., *Photonic Networks*, Springer-Verlag London Ltd., 1997, pp. 172–182.
13. T. Chikama, H. Onaka, and S. Kuroyanagi, Photonic networking using optical add drop multiplexers and optical cross-connects, *Fujitsu Sci. Tech. J.* **35**: 46–55 (July 1999).
14. N. Keil, H. Yao, C. Zawadzki, and B. Strebel, 4×4 polymer thermo-optic directional coupler switch at 1.55 μm, *Electron. Lett.* **30**(8): (April 1994).
15. N. Keil, H. Yao, C. Zawadzki, and B. Strebel, Rearrangeable nonblocking polymer waveguide thermo-optic 4×4 switching matrix with low power consumption at 1.55 μm, *Electron. Lett.* **31**(5): (March 1995).
16. C. Fernando et al., Thermo-optical switching in Si/Si1-xGex distributed Bragg reflectors, *Electron. Lett.* **30**(11): (May 1994).
17. T. Goh et al., Low loss and high extinction ratio 16×16 thermo-optic matrix switch using silica-based planar lightwave circuits, *Proc. Asia Pacific Conf. Communication'97*, Dec. 1997.
18. J.-C. Chiao, Liquid-crystal optical switches, *Proc. 2001 Optical Society of America Topical Meetings: Photonics in Switching*, June 11–15, 2001.
19. K. Noguchi, Transparent optical crossbar switch using liquid crystal optical light modulator arrays, *Integrated Opt. Opt. Fibre Commun.* (Sept. 1997).

20. Y. Fujii, Low-crosstalk 1×2 optical switch composed of twisted nematic liquid crystal cells, *IEEE Photon. Technol. Lett.* **5**(2): 206–208 (Feb. 1993).

21. A. Sneh and K. M. Johnson, High-speed continuously tunable liquid crystal filter for WDM networks, *J. Lightwave Technol.* (1996).

22. N. A. Riza and S. Yuan, Low optical interchannel crosstalk, fast switching speed, polarisation independent 2×2 fibre optic switch using ferroelectric liquid crystals, *Electron. Lett.* (June 25, 1998).

23. C. Mao et al., Liquid-crystal optical switches and signal processors, *Proc. 2001 Asia-Pacific Optical and Wireless Communications Conf.*, Nov. 2001.

24. J. L. Jackel and W. J. Tomlinson, Bistable optical switching using electrochemically generated bubbles, *Opt. Lett.* **15**(24): 1470 (1990).

25. J. E. Fouquet, Compact optical cross-connect switch based on total internal reflection in a fluid-containing planar lightwave circuit, *Proc. Optical Fiber Communications Conf. 2000*, 2000.

26. M. Makihara, M. Sato, F. Shimokawa, and Y. Nishida, Micromechanical optical switches based on thermocapillary integrated in waveguide substrate, *J. Lightwave Technol.* **17**: 14–18 (1999).

27. M. Sato et al., Thermo-capillary optical switch, *Hitachi Cable Rev.* (20): (Aug. 2001).

28. J. T. Gallo, B. L. Booth, C. A. Schuetz, and R. J. Furmanak, *Polymer Waveguide Components for Switched WDM Cross-Connects*, Optical CrossLinks, Inc. (online), *http://www.opticalcrosslinks.com*.

29. D. J. Bishop, C. R. Giles, and G. P. Austin, The Lucent LambdaRouter: MEMS technology of the future here today, *IEEE Commun. Mag.* **40**(3): 75–79 (March 2002).

30. P. B. Chu, S.-S. Lee, and S. Park, MEMS: The path to large optical crossconnects, *IEEE Commun. Mag.* **40**(3): 80–87 (March 2002).

31. P. De Dobbelaere et al., Digital MEMS for optical switching, *IEEE Commun. Mag.* **40**(3): 88–95 (March 2002).

32. T.-W. Yeow, K. L. E. Law, and A. Goldenberg, MEMS optical switches, *IEEE Commun. Mag.* **39**(11): 158–163 (Nov. 2001).

33. D. T. Neilson et al., Fully provisioned 112×112 micromechanical optical crossconnect with 35.8Tb/s demonstrated capacity, *Proc. Tech. Digest Optical Fiber Communications Conf.* (*OFC2000*), March 7–10, 2000 pp. 202–204.

34. R. Giles et al., Silicon micromachines in optical communications networks: Tiny machines for large system, in R. Rai-Choudhury, *MEMS and MOEMS Technology and Applications*, SPIE Press 2000, Chap. 6, pp. 301–329.

35. L. Y. Lin, E. Goldstein, and L. M. Lunardi, Integrated signal monitoring and connection verification in MEMS optical crossconnects, *IEEE Photon. Technol. Lett.* **12**(7): (July 2000).

36. L. Y. Lin, E. L. Goldstein, J. M. Simmons, and R. W. Tkach, High-density micromachined polygon optical crossconnects exploiting network connection symmetry, *IEEE Photonics Technol. Lett.* **10**: 1425–1427 (1998).

37. K. S. J. Pister, M. Judy, S. Burgett, and S. Fearing, Microfabricated hinges, *Sensors and Actuators* **33**(3): 249–256 (1992).

38. T. Akiyama and H. Fujita, A quantitative analysis of scratch drive actuator using buckling motion, *Proc. IEEE Workshop MEMS*, The Netherlands, Jan. 29–Feb. 2, 1995.

39. T.-W. Yeow, K. L. E. Law, and A. Goldenberg, Micromachined *L*-switching matrix, *Proc. IEEE ICC* (in press).

40. A. Neukermans and R. Ramaswami, MEMS technology for optical networking applications, *IEEE Commun. Mag.* **39**(1): 62–69 (Jan. 2001).

OPTICAL SWITCHING TECHNIQUES IN WDM OPTICAL NETWORKS

MYUNGSIK YOO
Soongsil University
Seoul, Korea

CHUNMING QIAO
SUNY at Buffalo
Buffalo, New York

1. INTRODUCTION

The major advances in optical technology have led to the development of optical communication systems and networks, beginning with long-haul communication systems to metropolitan-area networks, even to access networks, which are the final leg in communication systems, in the form of fiber to the curb (FTTC), fiber to the building (FTTB), and fiber to the home (FTTH).

There are a few reasons why *optical networks* are considered as a solution for transmission infrastructure:

1. Optical networks can provide vast bandwidth with low attenuation. Typically, optical systems use wavelengths in three ranges: previously 800–900 nm and 1280–1350 nm, and currently 1510–1600 nm. As a result of low attenuation, optical systems require fewer repeaters or amplifiers. The potential bandwidth of optical signals (wavelengths) is huge since the typical frequency is in the few hundred terahertz (10^{12}). This means that the data rate of optical systems can be much higher than that of communication systems using frequencies in the megahertz (10^6) or gigahertz (10^9) range. Currently, a single wavelength can operate at 10 Gbps (gigabits per second), and even at 40 Gbps. Furthermore, using dense wavelength division-multiplexing (DWDM) technology, it is possible to put many wavelengths into a single fiber. It is now possible to multiplex 80–100 wavelengths, creating Tbps capacity per fiber.
2. Optical networks provide a *transparency* to protocol, data format, and data rate. Thus, once a communication pipe is established between two client points, optical networks are easily able to accommodate any existing network protocols such as IP, ATM, and SONET/SDH. In addition, an optical network can carry data regardless of its format (e.g., analog or digital) and its data rate.
3. Because of WDM technology and transparency, optical networks can provide a cost-effective and futureproof way to building the transport network. As traffic demand grows, it is easy to upgrade

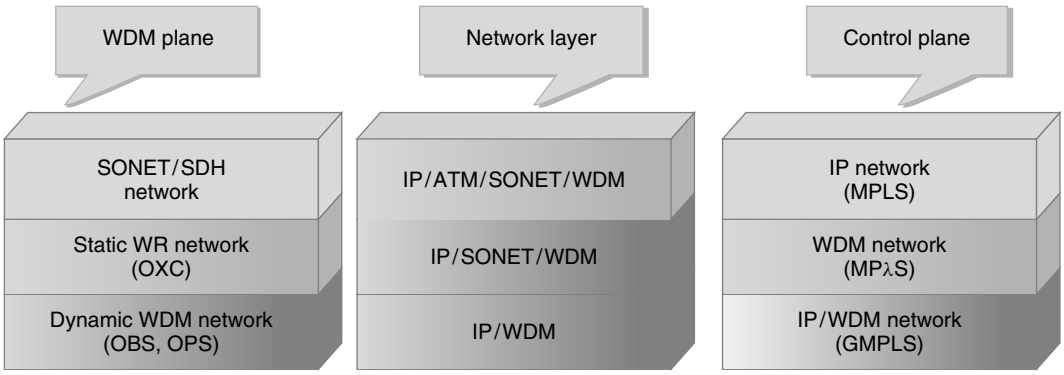


Figure 1. Evolution of optical networks.

network capacity using WDM technology. It is also easy to interface any existing or newly emerging network technology with optical networks due to service transparency.

4. Finally, optical networks become a more feasible solution as the internetworking technologies such as generalized multiprotocol label switching (GMPLS) [1] are actively developed and standardized.

Since it has been deployed in real systems (e.g., carrier systems), optical networking technology keeps evolving to seek better networking solutions in both cost and performance. Figure 1 shows the direction that optical networks have been evolving. We categorize the optical networking technology into three areas: WDM plane, network layer and control plane.

In the *WDM plane*, synchronous optical network (SONET)/synchronous digital hierarchy (SDH) systems were introduced as a first-generation (1G) optical network [2,3]. *SONET/SDH* are designed for carrying voice traffic over the optical fibers with very high transmission capacity. Currently, many carriers have buried many fibers and built SONET/SDH networks for transporting their voice traffic. Although SONET/SDH can be categorized as an optical network, it lacks the optical networking technology such as routing and switching. Thus, it can be said that SONET/SDH only takes advantage of the huge bandwidth of an optical transmission system. In order to enhance the networking capability in the optical network (or layer), *wavelength routing* (*WR*) *networks* [2,3] has been introduced. WR networks placed *optical cross-connect (OXC)* at the switching node, which switches individual wavelengths. In addition, WR networks perform a routing function in the optical domain, owing to the *control plane* such as MPλS (optical-domain MPLS) or GMPLS. On receiving a request from the clients (IP, ATM, SONET/SDH), WR networks set up an end-to-end *lightpath*, which is the process of assigning a wavelength to a particular path by routing and wavelength switching. Simply, WR networks provide a lightpath service to its client layers.

Depending on how frequently a network changes its virtual topology (or *lightpath topology*), an optical network can be classified either a static or dynamic. In a *static WDM optical network*, once established, a lightpath exists for a long period of time (e.g., years or months). This is the type of most optical networks deployed today. The drawback of a static network is that it results in inefficiency when traffic demand changes frequently. A *dynamic WDM optical network* can efficiently support bursty traffic by changing its virtual topology according to traffic demand. Thus, the lightpath in a dynamic WDM network reconfigures itself in much faster time scale. There are two optical switching technologies for dynamic WDM networks under active research and development (R&D), specifically, optical burst switching (OBS) [4–7] and optical packet switching (OPS) [8–10]. Both aim to switch optical packets (or optical bursts) in the optical domain as a conventional packet switching network does in the electronic domain.

In the network layer (layer in the OSI Reference Model), the evolution of optical networks is closely related to how to efficiently support IP traffic. Considering the unprecedented increase in Internet traffic since the mid-1990s, the network architecture should be optimized for the data traffic. Initially, the architecture includes ATM (asynchronous transfer mode) to carry IP packets due to its high-speed switching capability and QoS (quality-of-service) support. However, IP routers are improving their performance in capacity and forwarding speed, exceeding ATM's capability with the help of MPLS (multiprotocol label switching) technology. Thus, the architecture is simplified to IP over SONET over WDM, eliminating the ATM layer. In the next step, the IP layer is directly supported by the WDM layer without the SONET layer. Although the SONET layer provides fast restoration in the event of failure, the *IP over WDM architecture* without the SONET layer has few advantages: (1) SONET is designed for voice traffic, not for data traffic; (2) network control and management can be much simpler; and (3) it is more cost-effective, since SONET equipment increases in cost linearly with bit rate and the number of ports.

At the *control plane*, MPLS [11] was developed for IP networks. By introducing the concept of *label switching*, IP networks can forward packets much faster and overcome the shortcoming of connectionless service with a label-switched path (LSP). In addition, it is much easier to employ traffic engineering. While optical networks become a more attractive solution for transmission networks, MPλS was introduced, which is the application of MPLS

to the optical domain. The wavelengths and lightpaths in MPλS correspond to the labels and LSPs in MPLS, respectively. *GMPLS* [1] was developed for IP over WDM networks with a unified control plane. It is a generalized control plane in a sense that it encompasses packet-switch-capable (PSC), time-division-multiplex-capable (TDM), lambda-switch-capable (LSC), and fiber-switch-capable (FSC) interfaces. With GMPLS, it is possible to control different networks with a single unified control plane.

In the following discussion, we focus on optical switching techniques: wavelength routing, optical packet switching (OPS), and optical burst switching (OBS).

2. OPTICAL SWITCHING TECHNIQUES

Now, we look into the characteristics and issues in optical switching techniques. The general architecture of IP over WDM optical networks (optical Internet) is shown in Fig. 2 [12]. Multiple optical networks exist in the optical domain, where an ENNI (external network-to-network interface) is used for signaling between optical networks. A single optical network consists of multiple suboptical networks, where the INNI (internal network-to-network interface) is used for signaling between them. Again, a suboptical network has multiple optical nodes (e.g., OXCs or optical routers) interconnected with optical fibers. As clients, IP, ATM, and SONET networks are interfaced with optical networks via a UNI (user-to-network interface).

The optical switching techniques that we are interested in determine the service provided by optical networks to client networks. *Wavelength routing*, *optical burst switching*, and *optical packet switching* networks provide lightpath-level, burst-level, and packet-level services, respectively, to client networks.

2.1. Wavelength Routing Network

A *lightpath* is the service unit that the wavelength routing network provides. A *wavelength routing network* consists of multiple OXCs, which are interconnected with optical fiber links. At the network planning stage, for the estimated input traffic, network resources (e.g., number of wavelengths) are dimensioned so that the performance (e.g., blocking probability) can be satisfied.

When setting up a lightpath for a request from a client network, the edge node in the wavelength routing network sends out a lightpath request. In this process, the ingress edge node performs routing to find an optimal route to the egress edge node. The signaling mechanisms in GMPLS may be used in this routing process. Thus, the lightpath setup is done by the control plane. There are two ways of implementing the control plane: by centralized control or by distributed control. Both approaches have advantages and disadvantage. The interested reader may refer to two papers published in 1996 and 1997 [13,14].

When setting up a lightpath, OXCs in the wavelength routing network perform the key function: switching the wavelength to its destined port. The general *architecture of OXC* is shown in Fig. 3. There are M input fibers and M output fibers, each of which carries W wavelengths. The W wavelengths are demultiplexed and put into the optical switch, which has the dimension of $MW \times MW$. The OXC control is in charge of generating control signals to optical components such as optical switch and wavelength converters. The control signal is based on the decision made in the control plane where the lightpath setup requests are processed.

The OXC may have *wavelength converters* for the purpose of increasing performance. It is obvious that having wavelength converters decreases the blocking probability since without wavelength converters, the same wavelength should be used in every link (wavelength continuity), while with wavelength converters, any wavelength can be used. The wavelength conversion can be done either electrically or optically. If the conversion is performed electrically, the optical signal is terminated with O/E (optical to electrical) conversion, then transmitted over the different wavelength after E/O (electrical to optical) conversion. However, its cost linearly increases with the number of transponders per wavelength.

On the other hand, if the conversion is performed optically, the optical signal can be transparently transmitted without O/E/O conversion. However, the cost of optical wavelength converters is still high. Thus, it is desirable

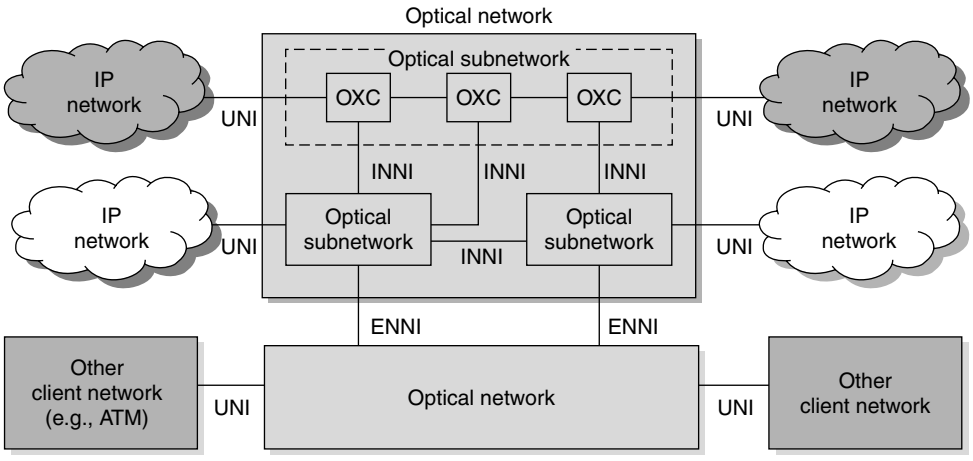


Figure 2. IP over WDM optical networks.

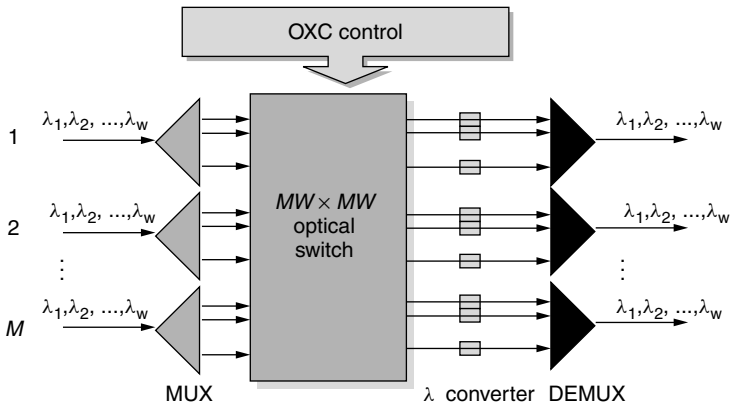


Figure 3. OXC architecture.

to keep the wavelength conversion capability minimum (no conversion or limited conversion). There are studies showing that the gain of full wavelength conversion over no wavelength conversion is a function of the number of wavelengths and the number of hops to traverse [15]. Thus, with proper network planning (e.g., short network diameter and enough wavelengths), it is possible to keep the conversion capability low.

Now, let us look at the characteristics of a wavelength routing network. A wavelength routing network is in the form of circuit switching, and thus we also call it an *optical circuit switching* (OCS) network. It requires *two-way reservation* for lightpath establishment. In other words, a lightpath is established when the acknowledgment comes back as a response to the setup request. This process introduces a setup delay, which is proportional to the round-trip propagation delay over the distance between the two points. Although the setup delay is ignored in a static OCS network with relatively long session time (e.g., years or months), the setup delay may affect the performance of a dynamic OCS network when the session time is of the same order of magnitude as the setup delay. It is another shortcoming of a wavelength routing network that the bursty traffic may result in poor performance due to the static nature of a wavelength routing network. Another important issue in a wavelength routing network is the routing and wavelength assignment (RWA). With an efficient RWA algorithm, it is possible to reduce the network resources for the given input traffic. However, in spite of intensive research on RWA [16], it is still a hard problem to solve, especially when RWA is performed online.

Although there are some disadvantages, a wavelength routing network is the feasible solution in the near term, due to immaturity of optical technology.

2.2. Optical Packet-Switching Network

In a *packet-switching network*, user data are segmented into packets. Each packet consists of two parts: a *header* containing the control information (e.g, routing information) and *data*. As an example, IP datagrams and ATM cells are the packets in IP and ATM networks, respectively. In a packet-switching network, there are two types of service: datagram service (or connectionless service) as in an IP network and virtual circuit service (or connection oriented service) as in an ATM network. While each packet takes the same path in the *virtual circuit service*, each packet may take a different path in a *datagram service*. In either case, each packet goes through intermediate packet switching nodes until it reaches its destination. The key functions performed by intermediate nodes is routing and forwarding, namely, deciding the next hop node and forwarding the packet. These are the characteristics of a packet switching network in the electrical domain.

The *optical packet switching* (*OPS*) *network* is to perform the same packet switching functions in the optical domain. Thus, an optical packet is transmitted and processed by the optical packet-switching nodes. The general architecture of an optical packet switching node is shown in Fig. 4.

An *OPS node* consists of the input and output processing units, a switching unit, a buffering unit, and control unit. The optical packets arrive at input processing unit, where synchronization and header extraction take place. If the system operates in time slots with a fixed size of packets, then synchronization is required for the alignment of multiple incoming packets, which may arrive at different times. The tunable delay, which can be implemented with fiber delay lines (FDLs) and 2×2 optical switches, may be used for synchronization.

When an optical packet arrives, its header is detached and sent to the control unit for processing. Currently, the implementation of the control unit using optical logic is very difficult. Thus, after being extracted, the header part goes through O/E conversion, while the data remain in optical form. To facilitate tapping the optical signal, optical packets are encoded using a technique such as subcarrier multiplexing (SCM). Once the header is decoded in the control unit, the routing process is performed to determine the output port. This process may be layer 3 IP routing or layer 2 label switching. In this process, the control unit generates the control signals to the switching unit (small and fast optical switch) and the buffering unit. The buffering unit resolves the contention problem when multiple packets are destined to the same output port. Since optical memory is not commercially available today, the buffering unit is usually implemented using FDLs, which provide only limited time of buffering.

Finally, the header part and data part are rejoined at the output processing unit before sending it out to the next switching node. If the information in the header needs to

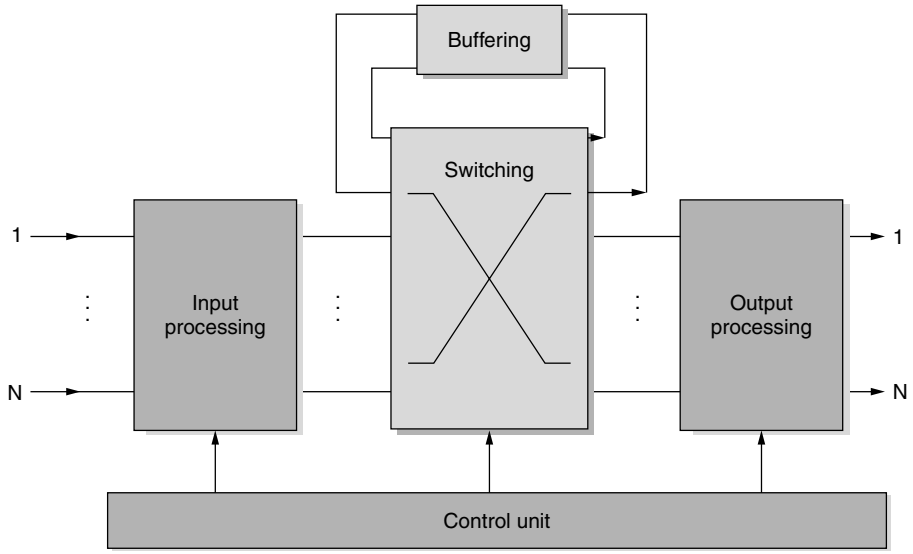


Figure 4. OPS node architecture.

be updated (e.g., by label update after label swapping), the rewriting process takes place.

Although an optical switching network overcomes some of the shortcomings of a wavelength routing network (e.g., inefficiency due to static nature), it is technically hard to implement packet-switching functions with current optical technology. The optical burst switching (OBS) technique attempts to effect balance between wavelength routing and optical packet switching by overcoming the disadvantages of both techniques, while preserving their merits.

3. OPTICAL BURST SWITCHING NETWORK

So far, we have described the characteristics of a wavelength routing network and an optical packet-switching network. Now, we focus on another alternative, *optical burst switching* (OBS), and look at its characteristics in some detail.

3.1. OBS Protocol

The distinction between OBS and OCS is that the former uses a one-way reservation while the latter uses a two-way reservation. It is called the two-way reservation when there must be a connection setup procedure before the data transmission takes place. It is called the *one-way reservation* when the data (which is called the data burst) follow the connection setup request immediately after waiting for some delay. This delay will be called an *offset time*, which will be explained later. Note that the connection setup request in the OBS will be called a *control packet* or a *burst control packet* (BCP).

Although OBS and OPS have similar characteristics (e.g., statistical multiplexing on the links), the distinction between the two is that the former has some unique features such as the offset time and the delayed reservation [4,5]. In addition, the payload in the OBS is much larger than that in the OPS. The payload unit in OBS networks will be referred as the data burst hereafter.

The operations of the *OBS protocol* are illustrated in Fig. 5a. When the source node S has a data burst to send, the burst control packet (BCP) is generated and transmitted first. At each node, the BCP takes δ time unit to be processed and makes the reservation for the following data burst. Meanwhile, the data burst, after waiting

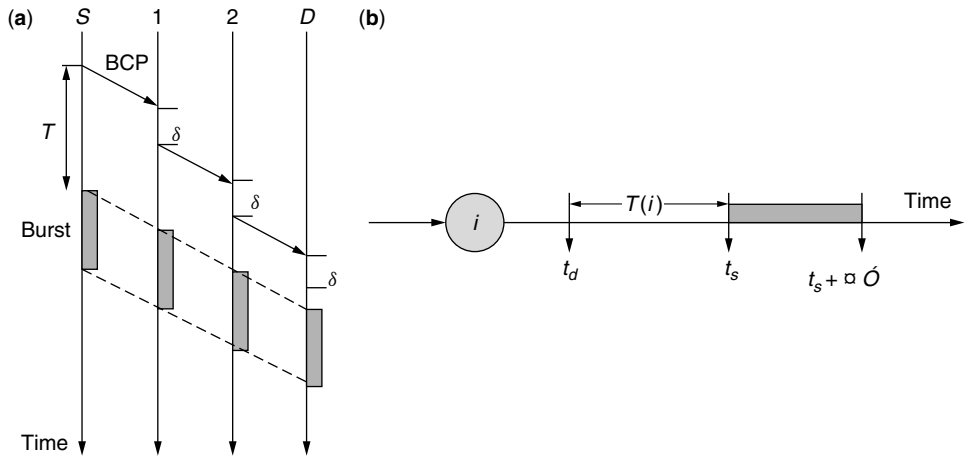


Figure 5. Offset time (**a**) and delayed reservation (**b**) in OBS.

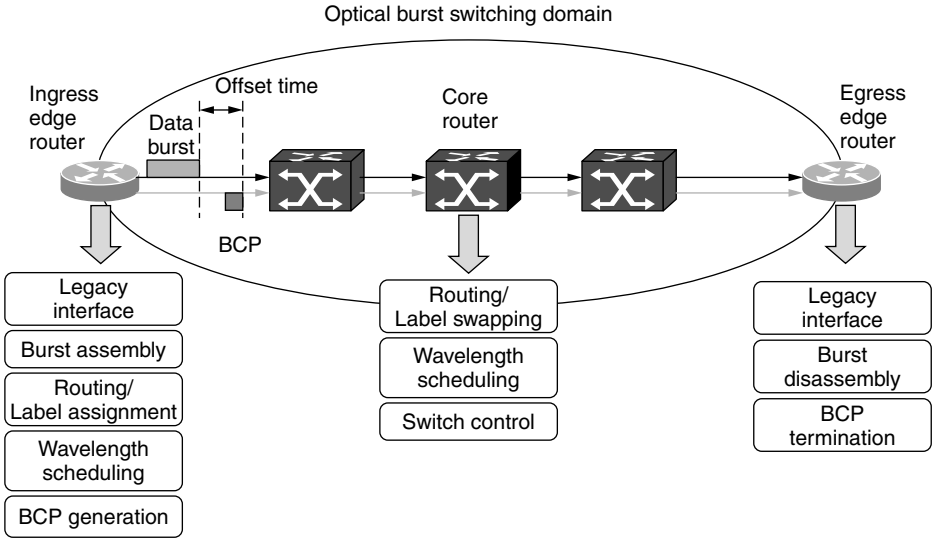


Figure 6. OBS network.

for an offset time, which is denoted as T, at the edge router, immediately follows the BCP. Since the BCP has already set up the necessary configurations, including the resource reservation, the data burst does not need to be delayed (buffered) for processing at the intermediate nodes (nodes 1 and 2), and just cuts-through to the destination node D. Thus, OBS networks can provide the data plane with the end-to-end transparency without going through any optical–electrical–optical (O/E/O) conversions.

The advantage of the offset time is to decouple the BCP from the data burst in time, which makes it possible to eliminate the buffering requirement at the intermediate nodes. In order to ensure that the data burst does not overtake the BCP, the offset time should be made long enough by considering the expected processing time at each intermediate node and the number of hops to be traversed by the BCP [4,5].

Another feature of OBS is the *delayed reservation* (DR), which makes it possible to statistically multiplex data bursts on the links. The DR is illustrated in Fig. 5b, where $T(i)$ is the offset time at node i, 1 is the transmission time of the data burst at node i, and t_a and t_s indicate the arriving time of the BCP and its corresponding data burst, respectively. According to DR, after being processed, the BCP reserves the resources at a future time (at the arrival time of the data burst), which can be obtained from the offset time $T(i)$. Also, the reservation is made for just enough time to finish the transmission (data burst duration 1). In OBS, two parameters — the offset time and the mean data burst size — need to be selected carefully in the design step since they have a great impact on performance.

3.2. OBS Network and OBS Routers

Now, we discuss architectural aspects of OBS [6]. For simplicity, we focus on a single OBS domain, where all nodes (or *OBS routers*) are well aware of the OBS protocols. Note that the terms *OBS domain* and the *OBS network* will be used interchangeably. Depending on the location in the OBS domain, the OBS routers are classified as *edge routers* and *core routers* as shown in Fig. 6. Note that the edge routers should be equipped with both capabilities as an *ingress edge router* (*IER*) and as an *egress edge router* (*EER*). Thus, the edge router functions as an ingress router when there are inbound data to the OBS domain, whereas the edge router functions as an egress router when there are outbound data from the OBS domain. In the following discussion, we describe the functions and general architectures for each type of OBS router.

The general *architecture of the IER* is shown in Fig. 7. The IER should provide the interface between the OBS network and other legacy networks. It also needs to provide the signaling interface between two different networks. When the IER receives incoming data (e.g., IP packets, ATM cells, or voicestreams) from the line interface cards, the burst assembly process takes place in which multiple packets are packed into a data burst. We will discuss the burst assembly process later. The arriving packets are switched to the appropriate assembly queues according to their destination and QoS. The first packet arrival in an assembly queue initiates the BCP generation where some BCP fields such as burst size, offset time, and label are to be determined and filled later when the *burst assembly* is completed.

When the burst is assembled long enough to meet the requirements, the BCP obtains the field information of burst size and offset time. On the basis of the routing decision, a label is assigned to establish the label-switched path (LSP). If the LSP already exists (has been set up by the previous BCP), the previously used label is assigned. Otherwise (i.e., if a new LSP needs to be set up or the existing LSP needs to be changed), a new label is assigned. The downstream nodes perform the label swapping, where the inbound label is mapped into the outbound label at the local node. The label information in the BCP should be updated accordingly. How the labels are distributed and assigned depends on the label distribution protocols such as RSVP-TE and CR-LDP [17,18].

According to OBS protocols, the BCP is transmitted to the core OBS router an offset time earlier than its corresponding data burst. The wavelength assignment

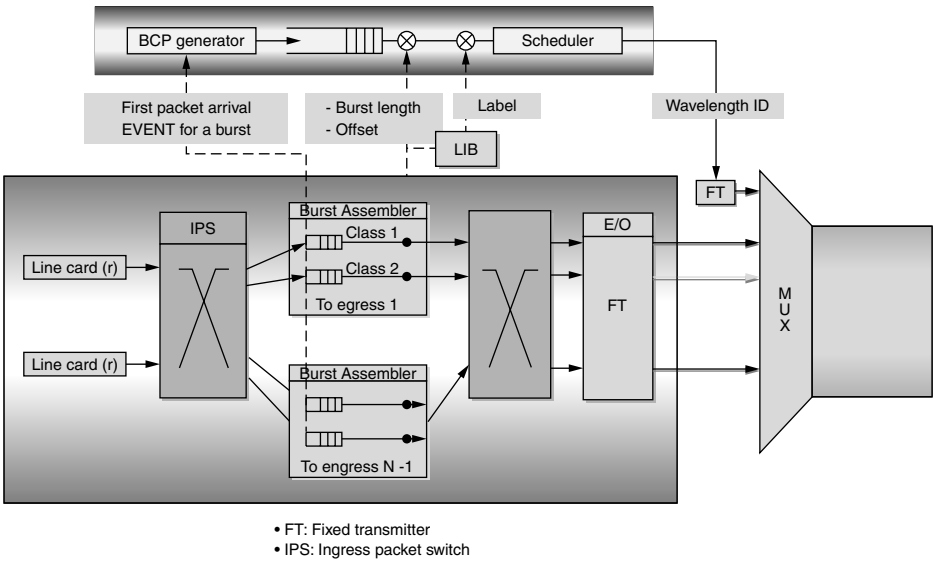


Figure 7. Architecture of ingress-edge router (IER).

and scheduling is required in this step. There are two kinds of wavelengths (or channels) in the OBS domain: the *control channels*, which carry the BCPs, called the *control channel group* (CCG); and the data burst channels, which carry the data bursts, called the *data burst channel group* (DCG). Thus, two wavelengths are assigned and scheduled for transmission of the BCP and its data burst. The wavelength scheduling on the DCG (for data bursts), which can enhance the utilization, is an interesting research area. There are a few scheduling algorithms proposed to date, such as first fit, horizon scheduling, and void-filling scheduling [6,19]. It is noted that while the labels carry the path (LSP) information, the wavelengths are meaningful only at the local node. In this way, the downstream nodes can assign the wavelengths dynamically, which combines with the label only for the duration of the data burst.

The *architecture for the EER* is shown in Fig. 8. The main functions of the EER are the BCP termination and the data burst disassembly. When a BCP arrives, the EER processes it and reserves the buffer space as required by the burst length field for the burst disassembly process. On arrival, the data burst, after being converted to an electrical signal, goes through the burst disassembly process. Then the disassembled packets are distributed to their destination ports.

The core router has the general architecture shown in Fig. 9. It consists of two parts: a switch control unit and a data burst unit. While the switch control unit is responsible for processing the BCPs on the CCG, the data burst unit is responsible for switching the data burst to the destined output port, which is controlled by the switch control unit. Most of the functions in the core router take place in the switch control unit, which includes the label swapping for establishing the LSP, local wavelength scheduling, burst contention resolution, and generation of the control signal to the data burst unit. The data burst unit consists roughly of demultiplexers, inlet fiber delay lines (FDLs) (for adjusting the timing jitter of offset time between the BCP and the data burst), optical switches, wavelength converters, FDL buffers (for

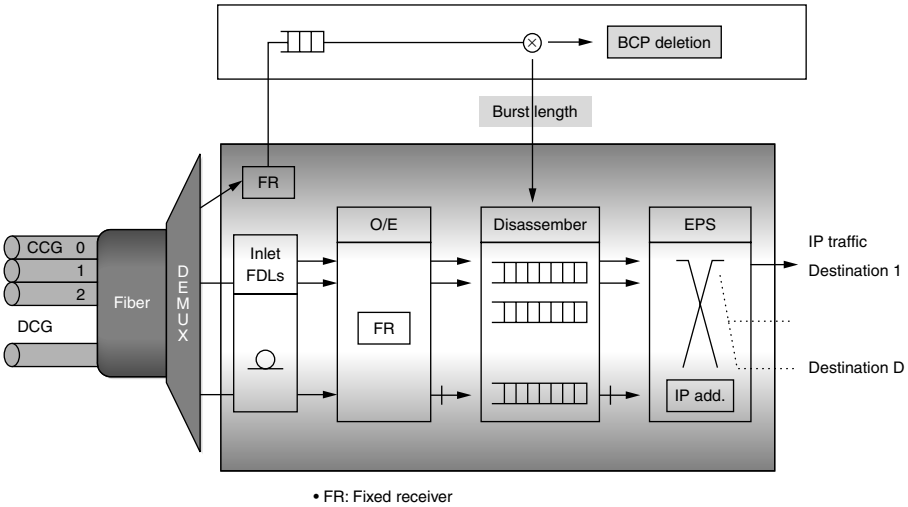


Figure 8. Architecture of egress edge router (EER).

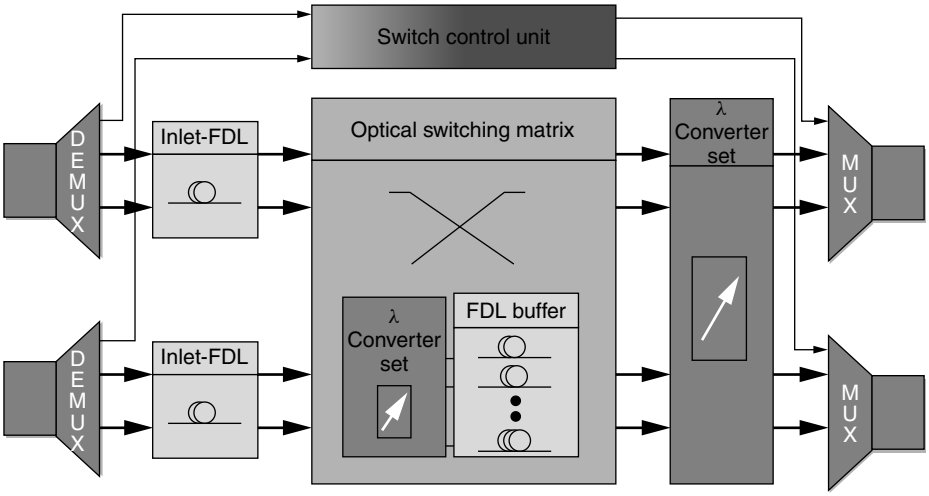


Figure 9. Architecture of core router.

contention resolution), and multiplexers. The components of the data burst unit are passively configured by the control signals from the switch control unit.

Before we conclude this description, of the architecture, it is worth mentioning some important design parameters such as the capacity of the core router, the switching speed of the optical switch, and the average burst size. The capacity of the router is determined by the number of incoming fibers and the number of wavelengths available in each fiber. For example, a core router of 10 Tbps capacity needs 32 incoming fibers, each of which has 32 wavelengths operating at 10 Gbps per wavelength. Of course, since some wavelengths are dedicated to the CCG, the router capacity is determined only by the number of wavelengths in the DCG.

The core router requires two switches; one is in the switch control unit, and the other is in the data burst unit. While the former switches the BCPs on the CCG, which can be implemented with a small electronic switch (depending on the number of wavelengths in the CCG), the latter switches the data bursts on the DCG, which can be implemented with an optical switch. The dimension of the optical switch depends on the number of wavelengths in the DCG. The architecture of the optical switch may be either a simple single stage if a large optical switch is available [e.g., MEMS (microelectromechanical systems) crossbar switch [20]] or multistage interconnection network (MIN) as in an ATM switch [21] if multiple small optical switches are used.

Next, consider the switching speed of an optical switch and the average burst size. The *switching time* of an optical switch is the delay that it takes to change from one state to another state. The switching speed imposes an overhead on the switch throughput. In other words, the slower the switching speed, the worse the throughput. However, the overhead caused by the switching speed can be reduced if the data burst is long compared to the switching time. Thus, for optimum switch throughput, the average burst size should be selected according to the switching speed of the optical switch in use. This average burst size is the requirement that the burst assembler should meet.

3.3. QoS in OBS Network

Given that some real-time applications (such as Internet telephony and videoconferencing) require a higher *QoS* than do non-real-time applications [such as electronic mail (email) and general Web browsing], the QoS issue should be addressed. Although QoS concerns related to optical (or analog) transmission (e.g., dispersion, power and signal-to-noise ratio) also need to be addressed, here, we focus on how to ensure that critical data can be transported in the OBS domain more reliably than noncritical data. Unlike the existing QoS schemes that differentiate the services using the buffer, the QoS scheme to be introduced in the following discussion takes advantage of the offset time, which was explained earlier. We call this an *offset-time-based QoS scheme*. For this purpose, we introduce a new offset time, which is called the *extra offset time*. Note that the offset time introduced previously, which is called the *base offset time* is different from the extra offset time.

We now explain how the offset-time-based QoS scheme works [22,23]. In particular, we explain how *class isolation* (or service differentiation) can be achieved by using an extra offset time in both cases with and without using fiber delay lines (FDLs) at an OBS node. Note that one may distinguish the following two different contentions in reserving resources (wavelengths and FDLs): the *intraclass contentions*, caused by requests belonging to the same class; and the *interclass contentions*, caused by requests belonging to different classes. In what follows, we focus on how to resolve interclass contentions using the offset-time-based QoS scheme.

For simplicity, we assume that there are only two classes of (OBS) services: classes 0 and 1, where class 1 has priority over class 0. In the offset-time-based QoS scheme, to give class 1 a higher priority for resource reservation, an extra offset time, denoted by t_o^1, is given to class 1 traffic (but not to class 0, i.e., $t_o^0 = 0$). In addition, we also assume that the base offset time is negligible as compared to the extra offset time, and will refer to the latter as simply the *offset time* hereafter. Finally, without loss of generality, we also assume that a link has only one wavelength for data (and an additional wavelength for control).

3.3.1. The Case Without FDLs. In the following discussion, let t_a^i and t_s^i be the arriving time and the service-start time for a class i request denoted by req(i), respectively, and let l_i be the burst length requested by req(i), where $i = 0, 1$. Figure 10 illustrates why a class 1 request that is assigned an (extra) offset time obtains a higher priority for wavelength reservation than does a class 0 request in the case of no FDLs. We assume that there is no burst (that arrived earlier) in service when the first request arrives. Consider the following two situations where contentions among two classes of traffic are possible.

In the first case as illustrated in Fig. 10a, req(1) comes first and reserves a wavelength using DR, and req(0) comes afterward. Clearly, req(1) will succeed, but req(0) will be blocked if $t_a^0 < t_s^1$ but $t_a^0 + l_0 > t_s^1$, or if $t_s^1 < t_a^0 < t_s^1 + l_1$. In the second case, as in Fig. 10b, req(0) arrives first, followed by req(1). When $t_a^1 < t_a^0 + l_0$, req(1) would be blocked had no offset time been assigned to req(1) (i.e., $t_o^1 = 0$). However, such a blocking can be avoided by using a sufficient offset time so that $t_s^1 = t_a^1 + t_o^1 > t_a^0 + l_0$. Given that t_a^1 may only be slightly behind t_a^0, t_o^1 needs to be larger than the maximum burst length over all bursts in class 0 in order for req(1) to completely avoid being blocked by req(0). With that much of offset time, the *blocking probability* of (the bursts in) class 1 becomes only a function of the offered load belonging to class 1, that is, independent of the offered load belonging to class 0. However, the blocking probability of class 0 is determined by the offered load belonging to both classes.

3.3.2. The Case with FDLs. Although the offset-time-based QoS scheme does not mandate the use of FDLs, its QoS performance can be significantly improved even with limited FDLs so as to resolve contentions for bandwidth among multiple bursts. For the case with FDLs, the variable B will be used to denote the maximum delay that a FDL (or the longest FDL) can provide. Thus, in the case of blocking, a burst can be delayed up to the maximum delay B.

Figure 11a,b illustrates class isolation at an OBS node equipped with FDLs where contention for both wavelength and FDL reservation may occur. In Fig. 11a, let us assume that when req(0) arrives at $t_a^0(t_s^0)$, the wavelength is in use by a burst that arrived earlier. Thus, the burst corresponding to req(0) has to be delayed (blocked) for t_b^0 units. Note that the value of t_b^0 ranges from 0 to B, and a FDL with an appropriate length that can provide a delay of t_b^0 will be chosen. Accordingly, if $t_b^0 < B$, the FDL is reserved for a class 0 burst as shown in Fig. 11b (the burst will be dropped if t_b^0 exceeds B), and the wavelength will

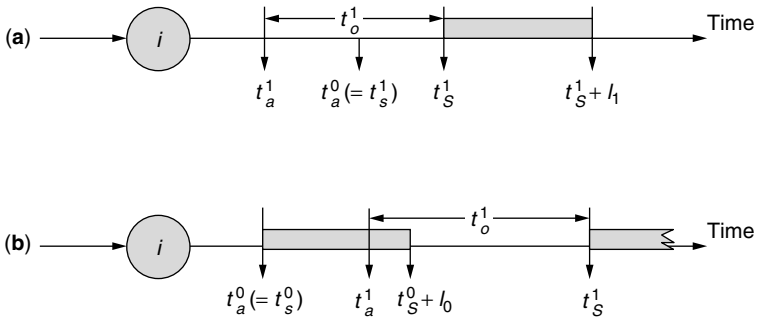


Figure 10. Class isolation in the case without FDLs.

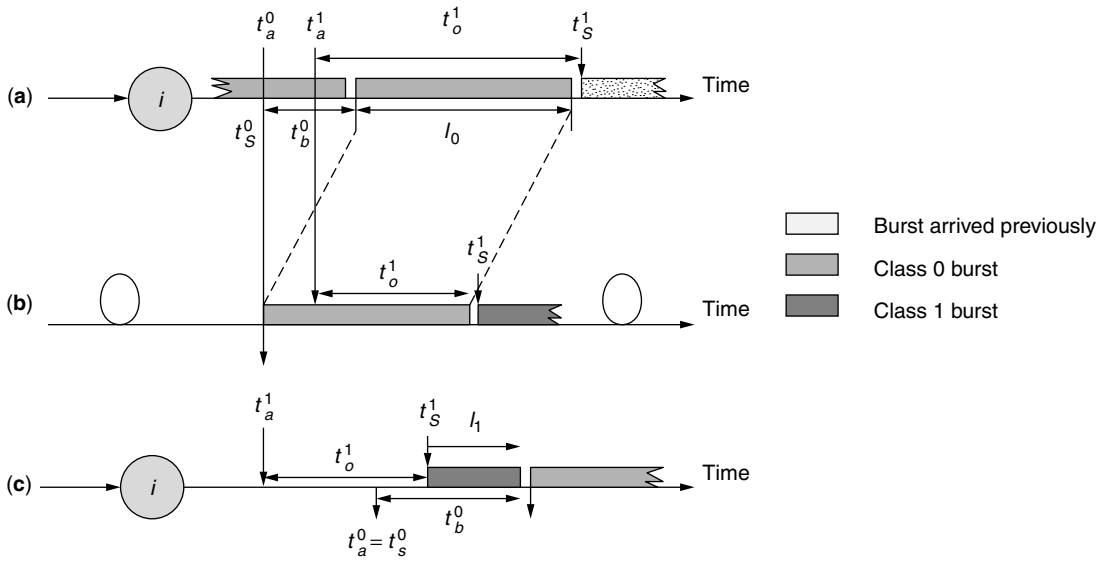


Figure 11. Class isolation in the case with FDLs.

be reserved from $t_s^0 + t_b^0$ to $t_s^0 + t_b^0 + l_0$ as shown in Fig. 11a. Now assume that req(1) arrives later at t_a^1 (where $t_a^1 > t_a^0$) and tries to reserve the wavelength. req(1) will succeed in reserving the wavelength as long as the offset time is so long that $t_s^1 = t_a^1 + t_o^1 > t_a^0 + t_b^0 + l_0$. Note that had req(1) arrived earlier than req(0) in Fig. 11a it is obvious that req(1) would not have interclass contention caused by req(0). This illustrates that class 1 can be isolated from class 0 when reserving a wavelength because of the offset time. Of course, without the offset time, req(1) would be blocked for $t_a^0 + t_b^0 + l_0 - t_a^1$, and it would be entirely up to the use of FDLs to resolve this interclass contention.

Similarly, Fig. 11b illustrates class isolation in FDL reservation. More specifically, let us assume that req(0) has reserved the FDLs as described earlier, and because t_o^1 is not long enough, req(1) would be blocked in wavelength reservation and thus needs to reserve the FDLs. In such a case, req(1) will successfully reserve the FDLs if the offset time is still long enough to have $t_s^1 = t_a^1 + t_o^1 > t_a^0 + l_0$. Otherwise (i.e., if $t_s^1 < t_a^0 + l_0$), req(1) would contend with req(0) in reserving the FDL and would be dropped.

As shown in Fig. 11c, if req(1) comes first and reserves the wavelength based on t_o^1 and delayed reservation (DR), and req(0) comes afterward, req(1) is not affected by req(0). However, req(0) will be blocked either when $t_a^1 < t_a^0 < t_s^1$ but $t_a^0 + l_0 > t_s^1$, or when $t_s^1 < t_a^0 < t_s^1 + l_1$. Similarly, if req(1) arrives first, it can reserve the FDL first regardless of whether req(0) succeeds in reserving the FDL. As mentioned earlier, this implies that class 1 can be isolated from class 0 in reserving both the wavelength and the FDL by using an appropriate offset time, which explicitly gives class 1 a higher priority over class 0. As a result of having a low priority on resource reservations, class 0 bursts will have a relatively high blocking and loss probability.

Although the offset-time-based QoS scheme does provide good service differentiation even when buffering is not possible, it has some disadvantages that need to be enhanced. For example, the offset-time-based QoS scheme introduces delay overhead caused by the extra offset time. Since the highest class suffers from the longest delay [22,23], the QoS provisioning will be strictly restricted in a given delay budget. It also lacks the controllability on the QoS, which can be enhanced by introducing the measurement based QoS provisioning and assigning the appropriate weight to each QoS class. In addition, it is worth considering how the offset-time-based QoS scheme can be integrated with the existing QoS domain such as DiffServ.

3.4. Burst Assembly

As we mentioned earlier, the *burst assembly* process takes place at the ingress edge router. The incoming data (e.g., IP packets) are assembled into a super packet, which is called the *data burst*. In the following discussion, we look into the issues of burst assembly.

It is obvious that IP packets would be the dominant traffic in future networks. Unlike the traffic from traditional telephone networks, Internet traffic is quite difficult to predict. This is because Internet traffic shows self-similarity [24,25]. *Self-similarity* means that even with a high degree of multiplexing, the burstiness of traffic still exists at all timescales, which makes the network resource dimensioning and traffic engineering harder.

One advantage of the burst assembly is that it may reduce the self-similarity of Internet traffic [26]. Figure 12 shows an example configuration of an OBS network for burst assembly [27]. IP routers in the IP domain inject IP packets into the OBS network. The OBS edge routers located at the boundary of the OBS network take IP packets and perform the burst assembly process.

The incoming IP packets are classified and queued into an assembly buffer according to their destination and QoS requirements. A simple burst assembly process is illustrated in Fig. 13. The burst assembler at the end of the assembly buffer runs a timer, which expires at a given time (i.e., the assemble time). Whenever the timer expires, the burst assembler takes the burst (multiple IP packets queued during a given period of time) out of the assemble buffer for transmission. The key parameter of the burst assembler is the *assemble time*, which controls the size of the data burst. The distribution of assemble time could be deterministic or random, in which case it is called either a constant assemble time (CAT) or a variable assemble time (VAT), respectively. The burst assembler waits for a constant time when using the CAT, whereas it waits for a random amount of time when using the VAT. Alternatively, the burst assembler adaptively controls the assemble time by monitoring both assemble time and the

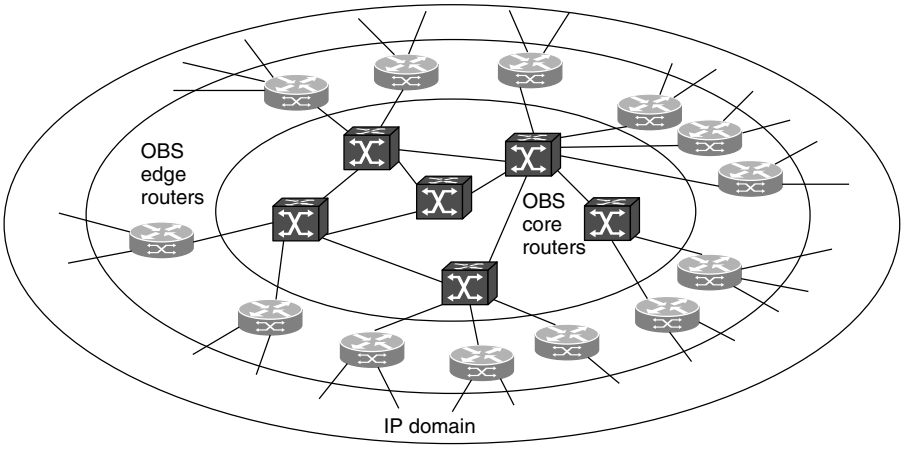


Figure 12. OBS edge routers for burst assembly.

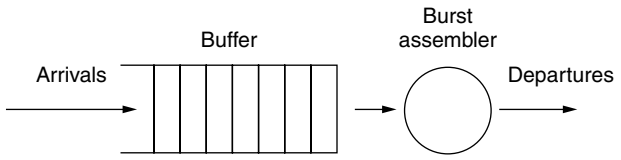


Figure 13. Burst assembly process.

size of the assembled IP packets. The optimization of the burst assembly process is a topic that requires further investigation.

4. SUMMARY

Because of its advantages over conventional communication networks, optical networks have been firmly positioned as a feasible solution for the next-generation networks. The optical networking technology keeps improving starting, from SONET to static wavelength routing network, and eventually, to dynamic optical networks. We have discussed optical switching techniques available today for optical networks: wavelength routing, optical burst switching, and optical packet switching.

Table 1 summarizes the qualitative comparison among these techniques. Wavelength routing is a simple and technically matured technology, but it may result in poor performance. On the other hand, optical packet switching may show its dynamicity in networking as in an electrical packet-switching network, but is technically immature for the implementation. Optical burst switching achieves balance between wavelength routing and optical packet switching by improving the inefficiency of wavelength routing and at the same time, by lessening the technical difficulty of optical packet switching. Thus, while the optical technology is mature enough to implement optical packet switching, optical burst switching appears to be a practical interim solution for optical networks.

Acknowledgments
This work was supported in part by the Korean Science and Engineering Foundation (KOSEF) through the OIRC project.

BIOGRAPHIES

Myungsik Yoo received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Korea, in 1989 and 1991, respectively, and the Ph.D. degree in electrical engineering from State University of New York at Buffalo (SUNY at Buffalo), Buffalo, New York (USA) in 2000. He was a Senior Research Engineer in Nokia Research Center, Burlington, Massachusetts. Since 2000, he has been an Assistant Professor in the School of Electronic Engineering, Soongsil University, Seoul, Korea. His current research interests are optical networks and optical Internet, including optical burst switching, protection/restoration, QoS support, and GMPLS.

Chunming Qiao is an Associate Professor at the University at Buffalo (SUNY). He has over 10 years of academic and industrial experience in optical networks. Dr. Qiao has published more than 100 papers in leading technical journals and conference proceedings, and several book chapters. He has given several keynote speeches, tutorials, and invited talks and is recognized for his pioneering research on optical Internet and in particular, the optical burst switching (OBS) paradigm. Dr. Qiao is the IEEE Communication Society's Editor-at-Large for optical networking and computing; an editor of several other journals and magazines, including *IEEE/ACM Transactions on Networking* (ToN), as well as a guest editor for *IEEE JSAC* and other publications. He has chaired and co-chaired many conferences and workshops on optical communications and networking, including the Optical Networks Symposium (ICC'03), and Opticomm 2002. Dr. Qiao is also the founder and Chair of the Technical Group on Optical Networks (TGON) sponsored by SPIE, and a Vice Chair of the IEEE Technical Committee on Gigabit Networking (TCGN).

Table 1. Comparison Between Optical Switching Paradigms

Switching Paradigm	Bandwidth Utilization	Implementation Difficulty	Switching Granularity
Wavelength routing	Poor	Low	Coarse
Optical packet switching	High	High (not mature)	Fine
Optical burst switching	Moderate	Moderate	Moderate

BIBLIOGRAPHY

1. B. Rajagopalan et al., A framework for generalized multi-protocol label switching (GMPLS), IETF Internet draft.
2. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Approach*, Morgan Kauffman, San Francisco.
3. W. Goralski, *Optical Networking & WDM*, McGraw-Hill, New York.
4. C. Qiao and M. Yoo, Optical burst switching (OBS)—a new paradigm for an optical Internet, *J. High Speed Network* **8**(1): 69–84 (1999).
5. C. Qiao and M. Yoo, Choice, features and issues in optical burst switching, *Opt. Network Mag.* **1**(2): 36–44 (May 2000).
6. Y. Xiong, M. Vandenhoute, and H. C. Cankaya, Control architecture in optical burst-switched WDM networks, *IEEE J. Select. Areas Commun.* **18**(10): 1838–1851 (Oct. 2000).
7. J. Turner, Terbit burst switching, *J. High Speed Network* **8**(1): 1–18 (1999).
8. S. Yao et al., All-optical packet switching for metropolitan area networks: opportunities and challenges, *IEEE Commun. Mag.* **39**(3): 142–148 (March 2001).
9. M. J. O'Mahony et al., The application of optical packet switching in future communication networks, *IEEE Commun. Mag.* **39**(3): 128–135 (March 2001).
10. X. Lisong et al., Techniques for optical packet switching and optical burst switching, *IEEE Commun. Mag.* **39**(1): 136–142 (Jan. 2001).

11. E. Rosen et al., *Multiprotocol Label Switching Architecture*, IETF RFC 3031.

12. B. Rajagopalan et al., IP over optical networks: A framework, IETF Internet draft.

13. R. Ramaswami and A. Segall, Distributed network control for wavelength routed optical networks, *IEEE/ACM Trans. Network.* **5**(6): 936–943 (Dec. 1997).

14. C. Qiao and Y. Mei, Wavelength reservation under distributed control, *IEEE/LEOS Broadband Opt. Network.* 45–46 (1996).

15. Models of blocking probability in all-optical networks with and without wavelength changers, *IEEE J. Select. Areas Commun.* **14**(5): 858–867 (June 1996).

16. R. Ramaswami and K. Sivarajan, Routing and wavelength assignment in all-optical network, *IEEE/ACM Trans. Network.* 489–500 (Oct. 1995).

17. P. Ashwood-Smith et al., Generalized MPLS signaling—RSVP-TE extensions, IETF Internet draft.

18. P. Ashwood-Smith et al., Generalized MPLS signaling—CR-LDP extensions, IETF Internet draft.

19. J. S. Turner, WDM burst switching for petabit data networks, *Proc. OFC'2000*, 2000, Vol. 2, pp. 47–49.

20. L. Y. Lin and E. L. Goldstein, MEMS for free-space optical switching, *Proc. LEOS'99*, 1999, Vol. 2. pp. 483–484.

21. R. Awdeh and H. T. Mouftah, Survey of ATM switch architecture, *IEEE Commun. Mag.* **27**: 1567–1613 (Nov. 1995).

22. M. Yoo, C. Qiao, and S. Dixit, QoS performance of optical burst switching in IP-over-WDM networks, *IEEE J. Select. Areas Commun.* **18**(10): 2062–2071 (Oct. 2000).

23. M. Yoo, C. Qiao, and S. Dixit, Optical burst switching for service differentiation in the next generation optical Internet, *IEEE Commun. Mag.* **39**(2): 98–104 (Feb. 2001).

24. V. Paxon and S. Floyd, Wide area traffic: the failure of Poisson modeling, *IEEE Trans. Network.* **3**(3): 226–244 (1995).

25. W. Leland et al., On the self-similar nature of Ethernet traffic (extended version), *IEEE Trans. Network.* **2**(1): 1–15 (1994).

26. A. Ge, F. Callegati, and L. Tamil, On optical burst switching and self-similar traffic, *IEEE Commun. Lett.* **4**(3): 98–100 (March 2000).

27. A. Detti, A. Eramo, and M. Listanti, Performance evaluation of a new technique for IP support in a WDM optical network: Optical composite burst switching (OCBS), *J. Lightwave Technol.* **20**(2): 154–165 (Feb. 2002).

OPTICAL SYNCHRONOUS CDMA SYSTEMS

Tomoaki Ohtsuki
Tokyo University of Science
Noda, Chiba, Japan

Iwao Sasase
Keio University
Yokohama, Japan

1. INTRODUCTION

Optical communication systems in the optical fiber play a main part of the digital communications in backbone networks, high speed local-area networks (LANs) using a fiber distributed data interface (FDDI), metropolitan-area network (MAN), and a next-generation subscriber system such as a fiber to the home (FTTH). The main advantages of the optical fiber communications are the high speed, large capacity and high reliability by the use of the broadband of the optical fiber. A desirable feature for future optical networks would be the ability to process information directly in the optical domain for purposes of multiplexing, demultiplexing, filtering, amplification, and correlation. Optical signal processing would be advantageous since it can potentially be much faster than electrical signal processing and the need for photon–electron–photon conversion would be obviated.

Asynchronous multiple-access methods where network access is random and collisions occur, such as token passing and carrier-sense multiple-access, are well suited to LANs with low traffic demand. However, these asynchronous access methods suffer from cumulative delay as the traffic intensity increases. Also, contention protocols generally proposed for low traffic demands are not suitable if traffic delay is a major issue, as, in networks where information must be transmitted simultaneously. On the other hand, synchronous accessing methods where transmissions are perfectly scheduled provide more successful transmissions than asynchronous methods. As a typical synchronous protocol, time-division multiple access (TDMA) is an efficient multiple-access protocol in networks with heavy traffic demands, since it can accommodate higher traffic demands and do not suffer from cumulative delay. However, in situations where the channel is sparsely used, TDMA is inefficient.

As an alternate optical multiplexing technique, there is wavelength-division multiple access (WDMA). The term WDMA rather than the popular wavelength-division multiplexing (WDM) is used to indicate the access, routing and switching functionality in addition to the transmission multiplex. Tuned lasers are used as the optical source for each transmitter, and the modulated data are transmitted within its assigned band. At the receiver, an optical filter is tuned to the desired band, while all others are filtered out. A photodetector and a decoder are followed to obtain the data. WDMA technique partitions the available spectrum to different users, and offers a means of increasing capacity at minimal cost within the existing optical fiber infrastructure. The fundamental disadvantage in WDMA is that sophisticated hardware such as wavelength-controlled tunable lasers and high-quality narrowband tunable filters for each channel is required. Although WDMA can be used as a degree of design freedom with respect to routing and wavelength selection, the usable wavelength might be limited because of the crosstalk caused by the nonlinearity within the optical fiber. Wavelength routing can offer the switching function for dense WDMA networks; however, it may cause the crosstalk problem in the cross-connects based on space and wavelength, and thus, network reliability and flexibility are restricted.

Code-division multiple-access (CDMA) is a multiple access protocol that is efficient with low traffic and has zero access delay. Especially, direct detection optical CDMA

systems have been investigated widely to apply for high-speed LAN, because they allow multiple users to access the network simultaneously. In the case of data transfer where traffic tends to be bursty rather than continuous, CDMA can be used for contention-free, zero delay access. Compared with TDMA, CDMA is attractive in other points. Channel assignment is much easier with CDMA. CDMA isolates irregular channels so that they do not influence other channels, while with TDMA, even one irregular channel, such as continuous emission from a transmitter, causes the failure of all other channels.

In optical CDMA, incoherent systems using narrow pulse laser sources are mainly implemented, since optical links have vast bandwidth and the optical components can produce very narrow pulse precisely in time and offer extrahigh optical signal processing. In the transmitter and receiver, low-cost devices with high cost performance and high reliability, such as Fabry–Perot laser diode and avalanche and pin photodiodes, are available. Thus, in optical CDMA, intensity modulation/direct detection (IM/DD) is mainly used. In IM/DD systems, other arriving pulse sequences having positive pulses happen to overlap a pulse of the desired sequence, and produce correlation crosstalk. In optical CDMA, multiple user interference called multiple-access interference (MAI) is dominant compared to photodetector shot noise, dark current and thermal noise. Thus, the elimination or suppression of MAI is the key issue in optical CDMA. Most published optical CDMA systems are based on discrimination in the time domain to reduce the effects of pulse overlaps. This time-encoding process is most commonly implemented by encoding each data bit with a high-rate sequence; that is, a pulse laser source is intensity-modulated by electrical (0,1) data bit and the narrow pulse is emitted in the first chip in a slot. Here, data are usually modulated in on/off keying (OOK) or pulse position modulation (PPM) formats, and a slot is divided into chips where the number of chips in a slot equals to the length of the spreading code consisting of 1 and 0 allocated for users. Then, in the time encoder, a narrow pulse is time-spread into several chips within the slot according to each user's unipolar signature code. Thus, the time-encoding process relies on a simple, intensity-based, pulse time addressing process, and the sequence encoder and decoder in the time domain can be easily and cost-effectively implemented by using tapped optical delay lines. At the receiver, optical incoherent direct detection is done by an optical delay-line decoder matched to the encoder at the transmitter. After decoding, unwanted signals are time-spread over much larger time intervals than is the desired user's signal, and the crosstalk from adjacent chips are rejected to some extent by this time-despreading process.

For optical CDMA systems, both asynchronous and synchronous systems have been studied. In an asynchronous optical CDMA system, the synchronization among users is not required, and optical orthogonal codes (OOC) with good correlation properties [1] are widely used. However, in asynchronous CDMA systems, the available number of signature sequence codes is very small; hence the number of users is very limited. To solve this problem, synchronous CDMA systems, in which all users are synchronized in frame, is considered. With synchronous CDMA, the available number of signature codes is larger than that of asynchronous CDMA systems, because the same code can be reused with different phases. The modified prime sequence codes are known as typical "signature codes for optical synchronous CDMA," in which time-shifted versions of the prime sequence code can be used [2]. The cross-correlation peak between two time-shifted versions of the sequence code is as high as the autocorrelation peak; however, it always occurs either delayed or ahead of the autocorrelation peak. Since in the synchronized CDMA the receiver can be synchronized to the expected position of the autocorrelation peak, the autocorrelation peak can be distinguished from adjacent cross-correlation peaks. For a given value of bit error rate, synchronous CDMA systems can accommodate more simultaneous users than asynchronous CDMA systems. Furthermore, synchronous CDMA can be efficiently used in conjunction with TDMA and WDMA on multimedia communication networks where multiple services with different traffic requirements are to be integrated.

In Section 2, a family of good optical unipolar pseudoorthogonal (non-zero-cross-correlation) codes suitable for optimal CDMA IM/DD system with OOK and PPM signaling is described. In Section 3, the IM/DD systems with OOK and PPM signaling are described as typical optical synchronous CDMA systems. Also, as an alternative optical CDMA system, a frequency-encoded spread-time optical CDMA system utilizing bipolar codes is briefly introduced. Since the performance of the optical CDMA is degraded by the multiple-user interference, the interference cancellation is the key to realize the practical optical CDMA system. In Section 4, we describe two typical interference cancellation techniques for optical synchronous CDMA, based on the use of properties of modified prime sequence codes and optical hard-limiter.

2. SEQUENCES FOR OPTICAL SYNCHRONOUS CDMA

Many classes of binary signature sequences that are suitable for radio CDMA have been studied. In most of these codes, a strong autocorrelation peak and zero-cross-correlation function can be obtained through bipolar $(-1,+1)$ sequences. However, optical IM/DD systems can only accommodate unipolar $(0,+1)$ sequences, since incoherent systems use only positive narrow pulses emitted from laser sources. Therefore, codes intended for communication systems in which both positive and negative levels are available, are not necessarily optimal in a fiberoptic environment using optical signal processing. Compared to conventional electronic bipolar $(-1,+1)$ codes such as maximum-length sequence codes and Gold codes, the cross-correlation function of unipolar codes is high and the number of codes in the family is very low. The minimum value of the cross-correlation that unipolar codes can achieve is limited to be one, since at least one pulse is overlapped for asynchronous unipolar sequences. Thus, sets of sequences having no more than one pulse overlap in the pairwise cross-correlation are often called as *pseudoorthogonal codes* in optical CDMA.

The characteristics needed for the unipolar codes suitable for optical CDMA are good autocorrelation and cross-correlation properties. Sharp autocorrelation property is needed to achieve the synchronization as well as the discrimination between the time-shifted sequence codes in optical synchronous CDMA where the time-shifted codes are assigned to other users. The minimum value of the cross-correlation function should be as small as possible to discriminate between the desired user and others as well as to mitigate multiple user interference. In optical CDMA IM/DD systems, other arriving pulse sequences having positive pulses that happen to overlap a pulse of the desired sequence produce correlation crosstalk called *multiple access interference* (MAI), which might degrade the decoding performance. The code length and weight (the number of "1"s) also affect the system performance. Long codes and sparse codes comprising very few ones and narrow pulses are preferred to support a large number of users and higher transmission capabilities, respectively. On the other hand, short codes and a large weight are preferred to increase data rate and improve signal-to-interference ratio, respectively.

In this section, as good optical unipolar codes suitable for optical CDMA IM/DD system with OOK and PPM signaling, the prime code family is described. Especially for optical synchronous CDMA systems, the family of modified prime sequences is known to have good correlation properties.

2.1. Prime Code

Prime codes are defined as follows [3]: from the Galois field $\mathrm{GF}(P) = \{0, 1, 2, \ldots, P-1\}$, where P is a prime number, a set of P prime sequences $\{S_i^P : i = 0 \text{ to } P-1, \text{ and } j = 0 \text{ to } P-1\}$ is first generated, each with P elements given by

$$S_i^P(j) = [i,j]_P \quad \text{for } i = 0 \text{ to } P-1, \text{ and } j = 0 \text{ to } P-1 \tag{1}$$

where $[\ ]_P$ denotes reduction modulo P. For example, the family of prime sequences for $P = 7$ is shown in Table 1. For each sequence S_i^P, a binary code C_i^P of length P^2 chips is then constructed using the following rules:

$$C_i^P(n) = \begin{cases} 1, & \text{for } n = jP + S_i^P(j) \quad \text{for } j = 0 \text{ to } P-1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

This requires that the code C_i^P be divided into P frames, each consisting of P chips. Within the jth frame, the chip shifted relative to the start of the frame by $S_i^P(j)$ is a "1"; all other chips are zero. The code C_i^P is therefore a time-mapped, binary version of the sequence S_i^P. The set of prime codes for $P = 7$ is shown in Table 2, where the frames have been slightly separated for clarity.

Table 1. Prime Sequences for $P = 7$

	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
Sequence S_0^7	0	0	0	0	0	0	0
Sequence S_1^7	0	1	2	3	4	5	6
Sequence S_2^7	0	2	4	6	1	3	5
Sequence S_3^7	0	3	6	2	5	1	4
Sequence S_4^7	0	4	1	5	2	6	3
Sequence S_5^7	0	5	3	1	6	4	2
Sequence S_6^7	0	6	5	4	3	2	1

The correlation functions arising in an IM/DD system, assuming on/off keyed data, are the aperiodic and periodic correlation functions, $C_{i,j}(l)$ and $\Theta_{i,j}(l)$, which are defined respectively as follows:

$$C_{i,j}(l) = \sum_{n=0}^{L-1} C_i^P(n) \cdot C_j^P(n+l) \tag{3}$$

$$\Theta_{i,j}(l) = \sum_{n=0}^{L-1} C_i^P(n) \cdot C_j^P([n+l]_L) = C_{i,j}([l]_L) + C_{i,j}([l]_L - L) \tag{4}$$

where $L = P^2$, $[\ ]_L$ denotes reduction modulo L, and $C_i^P(n) = 0$ for $n < 0$ and $n \geq L$, for all i. These are autocorrelation functions when $i = j$, and cross-correlation functions when $i \neq j$. The aperiodic form is generated at the matched filter output by an isolated "1" in the incoming datastream, while the periodic form is generated by adjacent "1"s; incoming "0"s produce no response. This periodic correlation is simply the number of positions where C_i^P and a cyclically shifted version of C_j^P both have "1"s. This means that the autocorrelation peak for any code (which occurs for $l = 0$, $i = j$) is equal to the number of "1"s it contains, or P in the case of a prime code:

$$\Theta_{i,j}(0) = C_{i,j}(0) = P \quad \text{for all } i \tag{5}$$

Note that the maximum number of coincidences of "1"s between two distinct prime codes C_i^P and C_j^P, having any

Table 2. Prime Codes for $P = 7$

	Frame 0	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6
Code C_0^7	1000000	1000000	1000000	1000000	1000000	1000000	1000000
Code C_1^7	1000000	0100000	0010000	0001000	0000100	0000010	0000001
Code C_2^7	1000000	0010000	0000100	0000001	0100000	0001000	0000010
Code C_3^7	1000000	0001000	0000001	0010000	0000010	0100000	0000100
Code C_4^7	1000000	0000100	0100000	0000010	0010000	0000001	0001000
Code C_5^7	1000000	0000010	0001000	0100000	0000001	0000100	0010000
Code C_6^7	1000000	0000001	0000010	0000100	0001000	0010000	0100000

relative shift, is 2, so that all periodic cross-correlation functions are bounded by

$$\Theta_{i,j}(0) \le 2 \quad \text{for all } l, \text{ and all } i, j \text{ such that } i \ne j \tag{6}$$

From Eq. (4), it is clear that $\Theta_{i,j}(l) \ge C_{i,j}(l)$ for all l (note that Θ and C are both positive functions); thus the above bound also applies to any interference contribution, regardless of the data it carries.

2.2. Quasiprime Code

Quasiprime codes are derived from prime codes as follows [4]: with the prime code C_i^P, a set of cyclically shifted versions of the code $\{C_{ik}^{SP} : k = 0 \text{ to } P-1\}$ is defined where C_{ik}^{SP} is obtained by cyclically shifting C_i^P left by k complete frames. The elements of C_{ik}^{SP} are thus given by

$$C_{ik}^{SP}(n) = C_i^P([n + kP]_L) \qquad \text{for } n = 0 \text{ to } P^2 - 1. \tag{7}$$

For each shifted code C_{ik}^{SP}, a quasiprime code C_{ik}^{SP} may then be defined for any positive integer Q, where

$$\begin{aligned} C_{ik}^{QP}(n) &= C_{ik}^{SP}([n]_L) \quad \text{for } n = 0 \text{ to } QP - 1 \\ &= C_i^P([n + kP]_L) \end{aligned} \tag{8}$$

Each quasi-prime code then comprises QP chips taken cyclically from a shifted prime code, and contains Q "1"s. For example, Table 3 shows two of the seven quasiprime codes in the set $\{C_{2k}^{87} : k = 0 \text{ to } 6\}$ together with prime code C_2^7. Both are derived from the same prime code C_2^7.

We have to take care to define correlation functions for quasiprime codes, because truncating or extending the shifted prime codes destroys the periodicity of the basic prime code, and this must be restored if the quasiprime codes are to show good periodic cross-correlation properties. Accordingly, when the length QP of the quasiprime codes lies between $\Lambda - 1$ and Λ lengths of the original prime codes, it is made up to Λ lengths by packing each code with "0"s. The aperiodic and periodic correlation functions are then defined as

$$C_{i,j}(l) = \sum_{n=0}^{\Lambda L - 1} C_{ik}^{QP}(n) \cdot C_{jl}^{QP}(n + l) \tag{9}$$

$$\Theta_{i,j}(l) = \sum_{n=0}^{\Lambda L - 1} C_{ik}^{QP}(n) \cdot C_{jl}^{QP}([n + l]_{\Lambda L}) \tag{10}$$

where $(\Lambda - 1)L < QP \le \Lambda L$, $[\ ]_{\Lambda L}$ denotes reduction modulo ΛL, and $C_{ik}^{QP}(n) = 0$ for $n < 0$ and $n \ge QP$, for all i, k.

Different quasiprime codes derived from the same prime code cannot act as distinct, orthogonal members of an asynchronous CDMA code set. This means that a quasiprime code set can contain a maximum of P codes (one for each code in the original prime code set). The correlation properties of such a set are as follows: first, each code contains Q "1"s, so that the autocorrelation peak is given by

$$\Theta_{i,j}(0) = C_{i,j}(0) = Q \quad \text{for all } i \tag{11}$$

In addition the periodic cross-correlation $\Theta_{i,j}$ between two distinct quasiprime codes C_{ik}^{QP} and C_{jl}^{QP} is bounded by

$$\Theta_{i,j}(l) \le 2\Lambda \quad \text{for all } l, \text{ and all } i, j, k, l \text{ such that } i \ne j \tag{12}$$

Considering (11) and (12), the interference probability obtained with quasiprime codes might be expected to be worse than that obtained with prime codes in general. For example, the number of interfering signals that can be accommodated without error is expected to be $\lfloor Q/2\Lambda \rfloor$, where $\lfloor x \rfloor$ is the integer part of x, and this is less than or equal to the prime code result $\lfloor P/2 \rfloor$. In fact, this is pessimistic, because when QP is only slightly greater than $(\Lambda - 1)L$, the bound $\Theta_{i,j}(l) \le 2(\Lambda - 1)$ can still apply, and in such cases the quasiprime codes would be able to achieve better interference probability.

2.3. 2^n Prime Code

Usually, the 2^n codes are defined as collections of binary N-tuples with weight 2^n [5]. Using the serial optical encoders, the distribution of the pulses in each generated codeword must be symmetric (i.e., the distribution of the current 2^m pulses highly depends on that of the previous 2^{m-1} pulses, where $1 < m \le n$) and results in a very restrictive pulse distribution constraint. Alternatively, it is sometimes more convenient to represent the pulse distribution in terms of delay distribution. Therefore, the constraint can be equivalently presented as a delay distribution constraint.

The delay distribution constraint functions as follows. For a given integer n, integers x, y, and z are assumed such that $x \ne y$, $0 \le x \le 2^n - 2$, $0 \le y \le 2^n - 2$, and $1 \le z \le n - 1$. If both x and y are divisible by 2^z, then adjacent relative cyclic delays $[t_0, t_1, \ldots, t_q, \ldots, t_{2^n-1}]$ of each codeword of the 2^n codes are related such that

$$t_{x \uplus (2^{z-1}-1) \uplus m} = t_{y \uplus (2^{z-1}-1) \uplus m} \tag{13}$$

for a given integer $m \in [0, 2^n - 1]$, where "$\uplus$" represents modulo-2^n addition. t_q denotes the adjacent relative cyclic delay (or simply the separation in chips) between the qth

Table 3. Quasiprime Codes for $Q = 8$, $P = 7$ Based on C_2^7

Code	Frame 0	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7
C_2^7	1000000	0010000	0000100	0000001	0100000	0001000	0000010	
C_{20}^{87}	1000000	0010000	0000100	0000001	0100000	0001000	0000010	1000000
C_{25}^{87}	0001000	0000010	1000000	0010000	0000100	0000001	0100000	0001000

and $(q+1)$th pulses. For the last delay t_{2^n-1}, the codeword is wrapped around to obtain the separation between the last and first pulses.

Example 1. Assuming that $n = 2$ (i.e., code weight of 4), we get $0 \le x \le 2$, $0 \le y \le 2$, $z = 1$, and $m \in [0, 3]$ from the above mentioned delay distribution constraint. The adjacent relative cyclic delays $[t_0, t_1, t_2, t_3]$ are related such that $t_0 = t_2$ for $m = \{0, 2\}$, or $t_1 = t_3$ for $m = \{1, 3\}$. Using sequence codes 100001010000100, 110000100010000, and 100010000100100 as examples, their corresponding adjacent relative cyclic delays are [5,2,5,3], [1,5,4,5], and [4,5,3,3], respectively. On the basis of the constraint, the first two are valid sequence codes, while the last one is not.

According to the symmetric property described above, the algebraic construction on the 2^n prime-sequence codes begins with Galois field $\mathrm{GF}(P) = \{0, 1, \ldots, P-1\}$ of a prime number P. As an example, the prime sequences in GF(13) are shown in Table 4 in a form different from that in Table 1. By inspection, the adjacent relative cyclic delays of the codeword generated by S_i^P can be found according to

$$t_j = \begin{cases} S_i^P(j+1) - S_i^P(j) + P, & \text{for } j = \{0, 1, \ldots, P-2\} \\ S_i^P(0) - S_i^P(j) + P, & \text{for } j = P-1 \end{cases} \tag{14}$$

for $i \in GF(P)$. Table 5 shows the adjacent relative cyclic delays for the prime sequences in GF(13).

The adjacent relative cyclic delays for each prime sequence S_i^P are then determined depending on whether the delay-distribution constraint is satisfied. If the constraint is satisfied, the prime sequence S_i^P will be modified: the elements $S_i^P(j)$ and $S_i^P(j+1)$ whose relative cyclic delay t_j satisfies the constraint are kept unchanged, while the remaining elements are replaced by Xs. Note that every X in the replaced prime sequences is simply mapped to P zeros. However, this S_i^P will be discarded if none of the delays satisfies the constraint.

Example 2. Using $2^n = 8$ and $i = 3$ as an example, the adjacent relative cyclic delays for S_3^{13} are [16, 16, 16, **16**, **3**, **16**, **16**, **16**, **3**, **16**, 16, 16, 3] as shown in Table 5. Those delays that satisfy the delay distribution constraint are boldfaced. From (13), S_3^{13} satisfies the two conditions $t_m = t_{m+2} = t_{m+4} = t_{m+6}$ and $t_{m+1} = t_{m+5}$ with $m = 3$. The remaining elements $S_3^{13}(0)$, $S_3^{13}(1)$, $S_3^{13}(2)$, $S_3^{13}(11)$, and $S_3^{13}(12)$ are then replaced by Xs as shown in Table 6, where the replaced prime sequences for $P = 13$ and $2^n = 8$ are tabulated. Note that the prime sequences S_5^{13} and S_8^{13} are discarded since the delay distribution constraint cannot be satisfied.

Finally, the codewords of the 2^n prime sequence codes are generated by mapping each replaced prime sequence S_i^P into a binary code sequence $C_i^P = (C_i^P(0), C_i^P(1), \ldots, C_i^P(k), \ldots, C_i^P(N-1))$ of length $N = P^2$ according to

$$C_i^P(k) = \begin{cases} 0, & \text{for } k = S_i^P(j) + jP \text{ and } S_i^P(j) \ne \mathrm{X} \\ 1, & \text{otherwise} \end{cases} \tag{15}$$

for $i, j \in \mathrm{GF}(P)$ and $k = \{0, 1, \ldots, N-1\}$.

2.4. Modified Prime Code

For prime sequence codes of length P^2, the number of sequence codes is limited to P; therefore, so is the number of total subscribers. In order to generate more sequence

Table 4. Prime Sequences in GF(13)

	j												
i	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9	10	11	12
2	0	2	4	6	8	10	12	1	3	5	7	9	11
3	0	3	6	9	12	2	5	8	11	1	4	7	10
4	0	4	8	12	3	7	11	2	6	10	1	5	9
5	0	5	10	2	7	12	4	9	1	6	11	3	8
6	0	6	12	5	11	4	10	3	9	2	8	1	7
7	0	7	1	8	2	9	3	10	4	11	5	12	6
8	0	8	3	11	6	1	9	4	12	7	2	10	5
9	0	9	5	1	10	6	2	11	7	3	12	8	4
10	0	10	7	4	1	11	8	5	2	12	9	6	3
11	0	11	9	7	5	3	1	12	10	8	6	4	2
12	0	12	11	10	9	8	7	6	5	4	3	2	1

Table 5. Adjacent Relative Cyclic Delays for the Prime Sequences in GF(13)

	j												
i	0	1	2	3	4	5	6	7	8	9	10	11	12
0	13	13	13	13	13	13	13	13	13	13	13	13	13
1	14	14	14	14	14	14	14	14	14	14	14	14	1
2	15	15	15	15	15	15	2	15	15	15	15	15	2
3	16	16	16	16	3	16	16	16	3	16	16	16	3
4	17	17	17	4	17	17	4	17	17	4	17	17	4
5	18	18	5	18	18	5	18	5	18	18	5	18	5
6	19	19	6	19	6	19	6	19	6	19	6	19	6
7	20	7	20	7	20	7	20	7	20	7	20	7	7
8	21	8	21	8	8	21	8	21	8	8	21	8	8
9	22	9	9	22	9	9	22	9	9	22	9	9	9
10	23	10	10	10	23	10	10	10	23	10	10	10	10
11	24	11	11	11	11	11	24	11	11	11	11	11	11
12	25	12	12	12	12	12	12	12	12	12	12	12	12

Table 6. Replaced Prime Sequences in GF(13) with $2^n = 8$

	j												
i	0	1	2	3	4	5	6	7	8	9	10	11	12
0	X	X	X	0	0	0	0	0	0	0	0	X	X
1	X	X	X	3	4	5	6	7	8	9	10	X	X
2	X	X	X	6	8	10	12	1	3	5	7	X	X
3	X	X	X	9	12	2	5	8	11	1	4	X	X
4	0	4	8	12	3	7	X	X	X	X	X	5	9
5													
6	X	X	X	5	11	4	10	3	9	2	8	X	X
7	X	X	X	8	2	9	3	10	4	11	5	X	X
8													
9	0	9	5	X	X	X	X	X	7	3	12	8	4
10	X	X	X	4	1	11	8	5	2	12	9	X	X
11	X	X	X	7	5	3	1	12	10	8	6	X	X
12	X	X	X	10	9	8	7	6	5	4	3	X	X

codes for the same length, that is, the same bandwidth expansion, at the expense of requiring synchronization among users, modified prime sequence codes have been proposed [2]. Modified prime sequence codes are time-shifted versions of prime sequence codes. Each original P prime sequence S_x^P is taken as a seed from which a group of new sequence codes can be generated. The sequence codes of the first group (i.e., $x = \{0\}$) are obtained by left-rotating the prime sequence code C_0^P. C_0^P can be left-rotated $P-1$ times before being recovered, so that $P-1$ new sequence codes can be generated from C_0^P. For the other $P-1$ groups (i.e., $x = \{1, \ldots, P-1\}$), the elements of the corresponding prime sequence S_x^P can be left-rotated $P-1$ times to create new prime sequences $S_{x,r}^P = (S_{x,r}^P(0), S_{x,r}^P(1), \ldots, S_{x,r}^P(P-1))$, where r represents the number of times S_x^P has been left-rotated. Therefore, P prime sequences per group are obtained. Each prime sequence $S_{x,r}^P$ is then mapped into a binary sequence code $C_{x,r}^P = (C_{x,r}^P(0), C_{x,r}^P(1), \ldots, C_{x,r}^P(i), \ldots, C_{x,r}^P(P^2-1))$ according to

$$\theta_{x,r}^P(i) = \begin{cases} 1, & \text{for } i = S_{x,r}^P(j) + jP, j = 0, 1, \ldots, P-1 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

The set of prime sequence S_x^P and their associated sequence codes C_x^P for GP(5) are tabulated in Table 7. The set of new prime sequences $S_{x,r}^P$ and their associated sequence codes $C_{x,r}^P$ for $GF(5)$ are tabulated in Table 8. Note that each new sequence code has P binary "1"s. Considering all groups, the total number of modified prime sequence codes is P^2. For a synchronous system, the cross-correlation between the modified prime sequence codes of the xth and the yth users can be written as [2]

$$\theta_{x,y} = \begin{cases} P, & x = y, \\ 0, & x \text{ and } y \text{ are in the same group} \\ 1, & x \text{ and } y \text{ are in the different groups} \end{cases} \tag{17}$$

The modified prime code has unique characteristics in that there is no correlation between any two users among the same group and the interference from other groups has the equal effect on the user in the same group. On the other hand, the optical orthogonal code (OOC) has the problem that it has to set the weights and sequence code independently and keep the number of spreading codes small to attain good correlation properties. Table 9 shows the OOCs with the code length $F = 32$ and weight $K = 4$. The total number of OOCs is given by the integer part of $(F-1)/(K^2-K)$, and there are only two codes when $K = 4$ and $F = 32$, because the maximum value of the cross-correlation should be 1. To satisfy this condition, the distance of any two "1"s should be different in all codes as shown in Table 9. Therefore, to increase the number of spreading codes in OOC, the frame length has to be increased or weights have to be smaller. In practice, in order to make 25 spreading codes with weight 5, the sequence code length needs to be 501. Therefore, OOC needs 20 times larger frame length compared to the modified prime sequence code, and the bit rate decreases when OOC is used. When the weights are decreased in OOC, the correlation property is degraded. Therefore, the modified prime sequence code whose weights are the same as those of OOC is more effective in a synchronous optical CDMA, because the modified prime sequence code can make more spreading codes with a shorter frame length.

Table 7. Prime Sequences S_x^P and Prime Sequence Codes C_x^P for GF(5)

	i		
x	01234	Sequence	Code
0	00000	S_0^5	$C_0^5 = 10000\ 10000\ 10000\ 10000\ 10000$
1	01234	S_1^5	$C_1^5 = 10000\ 01000\ 00100\ 00010\ 00001$
2	02413	S_2^5	$C_2^5 = 10000\ 00100\ 00001\ 01000\ 00010$
3	03142	S_3^5	$C_3^5 = 10000\ 00010\ 01000\ 00001\ 00100$
4	04321	S_4^5	$C_4^5 = 10000\ 00001\ 00010\ 00100\ 01000$

3. SYSTEM MODEL OF OPTICAL SYNCHRONOUS CDMA

3.1. Optical Time-Encoded CDMA Systems

In optical CDMA, intensity modulation/direct detection is used mainly in conjunction with on/off keying (OOK) and pulse position modulation (PPM) signaling formats. A pulse laser source is intensity-modulated by electrical (0,1) data bits. Data are modulated by emitting optical positive pulses at the first chips of the slots in OOK or PPM signaling. In optical time-encoded CDMA, a slot is divided into chips and the number of chips in a slot equals the spreading code length. In optical synchronous CDMA systems, synchronization between transmitter and receiver is required, and the synchronization is achieved by adding to the correlated signal the receiver clock signal delayed by a proper amount. The delay must be such that the peak of the autocorrelation function coincides with the optical clock pulse. Code synchronization can be realized through a two-stage process: a coarse alignment referred to as *code acquisition* and a subsequent fine alignment referred to as *code tracking*.

In OOK, a "1" information bit is transmitted by emitting a optical pulse at the first chip of the slot. When no pulse is emitted in chips within a slot, this means that "0" information bit is transmitted. Thus, one bit binary information is conveyed in a slot. At the decoder, threshold detection is used in OOK. Since the threshold value depends on the intensity level of received signal pulse, multiple-user interference, and noise, proper adjustment of the threshold level is required. In an M-ary PPM, a narrow pulse is emitted at the first chip of one of M slots in a PPM frame to represent data. Since the combination of selecting one slot among M slots in a frame is $\log_2 M$, $\log_2 M$ bits can be conveyed in a frame. This results in a low channel traffic in PPM compared to OOK in terms of the number of transmitted pulses, due to the pulse position multiplicity of PPM signaling. PPM can utilize maximum-likelihood detection in which the slot with largest intensity level in a frame is selected in maximum likelihood manner, and thus, no precise threshold adjustment is required.

Each user is assigned a signature sequence with length F, which serves as its address. In time encoding, the encoder consists of the tapped optical delay lines, and the

Table 8. Left-Rotated Prime Sequences $S^P_{x,r}$ and Modified Prime Sequence Codes $C^P_{x,r}$ for GF(5)

Group x	i 01234	Sequence	Code
0	00000	$S^5_{0,0}$	$C^5_{0,0} = 10000\ 10000\ 10000\ 10000\ 10000$
	44444	$S^5_{0,1}$	$C^5_{0,1} = 00001\ 00001\ 00001\ 00001\ 00001$
	33333	$S^5_{0,2}$	$C^5_{0,2} = 00010\ 00010\ 00010\ 00010\ 00010$
	22222	$S^5_{0,3}$	$C^5_{0,3} = 00100\ 00100\ 00100\ 00100\ 00100$
	11111	$S^5_{0,4}$	$C^5_{0,4} = 01000\ 01000\ 01000\ 01000\ 01000$
1	01234	$S^5_{1,0}$	$C^5_{1,0} = 10000\ 01000\ 00100\ 00010\ 00001$
	12340	$S^5_{1,1}$	$C^5_{1,1} = 01000\ 00100\ 00010\ 00001\ 10000$
	23401	$S^5_{1,2}$	$C^5_{1,2} = 00100\ 00010\ 00001\ 10000\ 01000$
	34012	$S^5_{1,3}$	$C^5_{1,3} = 00010\ 00001\ 10000\ 01000\ 00100$
	40123	$S^5_{1,4}$	$C^5_{1,4} = 00001\ 10000\ 01000\ 00100\ 00010$
2	02413	$S^5_{2,0}$	$C^5_{2,0} = 10000\ 00100\ 00001\ 01000\ 00010$
	24130	$S^5_{2,1}$	$C^5_{2,1} = 00100\ 00001\ 01000\ 00010\ 10000$
	41302	$S^5_{2,2}$	$C^5_{2,2} = 00001\ 01000\ 00010\ 10000\ 00100$
	13024	$S^5_{2,3}$	$C^5_{2,3} = 01000\ 00010\ 10000\ 00100\ 00001$
	30241	$S^5_{2,4}$	$C^5_{2,4} = 00010\ 10000\ 00100\ 00001\ 01000$
3	03142	$S^5_{3,0}$	$C^5_{3,0} = 10000\ 00010\ 01000\ 00001\ 00100$
	31420	$S^5_{3,1}$	$C^5_{3,1} = 00010\ 01000\ 00001\ 00100\ 10000$
	14203	$S^5_{3,2}$	$C^5_{3,2} = 01000\ 00001\ 00100\ 10000\ 00010$
	42031	$S^5_{3,3}$	$C^5_{3,3} = 00001\ 00100\ 10000\ 00010\ 01000$
	20314	$S^5_{3,4}$	$C^5_{3,4} = 00100\ 10000\ 00010\ 01000\ 00001$
4	04321	$S^5_{4,0}$	$C^5_{4,0} = 10000\ 00001\ 00010\ 00100\ 01000$
	43210	$S^5_{4,1}$	$C^5_{4,1} = 00001\ 00010\ 00100\ 01000\ 10000$
	32104	$S^5_{4,2}$	$C^5_{4,2} = 00010\ 00100\ 01000\ 10000\ 00001$
	21043	$S^5_{4,3}$	$C^5_{4,3} = 00100\ 01000\ 10000\ 00001\ 00010$
	10432	$S^5_{4,4}$	$C^5_{4,4} = 01000\ 10000\ 00001\ 00010\ 00100$

Table 9. Optical Orthogonal Codes: $F = 32, K = 4$

Number of Chips between the Subsequent 1s	OOC
9,3,15,5	10000000010010000000000000010000
4,7,19,2	10001000000100000000000000000010

laser pulse emitted in the first chip in a slot is time-spread in F chips within a slot corresponding to "1"s of the spreading codes. That is, ON information is transmitted as a sequence of F chipped pulses and OFF information is sent as an all-zero sequence. Note that the repetition rate of the light source limits the transmission rate, because the light source has difficulty to generate long successive optical pulse transmissions in time. Thus, time encoding is useful because there is no successive optical pulse transmission in adjacent chips. Optical pulse sequences transmitted from users are combined in the star coupler and then transmitted over the fiber to the desired destination. At the receiver, an optical incoherent passive filter-matched detection is done by an optical delay-line decoder matched to the encoder at the transmitter. It produces a peak in the correlation output for the intended bits. Data bits are discriminated in the chip duration using a photodiode followed by a threshold process.

The "near/far" problem is an essential issue in most CDMA systems, especially in wireless applications. Fortunately, the transceivers in an optical fiber networks are fixed, and to a certain extent, the optical path loss between the transmitter and the receiver can be predicted. Hence, a gain-clamped preamplifier can be used to compensate for optical power loss, so that all the transmitted optical signals originating from different locations can be received at similar optical power levels.

In the following subsection, the IM/DD systems with OOK and PPM signaling are described as typical optical synchronous CDMA systems. Also, as an alternative optical CDMA system, an optical frequency-encoded CDMA system utilizing bipolar codes is briefly introduced.

3.1.1. Optical Synchronous OOK CDMA. Figure 1, shows the transmitter block diagram of a direct-detection

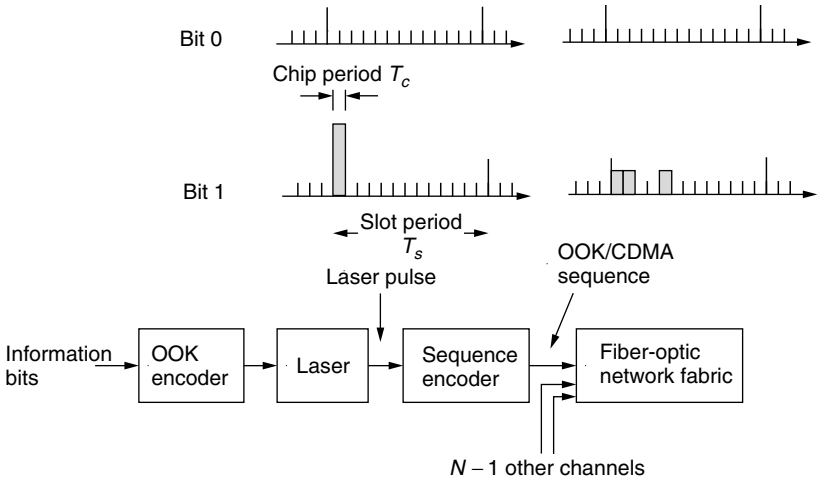


Figure 1. The transmitter block diagram of a direct-detection optical OOK CDMA system.

optical CDMA system. The output of the information source is fed into the OOK encoder where the information bitstream is directly converted into the OOK pulse sequences. When a logical "1" is to be conveyed, the laser is pulsed on at the first chip of the slot; when a logical "0" is to be conveyed, the laser is not pulsed on. Then the output laser pulse is converted into the assigned optical code sequence, that is, the signature sequence by a tapped optical delay line [6] shown in Fig. 2 that converts the initial laser pulse into a specific train of output pulses. When a logical "0" is to be conveyed, an all-zero sequence is transmitted. Light pulse sequences from all sources are combined in the fiber-optic network fabric and then transmitted over the fiber to the desired destination.

Figure 3 shows the receiver block diagram of the direct-detection optical OOK CDMA system. At the receiver, the matched optical correlator is used to recognize the arrival of the desired sequence. Figure 4 shows the optical correlator comprising a set of optical delay lines inversely matched to the pulse spacings. When the desired optical sequence passes through the correlator, the output light intensity traces out the correlation function of the sequence. At the last chip position, the sum of received optical intensity located in the same positions as the positions of "1" of the signature codes used for the desired channel is obtained. The correlator output is converted into an electrical signal by the photodetector and is then passed to the OOK decoder. The OOK decoder compares the correlator output voltage over the last chip position with the threshold level, then decides that "1" is sent if the output voltage is larger than the threshold level, and

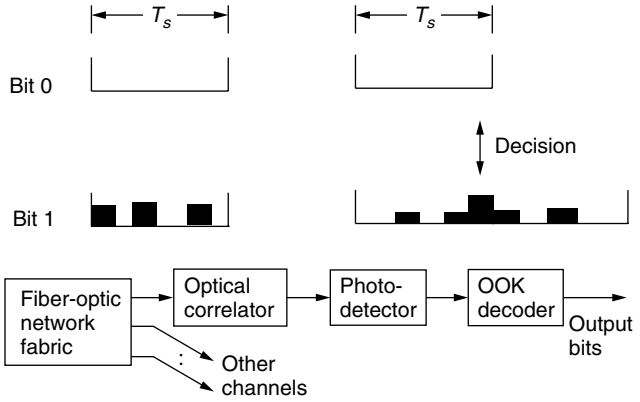


Figure 3. The receiver block diagram of a direct-detection optical OOK CDMA system.

"0" is sent otherwise. In this way each user can recover his own logical sequence.

3.1.2. Optical Synchronous PPM CDMA. Figure 5 shows the block diagram of an optical M-ary PPM CDMA transmitter and a signaling format. At the transmitter, a data bitstream is first blocked into words of length $\log_2 M$ bits in the PPM encoder and then each word is encoded into a M-ary PPM signaling format, that is, one of M slot positions. Every slot consists of F chips with the same chip period T_c, where F is the spreading factor of the CDMA signals. The laser is pulsed on at the first chip of the proper slot representing the word and the other slots

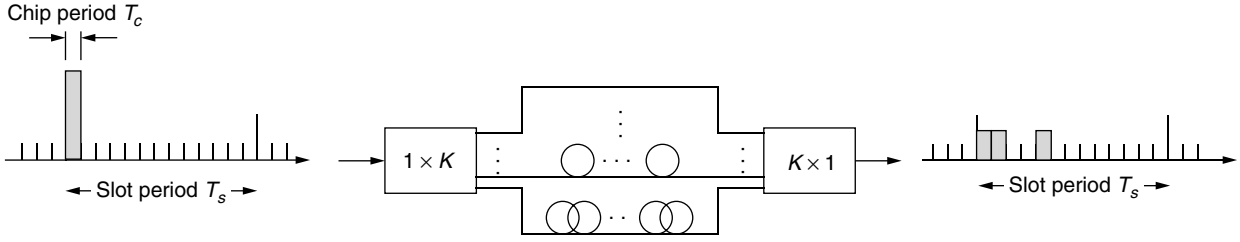


Figure 2. A sequence encoder consisting of tapped optical delay lines.

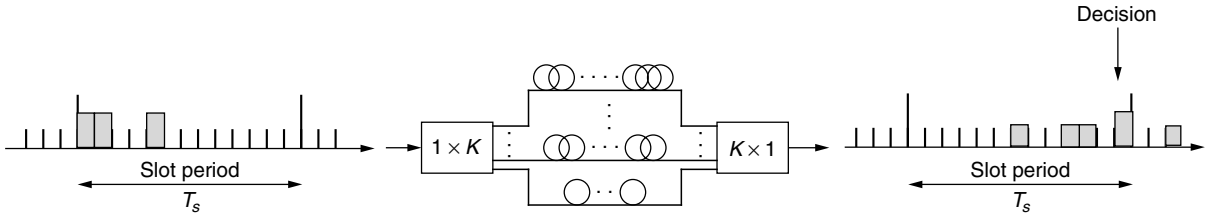


Figure 4. An optical correlator comprising tapped optical delay lines.

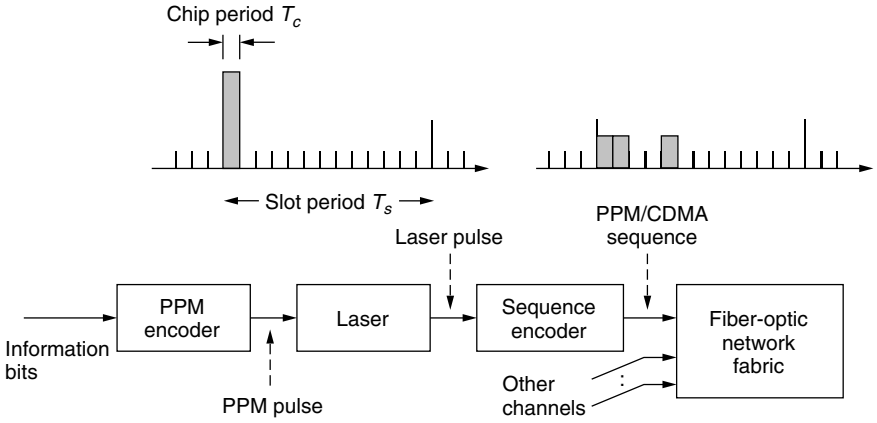


Figure 5. The transmitter block diagram of a direct-detection optical PPM CDMA system.

have no pulse. Then the output laser pulse is converted into the assigned optical code sequence, that is, the signature sequence by a set of tapped optical delay lines [6] that converts the initial laser pulse into a specific train of output pulses. The transmitted PPM CDMA signal in the pulsed slot have K pulses according to the assigned sequence code. Light pulse sequences from all the sources are combined in the fiberoptic network fabric and then transmitted over fiber to the desired destination.

The block diagram of an optical PPM CDMA receiver is shown in Fig. 6. At the receiver, a matched optical correlator is used to recognize the arrival of the desired sequence. The optical correlator is a set of optical delay lines inversely matched to the pulse spacings of the sequences. When the desired optical sequence passes through the correlator, the output light intensity traces out the correlation function of the sequence. In the photodetector the correlator output is converted into the electrical signal that is passed to the PPM decoder. The PPM decoder compares the output voltage over the last chips of all the slots and decides the slot having the highest voltage as the pulsed slot. Finally, the PPM decoder declares the corresponding word as the transmitted word.

There are two main advantages of synchronous M-ary PPM CDMA over OOK CDMA. The first advantage is that, under a bit error rate constraint, the maximum number

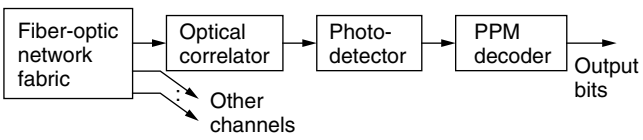


Figure 6. The receiver block diagram of a direct-detection optical PPM CDMA system.

of simultaneous users can be increased by increasing M and keeping the average power fixed. On the other hand, in the case of OOK CDMA, this number cannot be increased without increasing the average power. The second advantage is that any number of users can be accommodated by increasing M is the case of PPM CDMA. In the case of OOK CDMA, however, we may not be able to accommodate all the subscribers, even if the average power is increased. The reason is that, when the number of users is N, the average number of interfering optical pulses reduces to $(N-1)/M$ for PPM CDMA, whereas this number is equal to $(N-1)/2$ for OOK CDMA. Of course, these advantages of PPM CDMA over OOK CDMA are gained at the expense of increasing the system complexity [7].

3.2. Optical Frequency-Encoded CDMA System

In the previous two sections, we explained the time-encoded optical CDMA systems with OOK and PPM signaling and unipolar codes. Here, we introduce a frequency-encoded CDMA (FECDMA) system as an alternative optical CDMA using bipolar codes. Note that FE-CDMA can be used as both synchronous and asynchronous CDMA. The advantages of optical FECDMA are random-access, self-routing capability by the code itself, and the independence of the bit rate and processing gain, since coding and decoding are done by shifting phase without expanding the frequency band. In FECDMA, bipolar sequence codes such as pseudonoise (PN) sequences are assigned to each user, and the Fourier transform of the transmitted pulse for a given user is determined by encoding the phase of the desired transmitted spectrum by the user's PN sequence. The system model of the FE-CDMA is shown in Fig. 7. The FECDMA scheme is based on encoding and decoding of

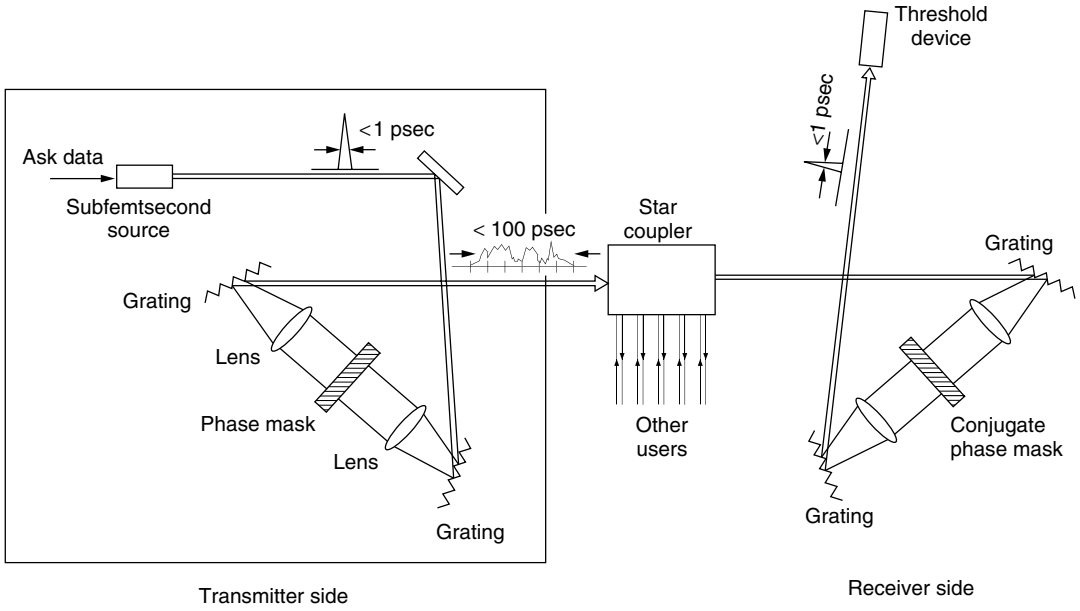


Figure 7. System model of FE-CDMA.

an ultrashort light pulse with duration τ_c and peak power P_0, and the operation is accomplished by using femtosecond pulseshapers. The pulseshapers offer high-resolution pulseshaping, programmability, and the flexibility to apply arbitrary phase codes of different code lengths. The operation of encoding and decoding is performed by using two fixed conjugate phase masks successively in the same pulse shaper. The liquid crystal modulator (LCM) is used to set the spectral phases to maximum-sequence phase. The LCM has a fully programmable linear array and individual pixels can be controlled by applying drive levels resulting in phase shifts (0 or π). By a phase mask, the dispersed bandwidth of a pulse is partitioned into N_c frequency chips, where each chip has the bandwidth W/N_c. Each chip is assigned a phase shift (i.e., 0 or π) depending on the user's PN sequence. The spectrum of the resulting pulse is reassembled by an inverse Fourier transform, and the encoded pulse is spread out within a time slot in synchronous CDMA as a low-intensity pseudonoise burst with average power P_0/N_c and duration T, as shown in Fig. 8. The transmitted data for a particular user can be recovered by sampling the output of a filter matched to the user's pulse. The phase mask and grating are implemented to perform the Fourier transform and matched filtering. At the receiver side, the spectral decoder consists of Fourier transformation of the time-windowed received signal followed by correlation with the PN sequence matched to the transmitter code. The spectral-phase code of the decoder is the complex conjugate of the encoder's spectral-phase code. The correctly decoded signal becomes a replica of the original short pulse with duration τ_c and peak power P_0, whereas the MAI signal remains a low intensity pseudonoise burst. The disadvantages of FECDMA are that the effects of MAI are large as the number of simultaneous user increases and a high-resolution phase mask as well as a narrow pulse are required to improve discrimination ability.

4. CHANNEL INTERFERENCE CANCELER FOR OPTICAL SYNCHRONOUS CDMA

Optical CDMA has several advantages over optical TDMA, including complete utilization of the entire time-frequency domain by each subscriber, flexibility in network design (because the quality depends on the number of active users), and security against interception. Synchronous CDMA has an additional advantage over asynchronous CDMA, where the number of available sequence codes (and in turn the number of subscribers) is much higher in the former under a given throughput constraint. The latter does not require, however, any time management as in the former. It follows that synchronous CDMA is suitable for very high speed networks with real time requirements (e.g., voice and digitized video). In contrast, asynchronous CDMA is suitable for bursty traffic with no stringent time requirements (e.g., data transmission). On the other hand, optical CDMA has a disadvantage over TDMA that is due to the multiple-user interference in the former. This leads in turn to a serious degradation in the bit error probability as the number of simultaneous users increases. This degradation cannot be overcome even for arbitrary high optical power. In fact, there will be an asymptotic error floor which limits the number of users that can communicate simultaneously and reliably. Several interference cancellation techniques have been proposed aiming at lowering these asymptotic error floors. These interference cancellation techniques are classified into two groups. One is based on the use of properties of modified prime sequence codes, and the other is based on the use of optical hard-limiters. The cancellation techniques using the properties of modified prime sequence codes have been proposed for both OOK CDMA and PPM CDMA systems [8–19]. These techniques estimate the amount of interference from a knowledge of some other users' sequence codes by using the correlation properties of modified prime sequence codes.

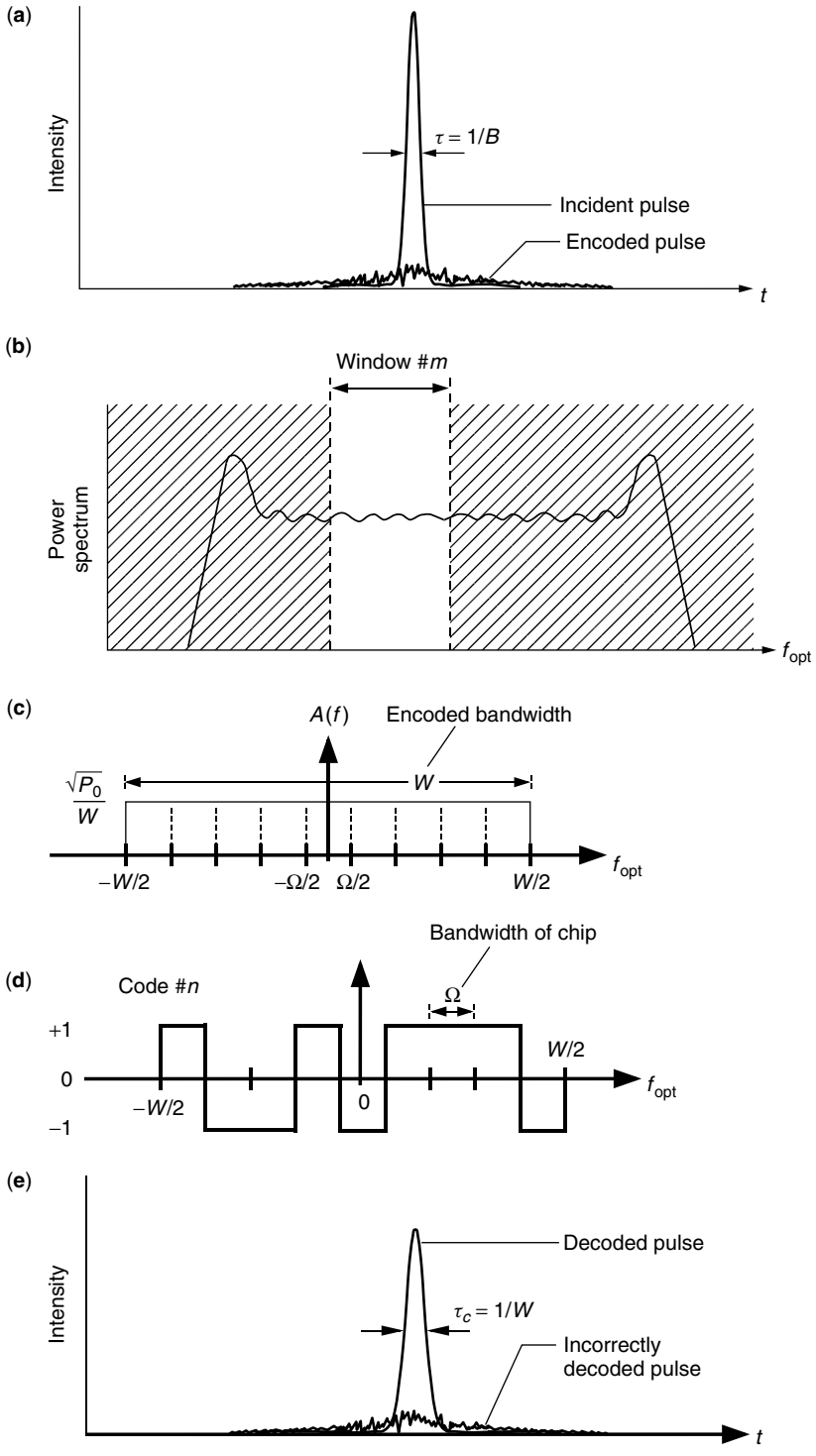


Figure 8. Frequency encoding.

As for the cancellation techniques based on the use of optical hard-limiters, Salehi and Brackett [20] used an optical hard-limiter that is placed before the optical correlator at the receiver side. This optical hard-limiter is shown to be able to remove some of the interference patterns. Ohtsuki et al. [21] proposed a synchronous optical CDMA system with double optical hard-limiters placed before and after the optical correlator. This system introduces an improvement in the performance over the system with a single optical hard-limiter as long as the number of users is not very large. In the case of asynchronous optical CDMA, Ohtsuki [22,23] showed that this improvement continues for all possible number of users. In Ohtsuki [24] was also able to reduce the error floor even lower than that of the system with double hard-limiters. Lin and Wu [25] suggested a synchronous optical CDMA system with an adaptive optical hard-limiter (or equivalently, a tunable optical attenuator) placed after

the correlator receiver. They were able to show that the performance can be improved as compared to the system with double hard-limiters.

We briefly review some cancellation techniques in the following sections.

4.1. Channel Interference Canceler Using Properties of Modified Prime Sequence Codes

We describe the channel interference canceller using a time-division reference signal [11,12] as an example of the channel interference cancellation technique using the properties of modified prime sequence codes.

Each user is assumed to be assigned a unique prime sequence code of length P^2 and weight P. In the system each user is allowed to access the network $P-1$ times out of P times, and $P-1$ users out of P users can access the network in each group simultaneously; that is, one user in each group is not allowed to access the network at each time; unallowable user's channel in each group at the time is used as a reference signal for other users in the same group at the time to cancel the effects of channel interference, because every code in the same group suffers the same amount of channel interference from other groups and every code in the same group does not interfere with each other.

Figure 9 shows an example of the access timing pattern for the first group in the system with the canceler where each user is not allowed to access the network at the marked time in the table. For instance, the first user is not allowed to access the network at time t_1, and the output of the first channel is used as a reference signal for other users in the same group at time t_1 to cancel the effects of channel interference. Notice that all the users in the same group suffer the same amount of channel interference from other groups. The bit rate of each user is thus

$$R_b = \frac{P-1}{PT} \tag{18}$$

At the same bit rate, the slot width of the proposed system is $(P-1)/P$ times as long as that of the system without canceler; thus the proposed system needs a slightly broader bandwidth.

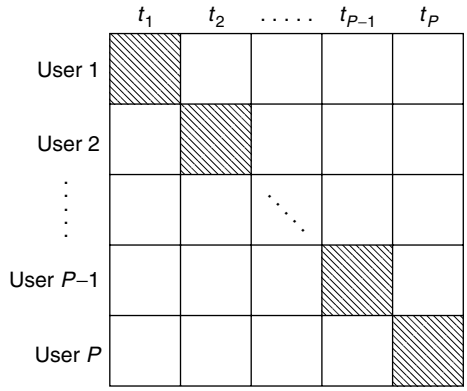


Figure 9. An example of the access timing pattern for the first group in the system with the canceler where each user is not allowed to access the network at the marked time in the table.

Figure 10 shows the receiver block diagram for the first user of the system. At the receiver, the matched optical correlators are used to recognize the arrival of the desired sequence. At the last chip position, the sum of the received optical intensity located in the same positions as the positions of "1" of the modified prime sequence code used for the desired channel is obtained. The correlator output is converted into an electrical signal by the photodetector. According to the table of the access timing pattern, the switch is connected to the reference channel that is not used at the time; the output of the reference signal is subtracted from the desired signal to cancel the effect of the channel interference. The signal after subtraction is passed to the OOK decoder. The OOK decoder compares the correlator output voltage over the last chip position with the threshold level and decides the data.

Figure 11 shows the bit error probability of OOK CDMA versus the received laser power P_W for some values of P where $N = P^2$, that is, the full-load case. As shown in Fig. 11, the received signal of each user is split into P

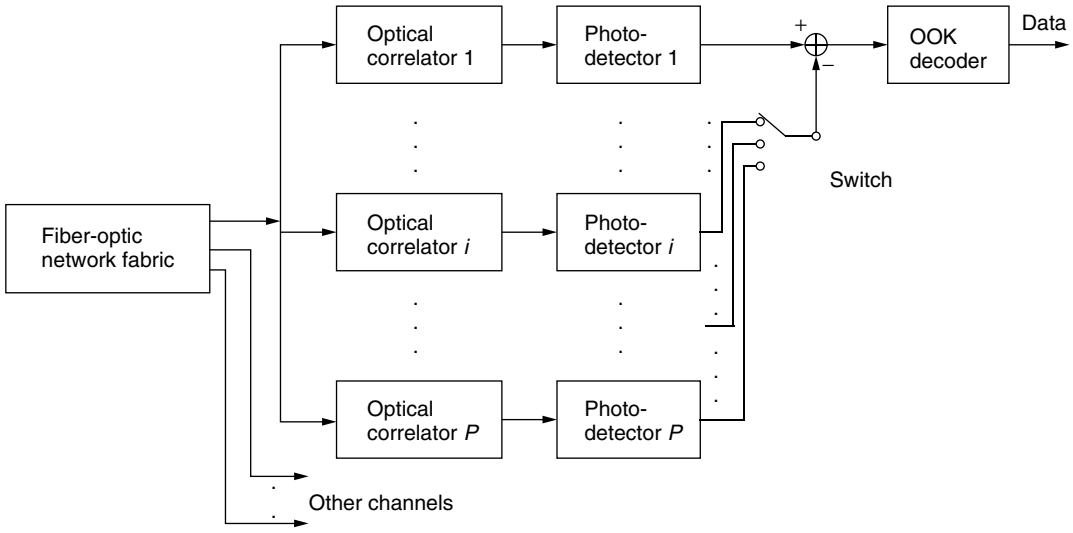


Figure 10. The receiver block diagram for the first user of the system with the canceler.

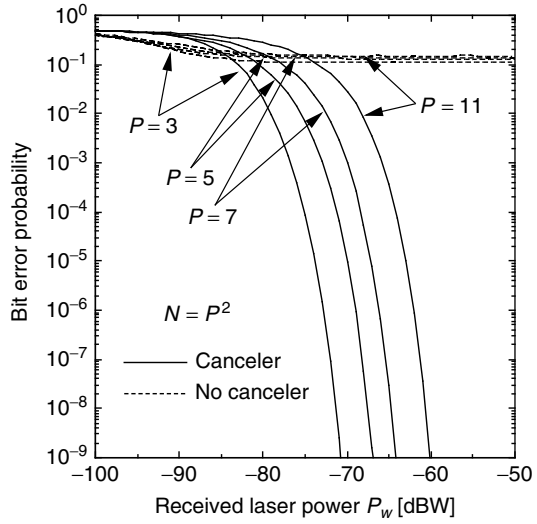


Figure 11. Bit error probability of OOK CDMA versus received laser power P_W for some values of P: $N = P^2$.

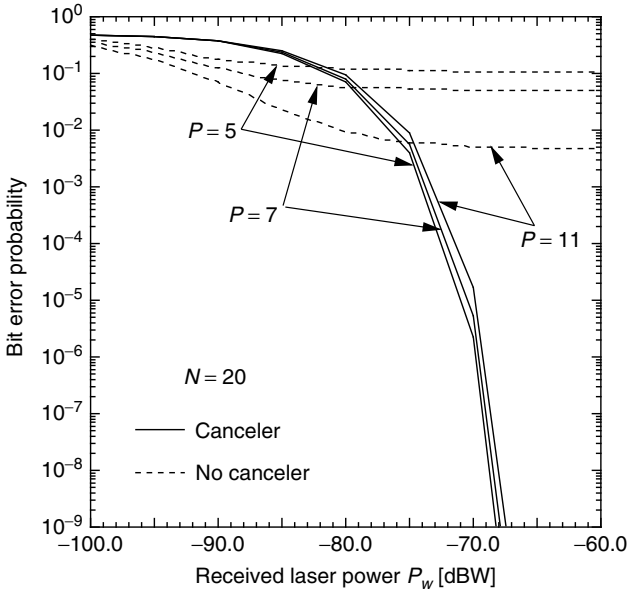


Figure 12. Bit error probability of OOK CDMA versus received laser power P_W for some values of P: $N = 20$.

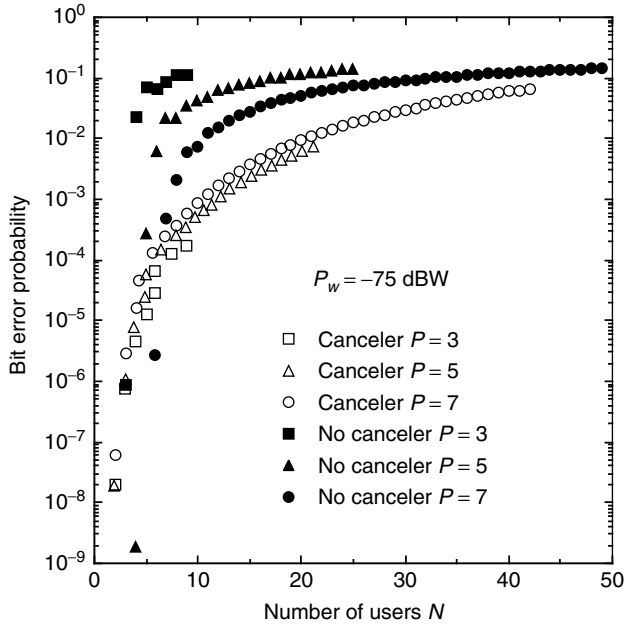


Figure 13. Bit error probability of OOK CDMA versus the number of users N for some values of P: $P_W = -75$ dBW.

branches at the receiver in the system with the canceler, and thus the unit received laser optical power in the delay-line of the optical correlator of the system with the canceler P_W is $1/P$ times as large as that of the system without the canceler. It can be seen that the system with the canceler has better performance than the conventional system without the canceler when P_W is not appreciably small; as P_W increases, the bit error probability of the system with the canceler is improved, while the error floor exists for the conventional systems without the canceler, because the effect of the channel interference is so large in the case of full load. Since the system with the canceler can reduce the effects of the channel interference, the error floor does not exist for the system with the canceler even in the case of full load. As P increases, the effect of the channel interference also increases in the case of full load, and thus the system with the canceler with larger P needs more power to have better performance.

Figure 12 shows the bit error probability of OOK CDMA versus the received laser power P_W for some values of P where $N = 20$. It can be seen that the system with the canceler has better performance than the conventional system without the canceller when P=5 and 7, and P_W is not appreciably small. Although the system with the canceler can reduce the effects of channel interference, it needs somewhat large power to have better performance than the system without the canceler, because the received signal of each user is split into P branches. It can be also seen that the system with the canceler has almost the same performance for any P, because when the number of simultaneous users is the same, the ratio of the signal power to the channel interference power is almost the same for any P. In addition, the system with the canceler can cancel the effects of the channel interference. Thus, the system with the canceler has almost the same performance for any P.

Figure 13 shows the bit error probability versus the number of users N for some values of P where $P_W = -75$ dBW. It can be seen that the system with the canceler has better performance when N is not appreciably small: when $N > 6$ for P=7, the system with the canceler has better performance. The effects of the channel interference are small when N is small. Thus the system with the canceler does not have better performance when N is small. It can be also seen that the system with the canceler has almost the same performance for any P, while the bit error probability of the conventional system is improved as P increases. In the conventional system without the canceler, the effect of channel interference is almost the same for any P at the same N, while the signal intensity becomes

larger as P increases even at the same N. In addition, the system with the canceler can reduce the effect of channel interference. Thus, the bit error probability of the conventional system is improved as P increases, while the system with the canceler has almost the same performance for any P.

4.2. Channel Interference Canceler Using Optical Hard-Limiters

Figure 14 shows the receiver block diagram of the direct-detection optical OOK CDMA system with a single optical hard-limiter [20]. In optical CDMA systems, the channel interference is the prime noise factor; it degrades the performance seriously and produces an asymptotic floor to the error probability. An optical hard-limiter is used to reduce the channel interference and to improve the system performance. An optical hard-limiter is defined as

$$g(x) = \begin{cases} v_f, & x \geq \text{Th} \\ 0, & 0 \leq x < \text{Th} \end{cases} \tag{19}$$

where v_f is the fixed value dependent on the signal intensity and Th is the threshold level. If an optical light intensity x is larger than or equal to the threshold level Th, the hard-limiter would clip the intensity back to v_f, and if the optical light intensity x is smaller than Th, the response of the optical hard-limiter would be zero. This optical hard-limiter would improve the system performance in the ideal link, because it would reduce the effect of the channel interference generated by some combinations of interference patterns. As shown in Fig. 14, the optical hard-limiter is placed before the optical correlator in the conventional optical CDMA systems. At the receiver, the matched optical correlator is used to recognize the arrival of the desired sequence. The optical correlator is a set of optical delay lines inversely matched to the pulse spacings. When the desired optical sequence passes through the correlator, the output light intensity traces out the correlation function of the sequence. At the last chip position, the sum of received optical intensity located in the same positions as the positions of "1" of the signature sequence code used for the desired channel is obtained.

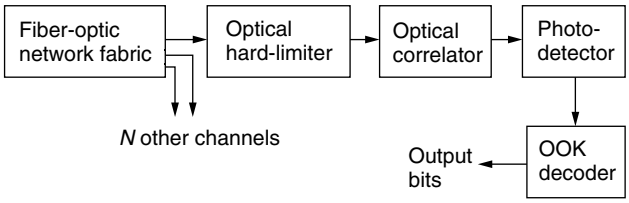


Figure 14. The receiver block diagram of a direct-detection optical OOK CDMA system with the single optical hard-limiter.

Figure 15 shows an example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends "1" where T_c is the chip duration: the second and the third marks of the desired user are hit by one undesired interferer's mark, respectively. The interference is removed by the first optical hard-limiter. Note that, however, the interference would contribute to the optical light intensity of the desired user when the desired user sends "1" if the optical hard-limiter is not used.

Figure 16 shows an example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends "0"; the first mark position of the desired user is hit by two undesired interferers' marks, and the second mark position of the desired user is hit by one undesired interferer's mark. As shown in this figure, there are some interference patterns that are not completely removed with the first optical hard-limiter when the desired user sends "0."

To improve the system performance by excluding some combinations of interference that cause incorrect bit decisions for "0" bit transmission, optical synchronous CDMA systems with double optical hard-limiters have been proposed. Figure 17 shows the receiver block diagram of the system with double optical hard-limiters. In the system, optical hard-limiters are placed before and after the optical correlator. The optical hard-limiters placed before and after the optical correlator are referred to as the *first* and the *second optical hard-limiters*, respectively.

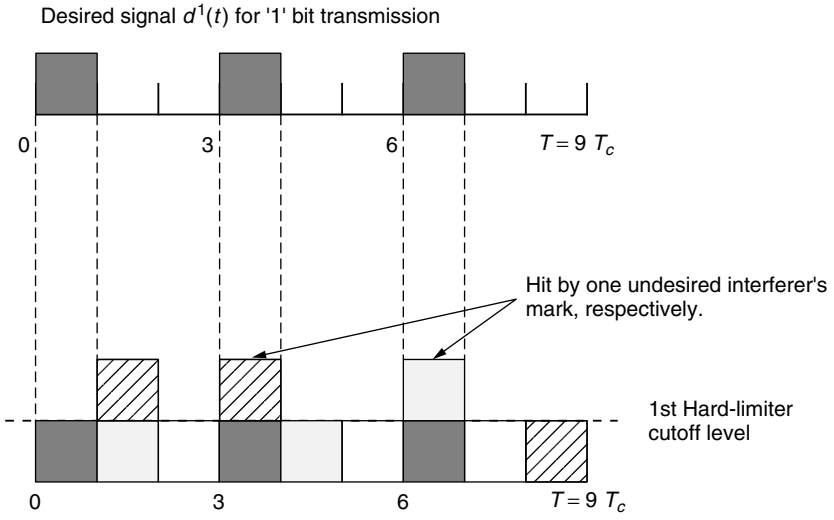


Figure 15. An example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends a "1"; the second mark of the desired user is hit by one undesired interferer's mark, and the third mark of the desired user is hit by one undesired interferer's mark.

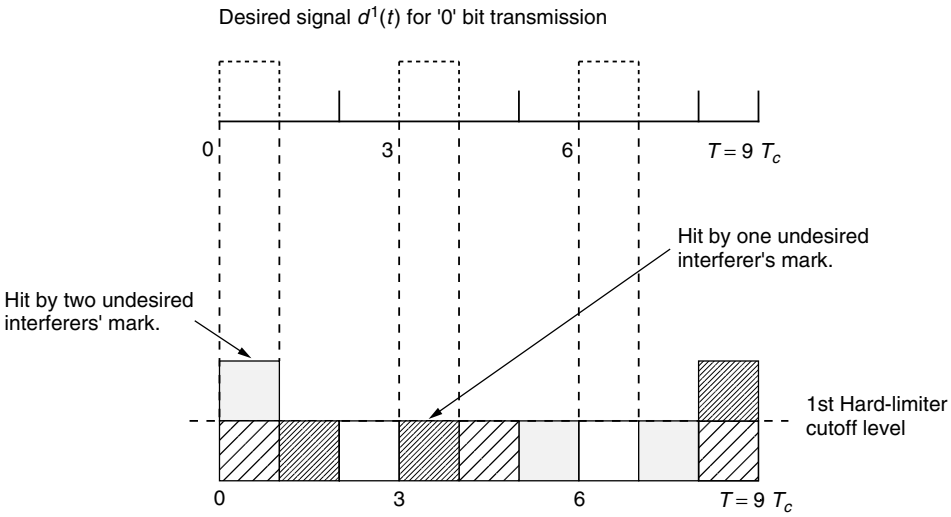


Figure 16. An example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends "0"; the first mark position of the desired user is hit by two undesired interferes' mark, and the second mark position of the desired user is hit by one undesired interferer's mark.

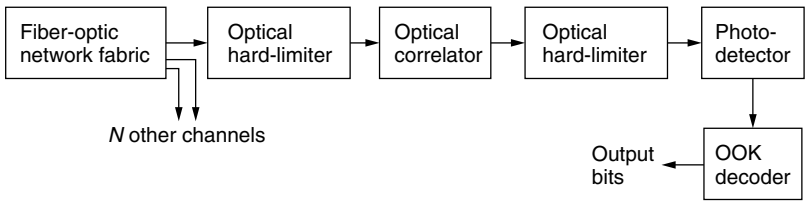


Figure 17. The receiver block diagram of a direct-detection optical CDMA system with double optical hard-limiters.

The first and the second optical hard-limiters are defined as the same as in the optical CDMA systems with the single optical hard-limiter given by Eq. (19). The first optical hard-limiter would clip the intensity back to v_f, and thus exclude or reduce the channel interference by other undesired interferers' marks. The second optical hard-limiter is used to exclude some interference patterns that are not completely removed with the first optical hard-limiter when the desired user sends "0" as shown in Fig. 16.

The second optical hard-limiter would clip the output intensity from the optical correlator back to zero if the optical light intensity x is smaller than Th. Therefore, the system using double optical hard-limiters would improve the system performance, because it would exclude the effect of the channel interference generated by some combinations of interference patterns as shown in Fig. 16.

Figure 18 shows the bit error probability versus the average received photocount in the last chip K_s for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with the double optical hard-limiters for some values of P. It can be seen that using the single optical hard-limiter placed before the optical correlator slightly degrades the performance of an optical synchronous CDMA system using modified prime sequence codes. Using the single optical hard-limiter excludes some combinations of interference patterns; however, it also excludes some combinations of interference patterns contributing to the optical intensity of the desired user. Thus, using the single optical hard-limiter slightly degrades the performance under the Poisson shot noise model for the receiver photodetector. It can be also seen that the

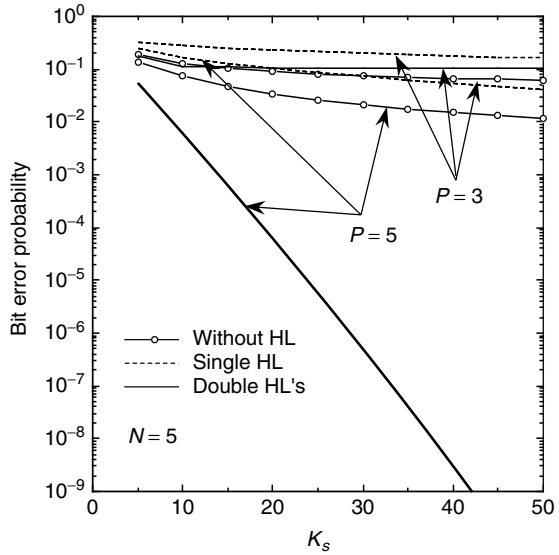


Figure 18. Bit error probability versus K_s for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with double optical hard-limiters for some values of P where $N = 5$.

optical CDMA system with double optical hard-limiters has better performance than do the other two systems when $P = 5$; when $P = 3$ using the system with double optical hard-limiters has a performance slightly worse than that of the system without the optical hard-limiter and slightly better that of than, the system with the single optical hard-limiter. The double optical hard-limiters can exclude combinations of interference patterns that the

single optical hard-limiter cannot exclude; all the mark positions of the desired user are not hit by undesired interference marks when the desired user sends "0." When $P = 3$ and $N = 5$, there are still some combinations of interference patterns that the double optical hard-limiters cannot exclude, while there is no combination of interference patterns that the double optical hard-limiters cannot exclude when $P = 5$ and $N = 5$. Thus, the error floor exists for the optical CDMA systems with double optical hard-limiters when $P = 3$ and $N = 5$ and does not exist when $P = 5$ and $N = 5$. Moreover, it can be seen that all the systems with $P = 5$ have better performance than do those with $P = 3$, respectively. This is because the probability that the undesired users' marks hit the desired user's mark positions is small when $P = 5$. Thus all the systems with $P = 5$ have better performance.

Figure 19 shows the bit error probability versus the number of users N for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with double optical hard-limiters with $P = 7$ and $K_s = 30$. It can be seen in the figures that the optical CDMA system with double optical hard-limiters has the constant low bit error probability when N is smaller than or equal to P: $N \leq 7$. This is because the optical hard-limiters can exclude all the combinations of interference patterns when $N \leq P$, that is, the number of interferers is smaller than P. It can be also seen that the performance of the optical CDMA system with double optical hard-limiters becomes worse than that of the optical CDMA systems without the optical hard-limiter when N is large. When N is large, the probability that all the mark positions of the desired user are hit by the undesired interference marks becomes high, and thus the number of combinations of interference patterns that the double optical hard-limiters cannot exclude becomes large and the performance is degraded. Therefore, using the double optical hard-limiters is effective in improving the performance of optical synchronous CDMA systems when the number of simultaneous users is not so large.

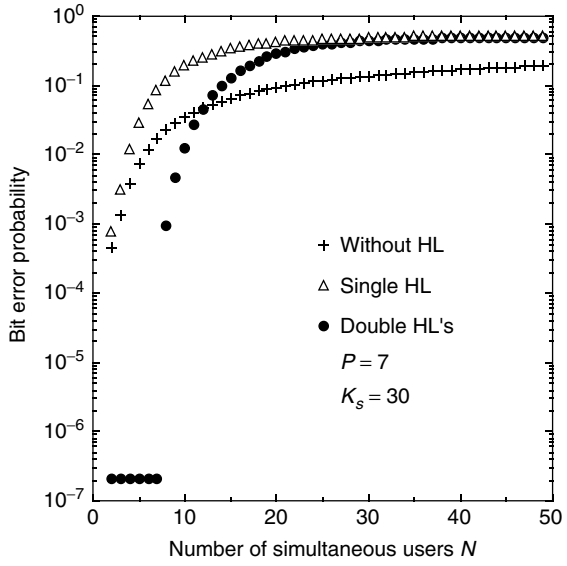


Figure 19. Bit error probability versus the number of simultaneous users N for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with double optical hard-limiters: $P = 7$ and $K_s = 30$.

BIOGRAPHY

Tomoaki Ohtsuki received his B.S., M.S., and Ph.D. degrees in electrical engineering from the Keio University, Yokohama, Japan in 1990, 1992, and 1994, respectively. From 1994 to 1995 he was a postdoctoral Fellow and a visiting researcher of electrical engineering at the Keio University. From 1993 to 1995 he was a special researcher of fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. He joined the Tokyo University of Science in 1995 as an assistant professor. Since 2000 he has been a lecturer tenured of the Tokyo University of Science. He has been working on optical communication systems, wireless communication systems, and information theory. He was the recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, the Erricson Young Scientist Award 2000, and the 2002 Funai Information and Science Young Scientist Award.

Iwao Sasase received his B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1979, 1981, and 1984, respectively. He is currently a professor of information and computer science at Keio University. His research interests include mobile, satellite and optical communications, and information networks. He has authored more than 170 journal papers and 260 international conference papers. He was the recipient of the 1984 IEEE Communications Society Student Paper Award, 1986 Inoue Memorial Young Engineer Award, 1988 Hiroshi Ando Memorial Young Engineer Award and 1988 Shinohara Memorial Young Engineer Award, in 2002. He is the IEEE Communications Society Satellite and Space Communications Technical Committee Chair and Asia Pacific Region vice director.

BIBLIOGRAPHY

1. F. R. K. Chung, J. A. Salehi, and V. K. Wei, Optical orthogonal codes: design, analysis, and applications, *IEEE Trans. Inform. Theory* **IT-35**: 595–604 (May 1989).
2. W. C. Kwong, P. A. Perrier, and P. R. Prucnal, Performance comparison of asynchronous and synchronous code-division multiple-access, *IEEE Trans. Commun.* **COM-39**: 1625–1634 (Nov. 1991).
3. P. R. Prucnal, M. A. Santro, and T. R. Fan, Spread spectrum fiber-optic local area network using optical processing, *IEEE J. Lightwave Technol.* **LT-4**: 547–554 (May 1986).
4. A. S. Holmes and R. R. A. Syms, All-optical CDMA using "quasi-prime" codes, *IEEE J. Lightwave Technol.* **LT-10**: 279–286 (Feb. 1992).
5. W. C. kwong, G. C. Yang, and J. G. Zhang, 2^n prime-sequence codes and coding architecture for optical code-division multiple-access, *IEEE Trans. Commun.* **COM-44**: 1152–1162 (Sept. 1996).

6. K. P. Jackson et al., Optical fiber delay-line signal processing, *IEEE Trans. Microwave Theory Tech.* **MTT-33**: 193–210 (March 1985).

7. H. M. Shalaby, Performance analysis of optical synchronous CDMA communication systems with PPM signaling, *IEEE Trans. Commun.* **43**(2–4): 624–634 (Feb.–April 1995).

8. H. M. H. Shalaby and E. A. Sourour, Co-channel interference cancellation in optical synchronous CDMA communication systems, *Conf. Rec. ISSSTA'94*, Oulu, Finland, July 1994, pp. 579–583.

9. Y. Gamachi, T. Ohtsuki, H. Uehara, and I. Sasase, Optical synchronous PPM/CDMA systems using co-channel interference cancellation, *Proc. IEEE Global Telecommunications Conf.* (*GLOBECOM'95*), Singapore, Nov. 1995, pp. 2161–2165.

10. Y. Gamachi, T. Ohtsuki, H. Uehara, and I. Sasase, Performance analysis of optical synchronous PPM/CDMA systems with interference canceller under number-state light field, *Trans. IEICE*, **E79-B**(7): 915–922 (July 1996).

11. T. Ohtsuki, Channel interference cancellation using time division reference signal for direct-detection optical synchronous CDMA systems, *Proc. IEEE Int. Conf. Communications* (*ICC'96*), Dallas, TX, June 1996, pp. 187–191.

12. T. Ohtsuki, Direct-detection optical synchronous CDMA systems with channel interference canceller using time division reference signal, *Trans. IEICE* **E79-A**(12): 1948–1956 (Dec. 1996).

13. H. M. H. Shalaby, M. A. Mangoud, and S. E. El-Khamy, A new interference cancellation technique for synchronous CDMA communication systems using modified prime codes, *Conf. Rec. 2nd IEEE Symp. Computers and Communications*, 1997, pp. 556–560.

14. T. Ohtsuki, M. Takeoka, and E. Iwahashi, Performance analysis of direct-detection optical synchronous CDMA systems with co-channel interference canceller, *Trans. IEICE* **E80-A**(12): 2260–2263 (Nov. 1997) (letter).

15. T. Ohtsuki, Optical CDMA canceller systems with tunable prime code decoder, *Proc. Int. Symp. Information Theory and Its Applications* (*ISITA'96*), Victoria, Canada, Sept. 1996, pp. 766–769.

16. Y. Gamachi, H. Uehara, T. Ohtsuki, and I. Sasase, Upper bound of optical synchronous PPM/CDMA systems with interference canceller using reference signal, *Proc. Int. Conf. Telecommunications* (*ICT'97*), Melbourne, Australia, April 1997, pp. 909–914.

17. H. M. H. Shalaby, Cochannel interference reduction in optical synchronous PPM-CDMA systems, *IEEE Trans. Commun.* **COM-46**: 799–805 (June 1998).

18. H. Sawagashira, K. Kamakura, T. Ohtsuki, and I. Sasase, Direct-detection optical synchronous CDMA systems with interference canceller using group information codes, *IEEE Global Telecommunications Conf.* (*GLOBECOM'00*), San Francisco, Nov. 2000, pp. 1216–1220.

19. H. Sawagashira, K. Kamakura, T. Ohtsuki, and I. Sasase, Direct-detection optical synchronous CDMA systems with interference canceller using group information codes, *Trans. IEICE* **E83-A**(11): 2138–2142 (Nov. 2000) (letter).

20. J. A. Salehi and C. A. Brackett, Code division multiple-access techniques in optical fiber networks — Part II: Systems performance analysis, *IEEE Trans. Commun.* **COM-37**: 834–842 (Aug. 1989).

21. T. Ohtsuki, K. Sato, I. Sasase, and S. Mori, Direct-detection optical synchronous CDMA systems with double optical hard-limiters using modified prime sequence codes, *IEEE J. Select. Areas Commun.* **14**(9): 1879–1887 (Dec. 1996).

22. T. Ohtsuki, Performance analysis of direct-detection optical asynchronous CDMA systems with double optical hard-limiters, *IEEE J. Lightwave Technol.* **15**(3): 452–457 (March 1997).

23. T. Ohtsuki, Direct-detection optical asynchronous CDMA systems with double optical hard-limiters: APD noise and thermal noise, *Trans. IEICE* **E81-B**(7): 1491–1499 (July 1998).

24. T. Ohtsuki, Channel interference cancellation using electrooptic switch and optical hard-limiters for direct-detection optical CDMA systems, *IEEE J. Lightwave Technol.* **16**(4): 520–526 (April 1998).

25. C. L. Lin and J. Wu, A synchronous fiber-optic CDMA system using adaptive optical hardlimiter, *IEEE J. Lightwave Technol.* **16**(8): 1393–1403 (Aug. 1998).

OPTICAL TRANSMITTERS, RECEIVERS, AND NOISE

PETER J. WINZER
Bell Laboratories, Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

Driven by the desire to meet the ever-growing bandwidth demand of our communication society while steadily reducing the cost per transmitted information bit, per-channel data rates in wavelength-division multiplexed (WDM) optical communication systems have continuously been increased, with 40-Gbit/s systems being commercially available today. Aggregate single-fiber transmission capacities on the order of 10 Tbit/s, as well as capacity-times-distance products exceeding several 10 Pbit/s km have been reported [1]. Conversely, tremendous advances in optical filter design and optical multiplexer technology have enabled channel spacings of some 10 GHz in dense WDM systems. Thus, 40 Gbit/s has become the data rate at which optics and electronics have met, and — for the first time in optical communications — has made *spectrally efficient modulation* a major issue.

As a consequence, the investigation of cost-effectively manufacturable transmitters for bandwidth-efficient optical modulation formats, as well as the optimization of high-speed optical receivers for dense WDM systems have become key topics of optical communications research and development. The quest is on for identifying combinations of modulation formats and receiver structures that can best cope with optical noise as well as with various linear and nonlinear signal-distortions accumulated along the fiber-optic transmission path [2], with the aim to trade high receiver sensitivities for longer transmission distances, relaxed component tolerances, or increased system margins.

Another field that asks for highly optimized optical transmitters and receivers is free-space optical communications. With applications both in terrestrial broadband access and in space-borne intersatellite links [3,5], free-space optical communications will enable future high-speed mobile data networking, bringing broadband data services to remote locations on the globe as well as to users on airplanes.

This section intends to open up the field of optical modulation and reception on an introductory level by discussing a selection of optical modulation techniques currently viewed as being most promising. Further, high-performance optical receiver structures and their noise properties are outlined, both for the fiber channel and for the free-space channel. Basic receiver design rules as well as important performance trade-offs are extracted. Frequently used concepts for quantifying receiver performance, such as *receiver sensitivity, quantum limit, Q-factor*, and *optical signal-to-noise ratio* (OSNR) are explained. To probe beyond the overview given in this chapter, and to acquire a more complete picture of the wide field of optical transmission, reception, and noise, the reader is kindly referred to the selection of excellent texts referenced at the end of the section.

2. OPTICAL MODULATION FORMATS AND THEIR IMPLEMENTATION

After giving a general classification of optical modulation formats, this section discusses the most important optical modulation techniques known today, with a particular view on their practical implementation by means of state-of-the-art high-speed opto-electronic components.

2.1. Classification

2.1.1. What to Modulate? The optical field[1] has three physical attributes that can be used to carry information: *Amplitude, phase* (including *frequency*), and *polarization*.

Depending on which of the three quantities is used to convey information, we distinguish between *amplitude-modulated, phase-modulated* (*frequency-modulated*), and *polarization-modulated* formats. Hybrid formats that simultaneously modulate two or more properties of the optical field (e.g., quadrature amplitude modulation (QAM)) have not yet made their way into high-speed optical communications. These formats are widely used in microwave communications, as well as in the related field of optical subcarrier-multiplexing, predominantly for cable-TV applications [6]. Here, several individually modulated signals on separate radio-frequency (RF) carriers are imprinted on an optical field by (linear) amplitude modulation.

[1] In optical communications, the term *optical field* is used to denote either one of the four electromagnetic field quantities observing the wave equation. It is usually expressed as a complex baseband quantity by eliminating the optical carrier frequency and is normalized such that its squared magnitude represents the optical power.

While amplitude and phase modulation have been widely used in high-speed optical communications, polarization modulation has received comparatively little attention so far [7]. This can primarily be attributed to the random polarization changes in optical fibers, necessitating active polarization control at the receiver. For amplitude- or phase-modulated formats and direct-detection receivers, however, polarization control is only required if polarization-mode dispersion (PMD) becomes an issue [8]. From a receiver point of view, this additional complexity would only be justifiable if polarization modulation offered significant baseline receiver sensitivity improvements over amplitude modulation, which it does not [9].

Note that our classification does not require a phase-modulated optical field to be constant-envelope, nor an amplitude-modulated field to have constant phase. It is the physical quantity from which information is extracted at the receiver that drives our classification. To give some examples: Differential phase shift keying (DPSK, cf. Section 2.3.1) is a phase-modulated format, regardless of whether it is transmitted constant-envelope or by means of phase-modulated optical pulses in the form of return-to-zero-DPSK (RZ-DPSK). Conversely, carrier-suppressed return-to-zero (CSRZ, cf. Section 2.2.3) is an amplitude-modulated format, regardless of the fact that the optical field's phase is additionally modulated in order to beneficially influence the spectrum.

2.1.2. How Many Symbols? The most widely used classes of optical receivers use *direct detection* (cf. Sections 3.2 and 3.6), that is they make use of the optical *power* $P = |E|^2$, the squared magnitude of the complex optical field amplitude E. If no optical phase-to-amplitude converting element is employed prior to detection, a direct-detection receiver is unable to distinguish between the two received symbols $E_{1,2} = \pm|E|$, since they both have the same optical power, $P_1 = P_2 = |\pm E|^2$. The additional degree of freedom gained by this ambiguity can be beneficially employed to shape the optical spectrum, or to make a format more resilient to distortions accumulated along the transmission line. Formats making use of this potential fall in the class of *pseudo-multilevel* or *polybinary* signals, depending on whether bit-correlations are introduced (as for duobinary formats, (M)DB, cf. Section 2.2.4) or not (as for carrier-suppressed return-to-zero, CSRZ, cf. Section 2.2.3). It is important to realize that these two classes of modulation formats use more than two symbols to encode a single bit of information, but transmit symbols at the bit rate R. For the formats of interest in optical communications today, the symbol alphabet $\{+|E|, -|E|, 0\}$ is used, which is mapped onto $\{0, |E|^2\}$ at the receiver. Pseudo-multilevel and polybinary signaling must not be confused with *multilevel* signaling, where $\log_2(M)$ bits are encoded on M symbols, and are then transmitted at a reduced symbol rate of $R/\log_2(M)$. Both multilevel amplitude shift keying [10] and (differential) quadrature phase shift keying (DQPSK, cf. Section 2.3.2) are multilevel optical modulation techniques. The difference between polybinary, pseudo-multilevel, and multilevel signaling is

Table 1. Symbol Encoding Examples for Multilevel, Pseudo-Multilevel, and Polybinary Signaling

Bit Sequence	0	0	1	0	1	1	1	0	0	1	0	1
pseudo-multilevel (CSRZ)	0	0	1	0	1	−1	1	0	0	−1	0	−1
polybinary (DB)	0	0	1	0	−1	−1	−1	0	0	−1	0	1
multilevel (DQPSK)	0		$+\pi/2$		π		$+\pi/2$		$-\pi/2$		$-\pi/2$	

visualized in Table 1, showing a data bit stream with three different symbol encodings.

2.1.3. Both Sidebands Needed? Apart from shaping (and compressing) the optical signal spectrum by means of (pseudo)-multilevel or polybinary signaling, it is possible for some modulation formats to additionally suppress half of their spectral content by appropriate optical filtering: Since the spectrum of real-valued baseband signals is symmetric around zero frequency, filtering out the redundant half of the spectrum (i.e., one of the two spectral '*sidebands*') preserves the full information content. This is exploited in *single-sideband* (SSB) signaling, where one sideband is completely suppressed, and in *vestigial-sideband* (VSB) signaling, where an optical filter with a gradual roll-off is offset from the optical carrier frequency to suppress major parts of one of the two sidebands, while at the same time performing some filter action on the other, desired sideband [11].

While broadband optical SSB is hard to generate in practice because of difficulties in implementing appropriate optical filter functions [10], optical VSB has been successfully demonstrated [12] on non-return-to-zero on/off keying (NRZ-OOK, cf. Section 2.2.1). Note that in fiber communications, VSB filtering is preferably done at the *receiver* instead of the transmitter, since if a sideband was suppressed at the transmitter, it would quickly reconstruct itself upon nonlinear fiber propagation. The advantage of using VSB comes from reduced WDM channel crosstalk for the desired sideband, if unequal channel spacings are employed. This situation is visualized in Fig. 1 [12], showing the composite spectrum of five wavelength-division multiplexed NRZ-OOK signals with alternating channel spacings of 1.2 and 1.7 times the data rate R. Severe WDM crosstalk is introduced for the sidebands on the closer-spaced sides (region A), making them useless for detection. Conversely, significantly less crosstalk is found for the sidebands on the larger-spaced sides (region B) than would be present if the channels were spaced on an equally-spaced frequency grid.

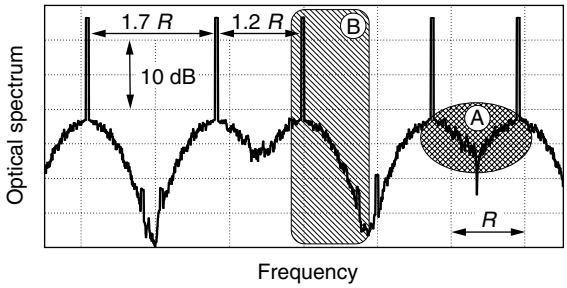


Figure 1. Vestigial sideband (VSB) transmission of optical signals on an unequally-spaced frequency grid to avoid WDM channel crosstalk [12].

2.2. Amplitude Modulation Formats

2.2.1. Non-Return-to-Zero On/Off Keying (NRZ-OOK). The simplest of all optical modulation formats is non-return-to-zero on/off keying (NRZ-OOK), often just called NRZ. This format imprints data on an optical carrier by switching light on and off. Historically, this was the first, and is still the most widely deployed optical modulation format. It has been used to *directly modulate*[2] both light-emitting diodes (LEDs) and lasers. Unfortunately, directly modulated laser light is highly chirped, that is, it exhibits strong residual phase modulation, which broadens the optical spectrum and degrades transmission performance through interaction with optical fiber dispersion in many important transmission scenarios. Thus, for modulation speeds above 2.5 Gbit/s and/or for long-haul fiber communication systems, *external modulation* has to be used. Here, the light of a continuously operating laser source is modulated by means of an external device, engineered for low chirp, or even designed for chirp-free operation. The two most important external modulators are semiconductor *electro-absorption modulators* (EAMs) and Lithium-Niobate ($LiNbO_3$) *Mach-Zehnder modulators* (MZMs). Both are commercially available for 40-Gbit/s modulation today.

EAMs [13] have the advantage of low-drive voltages (typ. 2 V), and are cheap in volume production. However, they still produce some residual chirp, have dynamic extinction ratios (maximum-to-minimum modulated light power) typically not exceeding 10 dB, and have limited optical power-handling capabilities (typ. 10 dBm). Their fiber-to-fiber insertion losses are about 10 dB, which has led to the integration with laser diodes, thus avoiding the input fiber-to-chip interface. *Electro-absorption modulated lasers* (EMLs) with output powers on the order of 0 dBm are widely available today. Another way of eliminating the high insertion losses of EAMs is the integration with semiconductor optical amplifiers (SOAs), which can even yield some net fiber-to-fiber amplification [14]. Figure 2a shows typical transmission characteristics of an EAM as a function of drive voltage. Note that the absorption of the EAM saturates at high drive voltages.

MZMs have excellent extinction performance (typ. 20 dB), can be made chirp free by balanced driving, and have lower insertion losses than EAMs (typ. 5 dB). The

[2] Using *direct modulation*, data are directly superimposed on a light-emitting device's drive current, which otherwise has biasing functionality only.

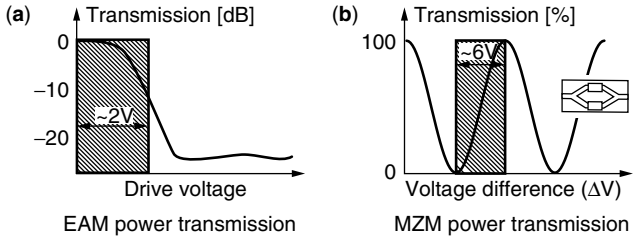


Figure 2. Typical optical power transmission characteristics of EAMs (**a**) and MZMs (**b**).

required (high-speed) peak-to-peak drive voltages of some 6 V, however, often represent serious practical problems. The MZM transfer characteristics is sinusoidal, owing to the Mach-Zehnder structure of the device (cf. inset to Fig. 2b): The incoming light is split into two paths at an input coupler. One (or both) paths are equipped with phase modulators that let the two fields acquire some phase difference relative to each other. Finally, the two fields interfere (destructively or constructively, depending on the modulated phases) at an output coupler. The optical field transfer function $T_E(V_1, V_2)$ thus reads

$$T_E(V_1, V_2) = \tfrac{1}{2}\{e^{j\phi(V_1)} + e^{j\phi(V_2)}\} = e^{j(\phi(V_1)+\phi(V_2))/2}\cos[(\phi(V_1) - \phi(V_2))/2] \quad (1)$$

where $\phi(V_{1,2})$ are the voltage-modulated optical phases of the two MZM arms. Since the phase modulation is a linear function of the drive voltage, the MZM *power* transfer function depends only on the drive voltage difference ΔV, $T_P(V_1, V_2) = |T_E(V_1, V_2)|^2 = T_P(\Delta V)$, which gives an additional degree of freedom in adjusting modulator chirp [15]. If the two modulator arms are driven by the same amount, but in opposite directions ($\phi(V_1) = -\phi(V_2)$), the phase term in Eq. (1) vanishes, resulting in purely real-valued transmission characteristics (i.e., in chirp-free operation). This driving condition is known as *balanced driving* or *push-pull operation*. Note that balanced driving cannot only be used to eliminate chirp, but also to reduce the output power requirements of the RF driver amplifiers by 6 dB (at the expense of having to use an additional amplifier, of course). Optical NRZ data signals are usually generated by driving the MZM from its minimum transmission to its maximum transmission, as visualized in Fig. 2b. Note that the non-linear parts of the MZM transfer function at high and low transmission can suppress overshoots and ripple on the electrical NRZ drive signal.

Figure 3 shows optical spectrum and optical eye diagram[3] of an idealized NRZ signal. The optical spectrum, defined as the bit-pattern-averaged squared magnitude of the optical field's Fourier transform, is composed of a continuous portion, which reflects the shape of the individual NRZ data pulses, and discrete tones at integer

[3] *Eye diagrams* are important means of visualizing the quality of digital signals. They are formed by plotting on top of each other copies of the same modulated bit pattern, shifted by integer multiples of the bit duration.

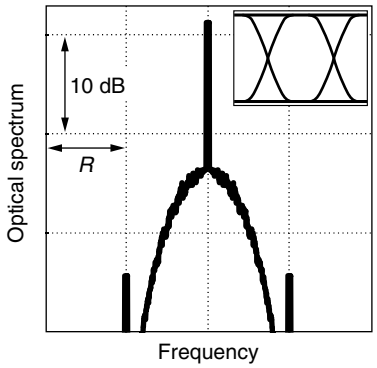


Figure 3. Optical spectrum and eye diagram of an NRZ signal.

multiples of the data rate. The weight of these tones is determined by the optical spectrum of the NRZ data pulses, and therefore depends on the NRZ rise/fall times. For ideal (rectangular) NRZ signals all tones vanish, apart from the one at zero frequency.

2.2.2. Return-to-Zero On/Off Keying (RZ-OOK). Regardless of the modulation format (phase or amplitude), NRZ often suffers from bandwidth-limitations, both at the transmitter and at the receiver, leading to the presence of intersymbol interference (ISI) in the bit sequence to be detected. ISI denotes the corruption of bits (most notably, of isolated '0'-bits) by their neighboring '1'-bits. In optical communications, ISI is particularly harmful, since detection noise often grows linearly with signal amplitude (cf. Sections 3.2, 3.4, and 3.6). Figure 4 shows typical NRZ and RZ electrical eye diagrams at the decision gate of a receiver for the same average optical input power and under the same filtering conditions, where the effect of ISI on NRZ becomes evident. In addition to ISI introduced by bandlimiting (optical or electrical) elements, NRZ formats degrade rapidly in many important fiber transmission scenarios. *Return-to-zero* (RZ, *impulsive coding*) coding mitigates these problems, and leads to enhanced system performance [16,17]. In the case of RZ-OOK, information is encoded on the presence or absence of optical pulses. By well centering the pulses in the bit slots, pattern effects coming from limited NRZ drive signal bandwidths are largely eliminated.

The advantages found for RZ coding come at the expense of higher optical transmission bandwidth requirements, as well as of more complicated transmitter structures, as shown in Fig. 5. Usually, RZ-OOK is generated from an optical NRZ signal by carving out pulses by

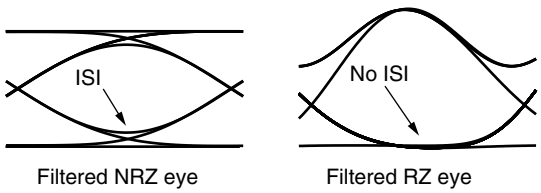


Figure 4. Eye diagrams for NRZ and RZ signals. Intersymbol interference (ISI) affects NRZ performance, while it is not seen for RZ.

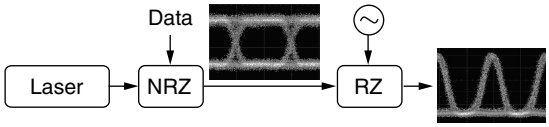


Figure 5. Structure of a typical RZ transmitter, consisting of a laser source, an external NRZ modulator, and a RZ pulse carver.

means of an additional modulator, termed *pulse carver*. Typically, pulse carvers are implemented as sinusoidally driven EAMs or MZMs.

Using an EAM, short (a few ps) optical pulses can be realized by biasing the modulator well in its absorption region, and letting only the peak portion of the sinusoidal drive signal reach appreciable transmission, as shown in Fig. 6. This technique is therefore widely used in optical time-division multiplexing (OTDM) transmitters [18,19].

If a MZM is used for pulse carving, three operating conditions have to be distinguished:

- Sinusoidally driving the MZM at the data rate between minimum and maximum transmission results in optical pulses with a full-width-half-maximum (FWHM) of 50% of the bit duration (a duty cycle of 50%), as shown in Fig. 7 (dashed). Decreasing the modulation swing while adjusting the modulator bias such as to still reach good extinction between pulses, the duty cycle can in principle be reduced to 36%, however, with significant excess insertion loss, since the modulator is then no longer driven to its transmission maximum. At a duty cycle of 40%, the excess insertion loss amounts to 2.2 dB. Increasing the drive voltage to reduce the pulse width (as can be done in the case of EAMs) is not possible with MZMs, owing to the periodic nature of the MZM transmission function.
- Sinusoidally driving the MZM at *half* the data rate between its transmission minima produces a pulse whenever the drive voltage passes a transmission maximum, as visualized in Fig. 7 (solid). This way, duty cycles of 33% can be realized, however, without the possibility for adjustments by varying the drive voltage. The doubled peak-to-peak drive voltage requirements usually pose little technical problems, since narrow-band RF amplifiers can be used.
- Sinusoidally driving the MZM at half the data rate between its transmission maxima results in pulses with 67% duty cycle and with alternating phase. The resulting format is called *carrier-suppressed RZ* (CSRZ), and will be discussed in Section 2.2.3.

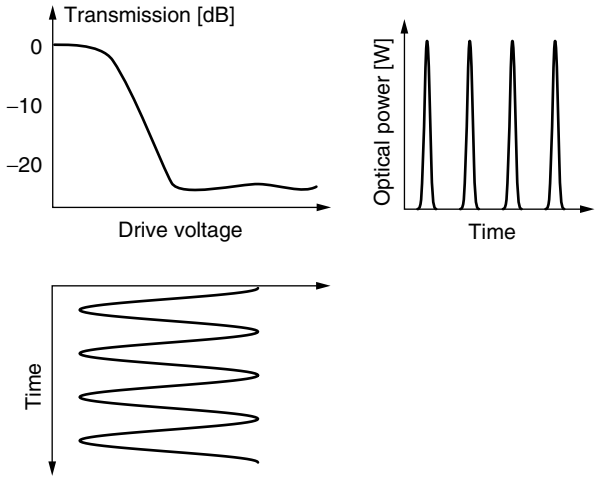


Figure 6. Sinusoidally driven EAM used as RZ pulse carver to attain short optical pulses, with duty cycles below 33%.

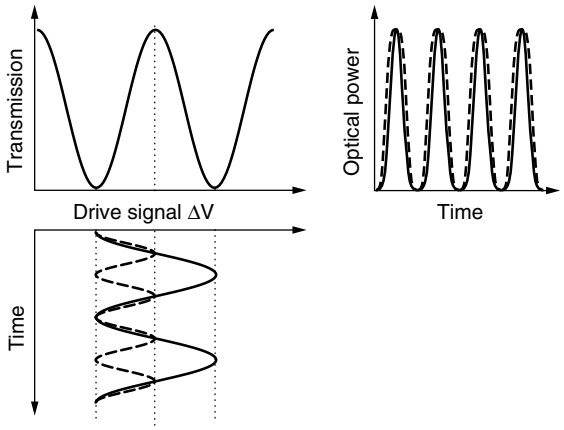


Figure 7. Sinusoidally driven MZM as pulse carver for 33%-duty-cycle RZ (solid) and 50%-duty-cycle RZ (dashed).

Other, less frequently used RZ-OOK modulation techniques include mode-locked lasers in combination with external NRZ-modulators to achieve very low duty cycle pulses for OTDM applications [20], single-step RZ-OOK modulation by means of an electrical RZ drive signal [21], and techniques employing the rising and falling edges of the electrical NRZ signal for RZ pulse generation [22,23]. Note that these methods make do with a single external optical modulator, without the need for a pulse carver.

Spectra and eye diagrams of 50% duty cycle RZ (gray) and 33% duty cycle RZ (black), as produced by a MZM in push-pull operation, are shown in Fig. 8.

2.2.3. Carrier-Suppressed Return-to-Zero (CSRZ). Carrier-suppressed return-to-zero (CSRZ) is a pseudo-multilevel modulation format, characterized by reversing

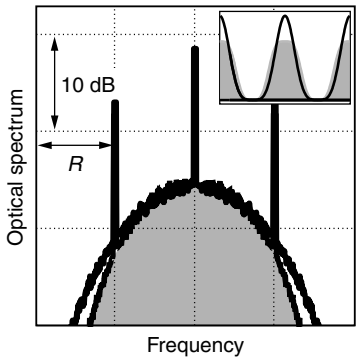


Figure 8. Optical spectra and eye diagrams for 50% duty cycle RZ (gray) and 33% duty cycle RZ (black), as produced by a MZM in push-pull operation.

the sign of the optical field at each bit transition. In contrast to the duobinary formats detailed in Section 2.2.4, the sign reversals occur at *every* bit transition, and are completely *independent* of the information-carrying part of the signal. CSRZ is most conveniently realized by sinusoidally driving a MZM pulse carver at half the data rate between its transmission maxima, as visualized in Fig. 9. Since the optical field transfer function $T_E(\Delta V)$ (dashed) of the MZM changes its sign at the transmission minimum (cf. Eq. (1) for push-pull operation), phase inversions between adjacent bits are produced. Thus, on average, the optical field of half the '1'-bits has positive sign, while the other half has negative sign, resulting in a zero-mean optical signal. As a consequence, the carrier at the optical center frequency vanishes, giving the format its name.

Using a MZM to generate CSRZ results in a duty cycle of 67%, which can be brought down to 50% at the expense of excess insertion loss by reducing the drive voltage swing. At a duty cycle of 55%, an excess insertion loss of 2 dB has to be accepted. It is important to note that, due to its most widely used practical implementation with MZMs, the duty cycle of CSRZ signals usually differs from the one of standard RZ. Thus, care has to be taken when comparing the two formats, since some performance differences result from the carrier-suppressed nature of CSRZ, while others simply arise from the different duty cycles.

Spectrum and eye diagram of 67%-duty cycle CSRZ, as generated by a MZM in push-pull configuration, are shown in Fig. 10. For comparison, the spectrum of a (hypothetical) 67%-duty cycle RZ signal is also given (gray). Note that the *only* difference between the two spectra is the location of the discrete tones.

2.2.4. Duobinary and Modified Duobinary (DB, MDB). Duobinary (DB) and modified duobinary (MDB) signals are polybinary signals, a subset of the partial response signaling format [11,24]. In optical communications they have also become known under the keywords *phase-shaped binary transmission* (PSBT) [25] and *phased amplitude-shift signaling* (PASS) [26]. Most conveniently, optical (M)DB signals, like CSRZ, employ

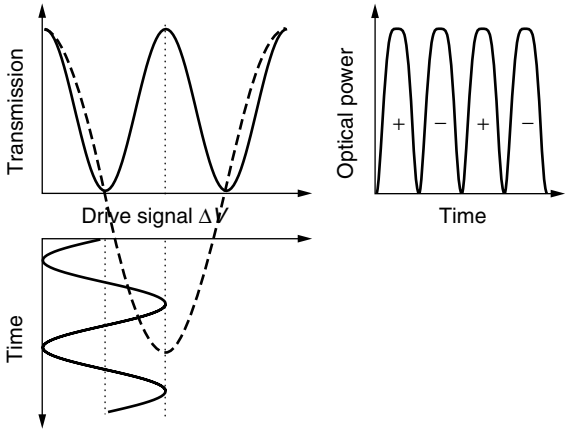


Figure 9. Sinusoidally driven MZM as pulse carver for 67%-duty-cycle CSRZ. The solid and dashed transmission curves apply for the optical power ($T_P(\Delta V)$) and field ($T_E(\Delta V)$), respectively.

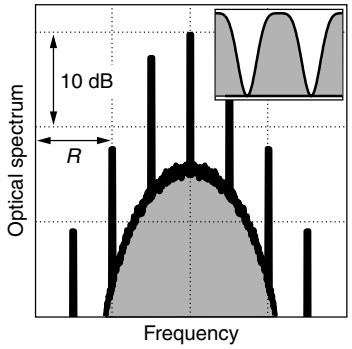


Figure 10. Optical spectrum and eye diagram of 67%-duty cycle CSRZ, as generated by a MZM in push-pull configuration (black). The spectrum of a 67%-duty cycle RZ signal without phase reversals is given for comparison (gray).

the signaling set $\{0, \pm|E|\}$, taking advantage of the power-detecting property of direct detection optical receivers that automatically converts the three optical symbols to the two electrical symbols $\{0, |E|^2\}$. However, unlike with CSRZ, the optical phases of the individual bits additionally depend on the bit pattern: For DB signaling, a phase change occurs whenever there is an odd number of '0's between two successive '1's, whereas for MDB the phase changes for each '1' (even for adjacent '1's), independent of the number of '0's inbetween (cf. also Table 1).

(M)DB signals are more tolerant than conventional binary signals with respect to chromatic dispersion, narrow-band optical filtering (thus allowing for closer WDM channel spacings), as well as to some non-linear transmission impairments. Explanations can be given both in the frequency domain [10,27] and in the time domain [25,26,28], the latter lending itself to a particularly intuitive interpretation: Consider the bit pattern ...0010100..., with the '1'-bits being represented by the tall, shaded pulses in Fig. 11. When transmitted through narrow-band optical filters or over dispersive optical fiber, the pulses broaden (hatched) to let some energy spill into the isolated '0'-bit. If the two pulses have the same optical phase, their optical fields add up constructively, leading to severe ISI (dashed). For (M)DB, however, two pulses separated by an isolated '0'-bit always have opposite phases, which lets them interfere destructively and thus reduces ISI (solid).

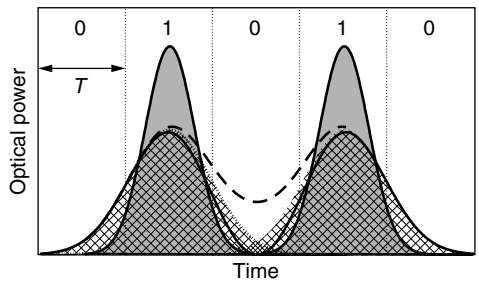


Figure 11. Initially narrow optical pulses (shaded) broaden through fiber dispersion or optical filtering (hatched). If the two pulses have the same optical phase, their optical fields add up constructively (dashed). For (M)DB, the two pulses have opposite phases, which lets them interfere destructively (solid) [25,26].

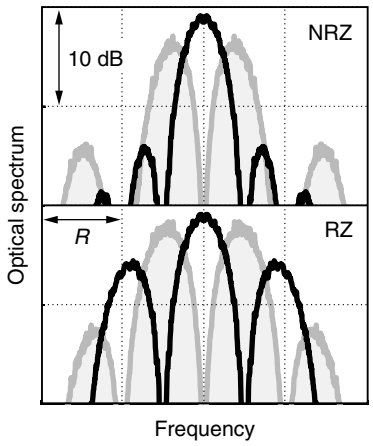


Figure 12. Optical spectra of duobinary (black) and modified duobinary (gray) signals in NRZ coding (upper) and RZ coding (lower).

Both DB and MDB can be implemented in RZ or NRZ format. Figure 12 shows the optical spectra of NRZ (upper) and RZ (lower) DB (black) and MDB (gray). A characteristic feature is the spectrally compressed main lobe as compared to NRZ-OOK (cf. Fig. 3). For NRZ-DB, the side lobes are filtered out for optimum performance [10]. Note that (M)DB spectra have no discrete spectral components, which helps to suppress stimulated Brillouin scattering (SBS) in optical fibers [29].

Duobinary transmitters are usually implemented using a three-level electrical drive signal $\{-1, 0, +1\}$ in combination with a MZM driven between its transmission maxima (like for CSRZ, Fig. 9, but using the data signal instead of a sinusoid). The methods resulting in chirp-free (M)DB signals operate the MZM in push-pull mode. As shown in Fig. 13a, the three-level electrical drive signal can be generated using analogue addition (DB) or subtraction (MDB) of the bit sequence with a 1-bit-delayed replica of itself, provided that appropriate precoding is performed on the data [11,27]. Since the required three-level (linear) RF driver electronics are hard to implement in practice, one usually resorts to method (b), taking a highly low-pass filtered version of the precoded data signal to drive the MZM. The filter bandwidth B has to be chosen on the order of one fourth the data rate R [27]. A third realization (c) uses a MZM as a phase modulator (cf. Section 2.3.1) to generate an intermediate DPSK signal, which is transformed to (M)DB using an optical delay interferometer [30]. By reducing the optical delay to values less than the bit duration T, variable duty cycle RZ-MDB can be generated without the need for a pulse carver [23,31].

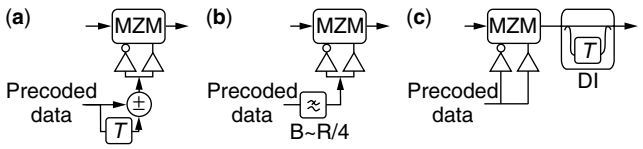


Figure 13. (Modified) duobinary signals are either generated by driving a MZM around its transmission minimum (cf. Fig. 9) using a 3-level electrical drive signal [(a) and (b)]. Alternatively, (M)DB can be generated using a MZM as phase modulator, and passing the resulting DPSK signal through a delay interferometer (DI) (c).

2.2.5. Chirped Return-to-Zero (CRZ). Chirped return-to-zero (CRZ) is predominantly used for ultra-long-haul fiber communication, as found in transoceanic (submarine) systems [32,33]. CRZ signals are generated by sinusoidally modulating the phase of a RZ signal at the data rate, using a separate phase modulator. The intentionally introduced chirp on the one hand beneficially influences nonlinear fiber transmission performance, but on the other hand broadens the signal spectrum. In WDM systems, the amplitude of the sinusoidal phase modulation has thus to be optimized by trading the gain due to enhanced nonlinear propagation performance against WDM channel crosstalk. Typically, the optimum phase modulation amplitude amounts to $\sim$1 rad [32].

The benefits of CRZ obviously come at the expense of more complex transmitter architectures, comprising a total of three external modulators whose drive signals have to be carefully synchronized. Integrated GaAs/AlGaAs modulators for CRZ, combining NRZ data modulator, RZ pulse carver, and CRZ phase modulator in one module, have been reported [34].

2.3. Phase Modulation Formats

The most widely used class of optical receivers employs direct detection, that is, the receiver is only sensitive to optical *power* variations (cf. Sections 3.2 and 3.6). To detect modulation of the optical field's *phase*, phase-to-amplitude converting elements therefore have to be inserted into the optical path prior to detection. Since these elements are unable to offer an absolute optical phase reference, the phase reference has to be provided by the signal itself: Each bit acts as a phase reference for another bit, which is at the heart of all *differential phase shift keying* formats.

2.3.1. Differential Phase Shift Keying (DPSK). Binary differential phase shift keying (BDPSK, or simply DPSK) encodes information on a binary phase change between adjacent bits. A logical '1' is encoded onto a π phase change, whereas a logical '0' is represented by the absence of a phase change. Thus, like for (M)DB, an appropriate precoding circuit has to be employed at the transmitter prior to modulation. Like OOK, DPSK can be implemented in RZ and NRZ format. The main advantage from using DPSK instead of OOK comes from a 3-dB sensitivity improvement at the receiver, provided that balanced detection is employed [9,35]. This enhanced sensitivity directly translates into increased transmission distance [36].

An optical (N)RZ-DPSK transmitter is shown in Fig. 14. The phase of the optical field of a narrow-linewidth laser source is flipped between 0 and π using the precoded (differentially encoded) version of the NRZ data signal, as visualized by the two bit patterns in the figure. If a straight-line phase modulator (PM) is used, the speed of the phase transitions is limited by the combined bandwidth of driver amplifier and phase modulator, while the

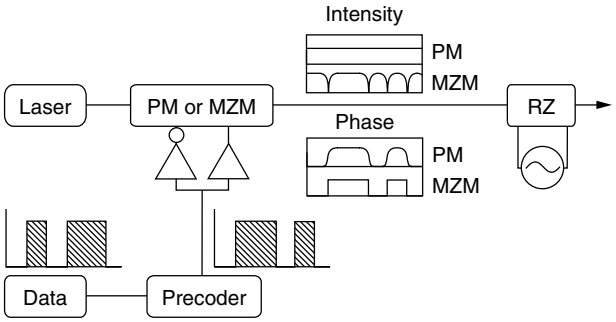


Figure 14. Setup of a RZ-DPSK transmitter. Phase modulation can either be achieved using a MZM, or by means of a straight-line phase modulator (PM), resulting in different amplitude and phase waveforms.

intensity of the phase-modulated light is constant. Instantaneous π phase jumps can be realized at the expense of some residual intensity modulation of the phase modulated light by using a dual-drive MZM, symmetrically driven around zero transmission [37], in analogy to the (M)DB transmitter of Fig. 13(a) and (b) and the CSRZ transmitter of Fig. 9. Typical intensity and phase waveforms of the two modulation techniques are shown at the output of the phase modulator in Fig. 14, where the upper traces apply to PM and the lower traces to MZM phase modulation. Like for OOK, a subsequent pulse carver converts the NRZ-DPSK signal to RZ-DPSK, if desired.

Figure 15 shows spectra and eye diagrams for RZ-DPSK (black) and NRZ-DPSK (gray), as generated by a MZM operated as a phase modulator. The carrier-free nature of the spectra, like for (M)DB, owes to the balance of $-|E|$ and $+|E|$ amplitude levels. Note the absence of a '0'-bit rail in the eye diagrams, which is characteristic for phase-coded formats. The deep amplitude dips between two bits

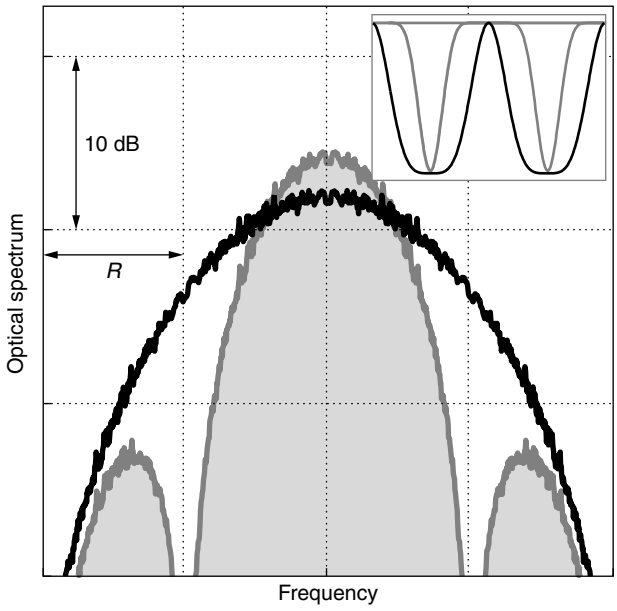


Figure 15. Optical spectra and eye diagrams of NRZ-DPSK (gray) and 33%-duty cycle RZ-DPSK (black), as produced by a MZM as phase modulator.

in the NRZ-DPSK eye represent the residual amplitude modulations of the MZM caused by the finite NRZ drive signal bandwidth.

Since DPSK cannot directly be received with direct-detection techniques, an optical delay interferometer (DI) is inserted in the optical path at the receiver to convert the differential phase modulation into amplitude modulation. As shown in Fig. 16, a DI splits the phase modulated signal into two paths, into one of which a delay equal to the bit duration T is introduced. At the DI's output coupler, the phase modulated optical field thus interferes with its one-bit-delayed replica to produce destructive interference at port $\mathcal{A}$ whenever there is *no* phase change, and constructive interference whenever there *is* a phase change, in agreement with the DPSK-coding rule described above. To exploit the 3-dB sensitivity advantage of DPSK over OOK, a *balanced receiver* has to be employed, where the second DI-output port $\mathcal{B}$, yielding the inverted data pattern, is also made use of, and the difference signal is detected [9,35].

2.3.2. Differential Quadrature Phase Shift Keying (DQPSK). Instead of using two phase levels (DPSK), one can use four phase levels $\{0, +\pi/2, -\pi/2, \pi\}$ to produce differential quadrature phase shift keying (DQPSK). DQPSK is a true four-level signaling format, transmitting symbols at *half* the aggregate bit rate (cf. Table 1). While the receiver sensitivity benefit over OOK that is gained for DPSK is largely lost for DQPSK, the transmission bandwidth is significantly reduced, potentially allowing for higher spectral efficiency in WDM systems, as well as for increased tolerance to chromatic dispersion and PMD [38].

A DQPSK transmitter is best implemented by taking advantage of the *exact* π-phase shifts produced by a MZM operated as a phase modulator (cf. Fig. 9). Figure 17 shows the corresponding transmitter setup [38], consisting of a continuously operating laser source, a splitter to divide the light into two paths of equal intensity, two MZMs operated as phase modulators, a $\pi/2$-phase shifter in one of the paths, and a combiner to produce a single output signal. While the modulated field E_1 of the upper path can take on the values $\pm E_0/\sqrt{2}$, the lower path produces $E_2 = \pm E_0 e^{j\pi/2}/\sqrt{2}$, leading to the four symbols $E_0/\sqrt{2} \cdot \{e^{j\pi/4}, e^{j3\pi/4}, e^{j5\pi/4}, e^{j7\pi/4}\}$ after the combiner. A pulse carver can optionally be added to yield RZ-DQPSK.

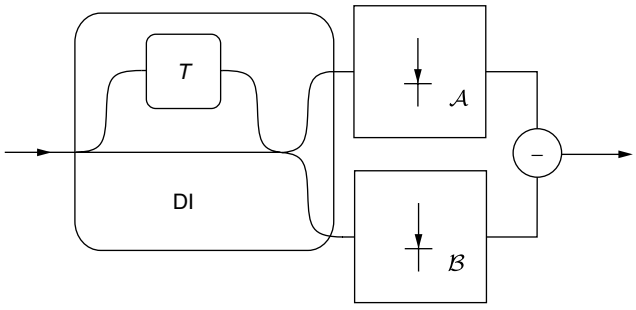


Figure 16. Balanced DPSK receiver using an optical delay interferometer (DI) to convert the phase modulation to amplitude modulation.

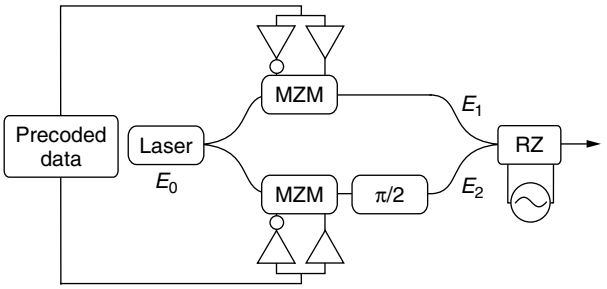


Figure 17. Structure of a DQPSK transmitter. Two MZMs are used as phase modulators, and the two separately modulated fields are combined with a $\pi/2$ phase shift [38].

The *shape* of the DQPSK optical spectra is identical to those of DPSK (Fig. 15). However, the DQPSK spectrum is compressed in frequency by a factor of two due to the halved symbol rate for transmission at the same bit rate.

At the receiver, the DQPSK signal is split, and *two* balanced receivers of the form depicted in Fig. 16 are used in parallel to simultaneously demodulate the two data streams contained within the DQPSK signal [38]. Note that the DI delay has to equal the *symbol* duration for DQPSK demodulation, which is *twice* the bit duration. Due to the DQPSK phase shifts of $\pi/2$ (instead of π for DPSK), the two DIs cannot simultaneously be operated to full destructive and to full constructive interference, which results in a reduced demodulated eye opening for DQPSK.

3. OPTICAL RECEIVER CONCEPTS AND NOISE

3.1. The *Q*-Factor

Before embarking on optical receiver concepts, we will briefly discuss the *Q-factor* as an important parameter that is widely used in optical communications to describe receiver performance. Although occasionally frowned upon by theoreticians, since its derivation relies on (sometimes hard to justify) approximations, the *Q*-factor allows for intuitive interpretations, reasonably accurate quantitative predictions, and has also become an indispensable tool for experimentalists [32].

The *Q*-factor was first introduced by Personick in 1973 [39] to relate mean and variance of the electrical signal at the receiver's decision gate to a bit-error ratio (BER), the quantity of ultimate interest when assessing the performance of digital communication systems. Leaving the derivation of the *Q*-factor to more comprehensive texts [35,40,41], we restrict ourselves to its definition,

$$Q = \frac{|s_1 - s_0|}{\sigma_1 + \sigma_0}, \tag{2}$$

where $s_{0,1}$ are the mean electrical signal amplitudes for a logical '0' and '1' at the decision gate, and $\sigma_{0,1}$ are the associated noise standard deviations. Under the assumption of Gaussian detection statistics, which is sufficiently accurate in most situations of practical interest [42], the BER is related to the *Q*-factor via

$$\mathrm{BER} = 0.5\ \mathrm{erfc}[\mathrm{Q}/\sqrt{2}], \tag{3}$$

where $\mathrm{erfc}[x] = (2/\sqrt{\pi}) \int_x^\infty \exp(-\xi^2)\, d\xi$ denotes the complementary error function. For $\mathrm{BER} = 10^{-9}$, which is often taken as a baseline for specifying receiver sensitivities, we have $Q \approx 6$.

Note that σ_0 and σ_1 may differ from each other, since many important noise terms encountered in optical communications are *signal-dependent*, that is, the noise variance is a function of the optical signal power. For purely *signal-independent* noise ($\sigma_1 = \sigma_0 = \sigma$), Eqs. (2) and (3) reduce to $\mathrm{BER} = 0.5\ \mathrm{erfc}[|s_1 - s_0|/(2\sqrt{2}\sigma)]$, a well-known expression in classical communication theory [11].

3.2. *Pin*-Receiver

The *pin*-receiver depicted in Fig. 18 is the simplest optical receiver structure. It consists of a *pin*-photodiode,[4] some postdetection electronic amplification and filtering with (single-sided) bandwidth B_e (electronics impulse response $h(t)$), and a sampling-and-decision device that restores the digital data. Detection of the filtered signal $s(t)$ is corrupted by two types of noise in a pin-receiver, *shot noise* and *electronics noise*.

Shot noise is a direct consequence of the quantum nature of light: Interactions of light and matter can only take place in discrete energy quanta (*photons*), governed by the rules of quantum statistics. Thus, a discrete, random number of electron-hole pairs is generated in a semiconductor diode when light impinges on it, causing the photocurrent to leave the diode in individual elementary impulses, each carrying the elementary charge $e \approx 1.602 \cdot 10^{-19}$. As, as visualized in Fig. 19. This fine structure of the electrical signal is perceived as shot noise [43,45]. On average, a fraction η of the incoming optical power $p(t)$ is converted to an electric current, leading to an average electrical signal amplitude of

$$\langle s(t) \rangle = (R_T) \cdot S(p * h)(t) \tag{4}$$

where $S = \eta e/(hf)$ [A/W] is the receiver's responsivity. The symbol * denotes a convolution, and $h(t)$ is normalized to let the low-frequency portion of its spectrum equal unity. Depending on whether the electrical signal $s(t)$ is specified in terms of current or of voltage, a resistance R_T has to be taken into account that converts the current-output of the *pin*-photodiode into a voltage. This resistance is

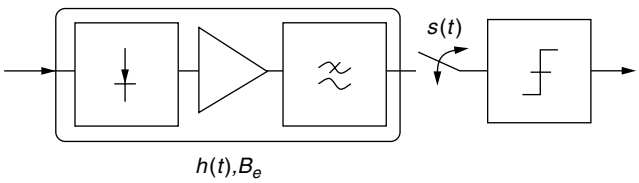


Figure 18. Setup of a *pin*-receiver, incorporating a *pin*-photodiode, an electrical preamplifier, electrical low-pass filtering, and a sampling-and-decision device.

[4] The abbreviation *pin* stands for *p-doped/intrinsic/n-doped*, and describes the basic layer structure of the associated semiconductor device.

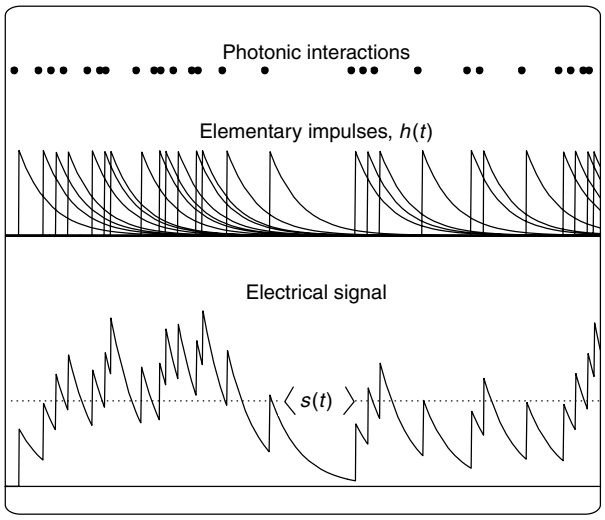


Figure 19. Photons arrive at random, dictated by quantum statistics. Each photonic interaction produces an elementary electronic impulse. The impulses add up to produce the overall electrical signal, whose fluctuations are known as shot noise [43].

frequently referred to as *transimpedance; hf* denotes the photon energy. ($h \approx 6.626 \cdot 10^{-34}$ Js is Planck's constant, and f stands for the optical carrier frequency of light.) The shot noise variance associated with photodetection is given by [43]

$$\sigma^2_{shot}(t) = (R_T^2) \cdot eS(p * h^2)(t) \approx (R_T^2) \cdot 2eSp(t)B_e \qquad (5)$$

The approximation in Eq. (5) applies for optical power variations that are slow compared to the speed of the electronics. Note that shot noise is a *nonstationary*, non-Gaussian noise process in general; its statistical parameters, most notably its variance, change with time.

Electronics noise is the sum of all stationary noise sources generated within the opto-electronic circuitry, *independent* of the optical signal, such as thermal noise, transistor shot noise, $1/f$-noise, or dark-current shot noise. The design of the receiver front-end electronics significantly impacts its noise performance, and is detailed in numerous excellent texts [39,46].

On a system level, electronics noise is often characterized by an *equivalent noise current density* $i_n[\mathrm{A}/\sqrt{\mathrm{Hz}}]$, which can be converted to an electronics noise variance σ^2_{elec} at the decision gate by

$$\sigma^2_{elec} = (R_T^2) \cdot i_n^2 B_e \qquad (6)$$

In the Gbit/s-regime, i_n is typically on the order of some 10 $\mathrm{pA}/\sqrt{\mathrm{Hz}}$. Alternatively, the noise performance of receivers can be specified using the *noise equivalent power* (NEP) $[\mathrm{W}/\sqrt{\mathrm{Hz}}]$, which is usually defined as the average optical power per square-root electrical receiver bandwidth that would be required to make the average electrical power equal to the electronics noise variance,

$$\sigma^2_{elec} = (R_T^2) \cdot S^2 \mathrm{NEP}^2 B_e \qquad (7)$$

3.3. Receiver Sensitivity and Quantum Limit

Instead of using equivalent noise current density or NEP, optical receiver front-ends are sometimes also characterized in terms of their *receiver sensitivity*, defined as the average optical power that is required at the receiver input to obtain a certain BER (typ.10^{-9}) at a certain data rate and for a certain modulation format (typ. NRZ-OOK). While the receiver sensitivity is undoubtedly of high interest in optical receiver design, it comprises not only the degrading effects of noise, but also encompasses essential properties of the received signal, such as extinction ratio, signal distortions and intersymbol interference (ISI), generated either within the transmitter or within the receiver itself. Thus, knowledge of the receiver sensitivity alone does not allow trustworthy predictions on how the receiver will perform for other formats (e.g., for RZ-OOK).

Although electronics noise usually dominates shot noise, it can in principle be engineered to zero. Shot noise, however, is fundamentally present. The limit, when *only* fundamental noise sources determine receiver sensitivity is called *quantum limit* in optical communications. The existence of quantum limits makes optical receiver design an exciting task, since there is always a fundamental measure against which practically implemented receivers can be compared, much like the Shannon-limit in information theory. Note, however, that each class of receivers in combination with each class of modulation formats has its own quantum limit.

The quantum limit for the *pin*-receiver using OOK is obtained by ignoring thermal noise, assuming a perfect ($\eta = 1$) receiver, and evaluating the BER for the Poissonian photon statistics of perfect laser light [43,44]. Leaving the derivation to more detailed texts [9,35,41], we merely cite the result,

$$\mathrm{BER} = 0.5 \exp[-2\overline{n}] \qquad (8)$$

where $\overline{n}$ is the average number of photons/bit at the receiver input. For $\mathrm{BER} = 10^{-9}$, the *quantum limited receiver sensitivity* of the *pin*-receiver is $\overline{n} = 10$ photons/bit (average power), or $n_1 = 20$ photons per '1'-bit. Note that specifying the receiver sensitivity in terms of *photons / bit* leads to more fundamental statements than specifying it in terms of an average optical power $\overline{P}$ ([W] or [dBm]), since both wavelength dependence and bit-rate dependence of receiver performance are eliminated. The two measures are related via

$$\overline{P} = \overline{n} hfR \qquad (9)$$

where R denotes the data rate. The intriguingly low-receiver sensitivity of *pin*-receivers, however, does not apply to practically implementable receivers, since in reality electronics noise by far dominates shot noise. As a consequence, receiver sensitivities achieved by *pin*-receivers are typically 20–30 dB off the quantum limit: Assuming an equivalent noise current density of 10 $\mathrm{pA}/\sqrt{\mathrm{Hz}}$ and a 10-Gbit/s receiver with some 7-GHz bandwidth operating at a wavelength of 1550 nm, the electronics noise variance amounts to $7 \cdot 10^{-13}\mathrm{A}^2$, while the '1'-bit shot noise variance going with the detection of 10 photons/bit comes to about $4 \cdot 10^{-17}\mathrm{A}^2$, four orders of magnitude below electronics noise. For realistic BER, the Q-factor is thus entirely determined by electronics noise, and for $Q = 6$ we arrive at a receiver sensitivity of some $\overline{n} \approx 5000$ photons/bit, 27 dB above the quantum limit.

To achieve higher receiver performance, more advanced receiver types must be employed. There are basically three ways to proceed: *Avalanche photodetection, coherent detection*, and *optically preamplified detection*. These rather diverse techniques, which will be discussed in the following sections, have still one common attribute: They all amplify the received signal before or at the stage of photodetection, while at the same time introducing additional noise. In the limit when the newly introduced noise terms dominate electronics noise, receiver performance becomes independent of electronics noise, leading to the respective quantum limits. In that limit, any further increase of the respective gain mechanism does *not* affect receiver performance any more. In contrast to *pin*-receivers, the quantum limits can be closely approached with these receiver types in experimental reality.

3.4. Avalanche Photodiode (APD) Receiver

An avalanche photodiode (APD) is the semiconductor equivalent to a photomultiplier tube. The incoming light generates primary electron-hole pairs (like in a *pin*-diode), which are then accelerated in a high-field region to launch an avalanche multiplication process through ionizing collisions [43]. The average number of the resulting secondary electron-hole pairs relative to the primary electron-hole pairs is called *avalanche gain* M_{APD}. The avalanche multiplication process is by itself a random process, since the exact number of secondary electron-hole pairs generated by a primary pair varies randomly. The unavoidable shot noise present for primary photodetection (cf. Fig. 19) is thus enhanced. This increase in detection noise is quantitatively captured in the APD's *noise enhancement factor* $F_{APD} > 1$ via the *multiplied shot noise* relationship [43]

$$\begin{aligned}\sigma^2_{shot,APD}(t) &= (R_T^2)\cdot eSM^2_{APD}F_{APD}(p * h^2)(t)\\ &\approx (R_T^2)\cdot 2eSM^2_{APD}p(t)F_{APD}B_e \end{aligned} \tag{10}$$

where the approximation, again, holds for optical power variations slow compared to the detection electronics' speed. In addition to multiplied shot noise, an APD also generates multiplied dark current shot noise through avalanche multiplication of dark current charge carriers, which is stationary and independent of the optical power, and can thus be added to the electronics noise variance.

In the desired limit when multiplied shot noise dominates electronics noise, the Q-factor for high-signal extinction ratios ($s_0 \ll s_1$) approaches[5]

$$\begin{aligned}Q &= \frac{(R_T)\cdot SM_{APD}P_1}{\sigma_{elec}+\sqrt{\sigma^2_{shot,APD}+\sigma^2_{elec}}}\\ &\xrightarrow{\sigma^2_{shot,APD}\gg\sigma^2_{elec}} \sqrt{\frac{\eta\overline{n}R}{F_{APD}B_e}} \sim \sqrt{2\overline{n}/F_{APD}}\end{aligned} \tag{11}$$

[5] Note that while Eq. (11) reveals general trends, care has to be taken with quantitative predictions, since the Gaussian assumption of detection statistics breaks down for shot-noise limited direct detection: Specializing equation (11) for *pin*-reception ($F_{APD} = 1$), we arrive at a quantum limit of $\overline{n} =$ 18 photons/bit, which is off its correct value by 2.6 dB.

with $P_1 = 2\overline{P}$ equal to the '1'-bit optical signal power for NRZ-OOK. Thus, the excess noise factor F_{APD} takes the role of a noise figure in degrading detection performance. Note that optimum performance of an APD receiver is in general *not* attained at the highest possible multiplication M_{APD}, since F_{APD} is a complicated and highly technology-dependent function of M_{APD}, necessitating joint optimization of M_{APD}, F_{APD}, and σ^2_{elec} [40,43,46].

Good APDs ($M_{APD} \sim 100, F_{APD} \sim 5$) for operation up to 1 Gbit/s are available in Silicon technology, which limits their operating range to wavelengths below $\sim$ 1100 nm. Receiver sensitivities of 200 photons/bit have been achieved at 50 Mbit/s [47]. InGaAs or InAlAs-based APDs for use in the 1550-nm wavelength region, however, exhibit fairly low multiplication ($M_{APD} \sim 10$) for 10 Gbit/s detection. Receiver sensitivities of 1000 photons/bit have been demonstrated at 10 Gbit/s [48].

3.5. Coherent Receiver

Another way of amplifying the signal and boosting the accompanying noise above the electronics noise floor is known as *coherent detection* [9,35,41,49]. A coherent receiver, as depicted in Fig. 20 a, combines the signal with a *local oscillator* (LO) laser by means of an optical coupler. Upon detection, the two fields beat against each other, and the average electrical signal reads

$$\begin{aligned}\langle s(t)\rangle = (R_T)\cdot S\{&\varepsilon P_s(t) + (1-\varepsilon)P_{LO}\\ &+ 2\mu\sqrt{\varepsilon(1-\varepsilon)}\sqrt{P_s(t)P_{LO}}\cos(2\pi f_{IF}t + \phi_s(t))\}\end{aligned} \tag{12}$$

where $P_s(t)$ and $\phi(t)$ denote the modulation-carrying received signal's power and phase, respectively, P_{LO} stands for the LO power, and f_{IF}, the beat frequency between signal and LO, is called intermediate frequency, since the IF signal is usually mixed down to baseband after photodetection, using standard microwave techniques. The parameter ε captures the splitting ratio of the optical coupler, which has to be chosen as high as possible such as not to waste too much signal power, and as low as acceptable to let sufficient LO power reach the detector to achieve shot-noise limited performance (see explanation below). The heterodyne efficiency μ accounts for the degree of spatial overlap as well as for the polarization match between LO field and signal field. If both LO and signal

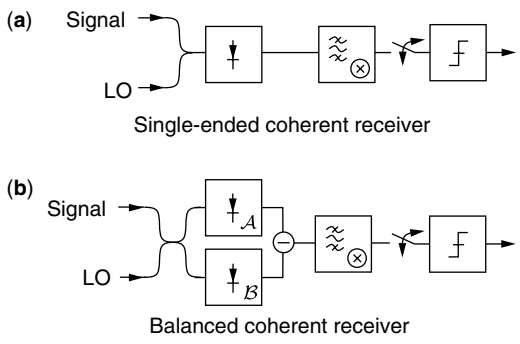


Figure 20. Single-ended (a) and balanced (b) coherent receiver. After photodetection, various kinds of electronic signal-processing can be performed.

are provided copolarized in single-mode optical fibers, μ equals unity.

If the frequency of the LO differs from the signal frequency, we speak of a *heterodyne* receiver. If LO and signal have the same frequency, such that $f_{IF} = 0$, we speak of a *homodyne* receiver. Homodyne detection strictly requires optical phase locking between the LO and signal optical fields, which implies significant technological effort.[6] In the desired range of operation, the LO power is chosen much stronger than the signal power $[(1-\varepsilon)P_{LO} \gg \varepsilon P_s(t)]$, and the first term in Eq. (12) can be neglected compared to the second and third. Filtering out the temporally constant second term in Eq. (12), we are then left with an exact replica of the received *optical field's amplitude* $\sqrt{P_s(t)}$ and *phase* $\phi_s(t)$ (as compared to the optical *power* accessible in direct-detection receivers). Thus, any amplitude or phase modulation scheme can directly be employed in combination with coherent receivers.

Due to the high LO power reaching the detector, the main noise contribution in a coherent receiver is the shot noise produced by the LO, $\sigma^2_{LO-shot} = 2eS(1-\varepsilon)P_{LO}B_e$. If this noise term dominates electronics noise ($\sigma^2_{LO-shot} \gg \sigma^2_{elec}$), optimum receiver performance is achieved. This limit is known as the *shot noise limit* in the context of coherent receivers. The highest receiver sensitivity with the potential of practical implementation that is known today can be achieved using homodyne detection of phase shift keying[7] (PSK), where the data bits are directly mapped onto the phase $\phi_s(t)$ of the optical signal, $\{0, 1\} \rightarrow \{0, \pi\}$. Without going into the derivations [9], the quantum limit for homodyne PSK can be shown to equal only 9 photons/bit, with a reported receiver sensitivity record of 20 photons/bit at 565 Mbit/s [51]. Using OOK instead of PSK, the sensitivity degrades by 3 dB. Going to heterodyne detection results in an additional loss of 3 dB in terms of receiver sensitivity. A detailed discussion of quantum limits for coherent receivers can be found in [9].

An alternative implementation of coherent receivers is shown in Fig. 20b. It makes use of *balanced detection* with $\varepsilon = 1/2$. While a balanced coherent receiver has *exactly* the same quantum-limited sensitivity as its single-ended equivalent, it offers the advantage of utilizing the full optical signal and LO power, and of being more robust to LO relative intensity noise (RIN).

Although still seriously considered for inter-satellite link applications due to the high achievable sensitivities, the interest in coherent receivers has vanished for fiber-optic systems with the availability of Erbium-doped fiber amplifiers (EDFA) in the early 1990s. To understand this evolution, let us look at the main advantages of coherent receivers, and how they have become outdated:

- Receiver sensitivities of coherent receivers by far outperform those achieved with pin-receivers and APDs, thus allowing for increased transmission distances in unamplified optical links. BUT: Optically preamplified receivers (cf. Section 3.6) exhibit similar receiver sensitivities to coherent receivers, are polarization-insensitive, and take less serious hits in performance if inline amplification is present.
- The possibility of correcting for chromatic dispersion in the microwave regime is offered in coherent detection, since both amplitude *and* phase of the optical field are converted to an electronic signal. BUT: Efficient and adaptive broadband phase corrections can only be performed on RF *bandpass* signals, asking for heterodyne detection. Since f_{IF} has to be chosen about 3 times the data rate [41], unrealistically high receiver front-end bandwidths would be needed for the high data rates used today. Conversely, all-optical dispersion compensators and adaptive optical filters are quickly advancing technologies [52], allowing for efficient phase corrections in the *optical* regime.
- Coherent receivers allow for the separation of closely spaced WDM channels by means of RF bandpass filters with sharp roll-offs. BUT: Optical filter technology has advanced dramatically, thus enabling channel spacings on the order of 10 GHz with sharp optical filter roll-offs, which opens up the possibility of *optical* channel filtering even for ultradense WDM applications.

But even if coherent reception is highly unlikely to reenter the high-speed fiber-optical communications market, the technique is far from being history: With coherent receiver terminals in their final product development phases [53], coherent receivers will soon find applications in nonfiber optical communications scenarios, most notably in free-space optical communications, where the enormous link distances of up to 80,000 km for geostationary intersatellite links ask for utmost receiver performance, and make homodyne PSK an attractive candidate.

3.6. Optically Preamplified Receiver

The historically youngest class of highly sensitive optical receivers uses *optical preamplification* to boost the weak received signal to appreciable optical power levels prior to detection, as shown in Fig. 21. At the same time, and fundamentally unavoidable, amplified spontaneous emission (ASE) with a power spectral density per (spatial and polarization) mode of

$$N_{ASE} = hfGF/2 \tag{13}$$

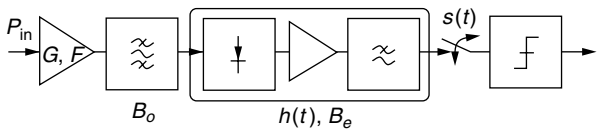


Figure 21. An optically preamplified receiver uses an optical amplifier in combination with an optical bandpass filter prior to detection with a *pin*-receiver.

[6] Using square-law detection of the electrical signal, one can also build a quasi-homodyne receiver without phase-locking and $f_{IF} \approx 0$ [50].

[7] Because the LO provides an optical phase reference, true PSK can be used instead of DPSK in direct detection receivers.

is introduced by the amplification process [43,54,55]; G and F denote the optical amplifier's gain and noise figure,[8] respectively. Today's high-performance Erbium-doped fiber amplifiers (EDFAs) exhibit gains between 15 dB and 30 dB, while having noise figures closely approaching their fundamental limit of 3 dB. These intriguing performance characteristics of EDFAs have made optically preamplified receivers the most important detection technique known today. Since the gain spectrum (and thus also the ASE spectrum) is much broader than the signal spectrum (typ. 30 nm in the 1550-nm wavelength band), an optical bandpass filter is employed to suppress out-of-band ASE. Upon detection, the random ASE field beats against the signal field (as well as against itself), leading to *signal-ASE beat noise* and *ASE-ASE beat* noise[9] after electrical filtering [42,16,17],

$$\sigma^2_{\mathrm{s-ASE}}(t) = (R_T^2)\cdot 2S^2 N_{ASE}\mathrm{Re}\left\{\iint_{-\infty}^{\infty} e(\tau)e^* \times(\tilde{\tau})r_n(\tau-\tilde{\tau})h(t-\tau)h(t-\tilde{\tau})\,d\tau d\tilde{\tau}\right\} \quad (14)$$

and

$$\sigma^2_{\mathrm{ASE-ASE}} = (R_T^2)\cdot M_{pol}S^2N_{ASE}^2\int_{-\infty}^{\infty}|r_n(\tau)|^2 r_h(\tau)\,d\tau \quad (15)$$

where $r_n(\tau)=\langle n^*(\tau)n(\tilde{\tau})\rangle$ is the autocorrelation of the optically filtered ASE field $n(t)$, and $r_h=\int h(\tau)h(\tau-t)\,d\tau$ is the autocorrelation of the detection electronics. The number of ASE modes reaching the detector is denoted M_{pol}. In a single-mode fiber system, M_{pol} usually equals 2, since polarization filtering typically is not done. The optically filtered signal field is denoted $e(t)$. In the limit of rectangular filters and constant input power P_{in} to the optical preamplifier, the above relations simplify to [56]

$$\sigma^2_{\mathrm{s-ASE}} \approx 4S^2GP_{in}N_{ASE}B_e \quad (16)$$

and

$$\sigma^2_{\mathrm{ASE-ASE}} \approx M_{pol}S^2N_{ASE}^2B_e(2B_o-B_e) \quad (17)$$

where B_o stands for the optical filter bandwidth. From Eqs. (14) through (17), we see that the signal-ASE beat noise variance is *nonstationary*, grows linearly with signal power, and is independent of B_o, as long as the optical filter does not significantly influence the signal spectrum. Also, we see that the ASE-ASE beat noise variance is *stationary*, grows linearly with B_o, and linearly depends on M_{pol}. Owing to the linear dependence of N_{ASE} on G (cf. Eq. (13)), both beat noise standard deviations as well as the detected electrical signal amplitude SGP_{in} grow linearly with G. Thus, the Q-factor becomes independent of G as soon as the beat noise terms starts to dominate electronics noise. This limit is called *optical noise limit* or *beat noise limit*, and forms the usual operating condition of optically preamplified receivers. Idealizing the beat noise limit ($F=3$ dB, $B_e=R/2$), we arrive at a quantum limit of 38 photons/bit for OOK, and of 20 photons/bit for DPSK [9]. Experimentally, sensitivities of 43 photons/bit at 5 Gbit/s [57] (52 photons/bit at 10 Gbit/s [58]) have been achieved for OOK, and 30 photons/bit at 10 Gbit/s [59] (45 photons/bit at 40 Gbit/s [36]) have been demonstrated for DPSK using balanced detection (cf. Fig. 16).

3.7. Bandwidth Optimization

Once the gain of the optical amplifier is chosen high enough to let optical beat noise dominate electronics noise, the main impact on receiver performance comes from optical and electrical filter characteristics. In the frame of a single-pulse theory (i.e., neglecting ISI), optimum receiver performance is achieved if the *optical* filter is matched to the data pulses,[10] and if the *electrical* filter is made sufficiently broadband to have no influence on the detected signal [41]. By constructing a receiver of this type, the same performance could in principle be achieved for NRZ and RZ signaling formats. However, the electrical bandwidth is usually upper-bounded by technological constraints and cost considerations, especially at data rates in the multi-Gbit/s regime, and ISI often *does* have a significant impact on detection. Thus, other than matched filter characteristics frequently turn out to be superior in practice, and the equality of NRZ and RZ formats is eliminated: NRZ formats usually take a noticeable hit in performance with respect to RZ '0'-bit ISI (cf. Figs. 4 and 23).

Figure 22 [60,61] shows a typical dependence of receiver performance on B_o and B_e for NRZ-OOK and 33%-duty cycle RZ-OOK. The contours give the dB-penalty relative to the quantum limit. RZ can be seen to have a better optimum sensitivity than NRZ, and, by the wider spacing of the contour lines, to be more tolerant to suboptimum choices of optical and electrical filters. For both formats, the performance decrease at larger-than-optimum filter bandwidths can be attributed to increased detection noise. At lower-than-optimum filter bandwidths, RZ is affected by pulse amplitude reductions due to filtering, while NRZ is predominantly affected by ISI. The different degrading effects for NRZ and RZ at low bandwidths become evident in Fig. 23, showing the dB-sensitivity penalty relative to the quantum limit for RZ and NRZ as a function of electrical filter bandwidth for fixed $B_o=2R$ (left) and as a function of optical filter bandwidth for fixed $B_e=0.8R$ (right). The solid curves apply for a pseudo-random bit sequence (PRBS) of length 2^7-1, and are section lines of Fig. 22. The dashed curves represent the results obtained for a single optical pulse, thus eliminating the effect of

[8] It has become common to call F a *noise figure*, although this terminology is somewhat sloppy, since it only considers the influence of signal-ASE beat noise [55].

[9] Note that the "beating"-picture is only correct in the frame of a classical consideration. Using quantum mechanical reasoning, the signal-ASE beat noise turns out to result from the fact that each photon is amplified by a random, integer number within the amplifier, similar to the random multiplication of electron-hole pairs in APDs [54].

[10] The impulse response of a *matched filter* is identical to the temporally reversed pulse to be detected [11].

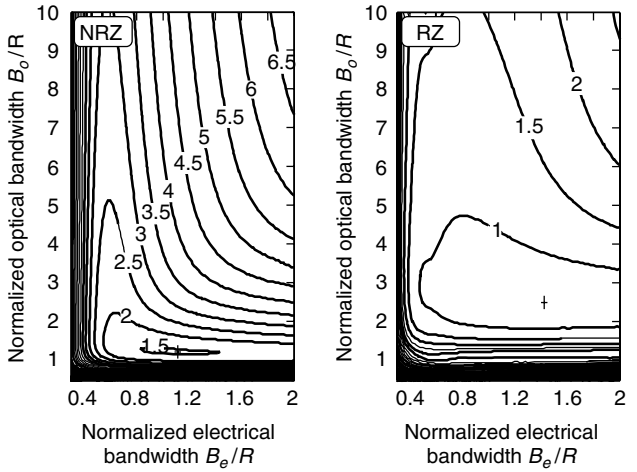


Figure 22. Dependence of receiver sensitivity on optical and electrical filter bandwidths for NRZ-OOK and 33%-duty cycle RZ-OOK. The contours are labelled in terms of dB-penalties relative to the quantum limit [60,61].

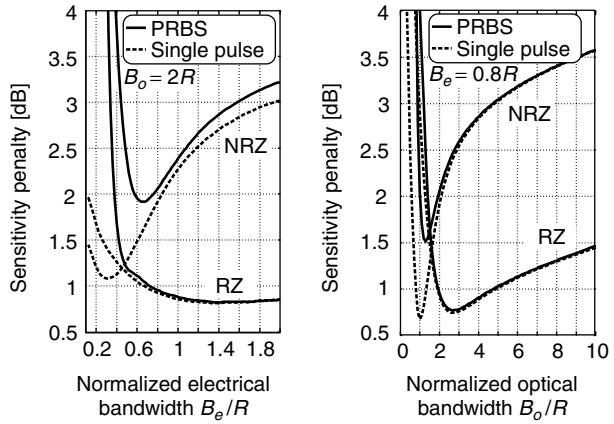


Figure 23. Sensitivity penalty to the quantum limit for NRZ and RZ as a function of electrical filter bandwidth (left) and optical filter bandwidth (right). Solid lines apply to ISI-corrupted detection, while dashed curves represent the ISI-free case [61].

ISI. For RZ, the ISI-free curves and the PRBS-curves run in parallel until well below the optimum bandwidth constellations, indicating that the RZ bandwidth optima are *not* influenced by ISI. For NRZ, however, the two curves depart for $B_e \lesssim 0.8R$ and $B_o \lesssim 1.5R$, which clearly shows that the optimum NRZ-receiver bandwidths are determined by trading ISI against detection noise.

In addition to the above bandwidth considerations, when optimizing operational, cost-effective receivers for WDM systems, one further has to take into account the effects of WDM channel crosstalk, optical source frequency offsets and drifts, filter concatenations effects due to a large number of optical add/drop multiplexers, as well as technological constraints on high-speed receiver bandwidths and receiver imperfections, such as jitter of the sampling phase.

3.8. Required Optical Signal-to-Noise Ratio (OSNR)

Specifying an optical receiver in terms of its *receiver sensitivity* dates back to pre-EDFA times, when the ultimate limit to fiber-optic link distances was given by the lowest possible receive power at which a specified receiver performance could still be guaranteed. With the deployment of in-line optical amplifiers this situation has changed, and optical signals can be transmitted over much longer distances through periodic optical reamplification. Since each amplifier fundamentally introduces ASE according to Eq. (13), it is now the *total* ASE N_{tot} accumulated along the transmission line per polarization mode rather than the received signal power level that sets limits on the maximum transmission distance, and the ability of a receiver to cope with ASE determines its performance in a system.

Figure 24 visualizes the situation of beat-noise limited detection in an in-line amplified transmission system. It shows the receiving end of a transmission line carrying a WDM signal with average per-channel power $\overline{P}_s^{\lambda_i}$, onto which the total ASE accumulated along the line is added. A WDM demultiplexer simultaneously acts to separate the WDM channels and to suppress out-of-band ASE. Comparing Fig. 24 to Fig. 21, we notice equivalence with

$$\overline{P}_s^{\lambda_i} \longleftrightarrow GP_{in} \quad \text{and} \quad N_{ASE} \longleftrightarrow N_{tot} \tag{18}$$

These substitutions are most conveniently captured in the definition of the *optical signal-to-noise ratio* (OSNR) as the ratio of the average optical signal power $\overline{P}_s^{\lambda_i}$ to the (unpolarized) ASE power within some reference bandwidth B_{ref},

$$\text{OSNR} = \frac{\overline{P}_s^{\lambda_i}}{2N_{tot}B_{ref}} \tag{19}$$

The bandwidth B_{ref} is typically (but not exclusively) chosen to be 0.1 nm at a wavelength of 1550 nm, that is, $B_{ref} \approx 12.5$ GHz. Using the relations (18), the OSNR required at a beat-noise limited receiver to attain a certain BER can be directly related to the input sensitivity of a preamplified receiver that is operated with a 'clean' input signal $\overline{P}_{in}$ (more precisely, with an input signal satisfying $GN_{in} \ll N_{ASE}$, where N_{in} denotes the ASE power spectral density at the optical amplifier input). It thus makes sense to also define the *quantum limited OSNR* as the minimum OSNR an ideal beat-noise limited direct detection receiver has to have at its input to produce a certain BER, for example, BER $= 10^{-9}$. The quantum-limited OSNR is then connected to the quantum limit of an optically preamplified receiver by

$$\text{OSNR} = \frac{\overline{n}R}{2B_{ref}}. \tag{20}$$

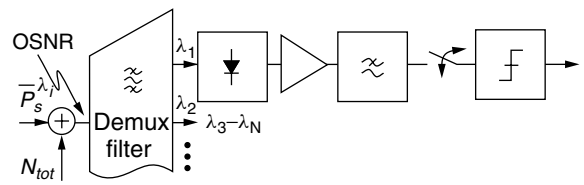


Figure 24. Beat-noise limited detection of optical WDM signals corrupted by noise accumulated along the transmission line through optical in-line amplification.

3.9. Photonic Integrated Receiver

A newly emerging class of optical receivers with a high potential for deployment is called *photonic integrated receiver* [62]. It is basically an optically preamplified receiver *without* optical bandpass filtering, and typically consists of a *pin*-photodiode following an on-chip integrated semiconductor optical amplifier (SOA). This receiver type is used to boost the optical signal power prior to detection to improve upon receiver performance, thus eliminating the need for an external optical preamplifier. Due to the absence of optical filtering, the bandwidth of the ASE generated by the SOA is solely determined by the amplifier's gain bandwidth, letting the ASE-ASE beat noise reach appreciable values. However, depending on the relationship of the ASE-ASE beat noise due to ASE from the SOA to the other receiver noise terms, significant improvements in receiver performance can be achieved. For example, if the signal-ASE beat noise originating from ASE produced along the transmission line is well above the ASE-ASE beat noise produced by the SOA, no receiver degradation will be noticed.

4. SUMMARY

In this article we discussed modulation formats with the potential of being used in high-speed fiber-optic communications. We distinguished between amplitude-modulated and phase-modulated formats, discussed the role of the number of signaling levels, and showed how the optical spectrum can be influenced to achieve high-spectral efficiency. We outlined optical receiver concepts, gave an introduction to their performance evaluation by means of frequently used performance measures, and discussed important receiver design trade-offs.

Acknowledgments
I would like to acknowledge many valuable discussions on modulation formats and receiver design with my present and former colleagues René-Jean Essiambre, Jake Bromage, Alan H. Gnauck, S. Chandrasekhar, Hoon Kim, Herwig Kogelnik, Martin Zirngibl, Klaus H. Kudielka, A. Kalmar, Martin M. Strasser, Martin Pfennigbauer, Martin Pauer, and Walter R. Leeb.

BIOGRAPHY

Peter J. Winzer was born in Vienna, Austria, in 1973. He studied electrical engineering/communications engineering at the Vienna University of Technology, and received his Dipl.-Ing. (M.S.) and Dr.techn. (Ph.D.) degrees in 1996 and 1998, respectively. His work, largely supported by the European Space Agency (ESA), was related to the analysis and modeling of noise in Doppler wind lidar and space-borne optical communication systems. Following his assistant professorship at the Vienna University of Technology, Dr. Winzer joined Bell Laboratories in Holmdel, New Jersey, in 2000, where he has since been working on fiber-optic communications, with an emphasis on Raman amplification, 40-Gbit/s optical transmitter and receiver optimization, and spectrally efficient optical modulation formats. Dr. Winzer has authored and co-authored some 60 papers and holds several patents. His present areas of interest include transmission and reception aspects in both fiber-optic and free-space optical communication systems.

BIBLIOGRAPHY

1. Record transmission distances and capacities are reported in the Post-Deadline Sessions of the annual conferences *Optical Fiber Communication* (OFC) and *European Conference on Optical Communication* (ECOC).
2. I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IV B*. Academic Press, 2002.
3. *Proc. Free-Space Laser Communication Technologies I* (1988) through *XIV* (2002), Proc. SPIE vols. 0885, 1218, 1417, 1635, 1866, 2123, 2381, 2699, 2990, 3266, 3615, 3932, 4272, and 4635; D. L. Begley, *Selected Papers on Free-Space Laser Communications I* (1991) and *II* (1994), Proc. SPIE vols. MS30 and MS100.
4. V. W. S. Chan, Optical space communications, *IEEE J. Selec. Top. Quantum Electron*. **6**: 959–975 (2000).
5. P. J. Winzer and W. R. Leeb, Space-borne optical communications — a challenging reality, *Proc. 15th Annual Meeting of the IEEE Lasers and Electro-Optics Society* (LEOS'02), 2002.
6. X. Lu and O. Sniezko, The evolution of cable TV networks, in [2], (2002).
7. A. S. Siddiqui et al., Dispersion-tolerant transmission using a duobinary polarization-shift keying transmission scheme, *IEEE Photon. Technol. Lett.* **14**: 158–160 (2002).
8. H. Kogelnik et al., Polarization-mode dispersion, in [2], (2002).
9. G. Jacobsen, *Noise in Digital Optical Transmission Systems*, Artech House, 1994.
10. J. Conradi, Bandwidth-efficient modulation formats for digital fiber transmission systems, in [2], (2002).
11. R. D. Gitlin et al., *Data Communications Principles*, Plenum Press, 1992.
12. S. Bigo et al., 10.2 Tbit/s (256×42.7 Gbit/s PDM/WDM) transmission over 100 km TeraLight^{TM} fiber with 1.28 bit/s/Hz spectral efficiency, *Proc. Optical fiber communication Conference* (OFC'01), paper PD25, (2001); W. Idler et al., Vestigial side band demultiplexing for ultra high capacity (0.64 bit/s/Hz) transmission of 128×40 Gb/s channels, *Proc. Optical fiber communication Conference* (OFC'01), paper MM3, 2001; Y. Frignac et al., Transmission of 256 wavelength-division and polarization-division-multiplexed channels at 42.7 Gb/s (10.2 Tb/s capacity) over 3×100 km of TeraLight (*TM*) fiber, *Proc. Optical fiber communication Conference* (OFC'02), paper FC5, 2002.
13. D. A. Ackerman et al., Telecommunication lasers, in I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IVA*, Academic Press, 2002.
14. A. Ougazzaden et al., 40Gb/s tandem electro-absorption modulator, *Proc. Optical fiber communication Conference* (OFC'01), paper PD14, 2001.
15. H. Kim and A. H. Gnauck, Chirp characteristics of dual-drive Mach-Zehnder modulator with a finite DC extinction ratio, *IEEE Photon. Technol. Lett.* **14**: 298–300 (2002).
16. L. Boivin and G. J. Pendock, Receiver sensitivity for optically amplified RZ signals with arbitrary duty cycle, *Proc. Optical*

Amplifiers and their Applications (OAA'99), paper ThB4, 106–109, (1999).

17. P. J. Winzer and A. Kalmar, Sensitivity enhancement of optical receivers by impulsive coding, *J. Lightwave Technol.* **17**: 171–177 (1999).
18. M. Suzuki et al., Transform-limited 14 ps optical pulse generation with 15 GHz repetition rate by InGaAsP electroabsorption modulator, *Electron. Lett.* **28**: 1007–1008 (1992).
19. R. -J. Essiambre, B. Mikkelsen, and G. Raybon, Pseudo-linear transmission of high-speed TDM signals: 40 and 160 Gb/s, in [2], (2002).
20. P. B. Hansen et al., 5.5-mm long InGaAsP monolithic extended-cavity laser with an integrated Bragg-reflector for active mode-locking, *IEEE Photon. Technol. Lett.* **4**: 215–217 (1992).
21. N. M. Froberg et al., Generation of 12.5Gbit/s soliton data stream with an integrated laser-modulator transmitter, *Electron. Lett.* **30**: 1880–1881 (1994).
22. J. J. Veselka et al., A soliton transmitter using a CW laser and an NRZ driven Mach-Zehnder modulator, *IEEE Photon. Technol. Lett.* **8**: 950–952 (1996).
23. P. J. Winzer and J. Leuthold, Return-to-zero modulator using a single NRZ drive signal and an optical delay interferometer, *IEEE Photon. Technol. Lett.* **13**: 1298–1300 (2001).
24. A. Lender, The duobinary technique for high-speed data transmission, *IEEE Trans. on Commun. Electronics* **82**: 214–218 (1963).
25. D. Penninckx et al., The phase-shaped binary transmission (PSBT): A new technique to transmit far beyond the chromatic dispersion limit, *IEEE Photon. Technol. Lett.* **9**: 259–261 (1997); D. Penninckx et al., Relation between spectrum bandwidth and the effects of chromatic dispersion in optical transmissions, *Electron. Lett.* **32**: 1023–1024 (1996).
26. J. B. Stark, J. E. Mazo, and R. Laroia, Phased amplitude-shift signaling (PASS) codes: Increasing the spectral efficiency of DWDM transmission, *Proc. European Conf. on Optical Communication* (ECOC'98): 373–374, 1998; J. B. Stark, J. E. Mazo, and R. Laroia, Line coding for dispersion tolerance and spectral efficiency: Duobinary and beyond, *Proc. Optical Fiber Communication Conference* (OFC'99), paper WM46, 1999.
27. T. Ono et al., Characteristics of optical duobinary signals in Terabit/s capacity, high-spectral efficiency WDM systems, *J. Lightwave Technol.* **16**: 788–797 (1998).
28. K. S. Cheng and J. Conradi, Reduction of pulse-to-pulse interaction using alternative RZ formats in 40-Gb/s systems, *IEEE Photon. Technol. Lett.* **14**: 98–100 (2002).
29. T. Franck, T. N. Nielsen, and A. Stentz, Experimental verification of SBS suppression by duobinary modulation, *Proc. European Conf. on Optical Communication* (ECOC'97): 71–74, (1997).
30. X. Wei et al., 40 Gb/s duobinary and modified duobinary transmitter based on an optical delay interferometer, *Proc. European Conf. on Optical Communication* (ECOC'02): paper 09.6.3, 2002.
31. Y. Miyamoto et al., S-band 3×120-km DSF transmission of 8×42.7-Gbit/s DWDM duobinary-carrier-suppressed RZ signals generated by novel wideband PM/AM conversion, *Proc. Optical Amplifiers and their Applications* (OAA'01), paper PD6, 2001.
32. N. S. Bergano, Undersea communication systems, in [2] (2002).
33. C. R. Menyuk et al., Dispersion managed solitons and chirped RZ: What is the difference?, in [2] (2002).
34. R. A. Griffin et al., Integrated 10 Gb/s chirped return-to-zero transmitter using GaAs/AlGaAs modulators, *Proc. Optical Fiber Commun. Conf.* (OFC'01), paper PD15, 2001.
35. G. Einarsson, *Principles of Lightwave Communications*, John Wiley & Sons, 1996.
36. A. H. Gnauck et al., 2.5 Tb/s (64×42.7 Gb/s) transmission over 40×100 km NZDSF using RZ-DPSK format and all-Raman-amplified spans, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FC2, 2002.
37. T. Chikama et al., Modulation and demodulation techniques in optical heterodyne PSK transmission systems, *J. Lightwave Technol.* **8**: 309–321 (1990).
38. R. A. Griffin and A. C. Carter, Optical differential quadrature phase-shift key (oDQPSK) for high capacity optical transmission, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper WX6, 2002; R. A. Griffin et al., 10Gb/s optical differential quadrature phase shift key (DQPSK) transmission using GaAs/AlGaAs integration, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FD6, 2002.
39. S. D. Personick, Receiver design for digital fiber optic communication systems, I, *Bell Syst. Tech. J.* **52**: 843–874 (1973).
40. G. P. Agrawal, *Fiber-Optic Communication Systems*, 3rd edition, John Wiley & Sons, 2002.
41. L. Kazovsky, S. Benedetto, and A. Willner, *Optical Fiber Communication Systems*, Artech House, 1996.
42. P. J. Winzer, Receiver noise modeling in the presence of optical amplification, *Proc. Optical Amplifiers and their Applications* (OAA'01), paper OTuE16, 2001; P. J. Winzer, Performance estimation of receivers corrupted by optical noise, in J. D. Minelly, and Y. Nakano, eds., *OSA Trends in Optics and Photonics* (TOPS) vol. 60, N. Jolley, 268–273, 2001.
43. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, John Wiley & Sons, 1991.
44. B. E. A. Saleh, *Photoelectron Statistics*, Springer-Verlag Berlin Heidelberg, New York, 1978.
45. L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, 1995.
46. S. D. Personick, Receiver design, in S. E. Miller and A. G. Chynoweth (eds.), *Optical Fiber Telecommunications* Academic Press, 1979; B. L. Kasper, Receiver design, in S. E. Miller and I. P. Kaminow (eds.), *Optical Fiber Telecommunications II*, Academic Press, 1988; K. Ogawa et al., I. P. Kaminow and T. L. Koch (eds.), Advances in high bit-rate transmission systems, in *Optical Fiber Telecommunications IIIA*, Academic Press, 1997; B. L. Kasper, O. Mizuhara, and Y. -K. Chen, High bit-rate receivers, transmitters, and electronics, in I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IVA*, Academic Press, 2002; T. V. Muoi, Receiver design for high-speed optical-fiber systems, *J. Lightwave Technol.* **2**: 243–267 (1984). J. N. Hollenhurst, Fundamental limits on optical pulse detection and digital communication, *J. Lightwave Technol.* **13**: 1135–1145 (1995); S. B. Alexander, *Optical Communication Receiver Design*, SPIE tutorial texts in Optical Engineering, vol. TT22, 1997.
47. G. Planche et al., SILEX final ground testing and in-flight performance assessment, *Proc. SPIE* **3615**: 64–77 (1999).

48. K. Sato et al., Record highest sensitivity of -28 dBm at 10 Gb/s achieved by newly developed extremely-compact superlattice-APD module with TIA-IC, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FB11, 2002.

49. S. Betti, G. De Marchis, and E. Iannone, *Coherent optical communication systems*, Wiley-Interscience, 1995; S. Ryu, *Coherent Lightwave Communication Systems*, Artech House, 1995.

50. L. G. Kazovsky, P. Meissner, and E. Patzak, ASK multiport optical homodyne receivers, *J. Lightwave Technol.* **5**: 770–790 (1987).

51. B. Wandernoth, 20 photon/bit 565 Mbit/s PSK homodyne receiver using synchronisation bits, *Electron. Lett.* **28**: 387–388 (1992).

52. C. R. Doerr, Planar lightwave devices for WDM, in I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IVA*, Academic Press, 2002.

53. K. Kudielka and K. Pribil, Transparent Optical Intersatellite Link Using Double-Sideband Modulation and Homodyne Reception, *Int. J. Electron. Commun.* (AE), **56**: 254–260 (2002).

54. E. Desurvire, *Erbium-Doped Fiber Amplifiers*, John Wiley & Sons, 1994.

55. H. A. Haus, *Electromagnetic Noise and Quantum Optical Measurements*, Springer Verlag, 2000.

56. N. A. Olsson, Lightwave systems with optical amplifiers, *J. Lightwave Technol.* **7**: 1071–1082 (1989).

57. D. O. Caplan, and W. A. Atia, A quantum limited optically-matchedcommunication link, *Proc. Optical Fiber Communication Conference* (OFC01), paper MM2, 2001.

58. M. M. Strasser, M. Pfennigbauer, M. Pauer, and P. J. Winzer, Experimental verification of optimum filter bandwidth in direct-detection (N)RZ receivers limited by optical noise, *Proc. LEOS 2001 Annual Meeting* (LEOS'01): 485–486, 2001.

59. W. Atia and R. S. Bondurant, Demonstration of return-to-zero signaling in both OOK and DPSK formats to improve receiver sensitivity in an optically preamplified receiver, *Proc. 12th annual meeting of LEOS*: 2244–225, 1999.

60. P. J. Winzer et al., Optimum bandwidths for optically preamplified RZ and NRZ reception, *J. Lightwave Technol.* **9**: 1263–1273 (2001).

61. M. Pfennigbauer et al., Performance optimization of optically preamplified receivers for return-to-zero and non return-to-zero coding, *Int. J. Electron. Commun.* (AE) **56**: 261–268 (2002).

62. B. Mason et al., 40Gb/s photonic integrated receiver with -17dBm sensitivity, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FB10, 2002.

OPTICAL TRANSPORT SYSTEM ENGINEERING

Milorad Cvijetic
NEC America
Herndon, Virginia

1. INTRODUCTION

In the time officially deemed as "the information era," we are witnessing the insatiable demand for high information capacity and distance-independent connectivity. Optical networking has been the most efficient solution in satisfying this ongoing demand for bandwidth and connectivity. Optical fiber has been laid down all the way to the curb, building, home, and desk. In general, all optical networks can be considered as part of a global optical network; they are all owned either by private enterprises or by telecommunication carriers.

Several logical parts in a global optical network can be identified, as illustrated in Fig. 1:

- *The core optical network*, which is a long-haul network interconnecting big cities or major communication hubs. Connections between big cities on different continents are made by submarine optical cables. The *core network* is a generic name, but very often we refer to the core network as a wide-area network (WAN) if it belongs to an enterprise, or as the interchange network if it is operated by a telecommunication carrier.
- *The edge optical network*, which covers a smaller geographical area, usually a metropolitan area. Again, we can refer to the edge network either as a metropolitan-area network (MAN) if it belongs to an enterprise, or as a local exchange network if it is operated by telecommunication carriers.
- *The access optical network*, which is the part of the network related to last-mile access and bandwidth distribution to individual end users (in corporate, government, medical, entertainment, scientific, and private sectors). Both the enterprise local-area networks (LANs) and the distribution part of the carrier network connecting the central office with individual users belong to the access network.

The physical network topology that best supports traffic demand is generally different for different parts of a global optical network, as presented in Fig. 1. It could vary between mesh, ring, or star topology. In spite of different network topologies, the main consideration of optical transport engineering is always an optical (lightwave) path, since an optical network is just the means of supporting an end-to-end connection via the lightwave path. Optical transport engineering is related to the physical layer of an optical network, and takes into account the optical signal propagation length, characteristics of optical elements used (fibers, lasers, amplifiers etc.), modulation bit rate, networking impact and transmission requirements.

In this article we will introduce fundamentals of optical transport system engineering. Although an optical signal can take on either a digital or an analog form, our focus will be on digital signals, since the majority of modern applications are related to digital signal transmission.

2. OPTICAL TRANSMISSION PARAMETERS

The simplest optical transmission system is a point-to-point connection on a single optical wavelength, which propagates through an optical fiber. An upgrade to this is the deployment of WDM (wavelength-division multiplex) technology, where multiple optical wavelengths are

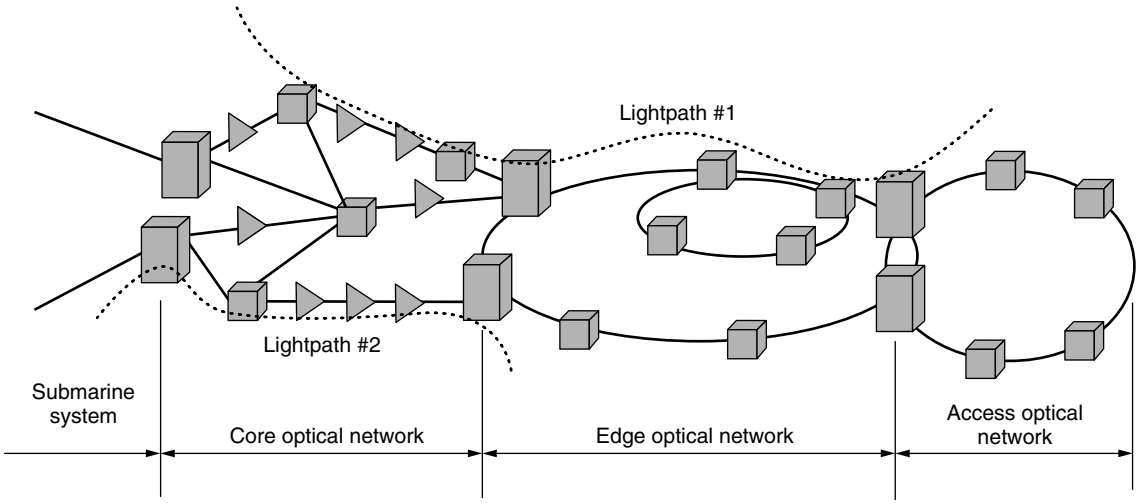


Figure 1. Optical networking structure.

combined to travel over the same physical path. WDM technology originally served to increase the bandwidth of already installed fiber, but it has quickly become the foundation of optical networking by combining optical signal transport over arbitrary distances with wavelength routing and optical protection.

The general scheme of an optical transport system is shown in Fig. 2. Several optical channels, carrying independent modulation signals, have been multiplexed by WDM technology, and sent to the optical fiber line. The aggregated signal is then transported over some distance before it is demultiplexed and detected (converted back to an electrical level). The optical signal transmission path can include a number of optical amplifiers, crossconnects, and optical add/drop multiplexers. The illustrated set of parameters, related either to enabling technologies or to transmission and networking issues, can be attached to Fig. 2.

Providing stable and reliable operation of an optical transport system over time requires proper design and engineering. Optical transport systems engineering involves accounting for all effects that can alter an optical signal on its way from the source (laser) on through photodetection by photodiode, and then to the threshold decision point. Different impairments will degrade and compromise the integrity of the signal before it arrives to the decision point to be recovered from corruptive additives (noise, crosstalk, and interference). The transmission quality is measured by the received signal-to-noise-ratio (SNR), which is defined as the ratio of the signal level to the noise level at the threshold point. The other parameter used to measure signal quality is the bit error rate (BER). BER is interrelated with SNR and defines the probability that a signal space (or a logic 0) will be mistaken for a signal mark (a logic 1), and vice versa. The main goal in optical signal transport is to achieve the required BER between end-to-end users, or between two specified points. Evaluating the BER requires determining the received signal level at the threshold point, calculating the noise power, and quantifying and including the influence of various relevant impairments.

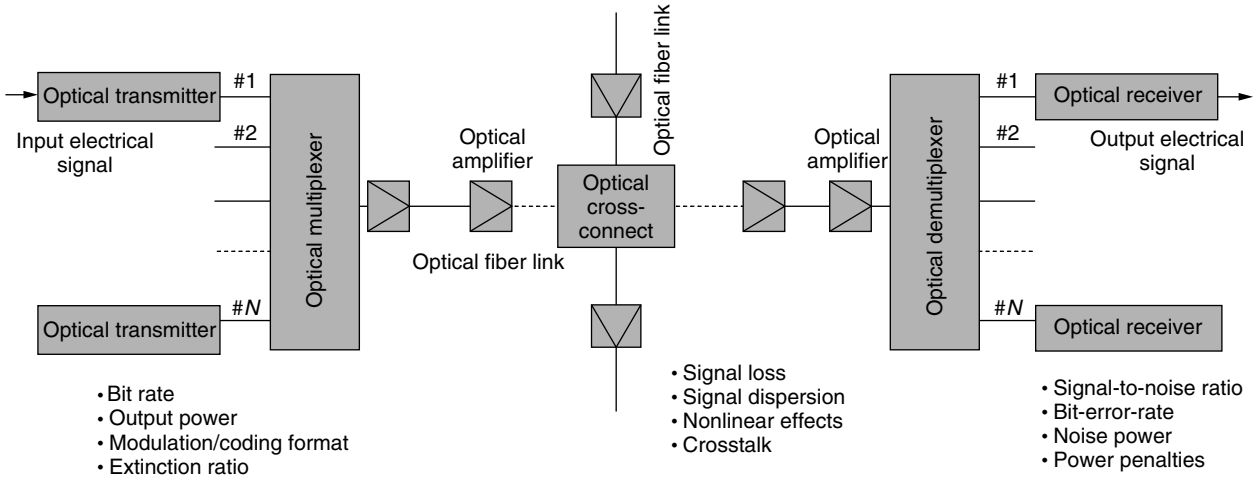


Figure 2. Optical DWDM system.

2.1. Optical Signal Path

The end-to-end signal path from Fig. 1 includes both the electrical and optical path portions. Conversion from the electrical to the optical level is done in the optical transmitter, while conversion from the optical level to an electrical signal takes place in the optical receiver. The key elements on an optical signal path are

1. *Semiconductor lasers* that convert an electrical signal to optical radiation. The bias current flows through the laser p-n junction and stimulates the recombination of electrons and holes, producing photons. If the current is higher than a certain threshold, recombination occurs in an organized way, with strong correlation in phase, frequency, and direction of radiated photons that form the output optical signal (called "stimulated emission of radiation.") Semiconductor lasers could either be directly modulated by an electrical signal or simply biased by a DC voltage and operate in combination with an external optical modulator. Each laser generates a specified optical wavelength, but some spectral linewidth is associated with the generated optical signal as well. These lasers are known as single-mode lasers (SMLs), characterized by a distinguished single longitudinal mode in the optical spectrum. If a set of separated longitudinal modes can be recognized under the optical spectrum envelope, the lasers are called *multimode lasers* (MMLs).
2. *Optical fibers* that transport an optical signal to its destination. The combination of low signal loss and extremely wide transmission bandwidth allows high-speed optical signals to be transmitted over long distances before regeneration becomes necessary. There are two groups of optical fibers. The first group, called *multimode optical fibers*, transfer light through multiple spatial or transversal modes. Each mode, defined through a specified combination of electric and magnetic field components, occupies a different cross-sectional area of the optical fiber core, and takes a slightly distinguished path along the optical fiber. The difference in mode pathlengths causes a difference in arrival times at the receiving point. This phenomenon is known as multimode dispersion and causes signal distortion and limitations in transmission bandwidth. The second group of optical fibers effectively eliminates multimode dispersion by limiting the number of modes to just one through a much smaller core diameter. These fibers, called *single-mode optical fibers*, do however introduce another signal impairment known as *chromatic dispersion*. Chromatic dispersion is caused by a difference in velocities among different wavelength components within the same pulse. There are several methods to minimize chromatic dispersion at a specified wavelength, involving either the introduction of new single-mode optical fibers, or the utilization of different dispersion compensation methods.
3. *Optical amplifiers* that amplify weak incoming optical signals through the process of stimulated emission, without conversion back to the electrical level. Optical amplifiers should provide enough gain to amplify a specified number of optical channels. There are different types of optical amplifiers currently in use, such as semiconductor optical amplifiers (SOAs), erbium-doped fiber amplifiers (EDFAs), or Raman amplifiers. Amplifier parameters are gain, gain flatness over amplification bandwidth, output power, bandwidth, and noise power. The noise generated in an optical amplifier occurs due to a spontaneous emission process that is not correlated with the signal. All amplifiers degrade the SNR of the output signal because of amplified spontaneous emission (ASE) that adds itself to the signal during its amplification. SNR degradation is measured by the noise figure. Optical amplifiers can take several positions along the optical path, as indicated in Fig. 2. The output power could be enhanced by a booster amplifier within a transmitter, on the transmission line (inline amplifier), or before the receiver (to act as a preamplifier to increase the receiver sensitivity).
4. *Photodiodes* that convert an incoming optical signal back to the electrical level through a process just opposite to the one that takes place in lasers. Photodiodes can be classified into PIN or avalanche photodiodes (APD). The process within the PIN photodiodes is characterized by *quantum efficiency*, which is the probability that each photon will generate an electron–hole pair. In the avalanche photodiode each primary electron–hole pair is accelerated in a strong electric field, which can cause the generation of several secondary electron-hole pairs through the effect of impact ionization. This process is random in nature and avalanchelike.

A more elaborate analysis of advanced optical transport systems can be found in the bibliography [1–4].

2.2. Optical Signal Parameters

There are a set of parameters along the lightwave path that determines the received signal power:

- *Output power* from the laser/modulator coupled to the fiber pigtail. Optical power is defined per individual wavelength and depends on the lasers/modulators used. The output optical power is usually expressed in decibels per milliwatt (dB_m), defined as $dB_m = 10\log(P)$, where the output power P is expressed in milliwatts.
- *The extinction ratio*, which is the ratio between the optical power related to a logic 1 (mark) to the power related to a logic 0 (space). By increasing the extinction ratio, the signal to noise ratio is increased as well, but at the cost of additional penalties in the modulation speed and laser chirp.
- *Optical amplifier gain*, which determines the level of an optical signal that is being amplified. Optical amplifier gain is correlated to noise parameters,

which means that higher gain will generate more noise and vice versa.

- *Photodiode responsivity*, which defines the ratio between the number of electrical carriers produced and the number of incoming photons.

2.3. Noise Parameters

The total noise under consideration in optical transport system engineering is generated along the lightwave path and during the photodetection process. There are some additive noise components (the noise components remaining even if the signal is not present), and some multiplicative noise components (which are produced only if the signal is present). The additive noise components are

- *Dark-current noise* generated in photodiodes due to the thermal process.
- *Amplified spontaneous emission* (ASE) noise generated by any optical amplifier along the lightwave path.
- *Crosstalk*, which occurs in multichannel systems. Components introducing crosstalk in WDM systems are optical filters, optical multiplexers and demultiplexers, optical switches, semiconductor optical amplifiers, and optical fibers through nonlinearities. Crosstalk can either be intrachannel (occurs when another signal of the same wavelength interferes with the signal in question), or interchannel (occurs when some portion of a neighboring channel has been spread out, and detected by the specified signal's receiver).
- *Thermal noise*, which is created in the resistive part of the input impedance of an electrical preamplifier that follows the photodiode. The noise created in the electronic amplification stages following the preamplification process is thermal noise in nature as well.

The multiplicative noise components are

- *Avalanche shot noise*, caused by the random nature of the amplification of primary electron–hole pairs through the effect of impact ionization in avalanche photodiodes.
- *Laser intensity noise*, which occurs as a result of microvariations in the laser output power intensity. This noise is characterized through the relative intensity noise (RIN) parameter, and is more relevant for analog transmission systems.
- *Laser phase noise*, which is related to microvariations in phase of generated photons. The output optical signal, as a collection of individual photons, exhibits finite nonzero spectral width.
- *Modal noise*, which arises in multimode fibers through the random process of excitation of transversal modes.

2.4. Impairment Parameters

Impairment parameters relevant to optical transport system engineering are either optical power related or optical wavelength related. They can also be constant or time-dependent. Each of them results in a *signal power penalty*, which means that a higher signal power is required at the receiver to keep the BER at a level that would exist if the impairment were negligible.

Optical power-related impairments are

- *Optical fiber attenuation*, or fiber loss, which is the ratio between the output and input optical power at the defined optical fiber section. Optical attenuation is characterized by an attenuation coefficient α, usually expressed in decibels per kilometer. The decibel is defined as $\mathrm{dB} = 10\log(P_2/P_1)$, where P_2 and P_1 are the output and input power respectively.
- *Insertion losses* in different optical components along the lightwave path, such as optical connectors, optical splices, optical couplers, optical multiplexers, and optical filters. These losses are sometimes added to the optical fiber loss and are considered together.

Impairments that are optical power related, but are also functions of time are

- *Polarization mode dispersion* (PMD), a stochastic process that appears in real fibers, caused by variations in the shape of their core along the fiber length. The light in an optical fiber can be considered as a superposition of two polarized components. If they travel at the same speed, no polarization dispersion occurs, but if they travel at different speeds, as in real fibers, the light will separate into its faster and slower components, leading to a difference in propagation of the two polarization states and to pulse spreading. Both mechanical stresses and temperature effects contribute to PMD.
- *Polarization-dependent loss* (PDL), which is similar in nature to PMD, but this time the difference between polarization states is in transmission losses, rather than in arrival times. These differential losses accumulate in the system, since there might be many components having polarization-dependent loss. Since polarization fluctuates with time, the SNR at the end of the lightwave path will fluctuate as well, causing a power penalty.

Impairments that are dependent on both the optical power and optical wavelengths are

- *Four-wave mixing* (FWM), a nonlinear effect where a new optical frequency is generated when three frequencies mutually interact. It causes crosstalk noise in channels, since the newly generated optical frequency might coincide with one of the original channels.
- *Stimulated Raman scattering* (SRS), a nonlinear effect that occurs when a propagating optical power interacts with glass molecules in the fiber undergoing a wavelength shift. The result is a power transfer from one wavelength to another, which causes crosstalk between channels.

Impairments that are dependent on optical wavelength, but are also functions of time are

- *Chromatic dispersion* caused by the dependence of the fiber refractive index on the wavelength. Since a laser is not an ideal monochromatic source, each pulse in its time domain contains different spectral components that travel at different velocities through an optical fiber. Chromatic dispersion induces pulse broadening when the neighboring pulses cross their allotted time slot borders, which can severely limit system transmission rates. Chromatic dispersion is also a cumulative effect that increases with optical fiber length.
- *Laser chirp*, which is the modulation of optical frequency (or wavelength) when the optical signal, is intensity-modulated by a specific electrical waveform. The change in the frequency causes laser spectral linewidth broadening and, in the interaction with chromatic dispersion, leads to optical pulse distortion and the intersymbol interference effect [see Eq. (4)]. Chirp can be reduced by decreasing the extinction ratio, but this decrease would introduce additional power penalties, requiring some compromise [5].

Finally, there are some impairments that are dependent on both the optical power and optical wavelength, and are also functions of time:

- *Stimulated Brillouin scattering* (SBS), a nonlinear effect that occurs when a high optical power reflects off the grating formed by acoustic vibrations, downshifts in optical frequency, and comes back. The SBS can cause signal attenuation if the launched power is higher than a certain threshold. Optical signal dithering with low frequencies helps to increase the SBS threshold and effectively suppress SBS effect.
- *The self-phase modulation* (SPM) *effect*, which results from the fact that a higher fiber refractive index causes wavelengths at the center of the pulse to accumulate phase more quickly than at the wings. This stretches the wavelengths at the leading edge (called "red shift") of the pulse and compresses wavelengths at the trailing edge ("blue shift"). This phase modulation effect broadens the spectrum, which causes pulse spreading. If combined with positive dispersion in the optical fiber under controlled conditions, it can lead to suppression of the chromatic dispersion effect. This is the basis for soliton transmission, where return-to-zero (RZ) soliton pulses propagate over very long distances without optoelectronic regeneration.
- *Cross-phase modulation* (XPM), which has the same nature as SPM, occurs following the interaction between multiple optical frequencies.

3. ASSESSMENT OF THE OPTICAL TRANSPORT LIMITATIONS AND PENALTIES

3.1. Attenuation

A silica-based optical fiber is the central point of an optical signal transmission, offering wider available bandwidth, lower signal attenuation, and smaller signal distortion than other wired physical media. The output power P_2 from an optical fiber can be calculated from the input power P_1 and the optical attenuation coefficient α. For the lightwave path with length L, $P_2 = P_1 \exp(-\alpha L)$. If parameters P_2, P_1, and α are expressed in decibels, then the relation becomes $P_2 = P_1 - \alpha L$.

Four low-attenuation bands can be recognized within the usable optical bandwidth of silica-based optical fibers. They are usually referred as U, S, C, and L bands, although this nomenclature is not standardized yet. The C band occupies wavelengths from 1530 to 1560 nm, while the L band includes wavelengths between 1580 and 1610 nm. Both these bands have been considered as the most suitable bands for high-channel-count WDM transmission. The S band (sometimes called the S+ band) and U band (sometimes referred to as the S− band) cover shorter wavelengths down to approximately 1230 nm, where optical fiber attenuation is slightly higher than in the wavelength region covered by the C and L bands.

3.2. Noise

Detected photocurrent is the sum of signal and noise contributions after the photodetection process has taken place. It can be expressed as $I_p = I + i_s + i_{\text{th}}$, where I is the signal current calculated as a product of incoming optical power P and photodiode responsivity R (R is expressed in amperes per watt). Noise components, expressed by currents $i_s + i_{\text{th}}$, correspond to the quantum (shot) and thermal noise, respectively. The power of the total noise that appears after photodetection is equal to the product of the sum of noise spectral density components and the noise electric bandwidth Δf. In the case where direct detection without optical preamplification takes place, the total noise power can be expressed as

$$\langle i^2 \rangle_{tot} = \left[2qM^2F(M)I + \frac{4kT}{R_L} \right] \Delta f \tag{1}$$

where the first term in brackets describes the quantum noise, while the second is related to the thermal noise generated at a load resistance R_L. In the previous equation q represents the electron charge ($q = 1.6 \times 10^{-19}$ C), M is the avalanche amplification factor, $F(M)$ is the avalanche excess noise factor, k is Boltzmann's constant ($k = 1.38 \times 10^{-23}$ J/K), and T is absolute temperature in kelvins. If the PIN photodiode is used, the factor $M^2F(M)$ becomes unity.

In case an optical amplifier precedes the photodiode the major noise contribution comes from ASE noise. The spectral density of ASE noise is

$$S_{\text{sp}} = \frac{(G-1)N_f h\nu}{2} \tag{2}$$

where G is amplifier gain, N_f is the amplifier noise figure, h is Planck's constant ($h = 6.63 \times 10^{-34}$ J/Hz), and ν is optical frequency in hertz. The total noise power in this case becomes

$$\langle i^2 \rangle_{\text{tot}} = 2qR[GP + S_{\text{sp}}B_{\text{op}}]\Delta f + 4R^2 \text{GPS}_{\text{sp}}\Delta f + 2R^2 S_{\text{sp}}^2 [2B_{\text{op}} - \Delta f]\Delta f + \frac{4kT}{R_L}\Delta f \tag{3}$$

Parameter B_{op} refers to the optical filter bandwidth. A standard deviation $\sigma = [\langle i^2 \rangle_{tot}]^{1/2}$ is usually used in signal-to-noise ratio and bit-error-rate calculations [see Eq. (12)].

3.3. Chromatic Dispersion

Recall that chromatic dispersion is the cause of pulse spreading and the occurrence of intersymbol interference. Pulse spreading is proportional to the fiber dispersion parameter D, expressed in picoseconds per nanometer and kilometer (ps/nm.km). The dispersion parameter is an ascending linearlike, wavelength-dependent function, characterized by its zero value cross-point and a dispersion slope. There are several fiber types that differ in their dispersion parameter profile [6]:

- *Standard single-mode fibers* (SMFs), where the parameter D has zero value at the 1310 nm wavelength and the dispersion slope of approximately 0.072 ps/km.nm^2. With this, the chromatic dispersion parameter reaches the value of 17–20 ps/nm.km in the wavelength region belonging to C and L bands.
- *Dispersion-shifted fiber* (DSF), where the parameter D has zero value at the 1550 nm wavelength and the dispersion slope of approximately 0.09 ps/km.nm^2. The chromatic dispersion parameter can take on both negative and positive values in the wavelength region belonging to C and L bands. This fiber type has been good for single-wavelength transmission, but is not suitable for WDM applications because of high penalties due to nonlinear effects.
- *Non-zero dispersion-shifted fibers* (NZDSF), where the parameter D has zero value shifted from the 1550 nm wavelength and the dispersion slope of approximately 0.03 ps/km.nm^2. The chromatic dispersion parameter has some minimal value in the wavelength region belonging to C and L bands, thus minimizing penalties due to nonlinear effects.

The influence of chromatic dispersion and the penalties related to it can be evaluated by assuming that the pulse spreading due to dispersion should be less than a fraction δ of the bit period T. For a 1-dB power penalty, $\delta = 0.306$; for a 2-dB penalty, $\delta = 0.491$. For a signal having a bit rate $B = 1/T$ and spectral linewidth $\Delta\lambda$, and transmitted over a distance L, this condition can be expressed as

$$\begin{aligned} \Delta\lambda |D| LB &< \delta \qquad \text{for direct modulation} \\ B\lambda [|D| L/2\pi c]^{1/2} &< \delta \qquad \text{for an external modulation} \end{aligned} \tag{4}$$

The influence of chromatic dispersion is a critical factor for higher bit rates and longer distances and should be suppressed by a proper dispersion compensation scheme. The dispersion compensation process is based on the following observation. While in single-mode optical fiber longer wavelengths impose more delay than shorter wavelengths, the dispersion compensating modules do just the opposite. As a result, signal delays over a specified wavelength band have been equalized. As for dispersion compensating modules, using dispersion compensation fibers (DCFs) with a negative dispersion coefficient is the most common method for dispersion compensation. Since there is an insertion loss introduced by DCF, the figure of merit, defined as the ratio of the absolute amount of dispersion divided by the insertion loss, is used to characterize DCF. Generally it is good if the figure of merit is better than 150 ps/nm/dB.

Optical fiber Bragg gratings can be used for chromatic dispersion compensation as well. The grating reflects different wavelengths at different points along its length, introducing different delays at different wavelengths. Delay introduced by the length of 10 cm is approximately 100 ps. Dispersion is inversely proportional to the bandwidth; that is, a large dispersion occurs over smaller bandwidth and vice versa. For example, 1000 ps/nm occurs over a 1 nm bandwidth, while 100 ps/nm occurs over a 10 nm bandwidth. Future applications, however, will require adaptive dispersion compensation modules that allow for adjustment of both the dispersion compensation value and the dispersion slope.

3.4. Polarization Mode Dispersion

Polarization mode dispersion (PMD) is characterized by two coefficients, D_{p1} and D_{p2}, reflecting so-called "first- and second-order" polarization mode dispersion, respectively. The extent of pulse broadening D_t is governed by the following relation:

$$D_t = D_{p1} L^{1/2} + D_{p2} L \tag{5}$$

where the coefficient D_{p1} presents the average differential group delay (DGD) between the two orthogonal states of polarization over length L, while D_{p2} measures the wavelength dependence of PMD. The contribution of D_{p2} is much smaller than the contribution of D_{p1}, and very often just the first term in Eq. (5) is considered. The value of the coefficient D_{p1} can vary from 0.01 ps/km$^{1/2}$ for new optical fibers to over 1 ps/km$^{1/2}$ for older fibers.

PMD is a stochastic process described by the Maxwellian distribution, which complicates the process of its control and compensation. The probability that actual delay will be 3 times larger than the average delay calculated by Eq. (5) is 4×10^{-5}. This is why we correlate the average delay expressed by Eq. (5) to the actual delay equal to three times the average delay. For differential delay equal to $0.3T$, the power penalty due to PMD will be less than 1 dB.

3.5. Nonlinear Effects

Nonlinear effects in an optical fiber are neither design nor manufacturing defects, but can occur regardless and can cause severe transmission impairments (unexpected loss and interference in the network). On the other hand, in some cases, they may be used to improve transmission characteristics (such as in soliton transmission).

Nonlinear effects are cumulative in nature and proportional to the lightwave pathlength L. Since signal power decreases with increasing lightwave pathlength, an effective length L_{eff} has been introduced to help with calculations. The effective length is defined as

$$L_{eff} = \frac{1 - \exp(-\alpha l)}{\alpha} M \tag{6}$$

where M is the number of fiber spans, each of length l. (Recall that one span is the distance between two amplifiers, therefore $l = L/M$.) In the wavelength region around 1.55 μm, and for links where $L > 1/\alpha$, L_{eff} is ~20 km (α is ~0.046 km^{-1}, or 0.2 dB/km). From the previous relation, it is clear that the effective length can be reduced by increasing the span length and by decreasing the number of amplifiers on the line. But what matters most is the product of the power launched from the amplifier, P, and the effective length L_{eff}. If amplifier spacing is increased, the launched power needs to be increased as well to compensate for additional fiber losses. This increase will be exponential: $P = \exp(\alpha l)$. Since the product increases with span length l, reducing the amplifier spacing can reduce the effect of nonlinearities.

The effects of nonlinearity are inversely proportional to the area of the fiber core. It is convenient to use an effective core area A_{eff} since the power is not uniformly distributed within the core section. This effective area is about 50 μm for a single-mode fiber with a core diameter of 8 μm, but for a dispersion compensating fiber (DCF) it is smaller (thus DCF tends to exhibit higher nonlinearities).

Nonlinear effects can be divided into two categories: the effects due to variations in the fiber refractive index, and the effects due to light scattering. Agraval has given a more detailed explanation of nonlinear effects [7].

Variations in the refractive index at high signal power are at the root of nonlinear effects classified as refractive-index phenomena: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). At low optical powers, an optical fiber's refractive index n is pretty constant [i.e., it is $n = n_1(\lambda)$ for specified wavelength λ]. Higher optical powers, however, cause a refractive index change as follows:

$$n(\lambda, E) = n_1(\lambda) + \frac{n_2 P}{A_{\text{eff}}} \tag{7}$$

where n_2 is the nonlinear refractive index ($n_2 \sim 3 \times 10^{-8}$ μm^2/W), and P is the optical signal power. Both SPM and XPM affect the optical signal phase in proportion to the nonlinear part of the refractive index and generate spectral broadening. Spectral broadening, in combination with chromatic dispersion, will contribute to signal distortion.

In *four-wave mixing* (FWM), new optical frequencies $\nu_{ijk} = \nu_i + \nu_j - \nu_k$ are generated whenever three wavelengths with frequencies ν_I, ν_j, and ν_k propagate through the fiber. The power of a resultant new wave is calculated as presented by Shibata et al. [8]

$$P_{ijk} = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2}\left[1 + \frac{4\exp(-\alpha l)\sin^2(\Delta\beta l/2)}{[1 - \exp(-\alpha l)]}\right] \times \left(\frac{2\pi \nu_{ijk} n_2 d_{ijk}}{3cA_{\text{eff}}}\right)^2 P_i P_j P_k L_{\text{eff}}^2 \tag{8}$$

where $P_{ijk}(i, j, k = 1 \cdots N)$ is the power of the generated wave, n_2 is the nonlinear refractive index, and d_{ijk} is the so-called "degeneracy" factor. The value $\Delta\beta = \beta_i + \beta_j - \beta_k - \beta_{ijk}$ defines a phase condition or relationship among the propagation constants of the optical waves involved (a propagation constant is defined as $\beta = 2\pi n\lambda/c$, where n is the refractive index, λ is the wavelength, and c is the speed of light in a vacuum).

The total crosstalk due to FWM in a given channel is the sum of all generated waves according to Eq. (8) and can be analyzed as interchannel crosstalk [see Eq. (17)]. To alleviate the penalty introduced by FWM, the following measures could be taken: using unequal channel spacing, increasing channel spacing, using dispersion, or reducing the power of interacting channels. The most effective means is to use some amount of dispersion; this clarifies the need to shift the zero dispersion point from the 1550-nm-wavelength region.

Nonlinear effects that occur due to light scattering include *simulated Raman scattering* (SRS) and *stimulated Brillouin scattering* (SBS). For SBS, the acoustic phonons are involved in an interaction that occurs over a very narrow linewidth $\Delta\nu_{\text{SBS}}$($\Delta\nu_{\text{SBS}} = 20$ MHz at 1.55 μm). There is no such interaction if channel spacing is greater than 20 MHz. The SBS process depletes the signal and creates a strong backward signal if the incident power per channel is higher than some threshold value P_{th} expressed as

$$P_{\text{th}} = \frac{21 b A_{\text{eff}}}{g_B L_{\text{eff}}}\left(1 + \frac{\Delta\nu_L}{\Delta\nu_{\text{SBS}}}\right) \tag{9}$$

where g_B is the SBS gain coefficient equal to approximately 4×10^{-11} m/W, $\Delta\nu_L$ is the laser linewidth, while parameter b takes on a value between 1 and 2 depending on relative polarization of pump and Stokes waves. The worst case leads to $P_{\text{th}} \sim 1.3$ mW, since $\Delta\nu_L$ is approximately 20 MHz. The SBS penalty can be reduced by either keeping the power per channel below the SBS threshold or broadening the linewidth of the source using signal dithering. This method is commonly deployed in high-bit-rate systems.

Stimulated Raman scattering is a broadband effect, and its gain coefficient is a function of wavelength spacing. The gain coefficient peak is $g_R \sim 6 \times 10^{-14}$ m/W, which is much smaller than the gain coefficient peak for SBS. Channels up to 125 nm apart will be coupled by SRS, possibly in both directions. SRS coupling occurs only if both channels are at a logic 1 at that moment. The fraction of the power leaking from a particular channel to all other channels is given by

$$P_{\text{SRS}} = \frac{g_R \Delta\lambda_s P L_{\text{eff}} N(N-1)}{4\Delta\lambda_c A_{\text{eff}}} \tag{10}$$

where $\Delta\lambda_s \sim 125$ nm, $\Delta\lambda_c$ is the optical channel spacing, P is the power per channel, and N is the number of channels [7]. The SRS effect is reduced by dispersion, since different channels travel with different velocities and the probability of an overlap between pulses at different wavelengths is reduced. The penalty introduced by SRS can be alleviated by proper channels spacing and/or postequalization of optical channel powers. Some special techniques, such as polarization interleaving between the neighboring optical channels, can help as well.

4. OPTICAL TRANSPORT SYSTEM ENGINEERING

4.1. BER, Signal-to-Noise Ratio, and the *Q* Factor

The bit error rate (BER) is the most important parameter for measuring a digital signal transmission quality. It is

defined as

$$\text{BER}\,(Q) = \frac{1}{2\pi}\int_{Q}^{\infty} e^{-y^2/2}\,dy \approx \frac{1}{Q\sqrt{2\pi}} e^{-Q^2/2} \tag{11}$$

The so-called Q factor, introduced above, corresponds to the electrical signal-to-noise ratio:

$$Q = \frac{R(P_1 - P_0)}{\sigma_1 + \sigma_0} = \text{SNR} \tag{12}$$

where P_0 is the optical power during a "space bit," P_1 is the optical output power during a "mark bit," R is the responsivity of the photodiode, while σ_1 and σ_0 are standard deviations of the noise current during the 1 and 0 bits, respectively. The following practical values are mutually related: $\text{BER} = 10^{-15}$ with $Q = 8$, $\text{BER} = 10^{-12}$ with $Q = 7$, and $\text{BER} = 10^{-9}$ with $Q = 6$. In optical transport systems with cascades of optical amplifiers along the lightwave path, the following important relation between the Q factor and an optical signal-to-noise ratio can be established:

$$\text{OSNR} = \frac{P_1 + P_0}{4S_{\text{sp}}B_{\text{opt}}} \approx \frac{2Q^2\Delta f}{B_{\text{op}}} \tag{13}$$

Equations (12) and (13) are the basic ones since the impact of various impairments is not included.

4.2. Power Penalty Handling

If there are impairments involved, signal, noise-related values will be Q', P_0', P_1', σ_1', and σ_0', rather than Q, P_0, P_1, σ_1 and σ_0 respectively. Each impairment will contribute to a power penalty to the transport system. The total optical power penalty can be calculated as

$$\Delta P = -10\log\left(\frac{Q'}{Q}\right) \tag{14}$$

The biggest contribution to the total power penalty comes from the nonideal extinction ratio, imperfect dispersion compensation, nonlinear effects, and crosstalk:

- The power penalty due to a nonfinite extinction ratio r, defined as $r = P_1/P_0$, can be calculated in decibels as
$$\Delta P_{\text{ER}} = 10\log\left[\frac{r+1}{r-1}\right] \tag{15}$$
- The power penalty due intrachannel crosstalk involving N interfering signals is
$$\Delta P_{\text{intra}} = C\log\left(1 - 2\sum_{i=1}^{N}\sqrt{\delta_i}\right) \tag{16}$$
where the coefficient C takes on the value 10 for direct detection, and the value 5 forAPD/preamp detection, while δ_i is the crosstalk portion divided by the power of specified channel signal. If we allow a 1-dB crosstalk penalty, then the intrachannel crosstalk level should be just 1%, or 20 dB, below the specified channel signal.
- The power penalty due interchannel crosstalk involving N interfering signals is
$$\Delta P_{\text{inter}} = C\log\left(1 - \sum_{i=1}^{N}\delta_i\right) \tag{17}$$
with the same coefficient C as in relation (17). If we allow a 1-dB crosstalk penalty, then the intrachannel crosstalk level should be 13.5 dB below the desired signal. A more thorough treatment of crosstalk can be found in the article by Zhou et al. [9].
- As for power penalties due to imperfect dispersion compensation or nonlinear effects (FWM and SRS), Eq. (17) can be applied. The portion δ for a particular case can be calculated by Eqs. (4), (8), and (10).

4.3. Noise Accumulation

Recall that in an optical transport system a lightpath contains a number of optical amplifiers spaced l km apart. The length l defines the span length. If fiber attenuation is α, the span loss between two amplifiers is $\alpha_{\text{span}} = \exp(-\alpha l)$. Each optical amplifier amplifies an incoming optical signal to compensate for the loss at the previous span. At the same time, however, it generates some spontaneous emission noise. Both the signal and the spontaneous emission noise are then amplified by the following optical amplifiers.

If the gain G of an optical amplifier is larger than the span loss α_{span} the signal power will increase gradually throughout the amplifier chain. However, the output power from an optical amplifier is physically limited to a saturated value P_{sat}, which means that as input power increases the amplifier gain drops. Consequently, after some number of spans, amplifiers will enter into the saturation regime and the total gain will drop from its initial value G to a saturated value G_{sat}. Further along the lightwave path, a spatial steady-state condition will be reached, in which both the saturated output power P_{sat} and the gain G_{sat} remain the same from span to span. If there are N optical channels, the saturated output power will be equally divided among them. Therefore, the output power per channel will be $P_{\text{out}} = P_{\text{sat}}/N$.

The OSNR gradually decreases along the chain, since the accumulated ASE noise gradually makes up a more significant portion of the limited total power from an amplifier. The steady-state gain will be slightly smaller than the span loss, due to added noise at each amplifier point. Thus, the best engineering approach is to choose a saturated gain that is very close to the span loss. If we prescribe the OSNR for a lightwave path with total length L, and M amplifiers on the line ($M = L/l$), the following relation can be established:

$$\text{OSNR} = P_{\text{out}} - \alpha l - \Delta P - 10\log(M) - 10\log(F_n h\nu B_{\text{op}}) \tag{18}$$

where all except the last two terms are expressed in decibels. If we did not need to worry about impairments, we would neglect the power penalty term and either maximize

the power or decrease the span length to increase the OSNR in Eq. (18). However, the story is different when impairments cannot be neglected, since a power margin equal to ΔP needs to be allocated in advance to compensate for impairment power penalties.

5. OPTICAL TRANSPORT ENABLING TECHNOLOGIES AND TRADEOFFS

5.1. Enabling Technologies

Enabling technologies will continue to provide the means of increasing both transmission capacity and lightwave pathlength. There is a number of enabling technologies that are helpful in resolving the before-mentioned issues in optical transport systems, and in approaching a transmission capacity predicted by Mecozzi and Shtaif [10]. These include optical amplifiers, forward error correction, advanced coding techniques, and advanced dispersion compensators.

Optical amplifiers should provide enough gain for a specified number of optical channels, which suggests that an aggregate optical power should be >22 dB for systems with more than 100 optical channels. Next, the noise figure should approach its theoretical value of 3 dB. In addition, the gain profile should be equalized along the entire wavelength band, and this gain equalization should be dynamically adjustable.

Fiber doped optical amplifiers, such as erbium-doped fiber amplifiers (EDFAs) and thulium-doped fiber amplifiers (TDFAs) can serve to cover the entire low-loss optical fiber bandwidth. In addition, Raman amplifiers can be used to cover a wide range of wavelengths as well. The Raman based amplification is a newer technology based on the SRS nonlinear effect. In the Raman amplifier, the pump light is launched into the fiber at inline amplifier sites (or optical receiver sites), opposite the signal direction. As a result, the forward-propagating optical signals get some energy, and a low-noise preamplifier has been created. By combining several pumps, a fairly flat gain profile over a wide range of optical wavelengths can be achieved.

Advanced dispersion-compensating techniques are necessary to take full advantage of the improved optical amplifiers. First, the chromatic dispersion-compensating device (DCM) should not only incorporate dispersion compensation ability but slope compensation as well. Secondly, polarization mode dispersion compensation will be needed more often, but an efficient PMD compensator still needs to be introduced.

Forward error correction (FEC) is needed to push systems reach even further in future optical transport networks. For now, so-called Reed–Solomon 239–255 coding scheme remains widely used, but some other coding methods have been introduced as well. The final result of FEC application is BER enhancement for a lower signal-to-noise ratio.

Advanced modulation/transmission methods are emerging for use in high-speed optical transport systems. Some of them are

- *Solitons* or return-to-zero distortionless pulses, where spectral disorder at the trailing and leading edges that occurs due to self-phase modulation is corrected by another disorder caused by chromatic dispersion. The technique has shown very promising results when used in combination with fibers having a "prescribed dispersion map."
- *Advanced coding*, such as duobinary coding, where special filtering reduces the spectral width. Thus, the slowest and the fastest components are eliminated; the rest is more resistive to chromatic dispersion influence.
- *Coherent detection* where the weak input optical signal is mixed with a much stronger signal from the local laser. With this method the signal current takes on the value $I = 2RP_sP_{\mathrm{LO}}\cos[|\omega_s - \omega_L|t + \phi(t)]$, where R, P_s, and P_{LO} relate to receiver responsivity, incoming optical power, and local laser power, respectively. Because of P_{LO} contribution, the coherent receiver sensitivity is considerably enhanced. The detection process can either be heterodyne or homodyne. In the heterodyne detection scheme, the incoming optical signal frequency ω_s and the frequency ω_L of the local laser are slightly different, while in homodyne detection scheme both the optical frequencies and the phases of the signal and the local laser are completely matched. The demodulation process applied to the signal current depends on the modulation format of the incoming optical signal; this format could be amplitude shift keying (ASK), frequency shift keying (FSK), or phase shift keying (PSK). The realization of the coherent detection scheme involves the necessity of having a stable relationship between the phases and frequencies of an incoming optical signal and the local laser signal, which brings an additional complexity to the system design.

5.2. Transmission System Engineering Tradeoffs

Proper design of a transmission system is a big challenge, and some tradeoffs must be done. Generally speaking, the longer the distance, the smaller the transmission capacity, and vice versa. Overall design tradeoffs include

- *Fiber type selection* is applicable just for new fiber deployment. It is obvious that single-channel transmission favors DSF fibers, while NZDSF fiber should be, generally speaking, favorable for long-distance DWDM systems. Standard single-mode fibers (SMFs) might be the best choice in certain cases where either chromatic dispersion is not critical, or where SMF-based systems are less vulnerable to the influence of nonlinearities. Even low-cost multimode optical fibers can be considered for some short-reach and/or lower-bit-rate applications.
- *Spectral efficiency* versus the total optical bandwidth occupied is an important design issue in some cases. By increasing spectral efficiency with dense wavelength spacing, designers expose the system to greater susceptibility to degradation from nonlinear

effects. On the other hand, increasing the optical bandwidth occupied would lead to system cost increase.

- *Chromatic dispersion management* is essential for high-speed optical transmission systems. A novel transmission line design with proper dispersion management will provide conditions for ultra-high-capacity systems.
- *Optical power level* per optical channel is dependent on the output power level from optical amplifiers, nonlinear effects, crosstalk, and safety issues. It is always desirable to increase the power level from the SNR point of view, but the power penalty due to an increase in nonlinearity and amplifier noise will limit the signal power level, leading to an optimal in-between value.
- *Optical pathlength* is very important in transmission systems engineering. Design is more complex on longer optical paths since optical networking issues, such as crosstalk, wavelength misalignment, and cascaded filter effects, accumulate with increasing pathlength. Optical signal powers and the SNR among different paths that come together at the input of an optical amplifier or receiver should be equalized.

BIOGRAPHY

Milorad Cvijetic received his Ph.D. degree in electrical engineering from University of Belgrade in 1984. Dr. Cvijetic has experience in both academia (teaching at University of Belgrade and Carleton University), and industry (work in the area of high-speed optical transmission systems and optical networks). His research work related to quasi-single mode optical fibers, BER evaluation in soliton-based systems, and system performance evaluation in high-speed optical systems with external modulation, has been widely recognized.

He currently serves as the chief technology strategist for Optical Network Products with NEC America, Herndon, Virginia. Previously, he has been with Bell Northern Research (later NORTEL Technologies) in Ottawa, Canada, working in the Advanced Technology Laboratory. Dr. Cvijetic has published more than 40 technical papers and two books titled *Digital Optical Communications* and *Coherent and Nonlinear Lightwave Communications*. He has taken part in numerous telecommunications conferences and symposiums, in some as a session/conference chairman, technical committee member, or invited speaker. He is a member of IEEE Communications Society and LEOS.

BIBLIOGRAPHY

1. I. P. Kaminov and T. L. Koch, eds., *Optical Fiber Telecommunications*, Academic Press, San Diego, CA, 1997.
2. J. A. Buck, *Fundamentals of Optical Fibers*, Wiley, New York, 1994.
3. E. Desurvire, *Erbium Doped Fiber Amplifiers*, Academic Press, New York, 1994.
4. M. Cvijetic, *Coherent and Nonlinear Lightwave Communications*, Artech House, Boston, 1996.
5. M. Cvijetic, Performance Evaluation of externally modulated high bit rate lightwave systems, *IEEE Photon. Technol. Lett.* **9**: 687–689 (1997).
6. *ITU-T Recommendations on Optical Fibers*, G.652/G. 653/G. 655, Geneva, 1993.
7. G. P. Agraval, *Nonlinear Fiber Optics*, 2nd ed., Academic Press, San Diego, CA, 1995.
8. N. Shibata, R. P. Brown, and R. G. Waarts, Phase mismatch dependence of efficiency of wave generation through four wave mixing in a single mode optical fiber, *IEEE J. Quant. Electron.* **QE-23**: 1205–1210 (1987).
9. J. Zhou et al., Crosstalk in multiwavelength optical crossconnect networks, *IEEE/OSA J. Lightwave Technol.* (Special Issue on Multiwavelength Technology and Networks) **JLT-14**: 1423–1435 (1996).
10. A. Mecozzi and M. Shtaif, On the capacity of intensity modulated systems using optical amplifiers, *IEEE Photon. Technol. Lett.* **13**: 1029–1031 (2001).

OPTICAL WIRELESS LASER COMMUNICATIONS: FREE-SPACE OPTICS

Dennis Killinger
University of South Florida
Tampa, Florida

1. INTRODUCTION

Free-space optics (FSO) communication involves the use of modulated optical beams to send telecommunication information through the atmosphere from one location to another location, and has been the subject of a series of several conferences on FSO communication [1–3]. The concept of FSO light communication is not new, having been used by the Romans to transmit information via mirror reflected optical beams from one hill to another during ancient times. Indeed, Alexander Graham Bell in his photophone patent dated 1880 showed the use of an intensity modulated optical beam to transmit telephone signals 200 m through the air to a distant receiver. More recently, however, the tremendous growth in Internet traffic due to the use of high-bandwidth optical fiber transmission networks and the development of low-cost and high-power diode lasers has greatly increased the utility of transmitting information on an optical laser beam from one location to another through the air. Since 1996, the use of free-space optics has grown exponentially in the commercial market since it offers the potential to help connect the millions of telecom users within the "last mile" and at extremely high bandwidths of 1–10 gigabits per second (Gbps) or more. While cable (coax) offers 1-Gbps capability, it must be shared in bandwidth among different users and channels within a neighborhood or hub. T1 lines into offices carry a 1 Mbps (megabits per second) bandwidth, but are usually fiberoptic-coupled to a main hub. The more recent RF wireless 802.11b capability to link office and home computers wirelessly

to a common hub provides bandwidths of 11 Mbps, but can become crowded in capacity when many (say, 20–100) notebook computers are being used at the same time, such as within a campus library room. As such, future need and growth is anticipated for the development of individual higher bandwidth (0.1–1 Gbps per user) connectivity within all offices and homes in a metro (metropolitan) market. Optical access and FSO in particular may offer the optimal technical solution for such connectivity in the future since only optics offers such large bandwidths for each individual user. As the National Academy/NRC Committee on Optical Science and Engineering (COSE) report recently stated, "The Tera-bit/s era for information technology . . . includes the need for cost-effective networks of virtually unlimited bandwidth with local area networks operating at tens of gigabits/s" [4].

1.1. Historical Background and Current Technology Perspective

Historically, a significant amount of laser telecommunication and laser atmospheric propagation studies were conducted in the 1970s and 1980s as part of the development of military electrooptic instruments, laser radar systems, and secure communication data links. Much of the optical and laser science underlying these systems can be found in current optical handbooks. [5,6] Early Department of Defense (DoD) work involved the detailed analysis of the attenuation and scatter of a laser beam transmitted through the atmosphere, the common physical and parametric analysis of different types of visible and infrared detectors, and the intensity modulation and wavelength control of a wide range of lasers. For example, several laser FSO communication systems were developed in the 1980s for secure ship-to-ship communication and ground-to-aircraft applications. In addition, since the early 1990s, several secure laser communication systems for use between the ground and satellite-to-satellite have been developed and launched [7–9]. Most of these early or DoD FSO systems were designed to be used for long-range (5–1000 km) communication links and often used high-power (1–200-W) 10-μm-wavelength CO_2 lasers, 1.06-μm Nd:YAG, 0.85 μm GaAs, or 1.5-μm diode/erbium:fiber amplifier lasers. They often involved complex tracking systems, multiple detector receivers or adaptive optics to compensate for atmospheric turbulence, external modulators for the high-power lasers to place the communication signal on the laser intensity bitstream, and seldom were considered to eye-safe. Many of these systems had development costs on the order of several millions of dollars, although vehicle mounted systems have typical costs on the order of $100,000. The number of systems deployed is relatively small, ranging from one-of-a-kind satellite-to-satellite or ground-to-airborne systems, to unit numbers in the hundreds for small-size mobile systems. These DoD FSO systems and laser atmospheric studies are emphasized here since they provide much of the scientific groundwork in laser, propagation, signal processing, and detector technology in the context of current commercial FSO systems and for the design of future optical wireless communication systems.

Since 1996 there has been an explosion in commercial development of FSO systems that has been driven in part by several considerations from both technology and business viewpoints [10]. First and most importantly, the demand for high-bandwidth Internet connectivity within the last-mile market has driven the use of FSO systems in places where optical fiber is too expensive to use, especially within the urban metro market. Here, the price to lay fiber from one building to another building just across the street may cost $300,000 and take 6 months time to obtain a permit, if allowed at all. Second, the advent of directly modulated, moderate power diode lasers and light-emitting-diodes (LEDs) that are inexpensive and compact have allowed the development of low–moderate-cost FSO systems for short to moderate ranges. Initially, FSO commercial systems developed in the 1990s used higher-power (1–10-W) lasers to transmit a 0.1–1-GHz bandwidth signal at distances of 5–10 km. The systems were designed to provide point-to-point connectivity over a long distance, often used active tracking to compensate for building sway and atmospheric turbulence, and cost upwards of $100,000 [1]. Since the mid-1990s, the design of many commercial systems has gravitated toward shorter-range and lower-cost systems that use moderate power (10–100 mW) diode lasers or LEDs, and operate at shorter ranges (100–500 m) [3]. There are about 15 commercial companies currently selling FSO systems, ranging in price from $1000 per unit for 10-Mbps systems to about $20,000–$100,000 per system for advanced capabilities (1–10 Gbps). Several of these companies have installed or sold nearly 5000 FSO systems each, while some are implementing a complete networking capability in a point-to-multipoint or mesh-net configuration [3]. It is interesting to note, however, that most of the current commercial FSO systems operate at wavelengths of 0.8 or 1.5 μm, and use laser or optical technology derived mainly from the fiberoptic telecommunication community as opposed to the previously mentioned DoD laser sensor and atmospheric propagation community. This is due to two reasons: (1) the development of inexpensive and reliable laser and receiver/detectors near 0.8 μm (fiberoptic telecommunication in the 1980s) and 1.5-μm diode lasers with erbium:fiber amplifiers (fiberoptic telecommunication in the 1990s), and (2) the wish to remain compatible in optical format to the burgeoning fiberoptics telecommunications field and the wavelength and information requirements imposed by the standards adopted by the industry. FSO technology and deployment in the commercial sector may be viewed as being just in its infancy, with several marketing studies indicating that the current $200 million/year market could grow to $2 billion/year before the year 2007. As such, the reader should remember that the current growth in FSO may continue with advances along the current wavelength and devices used, or it may branch out to other wavelength regions (say, for instance, 3.5 or 9 μm) according to the technical needs of the market. If the last-mile, metro, or home market becomes the main focus of the telecommunications field in the future (say, 10 years from this writing), and if FSO plays a major role in its development, then the optimization of FSO last-mile technology at other wavelengths may

be more important than the demands to interface and use optical technology from the fiberoptic network legacy. It is beyond the scope of this article to speculate on the future of this area in light of some of these considerations, but such thoughts focus on the importance of covering the basic optical physics and technology behind FSO in this so that the reader can appreciate the different optical tradeoffs, technical limitations, and capabilities of FSO.

1.2. Technical Emphasis of Tutorial Overview

This article presents a brief overview of the optical science and technology involved in free-space optics communication. The emphasis will be on the basic physics and engineering aspects of FSO design mainly with an eye toward point-to-point communication for the commercial market. As such, the physics of laser atmospheric propagation, laser specifications, detector criteria, telescope design, eye safety, laser-beam tracking, and atmospheric turbulence and beam wander will be discussed. These are the main criteria for the design of current FSO systems and will probably determine the design of future systems. Most FSO systems are "modulation-tolerant" or "protocol-agnostic," which means that they faithfully reproduce the input modulation codes of a communication link. As such, they can be placed inside a wide range of different communication networks (Ethernet, FDDI, SONET, etc.) without changes being required as the system communication schemes are modified and updated in future years. Although some traditional copper data link formats such as T-1 and T-3 may require some data conversion. In addition, their use within a traditional network ring or multipoint configuration may also be independent of other communication parameters, however, this may not be the case if the FSO is not an add-on to an existing net but develops as the main net or "mesh-net" component. Since many of these latter topics on communication network topology and modulation or coding schemes are covered elsewhere, they will not be covered in this article on FSO except only as needed. The interested reader is directed toward these topics within this encyclopedia. It should be noted that there are several excellent overviews and technical articles on FSO that the reader may want to review, including the book on free-space optics communication by Willebrand and Ghuman that covers technical, marketing, and network issues; in-depth papers presented at several SPIE conferences; and specific research papers [1–3,10].

2. OVERVIEW OF A TYPICAL FREE-SPACE OPTICS COMMUNICATION SYSTEM

A typical free-space optics (FSO) communication system is depicted Fig. 1. It often consists of a small laser source that can be directly modulated in intensity at fairly high data rates, a beamshaping–transmitting telescope lens to transmit the laser through the atmosphere toward a distant point, a receiving lens or telescope to collect and focus the intercepted laser light onto a photodetector, and a receiver amplifier to amplify and condition the received communication signal. Figure 1 also shows the transmitted laser beam passing through the atmosphere and being partially collected by the receiver telescope. The laser or optical beam can be absorbed, scattered, or displaced by the atmosphere, depending on the atmospheric conditions and wavelength/linewidth of the laser source. If the laser beam has to transverse distances less than about 200–500 m or so, then finite movement and sway of the local buildings attached to the system may move the transmitted beam away from the receiving telescope aperture and outside the angular acceptance angle of the system. In this case or the case of high atmospheric turbulence, an active tracking device may have to be used to align the beam onto the receiver using a small gimbal mirror, lens translation stage, or detector/laser translation stage; active tracking may be eliminated if sufficient power is available by expanding the divergence of the beam or if the building and alignment is stable.

To give the reader a perspective of the size and shape of a typical FSO system, Fig. 2 is a photograph of a FSO indoor unit from PlainTree, Inc. (model 340) that uses a low-power (40-mW) 0.85-μm-wavelength light-emitting diode (LED) nonlaser (noncoherent) source for optical communication at short to moderate ranges (100–200 m) [11]. The transmitter lens is about 6.5 cm in diameter, the receiver telescope lens is about 13 cm

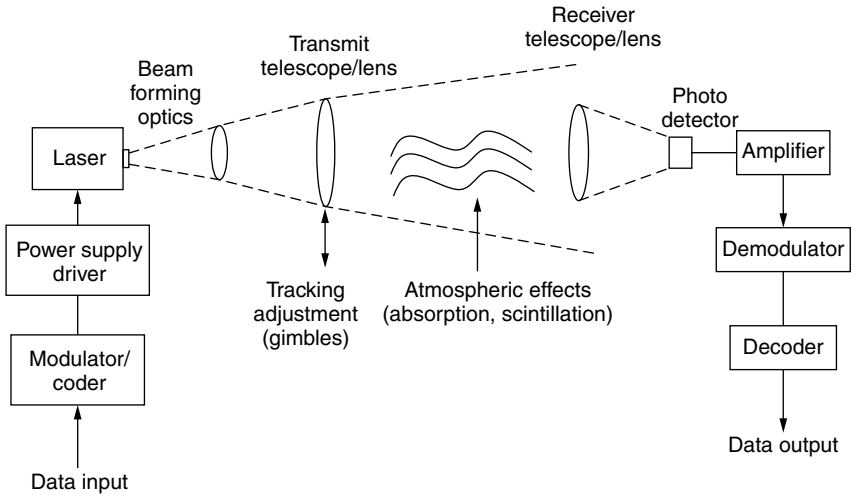


Figure 1. Pictorial schematic of free-space optics laser communication system for point-to-point applications. A typical FSO system has two of these optical channels for bidirection flow of the communication link.

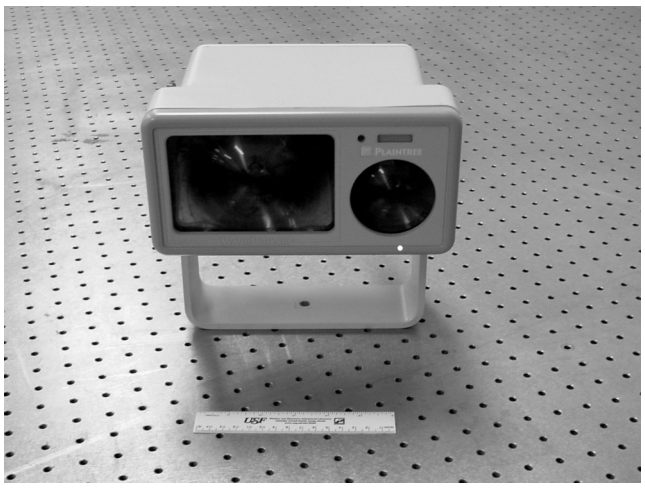


Figure 2. Photo of 0.85-μm noncoherent LED-based FSO system from PlainTree used for short/moderate indoor applications. Photo is that of a PlainTree Model 340 system.

Figure 3. Photo of 0.785-μm-wavelength diode laser FSO system from Optical Access that uses Four laser beams to reduce speckle fading and atmospheric turbulence effects.

in diameter, and the beam divergence is about 1° (0.0175 radians). The data rate is about 10 Mbps. Figure 3 shows a FSO system from Optical Access (model 155) that operates in the same near-IR wavelength region, but uses a 0.785-μm-diode laser and four laser beams to reduce fading and atmospheric turbulence effects. Here the laser output power is about 7 mW and the data rate is 155 Mbps [12].

Figure 4 shows a higher power and different wavelength system from fSONA, Inc. (model 622-M) that uses four separate 4-cm-diameter beams from 100-mW diode lasers operating at 1.55 μm that is able to transmit 622 Mbps information at ranges up to 2.5 km [13]. Here, the receiver telescope size is 20 cm and the four transmitted beams are used to reduce the effect of increased FSO signal fluctuations due to the effects of atmospheric turbulence at longer ranges and interference/speckle fluctuations associated with the use of a coherent laser. The four transmitter lasers also offer redundancy for the link.

Figure 4. Photo of 1.55-μm high-power diode laser FSO system from fSONA that uses Four laser beams to reduce signal fluctuations.

More sophisticated systems have also been developed that have been able to transmit data at rates beyond 40 Gbps at ranges of 5 km or more using multiple laser wavelengths, active atmospheric tracking, and multiple detectors and beams [14].

The detected FSO optical signal is usually converted to an electrical signal as shown in Fig. 1 and then sent to the communication network or individual hub. However, the optical detected signal can be redirected via mirrors to another location, or received by a router that will redirect it into a fiberoptic communication system. Of course, the optical signal received could also be amplified by another laser amplifier, as in the case of a 1.5-μm laser signal and an Er:YAG fiber amplifier. This is discussed later in Section 3 in this article.

As can be appreciated from Fig. 1, several important optics and laser spectroscopic issues need to be considered in the design and development of a FSO system. These include the availability and wavelength coverage of lasers and LED sources, the interaction and attenuation of the FSO optical beam as it traverses the atmosphere, the received light intensity collected by the receiver telescope, the sensitivity of the optical detectors, and the influence of atmospheric turbulence. These factors influence the relative SNR of the received laser signal and are depicted in the FSO range equation. These aspects are covered in the following sections.

3. LASER AND OPTICAL SOURCES FOR FSO

Most current FSO systems use either 0.8- or 1.5-μm lasers or LED sources. However, there are a wide range of other lasers and LED sources that have potential for use in a FSO system. For FSO applications, it is appropriate to look only at continuous-wave (CW) lasers as opposed to pulsed lasers since they can more often be intensity modulated at MHz or GHz rates. In this regard, Fig. 5 shows a plot of the output power of several CW lasers as a function of the wavelength covered [15–17]. As can

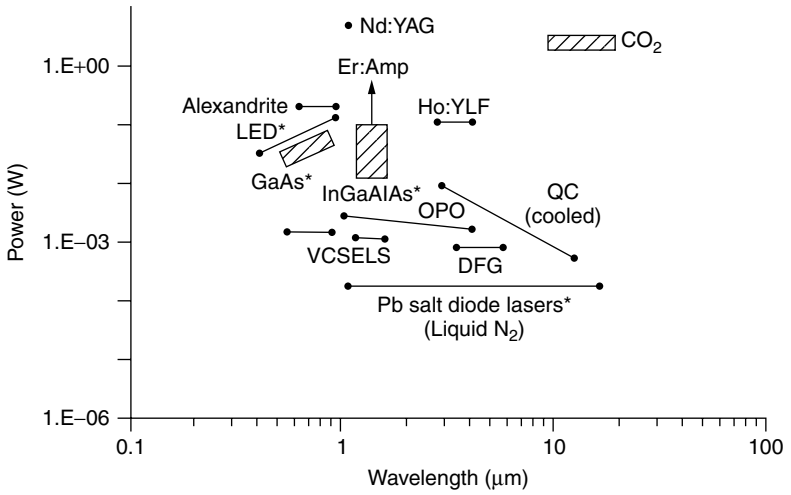


Figure 5. Typical laser output power in Watts as a function of wavelength for current lasers. Asterisks (∗) indicate direct modulation (10 MHz to 10 GHz); otherwise, external cavity modulation is used.

be seen, several high-power lasers such as the CO_2 and Nd:YAG laser operate only over a narrow wavelength range, while several lower-power lasers such as the optical parametric oscillator (OPO) and difference frequency generation (DFG) laser operate over a wider tuning range but at lower output power. The GaAs(Al) diode lasers operating near 0.8–0.9-μm wavelengths have CW power levels in the order of 0.01–0.1 W, while InGaAs(P) lasers near 1.5 μm operate with tens of mW power; the latter can be boosted by the use of Er:fiber laser amplifiers to levels of 1–10 W; governmental laboratory fiber lasers have reached levels of 100–200 W and higher, although problems with spectral mode hopping and spontaneous background emission require stringent cavity design and injection seeder laser-beam isolation. The vertical-cavity surface emitting lasers (VCSELs) are vertical layered semiconductor lasers that have output powers on the order of 1–10 mW and are tunable in some cavity arrangements. The quantum cascade (QC) laser offers potential for future FSO usage, especially with the development of 9-μm room-temperature lasers [18]. Also shown in Fig. 5 is the output power for a LED, which is a noncoherent light source but has wide utility as a FSO source. Of the lasers shown in Fig. 5, only the semiconductor diode lasers are directly modulated at rates up to 10 Gbps using the drive current of the laser or an internal loss material. The other lasers have to use external modulators (electrooptic, acoustooptic, or traveling-wave modulators) to reach Mbps–Gbps modulation rates. The LED modulation rate is generally 1–10 MHz, but newer models, including laboratory quasicavity LEDs, have modulation rates on the order of 100 MHz.

Some important output characteristics of these different CW lasers are tabulated in Table 1 [15–17]. As can be seen, the GaAs lasers and InGaAs lasers offer a significant combination of high output power and can be directly modulated via their drive currents. Some of the lasers have linewidths that are either single frequency (single longitudinal cavity mode) or consist of several laser modes within a group of lines, while an LED has a broad noncoherent emission spectrum. This can be seen in Fig. 6, which shows the spectral output measured by the author from three different optical sources as a function of wavelength or frequency. As can be seen in the figure, the output spectrum from the LED is broad, covering a range of 50 nm (about 500 cm^{-1} or 15,000 GHz), while the typical 1.33-μm diode laser shows multiple longitudinal modes covering a range of about 3 nm. The bottom portion of the figure shows the 50-kHz linewidth output spectrum of a single-frequency 1.55-μm distributed feedback (DFB) laser whose laser output has been controlled through use of Bragg reflection from imposed index variations along

Table 1. Output Characteristics of Several Currently Available CW Laser Sources[a]

Laser	λ (μm)	Power	Temperature	Modulation Rate	LineWidth[a]
GaAs	0.8	10–100 mW*	Room	100–500 MHz*	Multi/SF
VCSELs	0.8, 1.5	1–10 mW	Room	1 GHz*	SF
InGaAs	1.5	10–100 mW	Room/TE[b]	10 GHz*	Multi/SF
LED (nonlaser)	0.8	10–100 mW	Room	10–100 MHz*	50 nm
CO_2	10	1 W	Room	200 kHz*/20 GHz	$0.1cm^{-1}$/SF
Nd:YAG	1.06	1 W	Room	1 GHz	$0.1cm^{-1}$/SF
Quantum cascade	3–9	0.1–1 mW	77 K–room	?	$0.1cm^{-1}$/SF
Pb salt	2–10	0.1 mW	77 K	?	0.1–1 cm^{-1}
Ho:YAG	2.06	100 mW	Room	1 MHz ?	$0.1cm^{-1}$/SF

[a] Linewidths may have multilongitudinal modes or be single frequency (SF).
[b] Thermoelectrically cooled.
*indicates direct modulation of the laser intensity while other lasers use external cavity modulation.

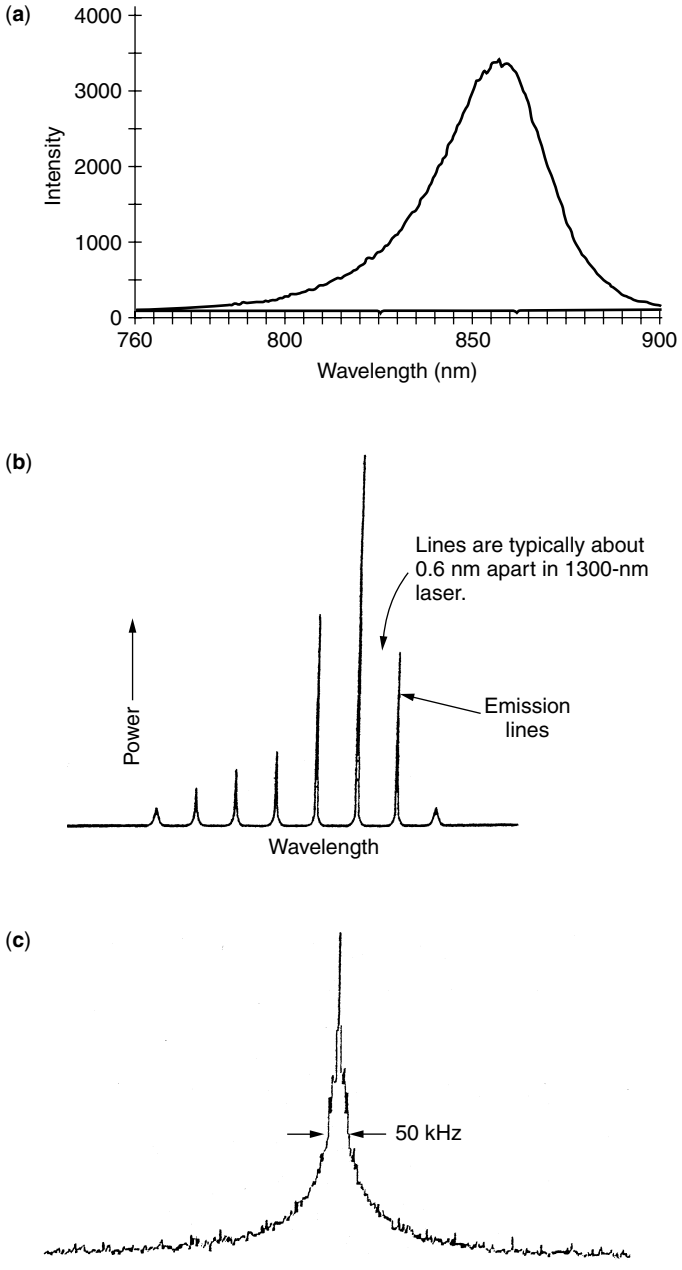


Figure 6. Spectral plots of the output power as a function of wavelength showing the spectral linewidth of a typical 0.86-μm LED source (**a**), 1.3-μm multimode diode laser (**b**), and a 1.55-μm single-frequency DFB diode laser (**c**).

the semiconductor laser cavity. Figure 6 is important in that the exact wavelength and linewidth of the laser can determine the absorption and scatter properties of a FSO laser beam as it propagates through the atmosphere. This will be seen more clearly in the next section.

It should be noted that many laser sources do not operate in a single spatial mode, but may have a divergence that is greater than the lowest-order (Gaussian) mode. A measure of this deviation is the M^2 parameter value, or mode structure parameter. For a Gaussian spatial beam, M^2 is equal to 1. However, for many short cavity lasers, the M^2 value may be closer to 2 or 3 for diode lasers, and as high as 20 to 50 for short cavity OPO lasers. An alternative to specifying the M^2 parameter value is to directly specify or measure the divergence of the laser in terms of milliradians. The divergence of the laser beam is equal to M^2 times the divergence of a Gaussian beam [19]. It is common in a FSO system for the divergence of the transmitted beam to be made larger than the diffraction minimum value so that the projected beam size is larger than the receiver optics and more tolerant of misalignment of the beam. This is discussed in Sections 6.1 and 6.2.

Finally, it should be added that Fig. 5 does not show an important third axis, which would be related to the cost of the laser system and modulation scheme. Such information is very important especially for commercial systems and influences the engineering trade-off design of the FSO system. To give the reader a ballpark value, typical costs range from a few dollars for a LED or low power GaAs laser to several $100K for an externally modulated CO_2 laser. Typical costs for moderate power 100 MHz 0.8 μm lasers are near $0.1K

to \$1K, with somewhat higher costs for 1.5 μm diode lasers, and approaching \$10K to \$20K for higher power Er doped amplifier lasers. Of course, these values are only approximate values amd will be reduced as technical progress is made in this area.

4. ATMOSPHERIC ATTENUATION AND SCATTER OF THE FSO BEAM

The attenuation of an optical beam as it propagates through a medium is given by the Beer–Lambert law as

$$I(x) = I_0 e^{-\alpha x} \tag{1}$$

where I_0 is the initial optical intensity in watts, $I(x)$ is the intensity after the beam has traveled a distance x meters, and α is the attenuation coefficient of the medium in reciprocal meters. The attenuation of the atmosphere can be due to several factors, including absorption of the beam via molecules in the atmosphere and scatter of the beam due to Rayleigh, Mie, and resonant scatter with molecules or aerosol particles in the air [20]. For most applications, the Mie Scatter (especially due to fog) is dominant.

The optical transmission of the normal atmosphere can be shown as in Fig. 7, which shows the low-resolution transmission spectrum of the atmosphere for a 2-km path near ground level [21]. The spectrum is that for a low-resolution spectrometer with a spectral resolution of about 20 cm^{-1}. As can be seen, there are several regions where water vapor and other gases absorb the optical beam, while there are large optical windows in the visible (0.4–0.7 μm) and at 1.5 μm and near 9–13 μm, where the beam is hardly absorbed. Of interest for FSO applications is the higher-resolution spectra for the atmosphere at regions that appear almost opaque in Fig. 7. Figure 8 is a calculated transmission spectrum of the atmosphere for a U.S. standard atmosphere for a path of 500 m over three different spectral regions of potential FSO interest near 0.85 μm, near 1.55 μm, and near 9 μm. As can be seen, the spectrums show individual absorption lines due to the vibrational–rotational absorption lines of water vapor, CO_2, CH_4, and other gases in the atmosphere. The individual lines all have a pressure-broadened linewidth of about 0.1 cm^{-1}, so that a tunable laser beam that has a linewidth on the order of 0.1 cm^{-1} or less can be absorbed if it is tuned online, or not absorbed if it is tuned in wavelength to the offline position. It is because of this close connection between the laser or optical source linewidth and wavelength and that of the atmosphere absorption lines, that careful selection of the laser wavelength can have a significant influence on the performance of a FSO system. It should be noted that the spectra in Fig. 8 were calculated using the HITRAN database and HITRAN-PC computer program since it has a spectral resolution better than 0.01 cm^{-1}, and usually produces spectral plots that have line centers with an accuracy of 0.001 cm^{-1} and line intensity accuracy of a few percent [22–24]. Other Air Force atmospheric spectral codes such as MODTRAN have a resolution of 2–20 cm-1 and may be valid for wide-linewidth LEDs and regions of little spectral absorption. However, in general, it is best to use the high-resolution capability of the Air Force FasCode program or HITRAN-PC, which uses the HITRAN spectral line database, and then convolute the overlap of the laser spectrum with that of the atmosphere. The HITRAN database has been developed by the U.S. Air Force since 1971 and is the compilation of over one million individual spectral lines of over 32 molecules in the atmosphere.

In addition to the absorption of molecules in the atmosphere, there is also the attenuation due to the scatter from aerosols and particles in the atmosphere. In this case, the fogs, clouds, and dust particles can add to the attenuation of the optical beam. For molecules and spatial scale changes in the index of refraction that are much smaller than the wavelength of light, the scatter is called Rayleigh scatter, named after Lord Rayleigh (1842–1919), who first quantified the effect. Rayleigh scatter attenuation coefficient can be given approximately for the standard atmosphere as [24,25]

$$\begin{aligned}\alpha_{\text{ray}} &= N\sigma_\pi = N\frac{8\pi}{3}\sigma_\pi \\ &= \frac{8\pi}{3}1.18 \times 10^{-8}\ [550\ \text{nm}/\lambda\ (\text{nm})]^4\ \text{cm}^{-1} \\ &= 1.1 \times 10^{-5}\ [550\ \text{nm}/\lambda\ (\text{nm})]^4\ \text{m}^{-1}\end{aligned} \tag{2}$$

where N is the number of molecules ($\sim 2.55 \times 10^{19}$ molecules/cm^3) in air and σ_π is the backscatter Rayleigh scatter cross section. Equation (2) is normalized to that for a wavelength of 550 nm. As can be seen, Rayleigh scatter increases in the short-wavelength or blue-wavelength regions, which is why sunsets appear red (most of the blue light has been scattered away from the viewing angle).

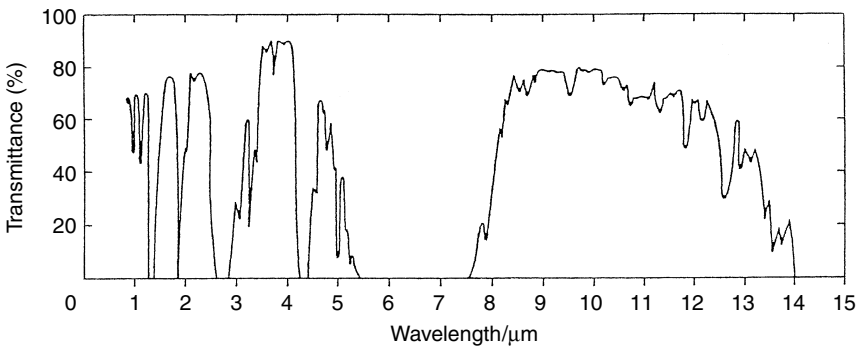


Figure 7. Calculated transmission spectrum of the standard atmosphere for a pathlength of 2 km. Strong absorption regions near 2.8μ m and 6–7 μ m are due mostly to water vapor.

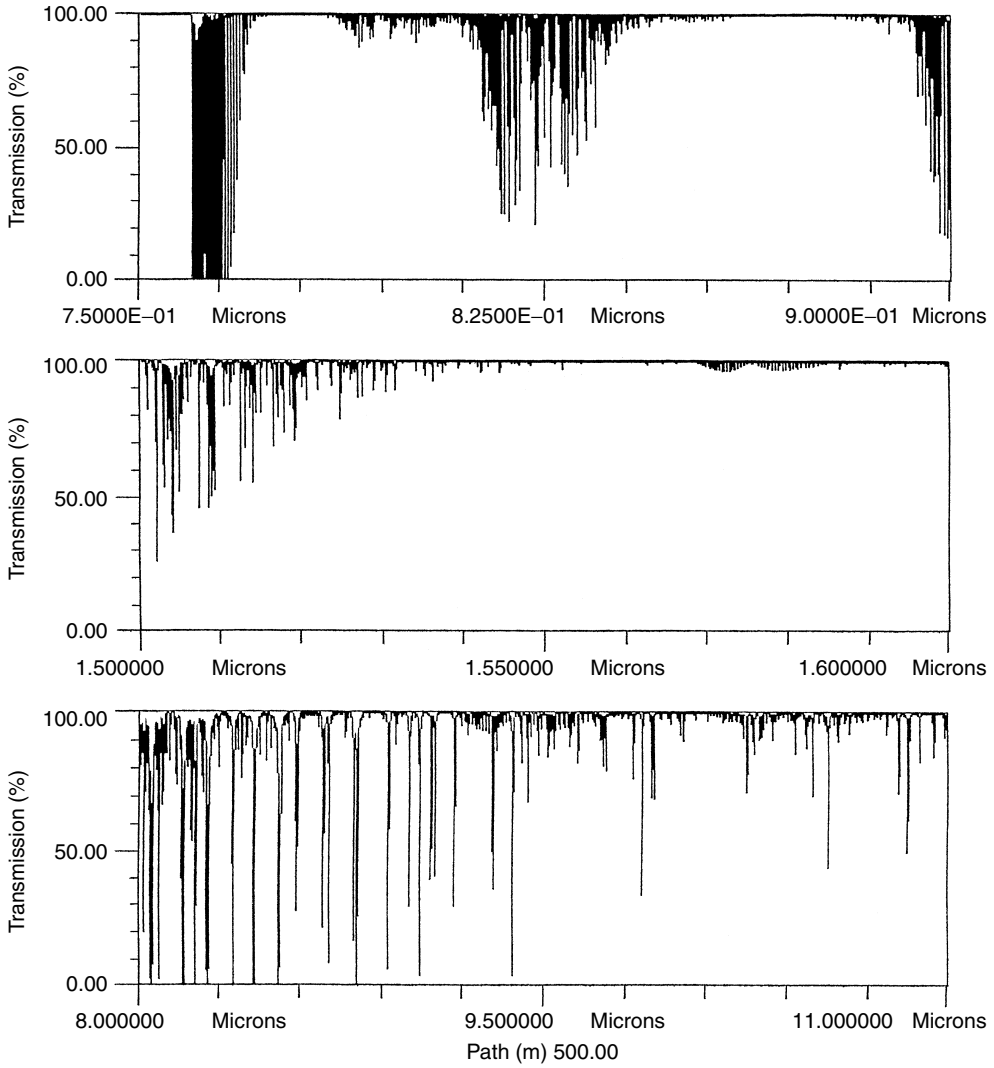


Figure 8. Calculated transmission spectrum of the atmosphere for a 500-m path near the wavelengths of 0.8, 1.55, and 9.5 μm. Most of the strong rotational–vibrational absorption lines seen are due to water vapor, CO_2, ozone, and oxygen.

When the scatter site is large or on the order of the wavelength of light then the scatter is a complex interference phenomenon with destructive and constructive interference lobes emanating outward from the particle. Such scatter is called *Mie scatter* and is highly dependent on angle, polarization, and wavelength/particle size. In theory, Mie scatter can be calculated for known particle sizes and orientation. However, it cannot be calculated a priori for complex shapes and orientations of particles such as those often found in the atmosphere. As such, the Mie scatter for the atmosphere is usually measured experimentally. Figure 9 shows the measured attenuation or extinction coefficient of the atmosphere as a function of wavelength for several different atmospheric conditions [25]. The values shown in Fig. 9 have error bars on the order of an order of magnitude dependent on atmospheric conditions.

Comparison of Figs. 8 and 9 suggests that at many laser wavelengths, the attenuation due to a strong absorption line in the atmosphere may be more dominant than that due to the normal background attenuation of the atmosphere such as that due to urban haze. In this case, FSO design is such that one chooses a laser wavelength that is offline of any strong absorption line in the atmosphere. After this choice, then the next dominant attenuation consideration is that due to clouds or heavy fog. For example, at a wavelength of about 1.51 μm, the extinction coefficient due to urban haze is about 0.9×10^{-4} m^{-1}, a value much smaller than that possible due to the molecular lines in Fig. 8. However, the attenuation due to absorption lines at 1.56 μm is negligible so that the Mie/Rayleigh or haze attenuation dominates.

Of more concern for a FSO system is the attenuation due to rain, snow, and fog. The Air Force MODTRAN and LOWTRAN computer programs have excellent attenuation calculations for rain and snow [22,23]. Under most cases with short ranges (<500 m), rain and snow attenuation may not be severe. However, fog can cause severe degradation in the signal. Figure 10 is a plot of the attenuation due to fog, rain, and snow as a function

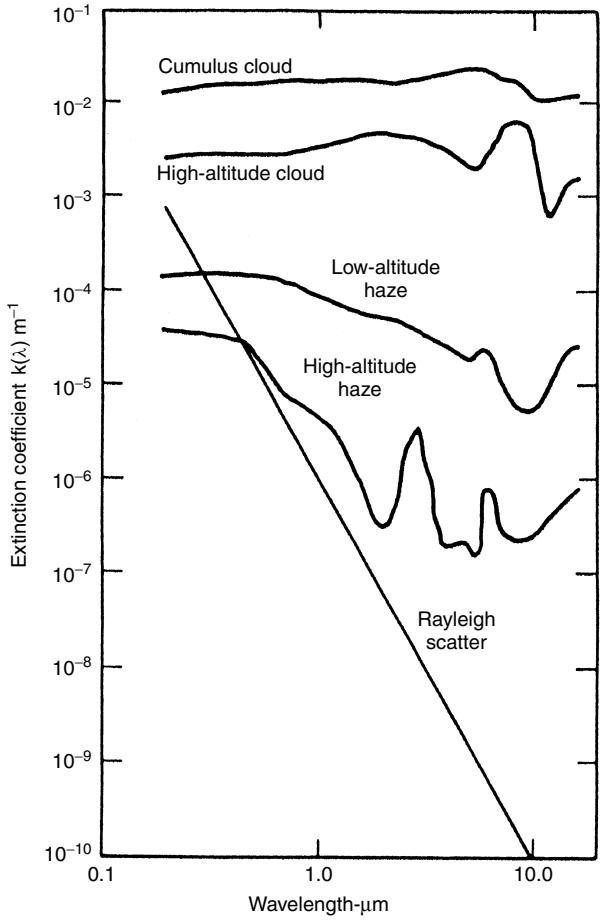


Figure 9. Approximate attenuation coefficient of the atmosphere as a function of wavelength for different atmospheric conditions. (Reproduced by permission of R. Measures, *Laser Remote Sensing*, John Wiley & Sons, New York, 1984.)

of the visibility [26]. As can be seen, thick fog can cause attenuation of up to 200 dB/km, or a reduction factor of 10^{-20} for a km path. Recent studies by Kim, McArthur, and Korevaar at 0.8 and 1.5 μm have indicated a refined equation for an approximation of the attenuation value, α, given by [26]

$$\alpha = \left(\frac{3.91}{V}\right)\left(\frac{\lambda}{550\text{ nm}}\right)^{-q}\ \text{km}^{-1} \qquad (3)$$

where V is the visibility (in km), λ is the wavelength in nm, and q depends on the size of the scattering particles, but is equal to 1.3 for average visibility and 0 for fog. In Eq. (3) α is given in units of km^{-1}.

It should be noted that in many reported studies, the attenuation values given in dB/km were extrapolated from short range studies on the order of several centimeters to meters, and were unable to take into account multipath scatter. The inclusion of multipath effects may decrease the overall attenuation value but also may spread out in time the modulated intensity waveform of the multiscattered beam [24].

Under normal conditions the first decision criterion for a FSO wavelength design is to reduce the atmospheric line spectra as shown in Fig. 8. Then the next consideration should be the reduction due to fog-type aerosols in the path. As can be seen in Fig. 9, the latter consideration may indicate that longer wavelengths near 9–10 μm may offer less attenuation due to fog and snow. As such, they may be considered for a backup system for a 0.8- or 1.5-μm FSO system, as opposed to the use of a microwave or RF backup system.

5. OPTICAL DETECTORS AND NOISE

Most current commercial FSO systems use the direct detection of the intensity-modulated laser beam. The optical detectors used are usually a small, high-bandwidth photodetector, either a Si photodiode or Si avalanche photodiode (APD) for wavelengths up to 1.1 μm, or a InGaAs photodiode or APD for the 1.5 μm wavelengths. To obtain the high speed required of 10 MHz up to 10 GHz, the size of the detector is kept small, on the order of 20–100 μm, to reduce capacitance and RC time constants. Table 2 shows a sampling of several optical detectors and some of their performance parameter values, including their size, electrical bandwidth, and noise equivalent power (i.e., minimum signal detected) [5,6,27]. As can be seen, they are small in size and have detection sensitivities ranging from a microwatt down to tens of nanowatts.

The overall science of optical detectors is covered in several excellent books, including that of R. Kingston, which covers photon counting, amplifier and background noise, and signal-dependent noise-limited performance of a FSO optical communication system [27,28]. In general, the detectors used in the visible to near-IR spectral region

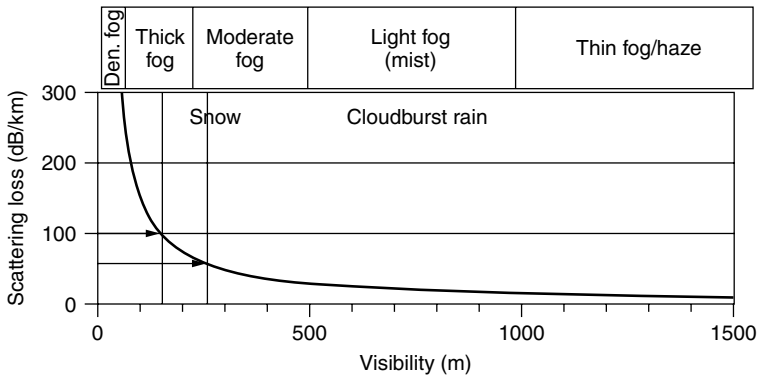


Figure 10. Attenuation/scattering loss as a function of visibility for different haze and fog conditions. (Reproduced by permission from I. I. Kim, B. McArthur, and E. Korevaar, SPIE Vol. 4214, 2001.)

Table 2. Typical NEP Values of Selected Detectors for Nighttime Conditions

	λ (μm)	Size (μm)	Bandwidth	Nighttime NEP (nW)
Si photodiode	0.9	2000	10 Mbps	200
Si APD	0.9	200	155 Mbps	20
InGaAS APD	1.5	50	2.5 Gbps	50
InGaAs APD	1.5	200	100 Mbps	20

[a]Daytime usage may require narrowband optical blocking filters to reduce background light. Typical daytime increase in NEP is ~2–8×.

are shot-noise-limited; that is, the dominant noise is the statistical fluctuations of the signal photons, which is essentially the square root of the number of photons in the signal. As such, the noise of the detector is usually stipulated in terms of a minimum detectable signal in decibels referenced to a milliwatt, or as a background current in the case of Johnson noise of the detector or amplifier combination. In the infrared spectral region, however, the background radiation or thermal emission of the 300-K world is often the dominant noise source. In the latter case, the detectors are background limited in their performance [i.e., background-limited infrared performance (BLIP)], and a different parameter related to the intrinsic sensitivity of the photodetector material is used to determine the system noise level. In this case, the detectivity, D^*, in units of Jones (i.e., cm $Hz^{1/2}$ W^{-1}), is a universal parameter for a particular material and is related to the noise equivalent power (NEP) in watts by

$$\text{NEP} = \frac{(A_D B)^{1/2}}{D^*} \tag{4}$$

where A_D is the area of the detector and B is the electrical bandwidth of the detector/amplifier combination [29]. The NEP is where the SNR in voltage equals 1 at the output terminals of the detector. The NEP is related to the term "sensitivity" as expressed in watts and used more often in the visible–near IR. "Sensitivity" is often used for a specified bit-error-rate and is about equal to 6 times NEP; this is explained in a later section.

With these concepts, one can compare different types of detectors at different wavelength ranges. Figure 11 shows a plot of the measured D^* for a range of infrared detectors along with that due to a S-20 photocathode photomultiplier tube (PMT), Si photovoltaic (pv) photodetector, and liquid nitrogen cooled HgCdTe for 10 μm wavelengths [30]. The influence of the 300-K background theoretical maximum thermal noise level shown by the doted line is given for both a photovoltaic detector and a photoconductor detector. As can be seen, the BLIP performance limit, due to the 300-K thermal background radiation that peaks near 8–10 μm in wavelength, is dominant for wavelengths greater than ~2 μm. Here, the D^* value was calculated for a bandwidth of 1 Hz for the PMT so that a comparison between the different detectors could be made. What are not shown in Fig. 11, however, are the signal bandwidth, lifetime, and cost of each detector. For example, many of the infrared detectors shown in Fig. 11 can operate at bandwidths

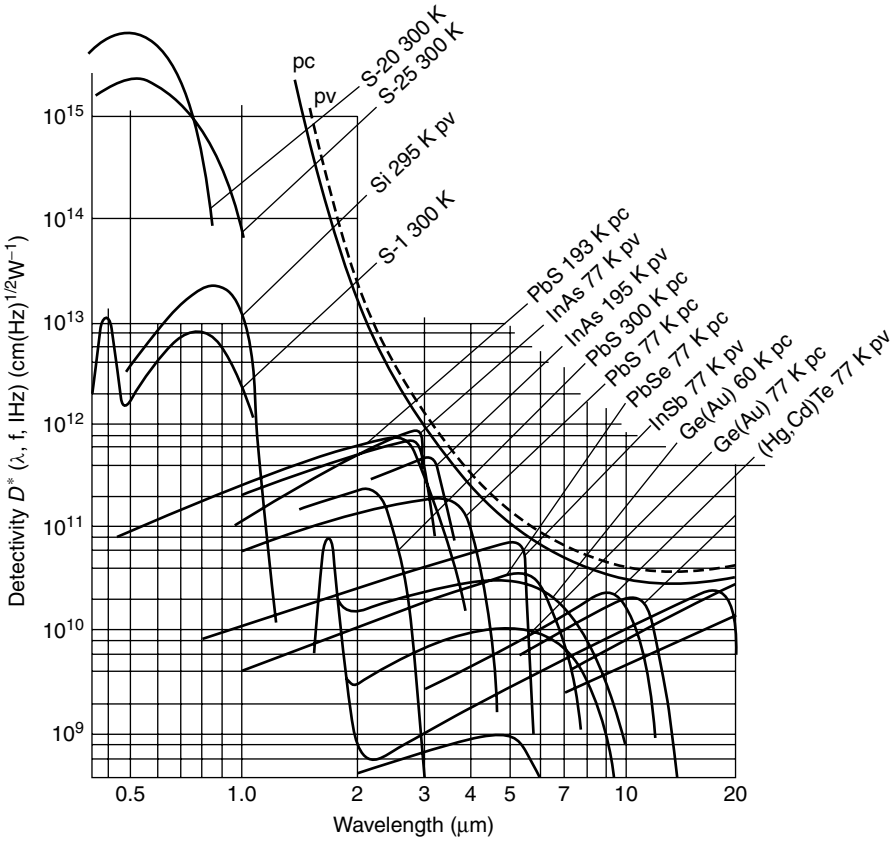


Figure 11. Detectivity, D^*, for selected detectors as a function of wavelength. (Adapted from P. R. Norton, *Handbook of Optics*, McGraw-Hill, New York, 1995, Chap. 15.)

of only 1–10 MHz or slower. On the other hand, cooled HgCdTe detectors have been operated at a wavelength of 10 μm in a heterodyne detection mode at speeds of ≤60 GHz.

Most communication systems use the bit error rate (BER) as a measure of the system sensitivity and level of signal-to-noise ratio in determining the probability of correctly decoding the bitstream in the signal [28,31]. The BER can be related to the signal-to-noise ratio (SNR) of the communication link and is a measure of the percentage of bits that are in error within a large ensemble of bits received. It is common for current FSO links to have a BER on the order of 10^{-9} to 10^{-10} for the case where no error correcting codes are used. The optical BER can be calculated as the integral of 1 minus the cumulative normal distribution function with argument $(\mathrm{SNR})_v/2$, where the SNR_v is the peak voltage SNR for a detector, which is the same as the returned peak power divided by the NEP of the detector. The average SNR power is half the peak SNR value. The "sensitivity" of a detector is defined for optical communication purposes as the average power required for a BER of 10^{-9}.

A plot of the BER value as a function of the voltage signal-to-noise ratio (returned optical power divided by NEP) is shown in Fig. 12 [28]. As can be seen, a BER of 10^{-9} requires a peak SNR of 12, which corresponds to a value of 6 for the average SNR value [28]. As such, the formula $\mathrm{SNR} = P_r/\mathrm{NEP} = 6$ is often used as the detection threshold for a FSO communication link; here, P_r is the average detected laser beam power by the receiver.

6. FSO RANGE EQUATION

The FSO range equation combines the attenuation and geometrical aspects of FSO in order to calculate the received optical power as a function of range and telescope aperture size. Before introducing the FSO range equation, some parameter values need to be defined and discussed.

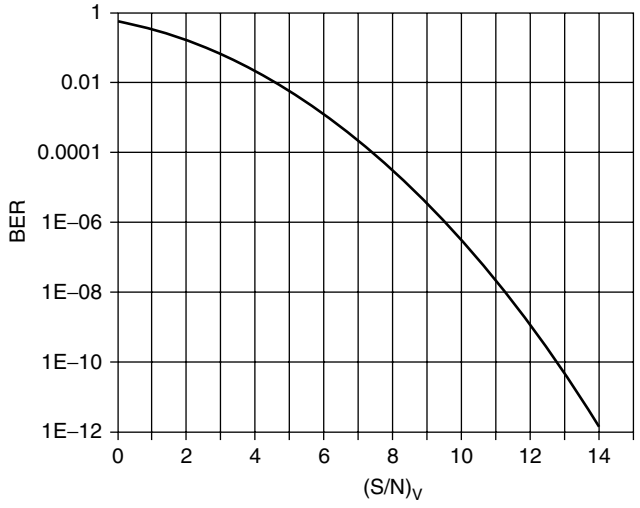


Figure 12. Bit error rate (BER) as a function of peak voltage signal-to-noise ratio (SNR), which is equal to the received power P_r, divided by the NEP of the detector. (Reproduced by permission of R. H. Kingston, *Optical Sources, Detectors, and Systems*, Academic Press, New York, 1995.)

6.1. Laser-Transmitted Beam Divergence

In the most simplistic case, the transmitted laser beam is divergent as a result of optical diffraction, where the angular spread, $\Delta\theta_1$ is equal to λ/D_1, where D_1 is the size of the initial laser beam. This is an approximate equation for the divergence of the lowest-order Gaussian spatial TEM_{00} mode for a Fabry–Perot laser cavity. Laser beams with higher-order spatial modes have greater diffraction than a Gaussian mode and will have a mode structure parameter, M^2, value greater than 1. In these cases, the divergence of the beam in one direction is equal to $M^2\Delta\theta_1$, or [19]

$$\Delta\theta_1 = \frac{(M^2)\lambda}{D_1} \tag{5}$$

The size of the projected laser beam at a distance of R meters will be equal to $D_1 + R\Delta\theta_1$. For example, a $M^2 = 10$ and 1 μm wavelength laser beam collimated through a 1 cm aperture will have an angular divergence of 10^{-2} radians (i.e., 0.57°) and will have a width of 10 m at a range of 1000 m.

Often, the beam divergence of a FSO system is made intentionally larger than the diffraction limit so that the projected beam size is larger than several times the size of the receiver telescope. This facilitates alignment of the two transmitter and receiver telescope optical axes. Beam divergence or beam spread is often made to be 0.1–1 mrad by slight defocusing of the transmitter telescope, as opposed to the normal Gaussian diffraction minimum of, say, 0.001–0.01 mrad.

Finally, the divergence of a semiconductor laser is often shaped by beamforming optics near the output facet of the diode laser. Normally, the output from a semiconductor laser has a beam divergence of, say, $3^\circ \times 15^\circ$. The beamshaping optics use a cylindrical lens to bring it down to a milliradian or so, in both axes.

6.2. Receiver Telescope Field of View

The receiver telescope field of view is the beam collection angle of the detector and telescope combination. Only light that falls or originates within this cone about the telescope optical axis will be focused onto the detector. The receiver telescope field-of-view angle, $\Delta\theta_r$, is given by

$$\Delta\theta_r = \frac{D_d}{f} \tag{6}$$

where D_d is the size of the detector and f is the focal length of the receiver telescope. Equation (6) does not usually influence the optical performance unless the optical axis of the receiver telescope–detector combination is aligned outside the receiver field of view, $\Delta\theta_r$. It should be noted that just because the receiver telescope intercepts a portion of the transmitted beam, the light collected would not be focused onto the detector unless the receiver telescope axis is pointed toward the transmitter location within $\Delta\theta_r$. For a typical detector size of 300 μm and a telescope focal length of 0.3 m, the field of view is about 10^{-3} radians.

6.3. FSO Range Equation Analysis

The FSO range equation can be given by inspection of Fig. 1, and the use of the Beer–Lambert law and Eq. (5). Under these simplifying assumptions, the FSO range equation is

$$P_R = P_T \frac{A_r}{(D_1 + R\Delta\theta_1)^2} T K e^{-\alpha R} \tag{7}$$

where P_R is the received optical signal power, P_T is the transmitted optical laser power, A_r is the area of the receiver telescope or collection lens, T is the transmission or efficiency of the receiver optical system, and the area of the beam at a range R is given by $(D_1 + R\Delta\theta_1)^2$. In Eq. (7) K is another loss factor that deviates from a normal value of 1 when a noncoherent light source is used, such as an LED. This latter parameter is equal to 1 for a laser source, and has a value equal to 1 or less for an LED source as

$$K = \frac{A_{\text{det}}}{A_{\text{LED}}} \tag{8}$$

if $A_{\text{det}} < A_{\text{LED}}$, and $K = 1$ otherwise, where A_{det} is the area of the detector and A_{LED} is the area of the LED source. The K factor takes into account the fact that a noncoherent optical source cannot be focused to an area smaller than that from which it originated due to thermodynamic reciprocity (brightness) considerations.

Equation (8) can be used to generate FSO SNR or power detection curves as a function of range. For example, Fig. 13 shows the calculated received power as a function of range for a case of a 10-Mbps bandwidth, low-power 0.85-μm LED FSO system with 40 mW power, 13 cm receiver, $T = 0.2$, and $K = A_{\text{det}}/A_{\text{LED}} = (0.28\text{ mm})^2/(0.5\text{ mm})^2 = 0.3$, divergence of $1^\circ = 0.0175$ radians, and NEP of the Si detector of 300 nW for daytime operation; these specifications are appropriate for a moderate power 0.85 μm LED FSO system and serves as a "strawman" for illustrative purposes only. Two atmospheric cases are shown for low α attenuation (10^{-4}/m, or 0.2 dB/km) due to low-altitude haze, and for moderate attenuation due to clouds (10^{-2}/m, or 20 dB/km) similar to light/moderate fog. As can be seen, the returned signal follows a $1/R^2$ dependence at close-in ranges, and follows the Beer–Lambert exponential decay at longer ranges. A threshold for the NEP of the detector at 300 nW is also shown, along with a value of the required SNR of 6 times greater than the NEP corresponding to a BER of 10^{-9}. As can be seen, the system should have good FSO communication capability at ranges out to 600 m for hazy conditions and out to 200 m under moderate/light fog conditions. These ranges would be even greater under nighttime conditions when the NEP of the detector would be less, or when operating under dry/no-fog conditions.

Another example is given in Fig. 14 for a 622-Mbps-bandwidth FSO "strawman" system appropriate for a higher power multi-beam 1.55 μm laser based FSO system. Here, the FSO parameters are approximately 1.55 μm wavelength, 400 mW power diode laser, $T = 0.2, K = 1$, 20 cm receiver telescope size, transmitted 1 mrad beam divergence, and a daytime NEP of 150 nW using two solar filters. Again, two attenuation cases are shown of 1.1×10^{-2}/m for light/moderate haze (cloud) and 0.7×10^{-4}/m for low-altitude haze. As can be seen the communication range is beyond 3 km for light haze and about 600 m under light/moderate fog. However, what is not shown in Fig. 14 are the deep fades due to constructive/destructive interference and atmospheric fluctuations in the beam. This will be mentioned in a later section of this article.

It should be added that the use of the FSO range equation is an approximation to give the reader some idea of the parameters and importance of there values. The calculated ranges are theoretical values, however, and are approximations subject to atmospheric conditions, which can cause errors in the attenuation as large as an order of magnitude or more. As such, it is important that the reader understand the usefulness and limitations in using the FSO range equation. Often, the values have to be modified by direct measurements under specific atmospheric conditions. This is why in so many cases extensive field tests of a FSO optical link have to be made under a wide range of weather conditions in order to accurately measure the system performance.

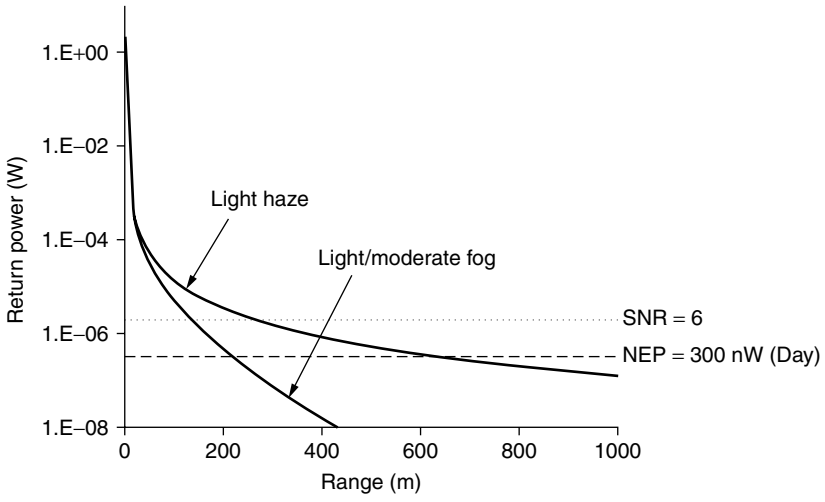


Figure 13. Calculated received optical signal as a function of range for "strawman" 0.85-μm LED FSO system. FSO range equation parameters included 40 mW power, 17 mrad beam divergence, and 13 cm telescope aperture.

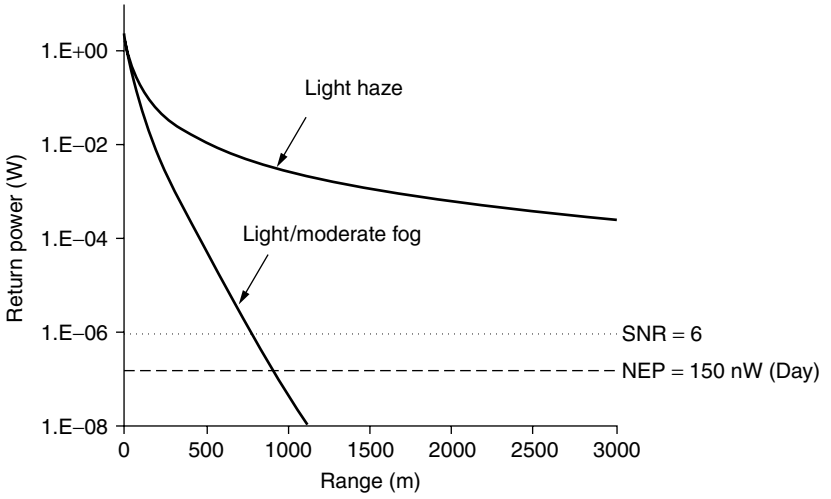


Figure 14. Calculated received optical signal as a function of range for "strawman" 1.55-μm high-power diode laser FSO system. FSO range equation parameters included 400 mW power, 1 mrad beam divergence, and 20 cm telescope aperture.

7. ATMOSPHERIC REFRACTIVE TURBULENCE

The most familiar effect of refractive turbulence in the atmosphere is the twinkling of the stars and the shimmer of the horizon on a hot day. The first of these is due to the random fluctuations in amplitude of the light, also known as *scintillation*. The second effect is the random change in the optical phase of the lightbeam that leads to a reduction in the resolution of an image. Other atmospheric effects are large-scale beam wander and breakup of the optical beam into smaller phase fronts or speckles. In the visible and near IR, these fluctuations are caused by small fluctuations in the temperature (0.01–0.1°) of the atmosphere on the spatial scale of 0.1 cm–10 m, which cause changes in the index of refraction of the atmosphere. These small-scale fluctuations can distort and break the laser beam into small turbulent cells. In the far IR spectral region, the influence of these temperature fluctuations is diminished, but large-scale spatial changes in the background absorption and concentration of water vapor can also cause large beam wander and fluctuations.

Extensive work by NOAA and DoD since the early 1960s, starting with the pioneering work of David Freed and Tatarskii, has been able to successfully clarify the phenomena of atmospheric refractive turbulence and yield predictive equations [24,32–34]. These atmospheric turbulence studies are valid for energy-conserving fluctuations in the atmosphere and have resulted in well-understood equations relating the optical beam fluctuations and the refractive-turbulence structure parameter, C_n^2. Figure 15 shows values of C_n^2 measured during a sunny day in Florida as a function of time using an optical beam intensity scintillometer instrument from NOAA [35]. As can be seen, the value of C_n^2 varies by several orders of magnitude, becoming largest during the middle part of the day.

The variation of C_n^2 with height above the ground is given approximately by [24]

$$C_n^2(h) = C_n^2(0)\, h^{-4/3} \tag{9}$$

where h is the height in meters above the ground and $C_n^2(0)$ is the value at ground level.

Fluctuations in the intensity or irradiance of the optical beam can be expressed approximately (for weak turbulence) as [24]

$$\sigma_r^2 = \exp(0.5\, k^{7/6} R^{11/6} C_n^2) - 1 \tag{10}$$

where σ_r^2 is the irradiance variance (normalized by the mean irradiance value), k is the optical wavenumber ($2\pi/\lambda$), and R is the range. This expression is modified slightly when the inner scale of the turbulence (smallest spatial fluctuation size) is greater than the square root of λR. There are several other expressions for σ_r^2 depending on the approach to saturation and the value of the inner

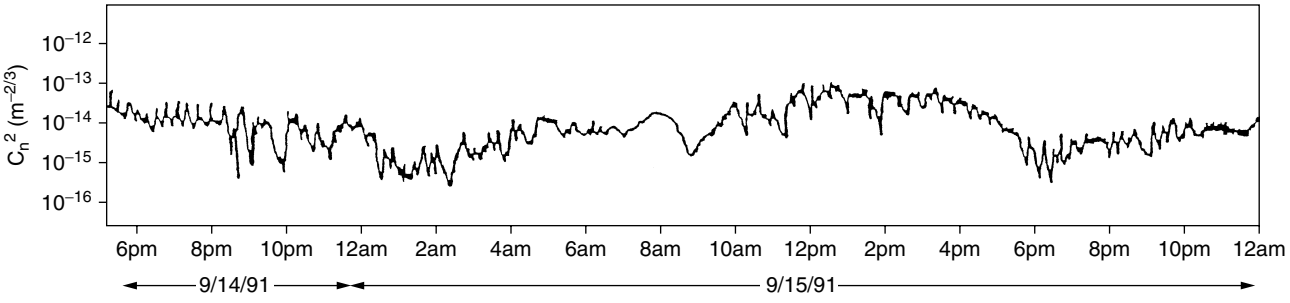


Figure 15. Measured atmospheric refractive-turbulence structure parameter, C_n^2, as a function of time throughout the day.

and outer scale values. The interested reader should consult the references for more detail [24,32–34].

The autocorrelation spectrum or power spectral density of the optical beam fluctuations gives the frequency or speed of the fluctuations. Experiments have shown that the decorrelation time of atmospheric refractive turbulence fluctuations is on the order of 1–10 ms, so that the frequency of the fluctuations is on the order of a few hundred hertz or less [36,37].

The influence of refractive turbulence is to break a Gaussian mode laser beam into smaller speckles or individual coherent wavefront groups. The effect can be given by the atmospheric turbulence field coherence length, ρ_0, which indicates the approximate size of the interference speckles within the laser beam, and is given by [24]

$$\rho_0 = (1.09\, k^2 R C_n^2)^{-3/5} \tag{11}$$

Equation (9) yields the size of the speckles within a beam front. It is an important parameter for a heterodyne detection or photon counting detection system since it places a limit as to the effective telescope aperture size (approximately $3\rho_0$) that can be used in a system [38]. However, for most current FSO systems that use direct detection of the beam, this will affect only the number of speckle modes to be aperture-averaged within the receiver, which will then affect the aperture-averaged SNR. The total system SNR will be a combination of all the fluctuation SNR values, power-limiting SNR considerations, and averaging effects within the communication decision time. Additional research is required in this area to better understand the tradeoffs in the area of signal averaging and atmospheric effects, including mitigation through the use of multiple wavelengths, beams, detectors, and temporal samples.

The beam wander due to refractive turbulence can be given by

$$\sigma_d^2 = 0.97 C_n^2 D^{-1/3} R^3 \tag{12}$$

where σ_d^2 is the variance in the displacement of the beam axis in m^2 and D is the diameter of the initial beam [39,40].

The preceding equations can be used to estimate the approximate level of fluctuations of a projected laser beam under controlled or laboratory conditions. Usually, the values for the standard deviation of the fluctuations σ are on the order of 0.05–0.7 (i.e., 5–70%), depending on the values of C_n^2 used. However, experience has shown that in many experimental and field-site cases complex windflow patterns exist (which are not energy-conserving) along with other atmospheric inhomogeneities that can cause beam drift and local absorption. As such, it is usually hard to accurately predict the fluctuation level in a particular setup. In this case, one has to resort to actual measurements of the fluctuation variance levels σ^2, in order to compare different experimental systems.

Under normal circumstances, the fluctuation variance level can be related to the information content signal-to-noise ratio (SNR) by [41]

$$\mathrm{SNR} = \frac{1}{\sigma^2} \tag{13}$$

where σ^2 represents the averaged or processed normalized variance measured over the time interval used to determine SNR of a decision period (possibly of one data bit or multiple bits for a coded word). Usually, if signal averaging is used, this has the effect of reducing the estimate of the variance by the square root of the number of samples averaged:

$$\mathrm{SNR}_n = \mathrm{SNR}_1\, n^{1/2} = \frac{n^{1/2}}{\sigma^2} \tag{14}$$

where n is the number of samples integrated, SNR_n is the SNR for n samples, and SNR_1 is the SNR for a single sample. Equation (14) is valid for an ergodic process that has random noise, but has to be reduced if nonrandom noise or processes are present [41]. This is true for large-scale attenuation processes in the infrared, where long-term temporal drifts in attenuation due to water vapor and other phenomena may be present. However, in the visible and near IR, the major noise sources are usually random.

Equation (14) is the general relationship for a signal detection process and helps relate the influence of increased fluctuation levels on SNR, which then directly affects the BER via Fig. 12. However, in the case of a FSO system, the fluctuation levels should more properly be measured only over the decision time of a bit. As such, the short-term variance measured over a time period of a data bit period may have to be used.

As can be seen in Eq. (14), the SNR can be improved through averaging multiple signals within the information decision period. This can be seen in an excellent research study by Kim et al., who studied the effect of refractive-turbulence fluctuations through the use of multiple laser beams [42]. Figure 16 shows the fluctuation levels measured over a 20-s time period for a 1.5-μm, 1.2-km FSO system that used one, two, and three separate laser beams. As can be seen, the use of several laser beams reduced the fluctuation levels that had the effect of increasing the measured SNR.

While the preceding equations indicate the level of fluctuations, it is customary in the FSO community for such an effect to be accounted for by introducing a fluctuation, fade, or link margin that is derived more from experimental measurements than from theory. Such a link margin may be on the order of 15–20 dB, although such a high value often includes the desired SNR for a specified BER compared to that for a SNR of 1. For example, for a system with a BER of 10^{-9} and a NEP of the detector of 20 nW, the BER value is related to a SNR of 6 [28]. In this case, a fade link margin could be 7 dB (factor of 5), so that the required SNR would be 6 times that due to 7 dB, specifically, a value of 30, or 15 dB. This is the equivalent of solving the range equation for a SNR of 1, but using a link margin of 15 dB. Both analysis types are used in the literature and are equivalent.

Finally, another atmospheric phenomena related to deep fades in the communication link has received considerable study [43,44]. For a FSO laser-based system transmitting over moderate to long ranges of 2–10 km, the fluctuations observed have the traditional refractive turbulence frequency fluctuations of up to a few hundred hertz, but long-term fades lasting 0.1 s to a few seconds are also observed. These are difficult to extract from similar effects due to tracking drifts or from building sway.

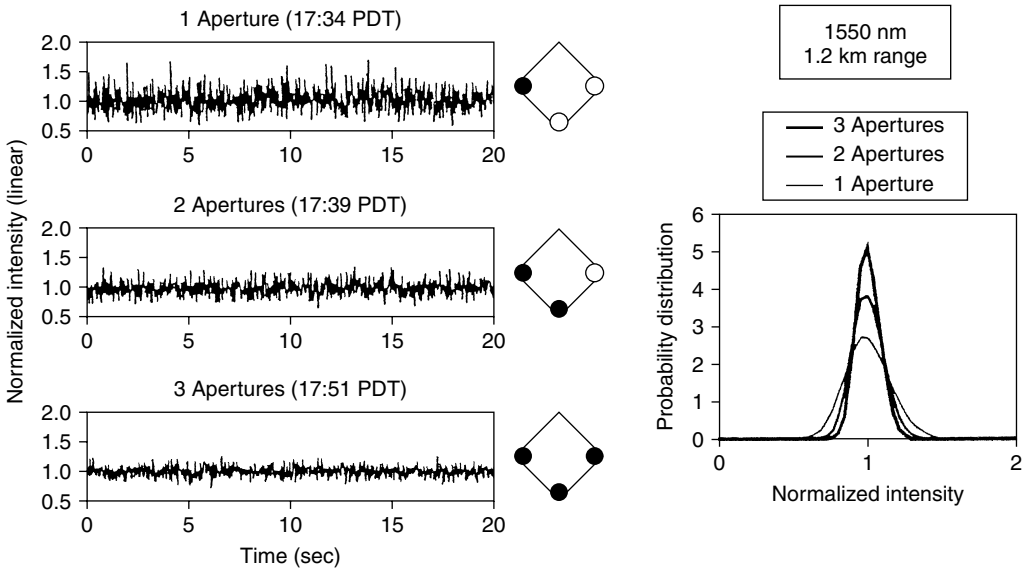


Figure 16. Intensity and fluctuation distribution measured for a 1.2-km FSO 1.55-μm system using one, two, and three laser beams. The reduction in signal fluctuations using three beams is easily seen. (Reproduced with permission of I. I. Kim, M. Mitchell, and E. Korevaar, SPIE Vol. 3850, 1999.)

However, some studies suggest that they may also be due to beam bending due to the presence of spatial localized concentrations of water vapor that move into the beam path. Such water vapor spatial clouds have been observed with high resolution Raman lidar (laser IR radar) systems, and indicate water vapor "clouds" of 10–100 m in diameter and movement times on the order of 1–100 s [45]. On the other hand, long-term fading can occur if the coherent beam produces a single speckle at the receiver and the optical alignment is such that destructive interference occurs (due to long-term building sway, beam wander, etc.). In this case, the intensity received is close to zero. Sometimes, these fades have values of up to 10–30 dB, suggesting partial destructive interference of the speckle. The effects of these fades can be mitigated through the use of multiwavelength, multispatial, or multitemporal samples within the link bit decision period in order to produce independent samples of the transmitted bit.

8. TELESCOPE DESIGN, TRACKING/ALIGNMENT DETECTORS, AND ENVIRONMENT

The telescope and receiver lens used in a FSO system is usually a compromise between the use of wide apertures for greater light gathering, short focal length for ease of handling, moderate field of view for ease of alignment, and optical coatings for narrow spectral filters for daytime use. This can be seen in Fig. 17, which shows several typical telescope configurations. Common configurations for a FSO system is either the use of a single lens and a

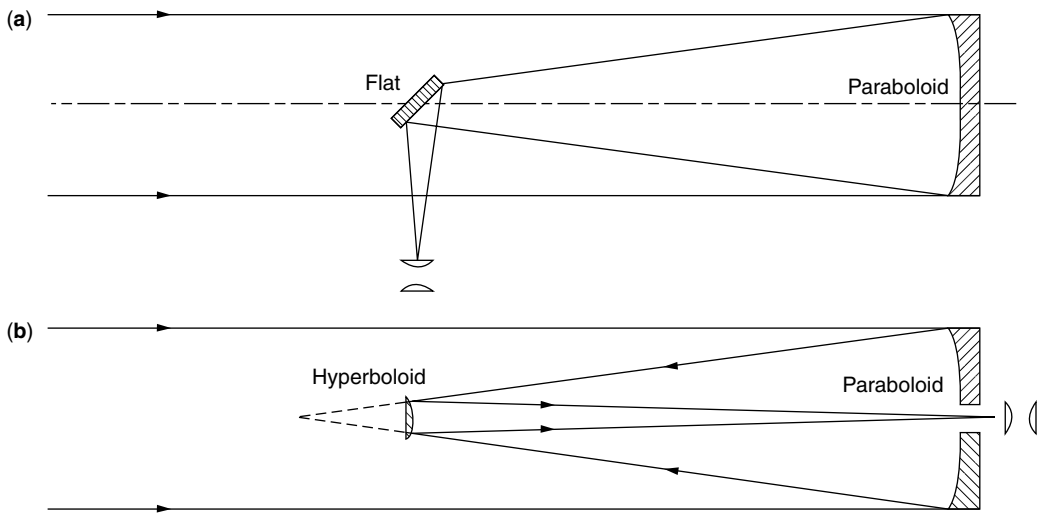


Figure 17. Schematic of Newtonian (**a**) and Cassegrainian (**b**) telescopes.

detector or the use of a Cassegrain telescope and a detector system that folds the beam path so that the telescope is shorter than that of a Newtonian configuration. An example of a short Cassegrain-type telescope used for FSO is one developed by Kaiser Electro-Optics in their f#/0.67 Hyperscope transceiver with an 8 in. aperture and 4 in. overall length. In this telescope a coated front reflecting window is used as the secondary mirror as in a Mangin mirror telescope configuration, which shortens the overall length of the telescope by a factor of ~2 [46,47].

Another method is to use a holographic lens or mirror arrangement to also ensure spectral and position placement that is similar to combining a diffraction grating with a lens [48]. Such a holographic lens is flat. Another possibility is that the holographic lens can be configured to displace or rotate the focused light around the perimeter of the lens as a function of input direction.

A common configuration for an inexpensive telescope systems is to use either a conventional optical lens or a flat Fresnel lens. In the latter case, the focal volume and resolution are not as good as with a conventional lens, but may be sufficient since in many cases the beam divergence and field of view of the transceiver have been enlarged to ease optical alignment of the system.

Alignment and tracking of the FSO system can be a difficult problem for longer-range systems. For short ranges under 200–500 m, the alignment and beam wander of the communication beam may be small enough that only coarse mechanical alignment is required during initial setup. For longer ranges, however, usually some form of active alignment and tracking is required. Two axis gimbals on the telescopes or the entire FSO unit are often used in this regard, as well as active positioning of the detector in the x–y image plane of the telescope. Fluctuation times on the order of 1–10 ms are required for atmospheric fluctuation, while building sway may have resonance times on the order of 1–10 s. Temperature- and wind-driven gross movement of buildings will have times on the order of hours. It is common to use quad (4) detectors and/or CCD cameras to monitor a separate alignment beam in order to actively track and compensate such movement.

Weather/environment protection of a FSO system is determined by the indoor or outdoor use of the system. Several systems are made for indoor use and require little environmental design. Some units are made for use behind the windows of an office and use the window and benign office conditions to separate them from harsher outside conditions. Outdoor use often entails hermetically sealing the FSO unit, with the placement of sunshields, rain canopies, and heaters on the telescope lens to reduce frost or condensation.

9. LASER EYE SAFETY

Any laser beam can cause damage to the human eye if it is operating with an irradiance (W/cm^2) that is above a certain level. The minimum permissible exposure level (MPE) is tabulated in the ANSI standards [49]. The standards are written for direct ocular view by the eye, and are given as a function of wavelength. This latter part is important because for wavelengths less than ~1.4 μm (~1400 nm), the optical radiation that enters the eye is focused and increased in irradiance onto the retina. For wavelengths longer than ~1.4 mm, the light is absorbed by the cornea and vitreous humor inside the eye. The reader should remember that it is possible for a coherent laser beam to be focused by a lens down to the diffraction limited spot size given by $f\#\,\lambda$, where the f-number, f#, of the lens is equal to the diameter of the lens (or pupil of the eye) divided by the focal length of the lens; as can be appreciated, for the human eye this spot size on the retina is close to the wavelength of light: about 1 μm in diameter. Such a focusing effect is already taken into effect by the ANSI standards.

While the reader should use the ANSI standards for each explicit situation, a general idea of the eye-safety values can be shown as in Table 3. Table 3 shows the direct ocular MPE values, and the approximate calculated maximum transmitted laser power levels for a FSO transmitter that delivers a 30 × 30-cm beam at a distant receiver where the ocular viewing would occur; this corresponds to the size of a 1° divergence laser beam after being transmitted 200 m through the atmosphere. The MPE is shown for a 10-s exposure. The maximum transmitted laser power is seen to be about 1 W for the 0.7-μm laser and about 90 W for the 1.55-μm system. Of course, these values are much lower if the eye is at a closer range and intercepts the beam where the beam is smaller and the irradiance is higher. In addition, the effect of atmospheric turbulence may increase the fluctuating intensity levels so that the values would need to be reduced appropriately. The eye safety for a LED is higher since it is a noncoherent source and will not be focused to a small diffraction-limited spot on the retina.

Most FSO systems are designed to be eye-safe, or to operate where a human will not intercept the beam. In the case where humans may intermittently intercept the beam, other warning systems such as the use of an inexpensive microwave radar system (e.g., marine radar) could be used to detect the presence of a human within the FSO beam path and shut down the system temporarily.

Finally, laser safety regulations and standards have been instituted as to the manufacturing classification of the laser, such as class IV or class IM, which covers the power levels and operational safety standards for the laser. The international organizations such as the International Electrotechnical Commission (IEC) and U.S.

Table 3. Approximate Maximum Permissible Exposure (MPE) Power in W/cm^2 for Direct Ocular (Eye) View for Several Different Wavelengths

λ (μm)	MPE (mW/cm^2)	P_t (Maximum Transmitted Power) – 200 m away (w)[a]
0.7	1	0.9
0.9	25	22
1.55	100	90
10	100	90

[a]The maximum transmitted laser power calculated is for the case of a beam size of 30 cm in diameter at the position of the eye (appropriate for a propagation distance of 200 m and beam divergence of 1 degree).

agencies (FDA, Laser Institute of America, and ANSI) have helped coordinate these classification schemes. For example, the IEC has expanded its coverage of TC 76/60825 part 12, Working Group 5 (WG5) to cover safety and transmission issues associated with laser and LED FSO communication [50]. The WG5 committee has been working on a draft titled *Part 12: Safety of Free Space Optical Communications Systems Using Directed Beams* to be released in 2003. The interested reader is encouraged to review this information in the references. They are not of concern for the scientific analysis of a FSO system, but are very important for the manufacturing and proper use of commercial FSO systems.

10. FSO SYSTEM AND ENGINEERING TRADE-OFFS

As can be seen from the above discussions, there are many scientific aspects of FSO design. In addition, there are a considerable number of system and engineering trade-offs that also have to be looked at. It is beyond the scope of this article to discuss this in detail, but some general aspects can be listed. Some of the trade-off parameters that need to be taken into account include (1) modulation of the laser or LED (direct modulation through the power supply or the need for an expensive external modulator), (2) detector bandwidth and cooling requirements in the case of IR detectors, (3) increased laser beam divergence and possible need to increase laser power versus increased cost of using active alignment of a narrow laser beam, (4) cost of laser or LED system at different wavelengths versus advantages of availability of cheaper detector components versus penetration of beam through fog or rain, and (5) eye safety versus laser beam size versus divergence of beam and beam size at detector telescope. As can be seen, there are a significant number of engineering trade-offs that have to be made in any FSO system. Although the above list of trade-offs seems formidable, it is not really as bad as it first appears. This is because there are many different ways to build a successful FSO system for a specified operating condition. As such, there is no "one" or ultimate maximized system, but rather several that are sufficient to provide the communication link required. The most basic, first-level trade-off studies are often conducted to provide a high level of reliability and link BER for the specified atmospheric conditions and system environmental conditions. Then, within these broad constraints, one finds that there are usually several approaches that will meet the requirements.

11. FUTURE TECHNICAL AND INDUSTRY CONSIDERATIONS

FSO is just starting to impact the Internet "last- mile" interconnectivity problem. It is felt that it may offer the unlimited bandwidth solution for this problem within the metro urban core involving downtown building-to-building communication, but may also be a major technology for home-to-home and office-to-office connectivity. As stated earlier, if the home/business last-mile connectivity becomes the main technical driving force for communication, then the optimization of the technical specifications for FSO may become more important than using laser wavelengths and transceivers as part of the current fiberoptic communication legacy. In that case, wavelength issues and tradeoffs between long-range point-to-point versus short-range mesh-net connectivity will be addressed and operational standards will be set by the industry. It is common to hear that FSO solves a technical problem at present, but that the industry does not yet recognize or fully understand the potential that FSO offers. As such, some of the current problems in the deployment of FSO is in the industry perception and marketing of this technology. FSO systems have now shown that they are reliable (99.9% to 99.999%) communication channels that have fast bandwidth, are easy to set up, and provide cost-effective solutions. The FSO community recognizes these concerns, and has launched the Free Space Optics Alliance organization [51]. The FSO Alliance currently consists of about 25 companies, and has a mission to educate and promote FSO technical information to the communication industry as a whole and the print and journal media sector. It is believed that through such industry wide education, that industry standards and proper growth of FSO technology will occur within the communication carrier industry.

It is believed by the author that FSO is on the verge of changing the basic communication medium and technology of the metro and last-mile network market. The challenges and importance of setting standards for FSO are becoming increasingly clear in order to help the field become a major component in the whole communication network. It is hoped that the reader has gained an appreciation of some of the technical challenges and opportunities that FSO offers in this regard.

Finally, the author would like to acknowledge several helpful discussions and suggestions from Drs. Issac Kim, David Rockwell, and John Schuster.

BIOGRAPHY

Dennis K. Killinger received the B.A. degree from the University of Iowa, M.A. degree from De Pauw University, and Ph.D. degree in Physics from the University of Michigan. He has conducted research on radar analysis and microwave atmospheric propagation while employed as a Research Physicist at the Naval Avionics Facility, and joined the research staff in Quantum Electronics at Lincoln Laboratory, Massachusetts Institute of Technology in 1978 conducting research in the development of new solid-state lasers and their application as spectroscopic lidar probes of the atmosphere. Since 1987 he has been a Professor of Physics at the University of South Florida and is a Distinguished University Professor and Director of the Laboratory for Atmospheric Lidar and Laser Communication Studies. Dr. Killinger is a Fellow of the Optical Society of America, Senior Member of the IEEE, past associate editor of *Applied Optics and Optics Letters*, past member of the NAS/NRC Committee on Optical Science and Engineering, and has served as chairman of several international conferences on lasers and applied spectroscopy. He has published over 200 technical papers, reports, and conference papers in laser remote sensing/lidar, applied laser spectroscopy, laser

physics, laser atmospheric propagation, and free-space optics (FSO) laser communication.

BIBLIOGRAPHY

1. E. J. Korevaar, ed., *Optical Wireless Communication II*, SPIE Vol. 3850, 1999.
2. G. S. Mecherle, ed., *Free-Space Laser Communication Technologies XII*, SPIE Vol. 3932, 2000.
3. E. J. Korevaar, ed., *Optical Wireless Communications III*, SPIE Vol. 4214, 2000.
4. *Harnessing Light: Optical Science and Engineering for the 21st Century*, Committee on Optical Science and Engineering (COSE) NRC Report, National Academy Press, Washington, DC, 1998.
5. W. Wolfe and G. Zissis, eds., *The Infrared Handbook*, 3rd ed., Environmental Research Institute of Michigan (ERIM) and SPIE, 1989.
6. M. Bass, E. Van Stryland, D. Williams, and W. Wolfe, eds., *Handbook of Optics*, 2nd ed., Optical Society of America, McGraw-Hill, 1995.
7. D. J. Petrovich, R. A. Gill, and R. J. Feldmann, US Air Force development of a high-altitude laser crosslink, in SPIE Vol. 4214, 2000, pp. 14–25.
8. I. I. Kim et al., Preliminary results of the STRV-2 satellite to ground lasercom experiment, in SPIE Vol. 3932, 2000, pp. 21–43.
9. S. Lee, J. W. Aleander, and M. Jeganathan, Pointing and tracking subsystem design for optical communications link between the international space station and ground, in SPIE Vol. 3932, 2000, pp. 150–157.
10. H. Willebrand and B. S. Ghuman, *Free-Space Optics: Enabling Optical Connectivity in Today's Networks*, Sams Publications, Indianapolis, 2002.
11. PlainTree, Inc., Ottawa, Ontario, Canada, *www.plaintree.com*.
12. Optical Access, Inc., San Diego, CA, *www.opticalaccess.com*.
13. fSONA, Inc., Richmond, BC, Canada, *www.fsona.com*.
14. G. Nykolak et al., A 40 Gb/s DWDM free space optical transmission link over 4.4 km, in SPIE Vol. 3932, 2000, pp. 16–20.
15. J. Hecht, *The Laser Guidebook*, 2nd ed., McGraw-Hill, New York, 1992.
16. *Laser Focus World Buyers' Guide*, Pennwell Publications, 2002; *www.laserfocusworld.com*.
17. *Photonics Spectra Buyers' Guide*, Laurin Publication, 2002.
18. G. Scamarcio et al., High power infrared (8 μm wavelength) superlattice lasers, *Science* **276**: 773–776 (1997).
19. P. Mamidipudi and D. Killinger, Optimal detector selection for a 1.55 micron KTP OPO atmospheric lidar, in SPIE Vol. 3707, 1999, pp. 327–335, and references cited therein; A. E. Seigman, *Lasers*, University Science Books, Mill Valley, CA, 1986.
20. R. M. Goody and Y. L. Young, *Atmospheric Radiation*, Oxford Univ. Press, 1989.
21. R. T. Menzies and D. K. Killinger, IR Lasers tune into the environment, *IEEE Circuits Devices* **10**: 24–29 (1994).
22. L. S. Rothman et al., The HITRAN molecular database: Editions of 1991 and 1992, *J. Quant. Spectrosc. Radiat. Transfer* **48**: 734 (1992).
23. HITRAN, FasCode, HITRAN-PC, and PCTRAN computer programs; ONTAR Corp., 9 Village Way, North Andover, MA 01845-2000; Website *www.ontar.com*.
24. D. K. Killinger, J. H. Churnside, and L. S. Rothman, Atmospheric Optics, in M. Bass, ed., *OSA Handbook of Optics*, 1995, Chap. 44.
25. R. Measures, *Laser Remote Sensing*, Wiley-Interscience, New York, 1984.
26. I. I. Kim, B. McArthur, and E. Korevaar, Comparison of laser beam propagation at 785 nm and 1550 nm in fog and haze for optical wireless communication, in SPIE Vol. 4214, 2001, pp. 26–37.
27. E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, Wiley, New York, 1996.
28. R. H. Kingston, *Optical Sources, Detectors, and Systems: Fundamentals and Applications*, Optics and Photonics; Academic Press, New York, 1995; R. H. Kingston, *Detection of Optical and Infrared Radiation*, Springer, New York, 1978.
29. J. S. Accetta and D. L. Shumaker, eds., *The Infrared and Electro-Optical Systems Handbook*, Environmental Research Institute of Michigan, Ann Arbor, MI, 1993.
30. P. R. Norton, Photodetectors, in *OSA Handbook of Optics*, McGraw-Hill, New York, 1995, Chap. 15, pp. 15–16.
31. B. R. Strickland, M. J. Lavan, E. Woodbridge, and V. Chan, Effects of fog on the bit error rate of a free-space laser communication system, *Appl. Opt.* **38**: 424–431 (1999).
32. D. L. Fried and J. B. Seidman, Laser beam scintillation in the atmosphere, *J. Opt. Soc. Am.* **57**: 181–185 (1967).
33. V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation*, Israel Program for Scientific Translations, Jerusalem, 1971.
34. L. C. Andrews and R. L. Phillips, *Laser Beam Propagation through Random Media*, SPIE Press, 1998.
35. W. E. Wilcox, Jr., *Diurnal measurements of atmospheric optical turbulence with application to coherent lidar*, master's thesis, Dept. Physics, Univ. South Florida, Tampa, 1991.
36. G. Nykolak et al., Update on 4 × 2.5 Gb/s, 4.4 km free-space optical communications link: Availability and scintillation performance, in SPIE Vol. 3850, 1999, pp. 11–19.
37. N. Menyuk and D. Killinger, Temporal correlation measurements of pulsed dual CO_2 lidar returns, *Opt. Lett.* **6**: 301–303 (1981).
38. K. P. Chan and D. K. Killinger, Enhanced detection of atmospheric-turbulence distorted 1 micron coherent lidar returns using a two-dimensional heterodyne detector array, *Opt. Lett.* **16**: 1219–1221 (1991).
39. J. H. Churnside and R. J. Lataitis, Wander of an optical beam in the turbulent atmosphere, *Appl. Opt.* **29**: 926–930 (1990).
40. I. I. Kim et al., Wireless optical transmission of fast Ethernet, FDDI, ATM, and ESCON protocol data using the TerraLink laser communication system, *Opt. Eng.* **37**: 3143–3155 (1998).
41. N. Menyuk, D. K. Killinger, and C. R. Menyuk, Limitations of signal averaging due to temporal correlation in laser remote sensing measurements, *App. Op.* **21**: 3377–3383 (1982).
42. I. I. Kim, M. Mitchell, and E. Korevaar, *Measurement of scintillation for free-space laser communication at 785 nm and 1550 nm*, in SPIE Vol. 3850, 1999, pp. 49–62.

43. P. Polak-Dingels, P. R. Barbier, D. W. Rush, and M. L. Plett, Long-term fading statistics measurements of an atmospheric optical communication channel, in SPIE Vol. 3850, 1999, pp. 40–48.

44. C. C. Davis et al., Characterization of a liquid filled turbulence simulator, SPIE Conf. Artificial Turbulence and Wave Propagation, July 1998.

45. D. N. Whiteman, S. H. Melfi, and R. A. Ferrare, Raman lidar system for the measurement of water vapor and aerosols in the earth's atmosphere, *Appl. Opt.* **31**: 3068–3082 (1992).

46. *Hyperscope FSO Telescope*, Kaiser Electro-Optics, Rockwell-Collins, *www.keo.com*.

47. T. Carbonneau and G. S. Mecherle, *SONAbeam optical wireless products*, in SPIE Vol. 3932, 2000, pp. 45–51.

48. *Holographic Rotating Telescope: HARLIE Lidar Technology Program*, NASA Goddard, Greenbelt, MD; *http://bll.gsfc.nasa.gov/harlie*.

49. *American National Standard for Safe Use of Lasers*, ANSI Z136.1 - 2000, Laser Institute of America, Orlando, FL, 2000; *http://www.laserinstitute.org*.

50. D. Britz, *Free Space Optical Communication: A New Broadband Access Technology with Implications for Laser Safety in the Public Sector*, Laser Institute of America Newsletter, Jan./Feb. 2002, pp. 1–7; *http://www.laserinstitute.org*.

51. Free Space Optics Alliance, *http://www.fsoalliance.com*.

ORTHOGONAL FREQUENCY-DIVISION MULTIPLEXING

LEONARD J. CIMINI JR.
LARRY J. GREENSTEIN
AT&T Labs-Research
Middletown, New Jersey

1. INTRODUCTION

The permissible data rate of a digital communications link is limited by the available bandwidth and also by power and noise. The data rate can also be limited by phenomena in the communications medium (channel) between the transmitter and the receiver, especially by intersymbol interference (ISI) caused by time dispersion of the transmission medium, such as occurs on the multipath radio channel and the frequency-selective telephone channel.

As a general rule, the effects of ISI are small as long as the time extent of the channel impulse response is significantly shorter than the duration of a transmitted symbol. This implies that the symbol rate transmitted over a dispersive channel is practically limited by the channel's memory. However, mechanisms exist for countering ISI and thus extending symbol rates. These include receiver equalization, transmitter preequalization, and some forms of radio diversity. All are aimed at permitting the transmission of datastreams with symbol periods comparable to, or even smaller than, the channel's memory.

An alternative approach employs multiple carriers. In multicarrier transmission, the datastream to be transmitted is split into multiple parallel datastreams of reduced rate, and each of them is transmitted on a separate frequency (or *subcarrier*). Each subcarrier is modulated at a rate so low (or, equivalently, has a symbol period so long) that dispersion does not cause a problem. A given subcarrier, with its associated data signal, constitutes a *subchannel*. Ideally, the bandwidth of a subchannel would be so narrow as to preclude any ISI. More realistically, there will be *reduced* ISI on each subchannel, which can be either tolerated or easily corrected. Since the system's data throughput is the sum of the throughputs of the parallel subchannels, the data rate per subchannel is only a fraction of that of a single-carrier system having the same throughput. Thus we see that multicarrier transmission permits high data rates while maintaining symbol durations much longer than the channel's memory.

At the same time, the subchannels must be spaced, and spectrally shaped, to ensure that they do not interfere with each other. Such precautions can limit spectral efficiency, defined as the total bit rate divided by the total bandwidth. *Orthogonal frequency-division multiplexing* (OFDM) [1–4] is a special case of multicarrier transmission that permits the subchannels to overlap in frequency without mutual interference. In addition to improved spectral efficiency, this technique exploits digital signal processing technology to obtain a cost-effective means of implementation. Our primary aim here is to detail the theory and practice of this form of multicarrier transmission.

Before proceeding with the basics, we pause to note the numerous communications systems, past and present, that have used some form of multicarrier transmission. The first systems using this technique were designed in the late 1950s and early 1960s for military high-frequency radio applications [5,6]. These included the Kineplex and Kathryn systems. Since these early systems, multicarrier transmission (and in particular OFDM) has been used over many different communications media. Practical interest has increased partly as a result of enabling advances in signal processing and microelectronics, and partly because of the demand for ever-higher data-rate services over dispersive channels. Multicarrier modems have been standardized in different parts of the world for both wireline and wireless data applications, including digital audio/video broadcasting (DAB/DVB) [7,8]; digital transmission over copper wire, for example, digital subscriber loops (DSLs) [9]; wireless local-area networks (WLANs) [10]; and have been proposed for mobile radio applications [11].

2. BASIC CONCEPTS

2.1. Multicarrier Transmission

There are several techniques for realizing a multicarrier link. In the conceptually simplest approach, the total signal frequency band is divided into N ideally nonoverlapping (band-limited) frequency subchannels, employing N independent transmitter–receiver pairs. A block diagram description of how this can be done is given in Fig. 1. In the transmitter (Fig. 1a), an input stream of data, at rate R bits per second (bps), is divided into N

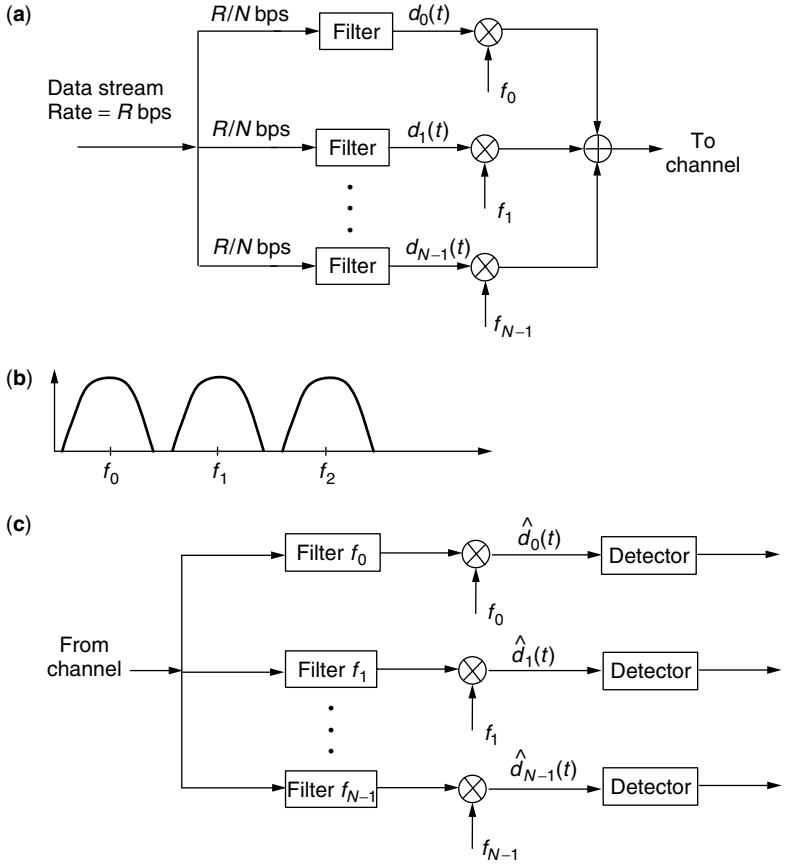


Figure 1. Multicarrier transmission system: (**a**) multicarrier transmitter; (**b**) transmit spectrum; (**c**) multicarrier receiver.

parallel substreams, each at data rate R/N bps. (The data values in the mainstream and the substreams are, in general, complex, and the real and imaginary components can be binary or multilevel.) Each substream is passed through a baseband pulseshaping circuit ("filter"), where we assume identical filters for all substreams. The kth filter output ($k = 0, 1, \ldots, N-1$) is then upconverted by a balanced mixer to frequency f_k. The result is a subcarrier with *quadrature amplitude modulation* (QAM). The N-QAM signals are combined (frequency-multiplexed) and sent over the channel. An example of the output signal spectrum is given in Fig. 1b. In the receiver, Fig. 1c, a set of bandpass filters centered on f_k, $k = 0, 1, \ldots, N-1$, is used to frequency-demultiplex the N subchannels, after which each subchannel is downconverted to baseband by a balanced mixer. Each substream is then applied to a detector, and the output data values are sent on for possible further processing. The spectral guard bands shown between subchannels in the figure are introduced so that easily realizable filters can be used in the receiver.

While the advantages of multicarrier transmission in terms of reduced sensitivity to dispersion are obvious, there are two major disadvantages to this particular realization. First, it is spectrally inefficient, since the signals must be sufficiently spaced in frequency to facilitate separation at the receiver. Second, a receiver with a large bank of filters may be prohibitive in terms of complexity and cost. The alternative approach (OFDM), using overlapping subchannels (to improve the spectral efficiency) and efficient digital signal processing techniques (to reduce the complexity and cost), is described next.

2.2. Basic OFDM

Orthogonal frequency-division multiplexing provides a solution to the disadvantages of conventional multicarrier transmission. In particular, a more efficient use of bandwidth can be obtained if the spectra of the individual subchannels are permitted to overlap, with specific orthogonality constraints imposed to facilitate separation of the subchannels at the receiver. Figure 2 shows the spectra for the two alternative forms of multicarrier transmission.

To analyze either form of multicarrier signal, we denote the symbol rate of the original data sequence by f_s, where $T_s = 1/f_s$ is the original symbol period. After serial-to-parallel conversion, there are N parallel data sequences, each with symbol rate f_s/N and symbol period $T = NT_s$. Thus, each subchannel is tolerant of N times as much time dispersion as would be the original datastream.

Now assume that, in a given symbol period $[0, T]$, the N subchannels carry data values $D_0, D_1, \ldots, D_k, \ldots, D_{N-1}$. We assume that D_k is two-dimensional, that is, $D_k = A_k + jB_k$, where A_k and B_k are real numbers representing the in-phase and quadrature data components, respectively. The set of discrete values possible for each component depends solely on the chosen data constellation, for example, $A_k = \pm 1$ and $B_k = \pm 1$ for four-level

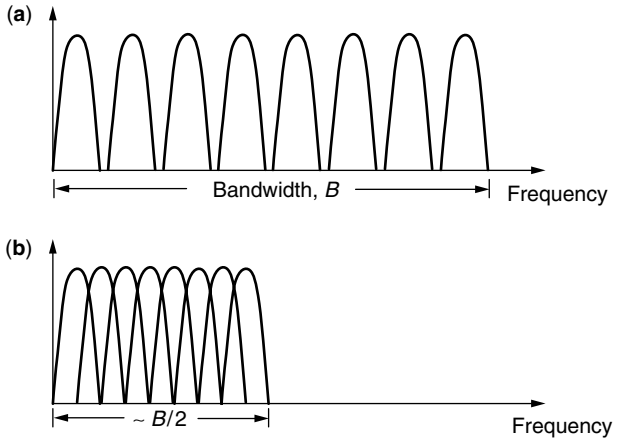


Figure 2. Transmit spectra of multicarrier and OFDM signals: (**a**) multicarrier spectrum; (**b**) OFDM spectrum.

quadrature amplitude modulation (4-QAM), also called *quadrature phase shift keying* (QPSK). Finally, assume initially that the data values are carried by rectangular pulses, that is, that the kth subchannel data value is carried by a pulse that is 1 on $[0, T]$ and 0 elsewhere. Then, the multicarrier signal transmitted on the given symbol interval can be represented as

$$s(t) = \mathrm{Re}\left\{\sum_{k=0}^{N-1} D_k e^{j\omega_k t}\right\}, \quad 0 \le t \le T \tag{1}$$

$$= \sum_{k=0}^{N-1} [A_k \cos \omega_k t - B_k \sin \omega_k t], \quad 0 \le t \le T \tag{2}$$

where the subcarrier radian frequency is $\omega_k = 2\pi f_k$, with $f_k = f_0 + k\Delta f$. The offset frequency, f_0, could represent the carrier frequency in a passband transmission system, such as one using a wireless channel, or could be adjusted for baseband transmission. Also, for baseband transmission, the data could be chosen in a symmetric fashion to guarantee a real output. This latter situation is discussed in Section 2.4. The parameter Δf represents the subcarrier spacing, which we discuss next.

The structure in Fig. 3 represents a general form of a multicarrier transmitter. For OFDM, the subchannels are permitted to spectrally overlap. To enable separation of these channels at the receiver, the data pulses for every pair of subchannels must be mutually orthogonal. For rectangular pulses, this can be achieved by relating the subcarrier spacing and the symbol duration via $\Delta f = 1/T$. Under these conditions, a simple correlation for each subchannel (i.e., multiplication by the appropriate waveform followed by integration over the symbol period) can separate out the subchannels. This receiver structure is shown in Fig. 4.

The power spectral density of the transmitted OFDM signal is the sum of the power spectral densities of N separate QAM signals at N subcarrier frequencies separated by the signaling rate. For rectangular symbol pulses, the Fourier transform of the symbol in each subchannel is a shifted version of $\sin x/x = \mathrm{sinc}(x)$, with nulls at the centers of the other subchannels. These and other spectral properties of an OFDM signal with rectangular symbol pulses are illustrated in Fig. 5. The Fourier transform of a single pulse in a single subchannel is shown in Fig. 5a; a set of Fourier transforms corresponding to eight subchannels ($N = 8$) is shown in Fig. 5b, and the OFDM power spectral density is shown in Fig. 5c for $N = 64$ and 256. (For convenience, we display the out-of-band portion in each case as the *envelope* of the actual lobe structure.) For large N, the total power spectral density is essentially flat in the bandwidth containing the subcarriers, and only the subchannels near the band edge contribute to the out-of-band power. Therefore, as the number of subcarriers becomes large, the spectral compactness approaches that of single-carrier modulation with rectangular bandpass filtering.

We now express the above ideas mathematically. Using the complex envelope of the transmitted signal, a single OFDM symbol can be represented as

$$S(t) = \sum_{k=0}^{N-1} D_k e^{j\omega_k t} \times \mathrm{rect}\left(\frac{t}{T}\right) \tag{3}$$

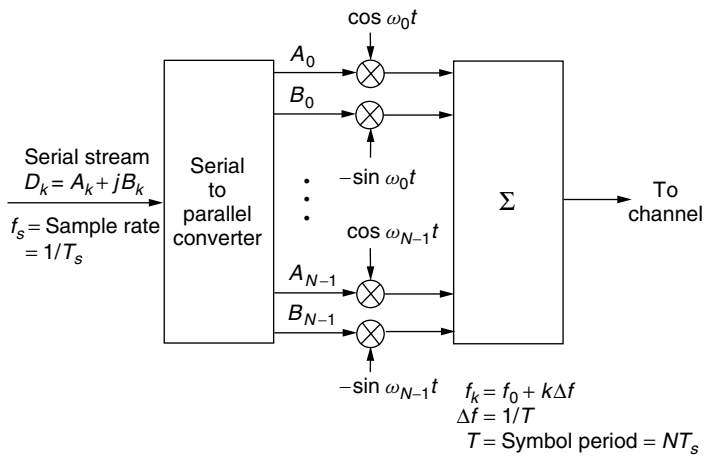


Figure 3. OFDM transmitter.

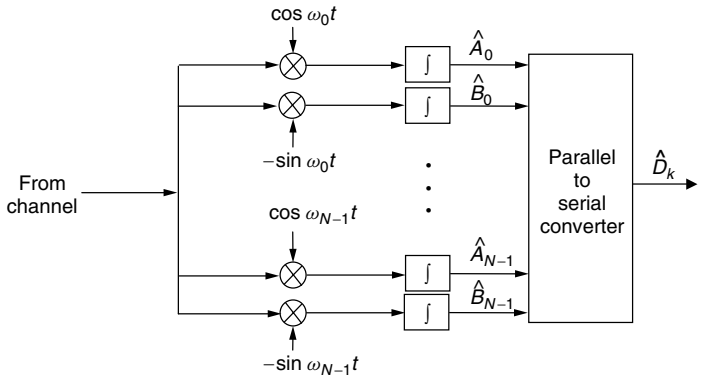


Figure 4. OFDM receiver.

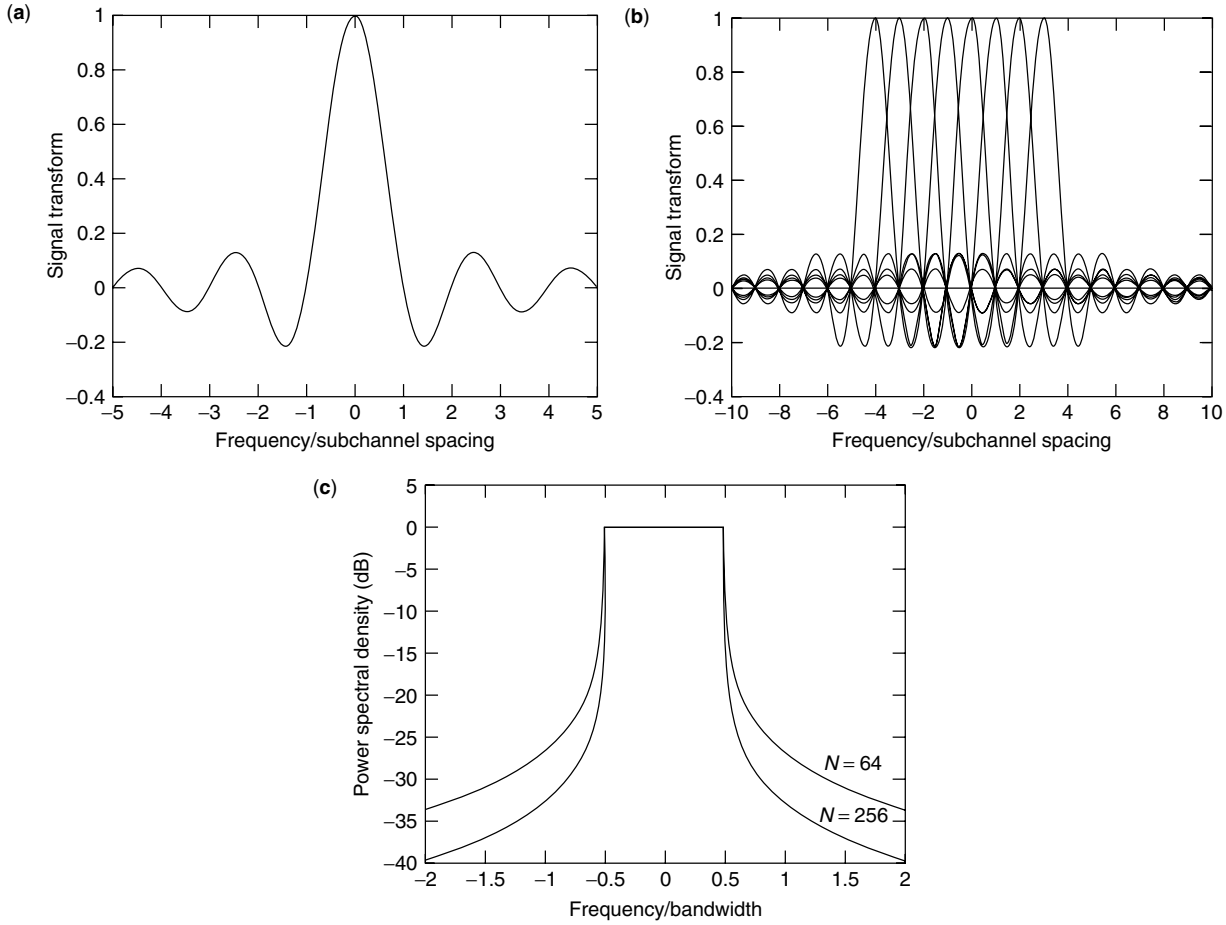


Figure 5. Spectral properties of an OFDM signal: (**a**) single channel; (**b**) eight individual subchannels; (**c**) Spectral power density (referred to center frequency).

where rect (x) is defined as having the value 1 on [0,1] and 0 elsewhere. The Fourier transform of $S(t)$ is thus given by

$$P(f) = T \sum_{k=0}^{N-1} D_k \frac{\sin \pi \left(\dfrac{f - f_k}{\Delta f} \right)}{\pi \left(\dfrac{f - f_k}{\Delta f} \right)} \tag{4}$$

If the data symbols are mutually independent (both among symbols and among subchannels), the power spectral density of the OFDM signal is $\overline{|P(f)|^2}/T$, where the overbar denotes averaging over the data. This formulation was used to obtain the results in Fig. 5c, where $f = 0$ corresponds to the center frequency of the OFDM spectrum. Note that the sharpness of the spectral falloff outside the main bandwidth increases with N.

We now show the orthogonality of the N transmitted pulses. Assuming, at this point, a perfect and noiseless channel, we can also regard $S(t)$ in Eq. (3) as the received signal. To detect the kth data value, D_k, $S(t)$ is multiplied by $e^{-j\omega_k t}$ and then integrated over $[0, T]$. The received data

symbols at the output of the kth correlator are

$$\hat{D}_k = \int_0^T S(t)e^{-j\omega_k t}\,dt = \sum_{l=0}^{N-1} D_l \int_0^T e^{j2\pi\Delta f(l-k)t}\,dt \tag{5}$$

For the case $\Delta f = 1/T$, it is easily shown that

$$\int_0^T e^{j2\pi\Delta f(l-k)t}dt = \delta(l-k) \tag{6}$$

where we use the Kronecker delta function, $\delta(l-k) = 1$ when $l = k$ and 0 otherwise. Therefore, $\hat{D}_k = D_k$, and, so, even though the subchannels overlap, they can be separated at the receiver with no interference among subchannels; that is, the subchannels are orthogonal.

We note that deriving the multicarrier transmitted signal from the data sequence, Eq. (5), and detecting that sequence from the received signal, Eq. (6), involves operations that resemble Fourier transforms. We will show more formally that the orthogonality that arises from setting $\Delta f = 1/T$ allows the use of the discrete Fourier transform (DFT) at both ends and thus the use of very efficient digital signal processing [12]. The combination of orthogonal pulses and efficient DFT processing constitutes the essence of OFDM. There are, of course, many details. The most important of these have to do with practical impairments in the transmission medium (notably, time dispersion and time variations) and in the system hardware (notably, frequency and timing errors in the receiver and amplifier nonlinearities in the transmitter). Discussions of basic implementation, channel and system impairments, and their remedies occupy most of the remaining sections.

2.3. DFT Implementation

We show here how the DFT and the inverse DFT (IDFT) can be used to implement OFDM. In most cases, these transforms can be done very efficiently by using the fast Fourier transform (FFT) algorithm. In this discussion, the number of subchannels and the FFT size are the same, N. (Later, we show why the FFT size is generally greater.) If N is a power of 2, the number of operations is on the order of $N\log_2 N$, as opposed to N^2 for conventional DFTs, leading to substantial savings for large N. For example, the number of FFT operations for $N = 1024$ is about 10^4, as opposed to about 10^6 with conventional processing, for a reduction of 100 to 1. Thus, completely digital implementations can be built around special-purpose hardware performing the FFT and its inverse (IFFT), replacing the banks of oscillators, mixers, and filters shown in Fig. 1.

Consider a discrete-time version of the complex envelope of the transmitted OFDM symbol in Eq. (3). Assuming $f_0 = 0$, without loss of generality, and sampling at times $t_n = nT_s$, Eq. (3) becomes

$$S_n = \sum_{k=0}^{N-1} D_k e^{j2\pi k\Delta f n T_s}, \quad 0 \le n \le N-1 \tag{7}$$

With the imposed orthogonality condition, $\Delta f = 1/T = 1/NT_s$, this becomes

$$S_n = \sum_{k=0}^{N-1} D_k e^{j2\pi kn/N}, \quad 0 \le n \le N-1 \tag{8}$$

which is simply the IDFT of the input data sequence, $D_0, D_1, \ldots, D_{N-1}$. With N suitably chosen, the transmitted signal samples can then be generated using the efficient IFFT algorithm.

The receiver correlation operations can also be performed in this fashion. Specifically, suppose that the block of received signal samples, $\{S_n\}$, is transformed in the receiver using a DFT. This yields

$$\begin{aligned}\hat{D}_k &= \frac{1}{N}\sum_{n=0}^{N-1} S_n e^{-j2\pi kn/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} D_l e^{j2\pi n(l-k)/N} \\ &= \frac{1}{N}\sum_{l=0}^{N-1} D_l \sum_{n=0}^{N-1} e^{j2\pi n(l-k)/N} \\ &= \frac{1}{N}\sum_{l=0}^{N-1} D_l N\delta(l-k) \\ &= D_k\end{aligned}$$

Thus, the correlation operations can also be efficiently implemented using the FFT algorithm.

We should mention at this point that there are several alternative forms of OFDM, that is, orthogonality can be achieved in various ways [13–15]. In particular, several of the early forms of OFDM were based on band-limited signaling, using specially designed pulses or special signaling patterns to guarantee orthogonality. Nevertheless, the form of OFDM described here is the most popular and the one proposed or implemented for all OFDM-based standards.

Finally, it is important to note how the sequence of IDFT samples in the transmitter, $\{S_n\}$, is converted into a continuous analog signal for transmission over the medium. The N samples, spaced in time by $1/N\Delta f = T_s$, are passed through a digital-to-analog converter (DAC) and then applied to a band-limiting filter. The spectrum of a discrete-time waveform such as the S_n sequence is periodic, with period $N\Delta f$. The purpose of the band-limiting filter is to pass one such period (the *primary spectrum*) and suppress all others. The burden on this filter can be severe if the primary spectrum has significant content at the band edges, as is generally the case (Fig. 5). This situation is avoided, and the filtering problem eased, by "padding" the original block of data values, $\{D_k\}$, with zeros before and/or after the actual data values; that is, zero-valued subcarriers are added to the data-carrying subcarriers. Thus, the band-limiting filter does not require as sharp a cutoff characteristic. This padding creates a difference between the number (N) of data-carrying subcarriers and the total number (N') of subcarriers

processed by the FFT. The FFT size, N', can readily be chosen to be a power of 2; we will assume hereafter that this is the case, so that the central transmitter and receiver processings are IFFTs and FFTs, respectively. Note that the bandwidth to be transmitted is still $N\Delta f$, but the samples in time, S_n, are now spaced by $1/N'\Delta f$ (oversampling) and the spectrum period is $N'\Delta f > N\Delta f$.

2.4. Baseband versus Passband Representations

For a given block of data values $\{D_k\}$, the complex envelope of the OFDM signal is $S(t)$ in Eq. (5). For passband transmissions (as, e.g., in wireless applications), a stage of modulation is needed to convert $S(t)$ to a real passband signal. For baseband transmission (as in wireline applications like DSL), no modulation is needed but $S(t)$ must be a real baseband signal. This can be done by converting the N-symbol stream $\{D_k\}$ to a stream of $2N$ symbols according to the following rule:

$$D'_k = \begin{cases} \mathrm{Re}(D_0) & k = 0 \\ D_k & k = 1, 2, \ldots, N-1 \\ \mathrm{Im}(D_0) & k = N \\ D^*_{2N-k} & k = N+1, \ldots, 2N-1 \end{cases} \tag{9}$$

where $*$ denotes complex conjugate. It is easy to show that a DFT or an IDFT applied to this sequence will produce a sequence of real numbers. This requires that D'_0 and D'_N be real, so these symbols are used in the above rule to carry the real and imaginary parts of D_0. Note that all the data are contained in the first $N+1$ terms (D'_0 through D'_N). The rest are used to ensure a real baseband signal at the IDFT output in the transmitter.

Since the input to the IDFT in the baseband case has twice as many samples, so must the output. For the same OFDM symbol length, T, this means the samples of $S(t)$ are now spaced by $T/2N$ (not T/N), thus requiring a baseband bandwidth of about $2N/T$ Hz. This is the same as the bandwidth required for the passband case.

2.5. Guard Interval Considerations

Even with a large symbol duration, channel time dispersion will cause consecutive symbols (also called *OFDM blocks*) to overlap, resulting in some residual ISI that could degrade performance. This residual ISI can be eliminated, at the expense of spectral efficiency, by using guard time intervals, between OFDM symbols, that are at least as long as the maximum extent of the channel impulse response. Samples of the received signal lying in the guard interval are discarded in the receiver and the demodulated OFDM symbol is generated from the remaining N samples.

The guard interval could be filled at the transmitter with null signal samples (zeros). However, in the case where there is dispersion, the receiver FFT processing will truncate the spread signal, so that each detected data value, $\hat{D}_k$, will consist of D_k plus interchannel interference (ICI) from the other data values.[1] In particular, if the signal has length NT_s and the impulse response of the channel is of length LT_s, the signal at the output of the channel is the linear convolution of the channel and the transmitted signal and is therefore of length $(N+L)T_s$. In the receiver, however, the FFT processes only N samples; this is the truncation that causes ICI. Putting it an alternative way, an FFT preserves orthogonality between tones only when the convolution in time is a *cyclic* convolution, rather than the linear convolution that occurs in a real channel.

One widely used solution to this problem is to cyclically extend the OFDM block by an amount longer than the expected time extent of the channel impulse response [16]. Specifically, to create a periodic received signal for the FFT to process (and thereby eliminate ICI), M' time samples are copied from the end of the original OFDM sequence and appended as a prefix; and, M'' time samples are copied from the beginning of the original OFDM sequence and appended as a suffix, where $M = (M' + M'') \geq L$. In some systems only a prefix is used ($M'' = 0$) and the processing window position is adjusted accordingly. An example is shown in Fig. 6. At the receiver, the samples of the cyclic extension are discarded before FFT processing. Clearly, the need for a cyclic extension in time-dispersive environments reduces the efficiency of OFDM transmissions by a factor of $N/(N+M)$. In most OFDM designs, a guard interval of not more than 10% to 20% of the symbol duration is employed.

2.6. Windowing

In some OFDM applications, compactness of the transmitted spectrum is important. A case in point is wireless systems, where spectrum is precious and multiple systems are closely spaced in frequency. The sharpness with which the signal spectrum falls off outside the allocated bandwidth is then of great interest.

In the OFDM systems described so far, a rectangular symbol pulse has been assumed. In other words, all samples of the IFFT output and the cyclic extension (if used)

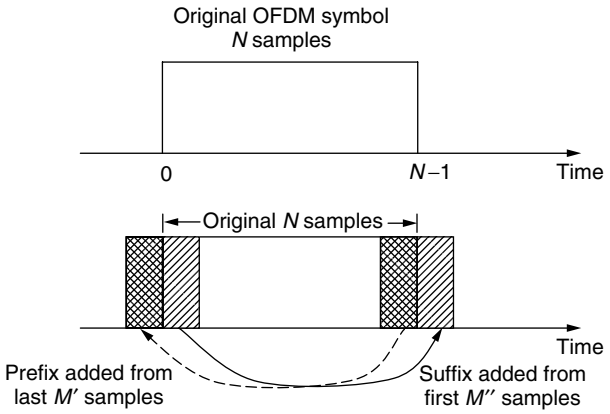


Figure 6. Cyclic extension.

between symbols in the same subchannel. Other causes of ICI, besides the one cited above, are channel time variations, frequency offset, and phase noise (Section 3).

[1] *ICI* refers, generally, to the interference between subchannels in the same symbol period. By contrast, *ISI* refers to interference

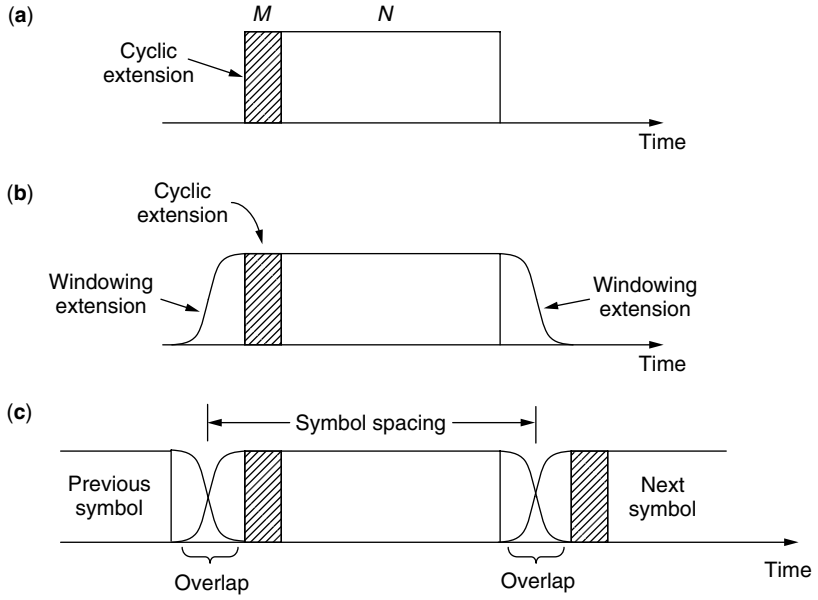


Figure 7. Windowing: (**a**) OFDM symbol; (**b**) windowed OFDM symbol; (**c**) sequence of windowed OFDM symbols.

are unweighted, which corresponds to having a rectangular symbol pulse, at each tone, of length $(N + M)T_s$. This is depicted in Fig. 7a. The spectral properties of the rectangular pulse shape (high sidelobes that decay slowly) lead to poor out-of-band spectral falloff [see Fig. 5c for $M = 0$, $N \leq 64$ or $(M + N) \leq 64$].

A simple way to improve the spectrum is to increase the periodic extension of $\{S_n\}$ even further and to taper the additional extension. This is called *windowing*. An example is shown in Fig. 7b. A commonly used shape is the cosine rolloff function. Although the total symbol duration is thus enlarged, the symbol *spacing* can be smaller than this duration, because adjacent symbols can overlap in the (unprocessed) rolloff region. This is shown in Fig. 7c.

To see the improvement possible with even a small extension, note the power spectral density plots of Fig. 8 for $N = 64$ ($M = 0$ for these computations). The parameter in the plots is the cosine rolloff factor, β, and the fractional

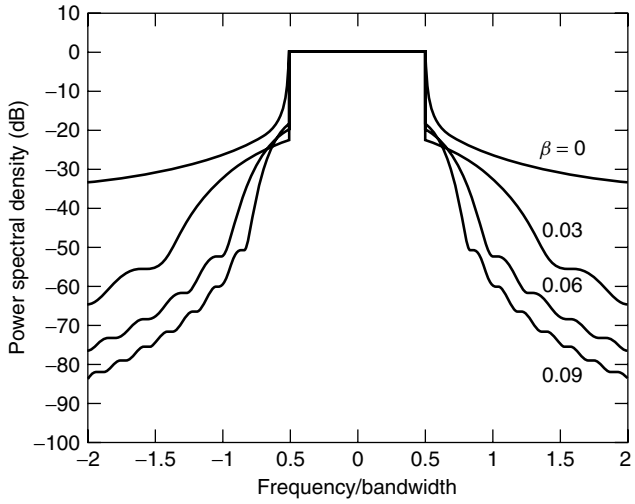


Figure 8. Power spectral density for windowed OFDM signal.

increase in symbol spacing can be shown to be $\beta/(1 - \beta)$. Thus, the curves show that an increase of only 3% in symbol spacing can produce dramatic benefits in out-of-band spectral falloff.

2.7. Coding

Because of the frequency-selective nature of the typical wideband channel (which is the main motivation for using OFDM), the OFDM subchannels generally have different received powers. Variations in the channel gain with frequency may cause some groups of received subcarriers to be much weaker than others, or even completely lost. Therefore, even though most subcarriers may be detected without errors, overall performance will be dominated by the performance of the few subcarriers with the lowest SNR (signal-to-noise ratio). As a result, satisfactory performance cannot be achieved without the addition of some form of error correction coding. By using coding across the subcarriers, errors in weak subcarriers can be corrected, up to a limit that depends on the code and the channel. Coding in OFDM systems has an additional dimension: it can be implemented in both the time and frequency domains, so that both dimensions can be utilized to achieve immunity against channel variations.

2.8. A Sample Design

The processing steps discussed above are all reflected in the simplified block diagrams of Fig. 9. We will present a

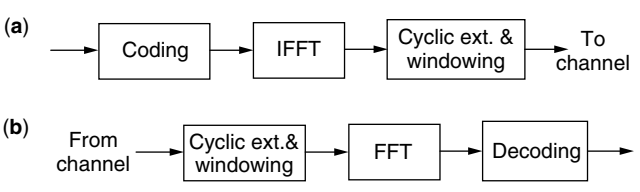


Figure 9. OFDM transmitter and receiver: (**a**) OFDM transmitter processing; (**b**) OFDM receiver processing.

sample design here, highlighting the factors influencing the choices of key parameters. It will be seen that OFDM system design involves tradeoffs among various, often conflicting, requirements. For example, to minimize the effects of time dispersion, a long symbol duration is required, meaning a large number of subchannels. However, if the channel is time-varying, as in a mobile radio environment, the variations during a long symbol period could be significant, causing possible ICI. The design parameters of interest are (1) the number of subcarriers, N; (2) the size of the FFTs, N'; (3) the guard time, T_g; the OFDM symbol duration, T; and (4) the subcarrier spacing, Δf. These are influenced by the assigned bandwidth, the desired bit rate, the time extent of the channel impulse response, and the rate of the channel time variations.

As an example, consider a wireless system that requires a bit rate of 1.2 Mbps (megabits per second) in a bandwidth of 800 kHz. Assume that the system must operate in an environment with a channel delay span of 20 μs, corresponding to wide-area transmission. A guard time, T_g, of 40 μs should be more than enough to guarantee that there is no ISI. (In this example, the guard time is assumed to be sufficient to handle the channel dispersion as well as any additional extension for windowing.) The OFDM symbol duration, T, is then chosen large enough to ensure that the efficiency loss due to the guard interval is small and to guarantee that the subchannel bandwidth is narrow enough to suffer only flat fading. In this case, we consider an OFDM symbol interval $T' = T + T_g = 200$ μs. This is five times the size of the guard interval, resulting in a 20% guard time overhead. The subcarrier spacing is then $\Delta f = 1/T = 6.25$ kHz. At a carrier frequency of 2 GHz and assuming a vehicle speed of no more than 100 km/h, the maximum Doppler spread is about 200 Hz, which is small enough compared to 6.25 kHz that ICI should be acceptably small. This choice of spacing allows for at most 128 subchannels in the 800 kHz bandwidth. Assuming QPSK modulation (i.e., 2 bits per symbol) and four guard subchannels at either end of the OFDM spectrum (to facilitate filtering), the resulting bit rate is

$$R_b = \frac{120 \text{ data subchannels} \times 2 \text{ bits per subchannel}}{200\ \mu s} = 1.2 \text{ Mbps}$$

With a half-rate code, this results in an information rate of 600 kbps. Finally, as discussed in Section 2.3, zero-valued subcarriers can be added to the data set $\{D_k\}$, to facilitate transmit filtering, so that the FFT size, N', is greater than the number of subcarriers, N. A typical choice for this design example might be $N' = 4N = 512$.

3. CHANNEL AND SYSTEM IMPAIRMENTS

3.1. Introduction

Noise and the channel frequency (or impulse) response largely determine the performance of an OFDM system. In addition, several phenomena can significantly degrade the performance. Time variations in the channel, as well as frequency offset, phase noise, and timing errors, can impair the orthogonality of the subchannels [17]. Also, the large amplitude fluctuations characteristic of a multicarrier signal can be a serious problem when transmitting through a nonlinearity, such as the transmit power amplifier. We discuss all these issues here.

3.2. Noise and Interference

The ultimate limit on system performance, even without other impairments, is the combination of thermal noise and interference. In the case of OFDM, we can assume that there are N independent subchannels, each with its own signal-to-interference-plus-noise ratio (SINR). The usual methods of analysis can be used to compute the performance (bit error rate, block error rate, etc.) of each subchannel as a function of SINR. Typical approaches for maximizing the performance of a given subchannel are power control and coding, as in other systems. The difference in OFDM is that power control and coding can be applied across subchannels as well as within subchannels.

In the case of an additive white Gaussian noise (AWGN) channel (no frequency or time selectivity), all subchannels have the same performance. Moreover, the total system performance is identical to that of a single-carrier system having the same modulation, bandwidth, and power.

3.3. Channel Time Dispersion

Channel time dispersion can produce deep fades at one or more subchannel frequencies, causing performance degradation. However, the problems of ISI and ICI due to dispersion can be avoided using guard times and a cyclic extension (see Section 2.5). To put this mathematically, let the channel impulse response be expressed in discrete-time form by the finite set $\{h_l, 0 \le l \le L\}$, where T_s is the spacing between samples and LT_s is the maximum delay. The channel response at the subcarrier frequency $f_k = k/N$ is

$$H_k = H\left(\frac{2\pi k}{N}\right) = \sum_{l=0}^{L} h_l e^{-j2\pi lk/N} \tag{10}$$

Assuming that the channel is time-invariant, each h_l is constant. Given suitable choices for the length of the OFDM symbol and the guard time, and the use of a cyclic extension to avoid ICI, the demodulated sequence may be expressed as

$$X_k = H_k D_k + \eta_k \tag{11}$$

where η_k is additive Gaussian noise in the kth subchannel. Note that the noise components for different subcarriers are generally uncorrelated, that is, $E[\eta_k \eta_l^*] = \sigma_k^2 \delta(k - l)$.

If the communication channel is time-invariant, its effect on each subchannel is seen to be represented by a single complex-valued coefficient. Therefore, correcting for the channel response can be accomplished by following the receiver FFT with a single complex gain adjustment at each subcarrier frequency. Estimation to correct for the channel is discussed in Section 4, along with the possibility of matching (adapting) the transmitted signal to the channel frequency response.

3.4. Channel Time Variations

Channel and system time variations over a symbol result in spectral spreading of the individual subchannels, which causes ICI. We now show this analytically for one type of variation. Specifically, assume that the composite effect of the channel and system time variations can be represented as a multiplicative complex factor, so that the received signal's complex envelope is $\gamma(t)S(t)$. The factor $\gamma(t)$ could represent a frequency-independent gain variation, as might be encountered in a narrowband mobile radio channel. Let the nth received sample be $R_n = \gamma_n S_n$. Then, we find the kth data value at the receiver output is

$$\begin{aligned}\hat{D}_k &= \frac{1}{N}\sum_{n=0}^{N-1}\gamma_n S_n e^{-j2\pi kn/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1}\gamma_n D_l e^{j2\pi n(l-k)/N} \\ &= \sum_{l=0}^{N-1} D_l \frac{1}{N}\sum_{n=0}^{N-1}\gamma_n e^{j2\pi n(l-k)/N} \\ &= D_k\Gamma_0 + \sum_{\substack{l=0\\ l\neq k}}^{N-1} D_l\Gamma_{l-k} \end{aligned} \tag{12}$$

where the sequence $\{\Gamma_l\}$ is the DFT of the sequence $\{\gamma_n\}$. If $\gamma_n = 1$ for all n (a time-invariant channel), then $\Gamma_{l-k} = \delta(l-k)$ and $\hat{D}_k = D_k$; otherwise, there is ICI, namely, a complex-weighted average of the other data values. In addition, the desired signal term is attenuated and rotated via the complex factor Γ_0. In a channel that is both time-dispersive and time-varying, the mathematics is more complicated, but the basic concept is the same.

3.5. Frequency Offset

Before an OFDM receiver can demodulate subcarriers, it has to perform at least two synchronization tasks: (1) it must locate the symbol boundaries and derive the optimal timing instants, so as to minimize the effects of ICI and ISI; and (2) it must estimate and correct for carrier frequency errors due to frequency offset and phase noise. We discuss these in the next three subsections, starting with frequency offset. A number of techniques have been devised for estimating and correcting timing and carrier frequency errors at the OFDM receiver, and these are discussed in Refs. 1 and 3.

The usual source of frequency offset in OFDM is a static frequency recovery error in the receiver. To analyze the impact, we can use Eq. (12), where $\gamma(t)$ can now be modeled simply as $e^{j2\pi\delta ft}$, with δf representing the difference between the transmitter and receiver carrier frequencies. In this case, the received data symbol again suffers from ICI, as in Eq. (12), with

$$\Gamma_0 = \frac{\sin\pi\left(\dfrac{\delta f}{\Delta f}\right)}{\pi\left(\dfrac{\delta f}{\Delta f}\right)} e^{j\pi\delta f/\Delta f} \tag{13}$$

and

$$\Gamma_{l-k} = \frac{\sin\pi\left(l-k-\dfrac{\delta f}{\Delta f}\right)}{\pi\left(l-k-\dfrac{\delta f}{\Delta f}\right)} e^{j\pi[l-k-\delta f/\Delta f]} \tag{14}$$

If the frequency error is a multiple, I, of the subcarrier spacing, then the received subcarriers are shifted in frequency by $\delta f = I\Delta f$. The subcarriers remain orthogonal in this case (all still have an integer number of cycles within the FFT processing window), but the recovered data have the wrong index values. This can be seen from Eqs. (12)–(14); if $\delta f = I\Delta f$, with $I \neq 0$, then Γ_0 will be 0 and so will every Γ_{l-k} except for $l = k + I$. Thus, the detected data for the kth subchannel will be $\hat{D}_k = D_{k+I}$, meaning that all data values are detected but are associated with the wrong subchannels. In general, *all* offsets of magnitude $\Delta f/2$ or more will lead to subchannel ambiguity, where the strongest component of $\hat{D}_k$ is that of a subchannel other than the kth. The first task of receiver frequency correction, then, is a coarse acquisition that brings δf within the range $\pm\Delta f/2$.

Assuming that δf lies within this range following initial acquisition, the number of cycles within the processing window will be a noninteger for all subchannels, and ICI will result, [Eq. (14)]. (This is analogous to the ISI in a single-carrier system caused by timing offset.) Also, the desired component will be reduced in magnitude by a factor $\text{sinc}(\delta f/\Delta f)$, as given by Eq. (13).

3.6. Phase Noise

A problem related to frequency offset is phase noise: a practical oscillator does not produce a carrier at exactly one frequency, but rather a carrier that is phase modulated by random noise. As a result, the receiver's recovered frequency, which is the time derivative of its phase, is never perfectly constant. Thus, phase noise produces a *dynamic* frequency error, whereas frequency offset is a static one. The result, in both cases, is ICI. The problem is more serious in OFDM than in a single-carrier system because the subchannels are so close in frequency and, in addition, their spectra overlap.

Although OFDM is more susceptible to phase noise and frequency offset than are single-carrier systems, there are techniques for keeping this degradation to a minimum. First, phase noise in the local oscillator is common to all subcarriers. If the oscillator linewidth (i.e., the spread of the oscillator tone due to phase noise) is much smaller than the OFDM symbol rate, which is usually the case, the common phase error is strongly correlated from symbol to symbol and from tone to tone; thus, tracking or differential detection (Section 4.2) can be used to minimize its effects. Second, the impact of phase noise grows monotonically with the ratio of the linewidth to the subcarrier spacing. Therefore, control of this ratio in choosing oscillators and subcarrier spacings can control the ICI.

3.7. Timing Errors

To achieve time synchronization (as well as frequency synchronization) with a minimum of processing at the

receiver, and also a minimum of redundant information added to the data signal, the synchronization process is normally split into an acquisition phase and a tracking phase. This is possible if the general characteristics of the timing (and frequency) errors are known. In the acquisition phase, an initial estimate of the errors is acquired, perhaps using more complex algorithms and more overhead; then, the follow-on tracking algorithms only have to correct for small short-term deviations.

With respect to timing offsets, OFDM is relatively robust; in fact, the symbol timing offset may vary over an interval equal to the guard time without causing ISI or ICI. This is because, for timing offsets smaller than the guard interval, the impact is just a phase shift; that is, for a timing offset δt, the received sample for the kth subcarrier is

$$\hat{D}_k = D_k e^{j2\pi f_k \delta t} \tag{15}$$

Thus, no ICI results; just a phase error which grows with f_k. If differential detection between tones is used, the impact of the timing error can be controlled by just ensuring that the root mean square (RMS) value of $\delta t/T$ is sufficiently small, say, 0.01 or less. The precise requirement depends on the modulation, the target bit error rate, and other such factors. Of course, if δt exceeds the guard time, the receiver's FFT window spans samples from two consecutive OFDM symbols and ISI occurs.

3.8. Transmitter Nonlinearities

An OFDM signal is the superposition of many modulated subcarrier signals and thus may exhibit a high signal peak relative to the average signal level. If the transmitter processing is not linear over the full range of the signal variation, nonlinear distortion will occur. This is manifested in two ways: (1) in-band intermodulation products, causing interference to each subchannel; and (2) out-of-band spectral spreading, potentially causing adjacent-channel interference (ACI) to other systems. Avoiding these problems requires a degree of transmitter linearity that can be costly.

One possible metric for characterizing signal peaking is the ratio of the peak signal power to the average signal power, or the *peak-to-average power ratio* (PAPR). This quantity can be taken over an OFDM symbol, in which case it varies from symbol to symbol, or over all time, in which case it is a single number. Either way, this metric must be used with care. The most extreme peaking occurs when all of the subcarrier signals line up in their peak amplitudes at the same time instant. It is easy to show that, for N subcarriers having equal average powers and using BPSK or QPSK (binary and quadrature phase shift keying) modulation, the PAPR defined as above (and taken over all time) is N. Thus, the PAPR would be 12 dB for $N = 16$ and 21 dB for $N = 128$. It may therefore seem that signal peaking progressively worse as N increases. However, worst-case signal peaking becomes less probable as N increases, so it is necessary to look at peaking in a *statistical* way. For N sufficiently large, the complex envelope converges to a complex Gaussian process, meaning that the squared envelope approximates an exponential variate. This approximation is used in Fig. 10, which shows the complementary cumulative distribution function for the ratio of the instantaneous power to the average power, taken over all time. If we redefine PAPR as the value not exceeded more than 0.001% of the time, the proper value to use is about 10.6 dB. This result, which holds for all realistic N and all modulations, is large enough to raise concerns about transmitter nonlinearities.

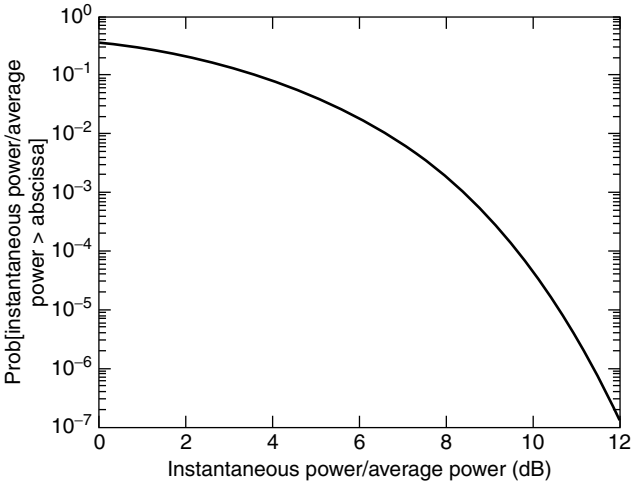


Figure 10. Complementary cumulative distribution function of PAPR of an OFDM signal.

To transmit signal peaks without major distortion, the transmitter's DAC must use a sufficient number of bits to accommodate these peaks, which is a cost/technology issue. More importantly, the power amplifier must remain linear over an amplitude range that includes the peaks, which leads to both high amplifier cost and high power consumption (low power efficiency). Several techniques have been proposed to mitigate the peaking problem, and they divide basically into three categories: (1) *signal distortion* techniques, which reduce the peak amplitudes by nonlinearly distorting the OFDM signal at or around the peaks. (e.g., clipping and filtering, peak windowing, peak cancellation); (2) *coding* techniques, involving special codes that exclude OFDM symbols with high peaking; and (3) *scrambling* techniques, that is, scrambling each OFDM symbol with different sequences and selecting the one that gives the least peaking. Details and the relative performances of these techniques can be found in Refs. 1, 2, 18, and 19.

4. OTHER MAJOR ISSUES

4.1. Introduction

We have seen that, to get the most value out of OFDM, special techniques have been devised such as cyclic extension and windowing. These relate primarily to how the signal is prepared at the transmitter to be sent over the channel. Equally important are methods of data detection and channel estimation at the receiving end and methods for adapting both transmission and reception to the frequency selectivity of the channel so as to maximize data efficiency. We explore these topics here.

4.2. Detection Techniques

In general, the data constellation of each subcarrier will show a random phase shift and amplitude change. These are caused by carrier frequency offset, timing recovery offset, and the frequency selectivity of the channel, as discussed in the previous section. To cope with these unknown changes, two classes of detection techniques exist. The first is *coherent detection*, using estimates of the channel response to derive reference values for the amplitude and phase correction for each subchannel. Spectrally efficient use of this approach requires reliable techniques for channel estimation that, at the same time, do not require excessive overhead, as discussed in the next section.

The second technique is *differential detection*, which does not require absolute reference values but accounts only for the phase and/or amplitude differences between two data symbols. In OFDM, differential detection can be done in the time domain or in the frequency domain. In the first case, each subcarrier is compared with the same subcarrier of the previous OFDM symbol; in the second case, each subcarrier is compared with the adjacent subcarrier within the same OFDM symbol. In contrast to coherent detection, differential detection does not require channel estimation, thereby saving complexity and gaining overhead efficiency. The cost is a degraded performance because of the noisy references that are effectively being used.

If differential detection is used within each subchannel, symbols must be highly correlated in time; performance can thus degrade if the channel response has significant time variation. Similarly, if differential detection is done between subchannels, symbols must be highly correlated in frequency; performance can thus degrade if the channel response has significant frequency variation.

4.3. Channel Estimation and Correction for Coherent Detection

In the kth OFDM subchannel, the data component appears at the detector input with a complex amplitude scaling, H_k, plus Gaussian noise, η_k, as in Eq. (11). Coherent detection of this sample amounts to comparing it against all points in the data constellation and choosing the point closest to it. To do this optimally, it is necessary to undo the amplitude scaling $|H_k|$ and the phase rotation $\mathrm{Arg}(H_k)$. Doing this individually for all frequency components of the FFT output is called *frequency-domain equalization*; broadly speaking, it consists of scaling each subchannel with a complex multiplier $1/\hat{H}_k$, where $\hat{H}_k$ is an approximation to H_k.

A conventional approach to implementing this equalization is to initially estimate the subchannel gains (e.g., by transmitting a known modulated sequence in each subchannel) and to then handle time variations via either periodic updates or decision-directed tracking. An alternative approach, ideally suited to OFDM, is pilot-aided estimation. Pilots are unmodulated tones, lasting for one or more symbols at a time, that are inserted by the transmitter and processed by the receiver to estimate channel gains. They can be distributed in time and frequency in any number of ways, one example of which is shown in Fig. 11. The two competing goals in placing pilots are that they should occupy a small fraction of the time–frequency slots, and their frequency of occurrence in each direction should be high enough to adequately sample the channel.

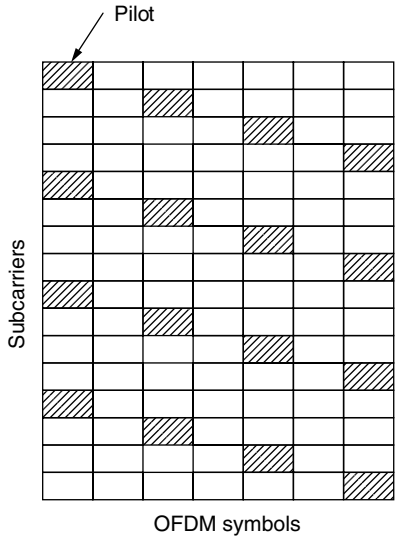


Figure 11. Typical pilot pattern.

Pilots are used for channel correction as follows. First, the receiver estimates channel gains at all times and frequencies where pilots have been transmitted. Then the channel gains at all other time–frequency positions can be estimated using two-dimensional interpolation filtering. Equalization then consists of setting the scaling to $1/\hat{H}_k$ for each data-carrying subchannel, where $\hat{H}_k$ is the gain estimate or some modification to account for additive noise.

To accurately interpolate the channel estimates from the available pilots, the pilot spacing in each dimension must satisfy the Nyquist sampling theorem. This means that there exist both a minimum necessary subcarrier spacing and a minimum necessary symbol spacing between pilots. To determine these spacings, two quantities must be known or estimated, namely, the double-sided bandwidth, $B_{\max}$, of the channel gain's time variations; and the full time extent, $\tau_{\max}$, of the channel's impulse response. The requirements for the pilot spacings in time and frequency, Δt_p and Δf_p, are then $\Delta t_p < 1/B_{\max}$ and $\Delta f_p < 1/\tau_{\max}$. In order to get a degree of noise reduction by filtering, the pilot symbol spacing should be smaller than half these values (oversampling) but not so small that the fraction of pilots is excessive.

Many solutions based on both pilot-aided estimation and decision-directed scaling are described in the references (e.g., see Refs. 1 and 3). The proper choice among pilots, training sequences, and decision-directed tracking, and the "best" design of whichever methods are used, depend on such factors as channel variability, type of traffic (e.g., continuous voice, packet data), and performance and cost objectives. For example, in the case of high-speed packet transmission to low-mobility users, as in wireless LAN applications, the most appropriate approach seems to be the use of a preamble consisting of one or more known OFDM symbols. The choice of the number of training

symbols is a tradeoff between short training time (better spectral efficiency) and good estimation performance.

4.4. Adaptive Loading

The frequency selectivity of an OFDM channel provides both challenges and opportunities. One way to address the problem of weak subchannels is to code across tones (thereby exploiting the frequency selectivity that causes the problem), as noted in Section 2.7. Another is to adaptively turn off weak subchannels, that is, to send no data at frequencies where the received SNR is below some threshold. To better exploit frequency selectivity and realize a spectral efficiency benefit, the data constellation used in each subchannel can be adaptively sized to its SNR [20]. This process, called *adaptive loading*, recognizes the fact that, in media where some subchannels are weaker than average, others are stronger. Thus, for example, each subchannel could use QPSK, 16-QAM, or 64-QAM (equivalently, 2, 4, or 6 bits per symbol), depending on the frequency response (gain) for that subchannel. This optimum form of OFDM is used for DSL applications and is called *discrete multitone* (DMT) [21]. A sample variation of the channel frequency response across the OFDM subchannels is illustrated in Fig. 12.

For the case of fixed transmit power, P, in each subchannel, the SNR in the kth subchannel would be

$$SNR_k = \frac{P}{N_0/T}|H_k|^2 \tag{16}$$

where N_0 is the noise power density at the receiver input. Assume that SNR is accurately measured in the receiver for every subchannel and communicated to the transmitter over a feedback channel. For a specified bit error rate, each such measurement can be used to select a data constellation size; specifically, the bits per symbol for the kth subchannel can be matched to the subchannel gain, $|H_k|^2$. To be effective, this approach requires an accurate SNR measurement in the receiver, a reliable feedback channel to the transmitter, and a means for changing constellations at the transmitter and efficiently notifying the receiver. If the same carrier frequency is used for both transmit and receive, as in Time-Division Duplexing, the over-the-air feedback channels is not required.

An additional degree of freedom is power control, that is, adaptively changing the transmit power, P_k, in the kth subchannel in accordance with its gain, subject to a total power constraint. If the power and constellation size are jointly distributed among subchannels in the most optimal way, the overall spectral efficiency of OFDM matches that of a single-carrier system with ideal decision feedback equalization [22].

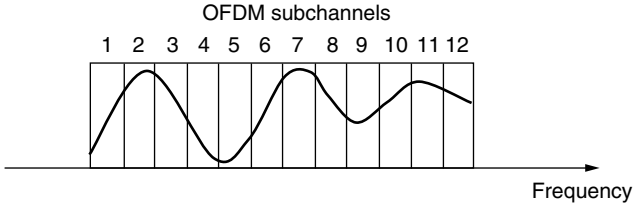


Figure 12. A channel response that justifies adaptive loading.

5. CONCLUSION

OFDM is a very attractive technique for high-bit-rate data transmission over a dispersive communication channel. However, to realize an OFDM system, many practical issues must be addressed, including high signal peaks, frequency offset, timing mismatch, and channel variation. Promising solutions have been devised for all these problems, but most solutions contribute to a "nibbling away" of the spectral efficiency. Even so, OFDM can attain the same spectral efficiency as an equalized single-carrier system and has the virtue of flexibility and a processing complexity that grows gracefully with channel dispersion.

Multicarrier modulation has been used in modems for both radio and telephone channels, as well as for digital audio and video broadcasting. The digital audio broadcasting (DAB) standard was, in fact, the first OFDM-based standard. The main reasons to choose OFDM for this system, which also applies to digital video broadcasting (DVB), are the possibility to provide a single frequency network and the efficient handling of multipath delay spread. A particularly suitable application of multicarrier modulation is in digital transmission over copper wire subscriber loops, as exemplified in high-speed digital subscriber loop (DSL) systems. In addition, the major high-bit-rate wireless LAN standards (IEEE 802.11a, HIPERLAN/2, and MMAC) use OFDM to overcome the bit-rate limitations caused by delay spread.

While much effort on OFDM is focused on critical implementation issues, there is also research on new variations and applications. Examples include *orthogonal frequency-division multiple access* (OFDMA); OFDM combined with *code-division multiple access* (CDMA); and OFDM combined with *multiple-input/multiple-output* (MIMO) antenna techniques. In OFDMA, multiple access is realized by providing each user with a fraction of the total number of subcarriers [23]. It is similar to conventional FDMA, except that it avoids the usual guard bands and exploits the power of FFT processing. OFDM-CDMA techniques (of which there are several variants) provide alternative ways to achieve multiple access while still combating frequency selectivity with moderate processing complexity [24]. OFDM-MIMO techniques exploit the power of array processing to increase wireless system capacity [25]. In all these applications, the need for time-domain equalization or RAKE reception is avoided because of the use by OFDM of narrow subchannels. Thus, powerful signal processing techniques like the FFT and adaptive arrays can be used instead to achieve high levels of performance.

BIOGRAPHIES

Leonard J. Cimini, Jr., received his B.S., M.S., and Ph.D. degrees in electrical engineering from the University

of Pennsylvania in 1978, 1979, and 1982, respectively. Over a 20-year AT&T career, starting at Bell Labs and then at AT&T Labs, he conducted research on lightwave and wireless communications systems. His main emphasis has been on devising techniques for overcoming the bit-rate limitations imposed by communications channels. In this context, he pioneered the application of Orthogonal Frequency Division Multiplexing to the emerging field of wireless communications. Dr. Cimini has been very active within the IEEE, including serving on several editorial boards and on the board of governors of the IEEE Communications Society. He was also the founding editor in chief of the IEEE J-SAC: Wireless Communications Series. Dr. Cimini is an adjunct professor in the Electrical Engineering Department of the University of Pennsylvania where he teaches a graduate-level course in wireless systems. He was elected a fellow of the IEEE in 2000 for contributions to the theory and practice of high-speed wireless communications.

Larry J. Greenstein received his B.S., M.S., and Ph.D. degrees in electrical engineering from Illinois Institute of Technology, Chicago, Illinois, in 1958, 1961, and 1967, respectively. From 1958 to 1970 he was with IIT Research Institute, working on radio frequency interference and anti-clutter airborne radar. He joined Bell Laboratories, Holmdel, New Jersey, in 1970. Over a 32-year AT&T career, he conducted research in digital satellites, point-to-point digital radio, lightwave transmission techniques, and wireless communications systems. For 21 years during that period (1979–2000), he led a research department renowned for its contributions in these fields. His research interests in wireless communications have included measurement-based channel modeling, microcell system design and analysis, diversity and equalization techniques, and system performance analysis and optimization. He recently retired from AT&T Labs–Research, Middletown, New Jersey, as a technology leader. Dr. Greenstein is an AT&T fellow and an IEEE fellow, has won two best paper awards, and has been a guest editor, senior editor, and editorial board member for numerous publications.

BIBLIOGRAPHY

1. R. Van Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*, Artech, 2000.
2. A. Bahai and B. Saltzberg, *Multi-Carrier Digital Communications — Theory and Applications of OFDM*, Kluwer, 1999.
3. L. Hanzo, W. Webb, and T. Keller, *Single- and Multi-carrier Quadrature Amplitude Modulation*, Wiley, 2000.
4. J. A. C. Bingham, *ADSL, VDSL, and Multicarrier Modulation*, Wiley, 2000.
5. M. L. Doelz, E. T. Heald, and D. L. Martin, Binary data transmission techniques for linear systems, *Proc. IRE* **45**: 656–661 (May 1957).
6. M. S. Zimmerman and A. L. Kirsch, The AN/GSC-10 (KATHRYN) variable rate data modem for HF radio, *IEEE Trans. Commun.* **COM-15**: 197–205 (April 1967).
7. M. Alard and R. Lasalle, Principles of modulation and channel coding for digital broadcasting for mobile receivers, *EBU Tech. Rev.* 168–190 (1987).
8. U. Reimers, DVB-T: The COFDM-based system for terrestrial television, *Electron. Commun. Eng. J.* **9**: 28–32 (Feb. 1997).
9. P. S. Chow, J. C. Tu, and J. M. Cioffi, A discrete multitone transceiver system for HDSL applications, *IEEE J. Select. Areas Commun.* **SAC-9**: 909–919 (Aug. 1991).
10. R. van Nee et al., New high rate wireless LAN standards, *IEEE Commun. Mag.* **37**: 82–88 (Dec. 1999).
11. L. J. Cimini, Jr., Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing, *IEEE Trans. Commun.* **COM-33**: 665–675 (June 1985).
12. S. B. Weinstein and P. M. Ebert, Data transmission by frequency-division multiplexing using the discrete fourier transform, *IEEE Trans. Commun. Technol.* **COM-19**: 628–634 (Oct. 1971).
13. U.S. Patent 3,488,445 (filed Nov. 14, 1966; issued Jan. 6, 1970), R. W. Chang, Orthogonal frequency division multiplexing.
14. B. R. Saltzberg, Performance of an efficient data transmission system, *IEEE Trans. Commun. Technol.* **COM-15**: 805–813 (Dec. 1967).
15. B. Hirosaki, An orthogonally multiplexed QAM system using the discrete fourier transform, *IEEE Trans. Commun.* **COM-29**: 982–989 (July 1981).
16. A. Peled and A. Ruiz, Frequency domain data transmission using reduced computational complexity algorithms, *Proc. ICASSP'80*, April 1980, pp. 964–967.
17. T. Pollet, M. van Bladel, and M. Moeneclaey, BER Sensitivity of OFDM systems to carrier frequency offset and Wiener phase noise, *IEEE Trans. Commun.* **43**: 191–193 (Feb.–April 1995).
18. X. Li and L. J. Cimini, Jr., Effects of clipping and filtering on the performance of OFDM, *IEEE Commun. Lett.* **2**: 131–133 (May 1998).
19. S. Müller and J. Huber, A comparison of peak power reduction schemes for OFDM, *Electron. Lett.* **33**: 3680–3689 (Feb. 1997).
20. I. Kalet, The multitone channel, *IEEE Trans. Commun.* **37**: 119–124 (Feb. 1989).
21. P. S. Chow, J. M. Cioffi, and J. A. C. Bingham, A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels, *IEEE Trans. Commun.* **43**: 773–775 (Feb.–March 1995).
22. N. A. Zervos and I. Kalet, Optimized decision feedback equalization versus orthogonal frequency division multiplexing for high speed data transmission over the local cable network, *Proc. ICC'89*, Sept. 1989, pp. 1080–1085.
23. M. Suzuki, R. Boehnke, and K. Sakoda, BDMA — band division multiple access — a new air interface for third generation mobile system, UMTS, in Europe, *Proc. ACTS Mobile Commun. Summit*, Oct. 1997, pp. 482–488.
24. N. Yee, J.-P. Linnartz, and G. Fettweis, Multi-carrier CDMA in indoor wireless networks, *Proc. IEEE PIMRC'93*, Sept. 1993, pp. 109–113.
25. G. G. Raleigh and J. M. Cioffi, Spatio-temporal coding for wireless communications, *IEEE Trans. Commun.* **46**: 357–366 (March 1998).

ORTHOGONAL TRANSMULTIPLEXERS: A TIME-FREQUENCY PERSPECTIVE

ALI N. AKANSU
HUSREV T. SENCAR
New Jersey Institute of Technology University Heights
Newark, New Jersey

1. INTRODUCTION

Orthogonality of carriers has been widely utilized in communications as the way to share available common resources by multiple users [1–3]. The most popular multiuser communication systems use one of the three modulation techniques. Namely, the frequency division multiple access (FDMA), the time division multiple access (TDMA), and code division multiple access (CDMA). The orthogonality of the user signature functions or carriers in a multiuser communication scenario is the underlying feature. The basic difference between these modulation techniques comes from the domain where the orthogonality conditions of the carrier functions are emphasized. In other words, an FDMA system aims to minimize the interaction of its multiple carriers in the frequency domain. Similarly, a TDMA system tries to reduce the time domain overlaps or correlations of its carrier or modulation functions. In contrast, a CDMA system prefers maximized overlaps of its signature functions both in the time and frequency domains while keeping their orthogonality features as the most vital requirement.

The fundamentals of signal and transform theories help us to better understand the multiple user or multicarrier communication systems where the time-frequency and orthogonality properties of the carrier functions are the defining issues. Therefore, we will briefly describe them in the following section.

2. MATHEMATICAL PRELIMINARIES

2.1. Time-Frequency Measures for a Discrete-Time Function

The time and frequency domain energy spread of a function has been a classical topic in signal processing field. The celebrated "uncertainty principle" states that no function can be concentrated simultaneously in both the time and frequency domains [4]. The time domain spread of a discrete-time function $\{h_0(n)\}$ is defined as [5],

$$\sigma_n^2 = \frac{1}{E}\sum_n (n-\overline{n})^2 |h_0(n)|^2 \tag{1}$$

The energy E and time center $\overline{n}$ of the function $\{h_0(n)\}$ are expressed as

$$E = \sum_n |h_0(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_0(e^{jw})|dw, \tag{2}$$

$$\overline{n} = \frac{1}{E}\sum_n |h_0(n)|^2 \tag{3}$$

where its Fourier transform is given as

$$H_0(e^{jw}) = \sum_n h_0(n)e^{-jwn}x \tag{4}$$

Similarly, we define the frequency domain spread of a discrete-time function with $\overline{w} = 0$ as follows

$$\sigma_w^2 = \frac{1}{2\pi E}\int_{-\pi}^{\pi} (w-\overline{w})^2 |H_0(e^{jw})|^2 dw \tag{5}$$

where its center in the frequency domain is given as

$$\overline{w} = \frac{1}{2\pi E}\int_{-\pi}^{\pi} w|H_0(e^{jw})|^2 dw \tag{6}$$

Figure 1 illustrates time-frequency properties of a discrete-time function $\{h_0(n)\}$ using spreading measures defined above. This representation is also called time-frequency tile of a function. The shape and location of the tile can be defined by properly designing the time and frequency features of the function under construction. This interpretation of functions can be further extended in the case of orthogonal basis design. In addition to shaping time-frequency tiles, the orthogonality requirements are also imposed on the basis functions.

For any real signal with $\overline{w} = 0$ and $\overline{n} = 0$ the lower bound for the product of *time-frequency spread* $\sigma_n\sigma_w$ is given as [6]

$$\sigma_n\sigma_w \geq \frac{|1-\mu|}{2} \tag{7}$$

where

$$\mu = \frac{|H_0(e^{jw})_{w=\pi}|^2}{E} \tag{8}$$

Similar time-frequency spreading measures for band-pass signals with peak frequency responses $\overline{w} \neq 0$ are also introduced in Ref. 6.

2.2. Orthogonal Function Sets

2.2.1. Orthogonal Block Transforms. Orthogonal block transforms like discrete Fourier transform (DFT) and

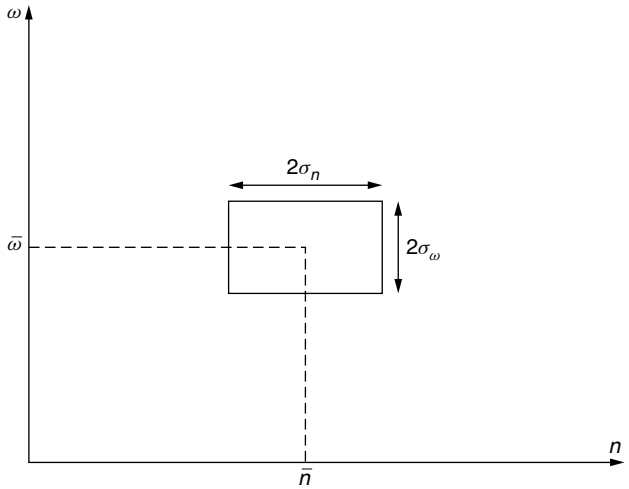


Figure 1. Time-frequency plane illustrating time-frequency properties of discrete time function $\{h_0(n)\}$.

discrete cosine transform (DCT) have been widely used in many engineering applications. The basis of an orthogonal block transform consists of functions $\{h_k(n)\}$ with the orthogonality property as

$$\sum_{n=1}^{N} h_k(n)h_l(n) = \delta_{k-l} \tag{9}$$

where δ_{k-l} is the Kronecker delta sequence given as

$$\delta_{k-l} = \begin{cases} 1, & k-l=0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Note that the length of basis functions and number of functions in the set are the same in block transforms. Therefore, the main emphasis in block transforms has been the orthogonality requirements along with implementation efficiency since there is not much freedom in the design to adjust the time and frequency properties of the functions in the set.

Figure 2 displays DCT basis functions in the time and frequency domains for $N = 8$. Due to the short time durations of these functions it is observed that their frequency selectivity is somehow limited and they overlap significantly. They perform like a filter bank with poor frequency selectivity. The time-frequency measures of these functions are presented in Table 1.

The only way to improve the frequency localizations of orthogonal basis functions is to increase their durations in the time domain. Due to the time-frequency duality property of functions, the localization of a function in one domain can be improved at the expense of its localization in the other domain. This property paved the way for filter banks and subband transforms that we introduce in the next section.

One can use orthogonal transforms for analysis of a given function or signal through an operation called forward transform. Let real orthonormal sequences $\{h_r(n)\}$ be the rows of a transformation matrix, $H(r, n)$,

$$H = [H(r, n)], \quad r, n = 1, \ldots, N \tag{11}$$

Orthonormality of the matrix H assures that its inverse matrix

$$H^{-1} = H^T \tag{12}$$

where ^{T} indicates a matrix transpose. Hence,

$$HH^T = I \tag{13}$$

where I is an $N \times N$ identity matrix.

The forward transform of an input vector $\underline{x}$ of size N is written in a matrix notation as

$$\theta = Hx \tag{14}$$

where θ is the transform coefficient vector of size N. Therefore, x can be perfectly reconstructed through the inverse transform operation as

$$x = H^T\theta \tag{15}$$

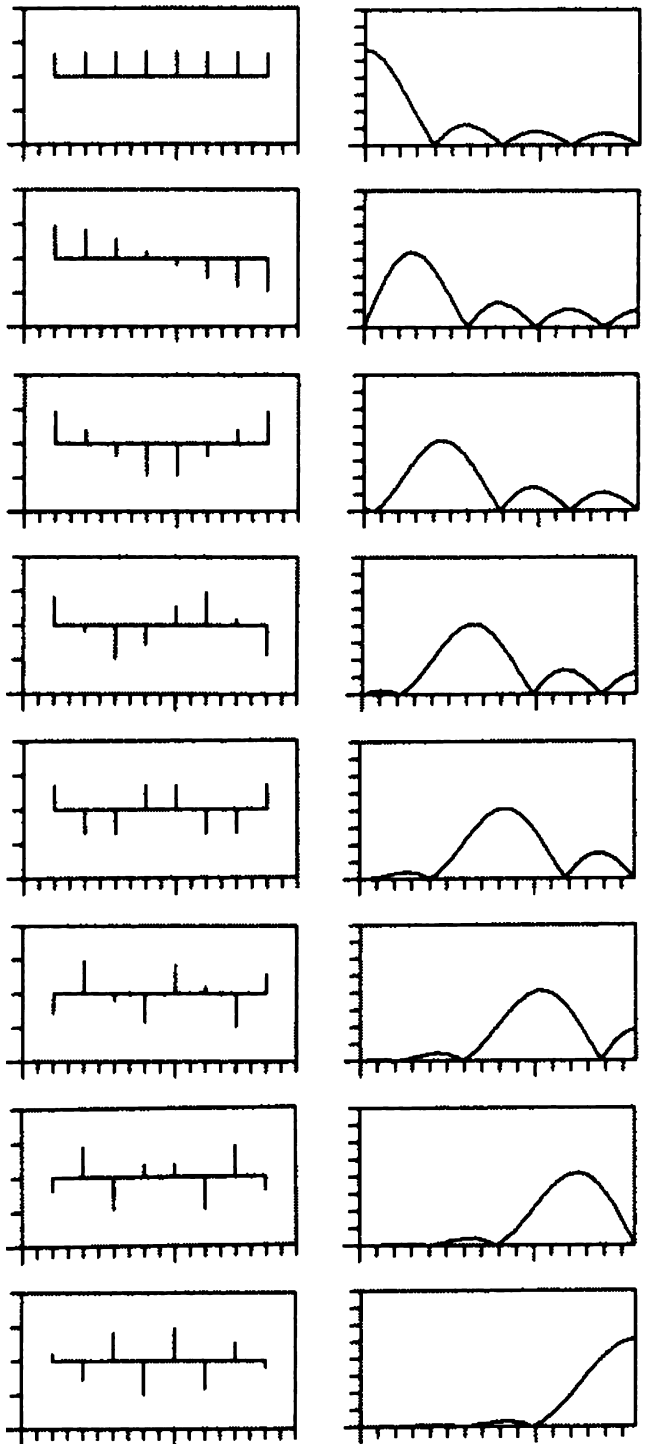


Figure 2. DCT basis functions in time and frequency domains for $N = 8$.

This defines a pair of forward/inverse transform operators where the input x is mapped onto the transform space as vector θ and perfectly recovered from θ through the inverse transform operator.

In contrast, one could synthesize the signal vector x by transforming the given input signal θ onto the inverse transform space as

$$x = H^T\theta \tag{16}$$

Table 1. Time-Frequency Localization of DCT for $N = 8$

$\overline{w}$	$\overline{n}$	σ_w^2	σ_n^2	$\sigma_w\sigma_n$
0	3.5	0.3447	5.25	1.3452
0.74	3.5	0.3021	8.4054	1.5935
1.02	3.5	0.2413	5.9572	1.1989
1.36	3.5	0.1957	5.4736	1.0350
1.71	3.5	0.1488	5.25	0.8839
2.08	3.5	0.1206	5.0263	0.7786
2.45	3.5	0.0797	4.5428	0.6017
π	3.5	0.1388	2.0955	0.5393

where H^T is the inverse transform matrix. It is a straightforward operation to perfectly reconstruct θ from x through a forward transform operation on x as

$$\theta = Hx \tag{17}$$

This is a sequence of inverse/forward transform operators that serves as the foundation for *orthogonal transmultiplexers* in multiple access communications. Fourier transform basis has been widely used in telecommunication applications for many decades utilizing the concept of transmultiplexers. As mentioned earlier, the frequency selectivity of these carrier functions, DFT basis, are not very good although their implementation in a real-time transmultiplexer structure is efficient. Therefore, they have been quite popular [7].

2.2.2. M-Band Filter Banks with Perfect Reconstruction. A maximally decimated M-band finite impulse response (FIR) perfect reconstruction quadrature mirror filter (PR-QMF) bank analysis/synthesis configuration is displayed in Fig. 3a. The output of this filter bank is the delayed version of its input as

$$y(n) = x(n - n_0) \tag{18}$$

where n_0 is a delay constant. In a paraunitary filter bank solution, the synthesis and analysis filters are related as

$$g_r(n) = h_r^*(p - n) \tag{19}$$

where p is a time delay. Hence, the PR-QMF conditions can be imposed on the analysis filters in the time domain

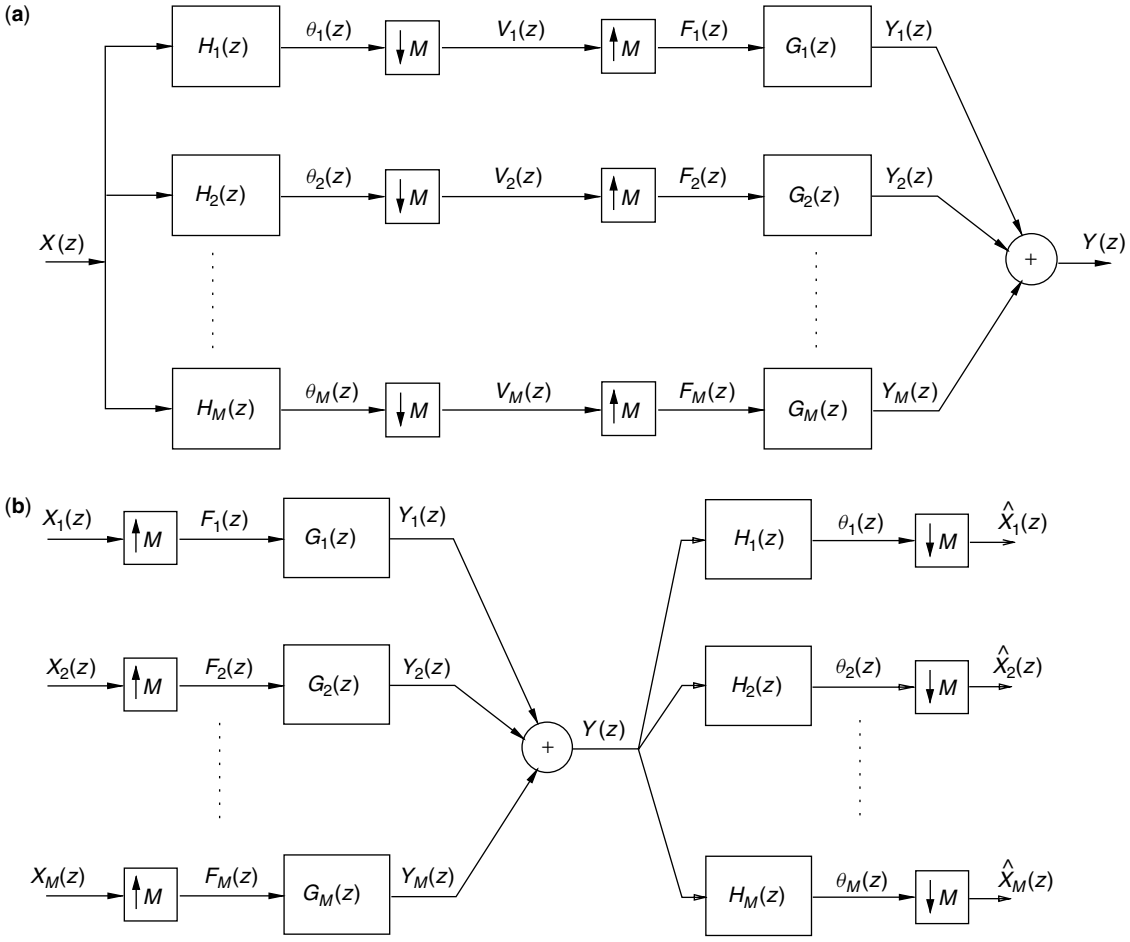


Figure 3. (**a**) Maximally decimated M-band FIR PR-QMF bank structure (analysis/synthesis filter bank configuration). (**b**) M-band Transmultiplexer structure (synthesis/analysis filterbank structure).

as [6]

$$\sum_n h_r(n)h_r(n+Ml) = \delta(l), r = 1, \ldots, M \quad (20)$$

$$\sum_n h_r(n)h_s(n+Ml) = 0, \text{ for all } l \quad (21)$$

Analysis/synthesis filter bank configurations are widely used in image or video processing, speech or audio processing, interference cancellation, and other applications [8–9].

In contrast, Fig. 3b depicts a synthesis/analysis filter bank where there are M inputs and M outputs of the system. It is shown that if the synthesis $\{g_r(n)\}$ and analysis $\{h_r(n)\}$ filters satisfy the PR-QMF conditions of Eqs. (19)–(21), synthesis/analysis filter bank configuration gives equal input and output for its all branches as

$$\hat{x}_r(n) = x_r(n - n_0) \quad r = 1, \ldots, M \quad (22)$$

where n_0 is a time delay. The synthesis/analysis PR-QMF bank with equal inputs and outputs at all branches has been used as orthogonal transmultiplexers in communications applications for single user and multiuser scenarios [6,10].

3. COMMUNICATION APPLICATIONS OF ORTHOGONAL TRANSMULTIPLEXERS

The unified framework for orthogonal transmultiplexers along with time-frequency tools allowing some design flexibilities for the application at hand was given in the previous sections of the article. The main engineering challenge in this context is to design the most suitable transform basis $\{h_r(n)\}$ for a given application. Applications using orthogonal multiplexers vary from single-user communication scenarios to multiuser communication systems. These applications might require to utilize frequency localized or time-localized orthogonal carriers depending on the system requirements including channel properties. The orthogonal block transforms like DFT has been widely used in a synthesis/analysis filter bank configuration (inverse/forward block transform sequence of operators) as a transmultiplexer in multicarrier (single and multiuser) communication systems. Although the frequency selectivity of DFT basis functions is not very good, except at the bin frequencies of the orthogonal carrier functions, the ease of its implementation has been very attractive for real-time applications.

An examination of M-band orthogonal transmultiplexer structure displayed in Fig. 3b helps us to interconnect time-frequency properties of carrier (modulation) basis with the type of communication system under consideration. The most popular types are FDMA, TDMA, and CDMA. We discuss these orthogonal modulation types further from a time-frequency perspective in the following sections.

3.1. FDMA

Figure 4a displays an ideal filter bank that consists of brick-wall frequency functions without any interbrand energy leakage. The frequency localizations of these orthogonal carriers are perfect although their time-localization is extremely poor. Since they are noncausal functions with infinite time durations they are not implementable. In practice, finite length (FIR) orthogonal carriers are used. Therefore, interbrand (cross-carrier) energy leakage (interference) is of a great concern in communication applications. As mentioned earlier, DFT basis has been used in many applications including the popular digital subscriber line (DSL) communications. The other applications like digital audio broadcasting (DAB) and low probability of intercept (LPI) communication also employ orthogonal transmultiplexers with properly selected filter banks or carrier basis [8,9].

3.2. TDMA

Similarly, Fig. 4b displays the ideal orthogonal carriers (basis functions) for a TDMA configuration. Note that each carrier is a unit sample function in the time domain with a perfect localization. In contrast, those functions completely overlap in the frequency domain. In this case, a perfect orthogonality is imposed in the time domain. Practical TDMA systems use nonideal time pulses or symbols where intercarrier energy leakage (interference) is inevitable.

3.3. CDMA

CDMA is a marked departure from the traditional FDMA and TDMA systems where the spreading of orthogonal carrier functions (signatures) in both domains (time and frequency) is aimed. Note that the orthogonal transmultiplexer configuration of Fig. 3b is still applicable even for the CDMA systems. The optimal code design problem for CDMA can also be handled by the PR-QMF requirements of Eqs. (19)–(21) with the addition of maximizing the joint time-frequency spread, $\sigma_n\sigma_w$, of the codes in the design. Figure 5 displays frequency spectra of a 32-length spread spectrum PR-QMF along with 31-length Gold Codes. It is observed from this figure that the functions used in orthogonal transmultiplexers for CDMA are not selective in either domain. Although filter banks have been mostly used for spectral analysis/synthesis problems the underlying theory is also applicable for any time-frequency shaped function sets.

4. CONCLUSIONS

The synthesis/analysis configurations of filter banks (transmultiplexers) have been widely employed in many popular communications applications. Conversely, the time-frequency shaping of orthogonal functions has been well tied to the optimal filter bank design in the signal processing literature. We highlighted those developments in the context of orthogonal transmultiplexers that serve as the foundation in single and multiple user communication systems.

BIOGRAPHIES

Ali N. Akansu received the B.S. degree from the Technical University of Istanbul in 1980 and the M.S. and Ph.D.

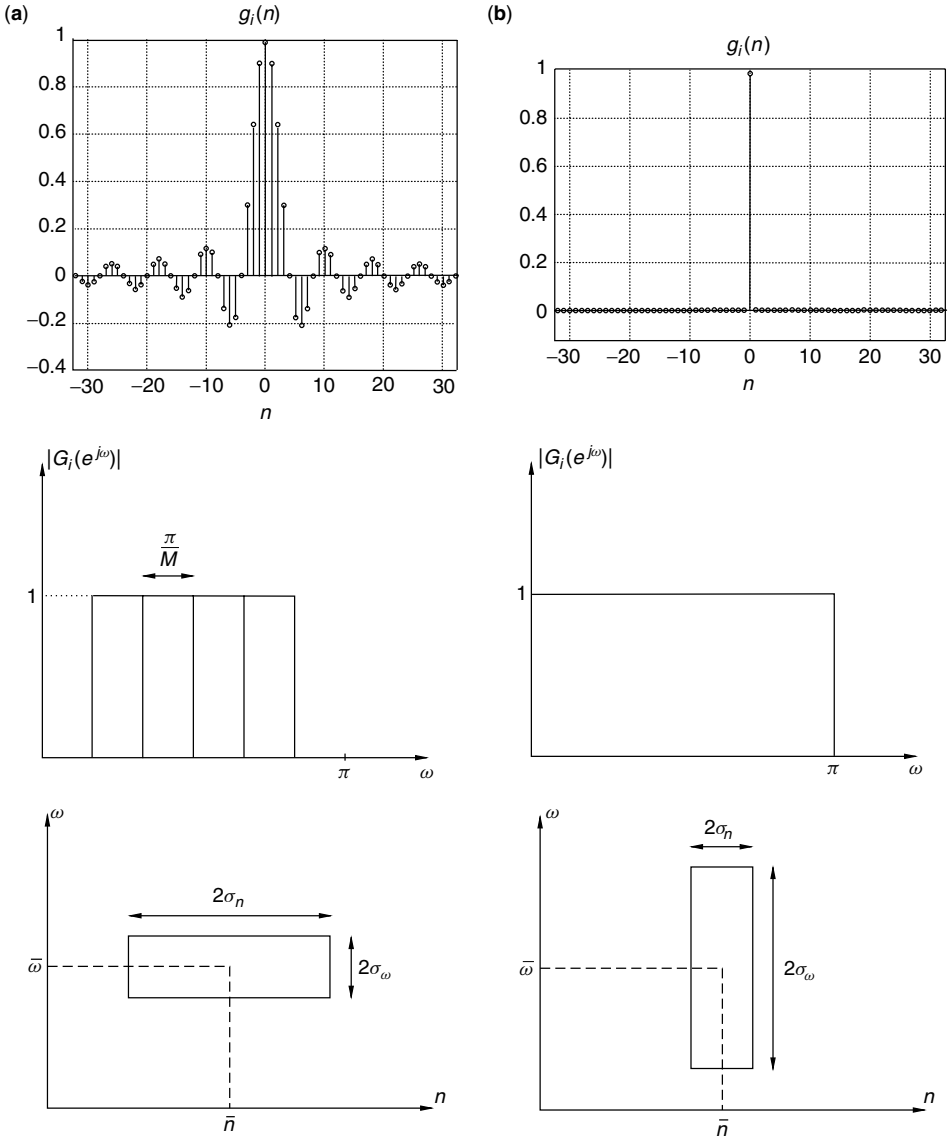


Figure 4. Ideal filter banks (orthogonal carriers) for the cases of (**a**) FDMA (brick-wall shaped in frequency) and (**b**) TDMA (unit sample function in time) scenarios.

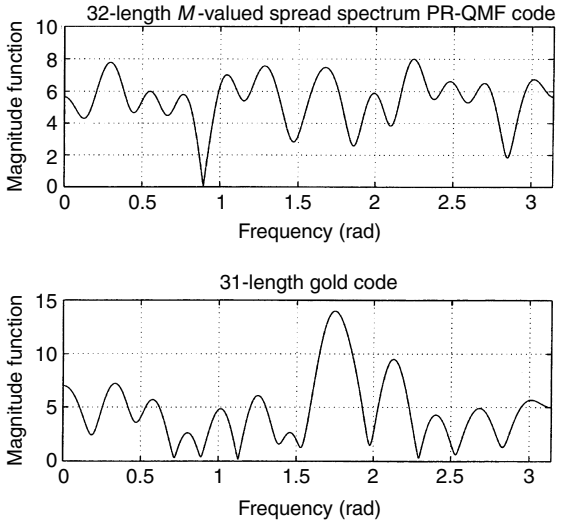


Figure 5. Frequency spectra of a 32-length spread spectrum PR-QMF and 31-length Gold codes.

degrees from the Polytechnic University in 1983 and 1987, respectively, all in electrical engineering. Since 1987, he has been with the New Jersey Institute of Technology, where he is a professor of electrical and computer engineering. He was a co-founder and director of the New Jersey Center for Multimedia Research between 1996 and 2000. He was the vice president for R&D of the IDT Corporation from 2000 to 2001. He has been the founding president of PixWave Inc. His industrial affiliations also include his visits to IBM T.J. Watson Research Center and GEC-Marconi Electronic Systems Corp. during the summers of 1989 and 1996, and 1992, respectively. He serves as a consultant to the industry and sits on the boards of a few Internet startup companies. He co-authored and co-edited three books and many papers on his research. His current research is on signal and transform theories with applications in multimedia and communications.

Husrev T. Sencar received the B.Sc. and M.Sc. degrees from Middle East Technical University, Ankara,

Turkey, and Baskent University, Ankara, Turkey, in 1996 and 1998, respectively, all in electrical and electronics engineering. He currently pursues a Ph.D. in electrical engineering at New Jersey Institute of Technology, Newark, New Jersey. His research interests include signal processing for communications and multimedia with emphasis on information hiding and video/image processing.

BIBLIOGRAPHY

1. B. R. Saltzberg, Performance efficient parallel data transmission system, *IEEE Trans. Commun.* **Com-15**: 805–811, (Dec. 1967).
2. *Special Issue on Transmultiplexers*, *IEEE Trans. Commun.* **Com-30**: (7, July 1982).
3. *Special Issue on Multicarrier Communications*, *Wireless Personal Communications* **2**(1–2): (July 1995).
4. A. Papoulious, *Signal Analysis*, McGraw-Hill, New York, 1977.
5. R. A. Haddad, A. N. Akansu, and A. Benyassine, Time-frequency localization in M-band filter banks and wavelets: A critical review, *J. Opt. Eng.* **32**: 1411–1429 (July 1993).
6. A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets*, 2nd Ed., Academic Press, 2001 347–349.
7. M. G. Bellanger and D. L. Daguet, TDM-FDM multiplexer digital polyphase and FFT, *IEEE Trans. Commun.* **Com-22**: 1199–1205 (Sep. 1974).
8. A. N. Akansu and M. J. T. Smith (eds.), *Subband and Wavelet Transforms: Design and Applications*, Kluwer, 1995.
9. A. N. Akansu and M. J. Medley (eds.), *Wavelet, Subband and Wavelet Transforms in Communication and Multimedia*. Kluwer, 1999.
10. *Special Issue on Theory and Application of Filter Banks and Wavelet Transforms*, *IEEE Trans. Signal Processing* **SP-46**(4): (Apr. 1998).

P

PACKET-RATE ADAPTIVE RECEIVERS FOR MOBILE COMMUNICATIONS

STELLA N. BATALAMA
State University of New York
at Buffalo
Buffalo, New York

1. INTRODUCTION

The effectiveness of a receiver designed for a rapidly changing multiple-access (multiuser) communications environment depends on the following design attributes: (1) system adaptivity under limited data support, (2) multiple-access-interference resistance, and (3) low computational complexity. Short-data-record adaptive designs appear as the natural next step for a matured discipline that has extensively addressed the other two design objectives, 2 and 3, in ideal setups (perfectly known or asymptotically estimated statistical properties). System adaptivity based on short data records is necessary for the development of practical adaptive receivers that exhibit superior signal-to-interference-plus-noise ratio (SINR) or bit error rate (BER) performance when they operate in rapidly changing communications environments that limit substantially the input data support available for adaptation and redesign.

In modern packet data transmission systems where the basic information flow unit is the packet (a group of bits that includes the actual information bits as well as other coding and network control bits), the main measure of link quality is the throughput (either packet throughput or information throughput) that which is directly related to the packet error rate (PER). Real-time voice communications impose stringent delay constraints and require a guaranteed upper bound on PER of about 10^{-2}. On the other hand, data packets can tolerate reasonable delays but may require a lower PER bound [1,2]. Packet throughput improvements can be achieved as a result of BER improvements. On the other hand, BER improvements can be achieved by means of advanced receiver designs that exploit both the characteristics of the transmitted signal and the current state of the environment (these "raw" BER values can be further improved through channel coding (forward error correction)). In dynamic environments, adaptive receiver designs can react to variations as opposed to static receivers that remain unchanged regardless of the changes in the environment. Inherently, a major consideration in the design of successful adaptive receivers is the fact that their adaptation rate must be commensurate to the rate of change of the environment.

An example of a system that can benefit from modern advanced adaptive receiver technology is the direct-sequence code-division multiple-access (DSCDMA) radiofrequency (RF) system. In such a system, the transmitted signal is a spread-spectrum (SS) signal obtained by multiplying each information bit by a unique code (or signature) waveform dedicated to each user. The SS characteristics of the transmitted signal allow intelligent temporal (code) processing at the receiver (unmasking of the signature). During RF transmission, the signal in general undergoes a process known as *multipath-fading* dictated by the physical characteristics of the communication channel. As a result, the received signal consists of multiple faded and delayed copies of the transmitted signal. At the receiver, the multiple copies, instead of being discarded as interference, can be processed in an advantageous manner (a procedure known as RAKE processing). Further performance improvements can be obtained by exploiting the spatial characteristics of the transmitted signal; such processing requires that antenna-array ("smart antenna") technology is employed at the receiver. DSCDMA systems equipped with antenna arrays offer the opportunity for jointly effective spatial (array) and temporal (code) noise suppression. Primary noise sources include additive white Gaussian noise (AWGN) usually due to the receiver front-end electronics as well as interference from other users who transmit similar signals at the same time and in the same frequency spectrum [CDMA systems allow such channel accessing as opposed to time-division multiple-access (TDMA) or frequency-division multiple-access (FDMA) systems]. This general DSCDMA signal model example will be revisited many times throughout our discussion, and a complete adaptive antenna-array DSCDMA receiver will be developed as an illustration.

Returning to the main topic of our discussion, an adaptive receiver consists of a set of building blocks that are a reevaluated (or estimated) every time there is a significant change in the statistics of the environment. The design of each building block is initially formulated mathematically as a solution to an optimization problem under the assumption that all statistical quantities are perfectly known. This is known as the *ideal* or *optimum* solution. Then, the statistical quantities that are present in the optimum solution are substituted by corresponding estimates that are based on the actual received data (observations). This is known as an *estimate* of the optimum solution. It is the latter estimates that need to be adapted according to the changes of the environment, justifying this way the term "adaptive receiver." For example, a popular class of DSCDMA receivers utilizes minimum mean-square-error (MMSE) linear (discrete-time) filters as a means to suppress multiple-access interference (MAI) and AWGN. In other words, the receiver consists of a linear filter that operates on a discrete sequence of spacetime (ST) received signal samples and the optimum filter solution is found by minimizing the mean-square error between the output of the linear filter and a pilot information bit sequence. Several adaptive MMSE filtering algorithms are known and include the sample matrix inversion (SMI), the least-mean-square (LMS) and the recursive-least-square (RLS) algorithms, which will be discussed in detail in subsequent sections. As a general comment, these algorithms/adaptive

filters outperform significantly the popular static ST RAKE filter when a sufficiently large number of data is made available to them [4–9]. Unfortunately, the time-varying nature of the channel frequently necessitates fast (short-data-record) adaptive ST optimization through the use of small input data sets that can "catch up" with the channel variations.

To motivate the developments presented in this article, let us consider a DSCDMA system with 5-element antenna-array reception, system processing gain 64, and, say, 3 resolvable multipaths for the user signal of interest (usually the number of resolvable multipaths is between 2 and 4 including the direct path if any [10]). For such a system, we will see later that jointly optimal S-T processing at the receiver under the MMSE criterion requires processing in the $5(64+3-1)=330$ space-time product space. That is, filter optimization needs to be carried out in the complex $\mathbb{C}^{330}$ vector space. We know that adaptive SMI implementations of the MMSE filter solution require data samples many times the space-time product to approach the performance characteristics of their ideal counterpart (RLS/LMS implementations behave similarly) [11,12]. In fact, theoretically, system optimization with data samples less than the spacetime product may not even be possible, as we will explain later in our discussion. With CDMA chip rates at 1.25 MHz [10], processing gain 64, and typical fading rates of $\geq$70 Hz for vehicle mobiles [13], the fading channel fluctuates decisively at least every 280 data symbols. In this context, conventional SMI/RLS/LMS adaptive filter optimization in the $\mathbb{C}^{330}$ vector space becomes an unrealistic objective.

The goal of our presentation is to first introduce and then elaborate on the underlying principles of short-data-record adaptive filter estimation. Through illustrative examples from the mobile communications literature, we will observe that short-data-record (e.g., packet-rate) filter estimation results in improved channel BER, which translates to higher packet success probability and higher user capacity for a given PER upper bound quality-of-service (QoS) constraint. This ensures an improvement in terms of packet throughput and delay characteristics of a network system that satisfies the QoS constraint. Additional performance improvements can and must be pursued through synergistic use of channel coding (FEC).

While our target applications are all time-critical communications problems, the theoretical developments that will be presented herein may touch many aspects of multidisciplinary engineering that are hampered by the "curse of dimensionality" and could benefit from adaptive filtering and/or adaptive system optimization through limited input data.

2. BASIC SIGNAL MODEL

For illustration purposes, we consider throughout this presentation a multiuser communications system where binary antipodal information symbols from user0, user1, ..., user $Q-1$ are transmitted at a rate $1/T$ by modulating (being multiplied by) a signal waveform $d_q(t)$ of duration T, $q=0,\ldots,Q-1$, that uniquely identifies each user and is assumed to be approximately band-limited or have negligible frequency components outside a certain bandwidth. If H_1 (H_0) denotes the hypothesis that the information bit $b_0=+1$ ($b_0=-1$) of the user of interest, say, user 0, is transmitted during a certain bit period T, then the corresponding equivalent lowpass composite received waveform over the bit interval T may be expressed in general as

$$\begin{aligned} H_1&: x(t)=(+1)\sqrt{E_0}v_0(t)+z(t) \quad \text{and} \\ H_0&: x(t)=(-1)\sqrt{E_0}v_0(t)+z(t), \quad 0\leq t\leq T \end{aligned} \tag{1}$$

With respect to the user of interest, user 0, E_0 denotes transmitted energy per bit, $v_0(t)$ represents the channel processed version of the original waveform $d_0(t)$ [w.l.o.g. the signal waveform $d_0(t)$ is assumed to be normalized to unit energy over the bit period T], and $z(t)$ identifies comprehensively the channel disturbance and includes one, some, or all of the following forms of interference: (1) MAI, (2) intersymbol interference (ISI), and (3) additive Gaussian or non-Gaussian noise.

The continuous-time waveform $x(t)$ is "appropriately" sampled and the discrete samples are grouped to form vectors of "appropriate" length P (both the sampling method and the length value P are pertinent to the specifics of the application under consideration). Let $\mathbf{x}$ denote such a discrete-time complex, in general, received signal vector in $\mathbb{C}^P$:

$$\begin{aligned} H_1&: \mathbf{x}=+\sqrt{E_0}\mathbf{v}_0+\mathbf{z} \quad \text{and} \\ H_0&: \mathbf{x}=-\sqrt{E_0}\mathbf{v}_0+\mathbf{z}, \quad \mathbf{x},\ \mathbf{v}_0,\ \mathbf{z}\in\mathbb{C}^P \end{aligned} \tag{2}$$

where P identifies the dimension of the discrete-time complex observation space, $\mathbf{v}_0$ is the signal vector that corresponds to $v_0(t)$, and $\mathbf{z}$ denotes the discrete-time comprehensive disturbance vector [14]. Our objective is to detect b_0 (i.e., to decide in favor of H_1 or H_0) by means of a linear filter $\mathbf{w}$ as follows:

$$\hat{b}_0=\text{sgn}(\text{Re}\{\mathbf{w}^H\mathbf{x}\}) \tag{3}$$

where $\text{sgn}(\cdot)$ is the ±1 hard-limiter, $\text{Re}\{\cdot\}$ extracts the real part of a complex number, and $(\cdot)^H$ denotes the Hermitian operation. In other words, we fix the structure of the receiver to that given by Fig. 1. Our discussion will be focused on the design of the linear filter $\mathbf{w}$ according to the MMSE or minimum-variance distortionless response (MVDR) optimization criteria we present in the section that follows.

The specific illustrative example of a multipath fading AWGN DS-CDMA packet data communication link with narrowband linear antenna-array reception that we considered earlier, is certainly covered by the above general basic signal model. The transmitted signal waveform of a particular user is obtained as follows. The user is assigned a unique binary antipodal signature (code)

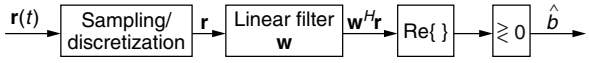


Figure 1. General receiver structure for the (one-shot) detection of binary antipodal information symbols of the user of interest.

sequence, that is a sequence with elements $+1$ or -1 of length L (L is also called *system processing gain*). The bits of the user code multiply a basic signal pulse (e.g., square pulse or raised cosine) of duration T_c, known as *chip*. This way we obtain the signature waveform $d_q(t)$ of duration $T = LT_c$. The transmitted signal waveform that corresponds to a single information bit is the product of the information bit itself and the signature waveform. The corresponding received waveform is the convolution of the transmitted waveform and the impulse response of the multipath fading channel (when the latter is modeled as a linear filter) and is assumed to be band-limited by the chip rate. Discretization of the continuous received waveform at each antenna element of the array can be achieved by chip-matched filtering of the received waveform and sampling at the chip rate (or by lowpass filtering, commensurate Nyquist sampling, and chip-rate accumulation) over the multipath-extended symbol period. The discrete vector outputs from all antenna elements are stacked together (one on top of the other) to create a supervector known as the *spacetime* (ST) received vector. In this way the data are prepared for processing by the linear filter $\mathbf{w}$, extraction of the real part of the filter output and finally sign detection as shown in Fig. 1, a process termed "one shot" detection: detection on a symbol-by-symbol (information bit) basis, as opposed to simultaneous detection of all information bits of the user of interest. If M is the number of antenna elements of the array, L is the length of the signature (code) vector, and N is the number of resolvable multipaths (w.l.o.g. we assume that N is the same for all users) then the discretetime, ST complex received vector $\mathbf{x}$ is of dimension $M(L+N-1)$; where $L+N-1$ is exactly the length of what we referred to earlier as the multipath-extended symbol period. "Inside" $\mathbf{x}$ in (2), $\mathbf{v}_0$ corresponds to the channel-processed (also known as "effective") ST signature vector of the user of interest, while $\mathbf{z}$ corresponds to the discrete-time disturbance vector that accounts for MAI, ISI, and AWGN. Specifically, $\mathbf{v}_0$ can be expressed as a function of the transmitted signal, channel and receiver structure parameters:

$$\mathbf{v}_0 = \sqrt{\frac{E_0}{L}} \sum_{n=0}^{N-1} c_{0,n} \left[\underbrace{0 \ldots 0}_{n} \quad \mathbf{d}_0^T \quad \underbrace{0 \ldots 0}_{N-n-1} \right]^T \odot \mathbf{a}_{0,n} \tag{4}$$

where $c_{0,n}, n = 0, \ldots, N-1$, denote the path coefficients of the channel of the user of interest. The coefficients $c_{0,n}, n = 0, \ldots, N-1$, are frequently modeled as independent zero-mean complex Gaussian random variables (that exhibit Rayleigh distributed amplitude and uniformly distributed phase that fits experimental measurements) and are assumed to remain constant over the entire packet duration. In a realistic environment the coefficients may vary approximately every 300 symbols [15]. Thus, keeping the packet size less than 300 validates the assumption of constant multipath coefficients over the duration of a packet. In (4), $\mathbf{d}_0 = [\mathbf{d}_0[0], \ldots, \mathbf{d}_0[L-1]]^T$ is the binary signature vector (spreading sequence) of the user of interest, $\mathbf{d}_0[l] \in \{\pm 1\}, l = 0, \ldots, L-1$, $\mathbf{a}_{0,n}$ is the array response vector that corresponds to the nth path of the user of interest, and $\odot$ denotes the Krönecker tensor product. The array response vector of the nth path of the user of interest is defined by

$$\mathbf{a}_{0,n}(m) = e^{j2\pi(m-1)\frac{d}{\lambda}\sin\theta_{0,n}}, \quad m = 1, 2, \ldots, M \tag{5}$$

where $\theta_{0,n}$ identifies the angle of arrival of the corresponding path, λ is the carrier wavelength, and d is the element spacing of the antenna array (usually $d = \lambda/2$). More details on the DSCDMA ST received signal model in (4) and the operational characteristics of an antenna-array system can be found elsewhere in the literature [5,16]. Finally, the noise vector $\mathbf{z}$ represents the comprehensive disturbance effect of AWGN and all other user signal contributions that are again of the form of (4), yet with different in general energy values, signature vectors, multipath coefficients, and angles of arrival.

3. FILTERING WITH KNOWN INPUT STATISTICS

3.1. Optimum MMSE/MVDR Filter

Minimum-variance distortionless response (MVDR) *receiver design* refers to the problem of identifying a linear finite-impulse response filter that minimizes the variance at its output, while at the same time the filter maintains a "distortionless" response toward a specific input vector direction of interest. In mathematical terms, if $\mathbf{x}$ is a random, *0*-mean (without loss of generality) complex input vector of dimension P, $\mathbf{x} \in \mathbb{C}^P$, that is processed by a P-tap filter $\mathbf{w} \in \mathbb{C}^P$, then the filter output variance is $E\{|\mathbf{w}^H\mathbf{r}|^2\} = \mathbf{w}^H\mathbf{R}\mathbf{w}$, where $\mathbf{R} = E\{\mathbf{x}\mathbf{x}^H\}$ is the input autocorrelation matrix ($E\{\cdot\}$ denotes the statistical expectation operation). The MVDR filter minimizes $\mathbf{w}^H\mathbf{R}\mathbf{w}$ and simultaneously satisfies an equation of the form $\mathbf{w}^H\mathbf{v}_0 = \rho$, where $\mathbf{v}_0$ is the given input signal vector direction to be protected. In this setup, MVDR filtering is a standard linear constraint optimization problem and the conventional Lagrange multipliers constraint optimization technique leads to the solution (the Lagrange multipliers optimization technique is presented in detail elsewhere [16])

$$\mathbf{w}_{\text{MVDR}} = \rho^* \frac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0} \tag{6}$$

where $(\cdot)^*$ denotes conjugation. Extensive tutorial treatments of MVDR filtering can be found in many sources [e.g., 16,17], along with historical notes on the early work by Capon [18] and Owsley [19].

MVDR filtering has long been a workhorse for blind (unsupervised) communications and signal processing applications where a desired (pilot) scalar filter output $y \in \mathbb{C}$ cannot be identified or cannot be assumed available for each input $\mathbf{x} \in \mathbb{C}^P$. Prime examples include radar and array processing problems where the constraint vector $\mathbf{v}_0$ is usually referred to as the "target" or "look" direction of interest. It is interesting to observe the close relationship between the MVDR filter and the MMSE ("Wiener") filter. Indeed, if the constraint vector $\mathbf{v}_0$ is chosen to be the statistical cross-correlation vector between the desired output y and the input vector $\mathbf{x}$; that is, if $\mathbf{v}_0 = E\{\mathbf{x}y^*\}$,

then the MMSE filter obtained by minimizing the mean-square (MS) error between the filter output $\mathbf{w}^H\mathbf{x}$ and the desired output y is given by

$$c\mathbf{R}^{-1}\mathbf{v}_0, \quad c > 0 \tag{7}$$

that is, the MMSE filter becomes a positive scaled version of the MVDR filter and exhibits identical output SINR performance. For this reason in the rest of our discussion we refer comprehensively to both filters as MMSE/MVDR filters as [16,17].

Conventionally, the computation of the MMSE/MVDR filter in (6) or (7) begins with the calculation of the inverse of the ideal input autocorrelation matrix $\mathbf{R}^{-1}$ (assuming that the Hermitian matrix $\mathbf{R}$ is strictly positive definite, hence invertible). The calculation of the inverse is usually based on numerical iterative diagonalization linear algebra procedures [20]. Then, the matrix $\mathbf{R}^{-1}$ is used for the linear transformation (left multiplication) of the constraint vector $\mathbf{v}_0$, followed by $\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0$ normalization, as necessary.

Linear transformations that involve the inverse of a high-dimension matrix are computationally intensive. In addition, and most importantly, severe complications arise at the adaptive implementation stage when the estimate of such a high-dimension matrix is inverted (particularly when the estimate is based on a small set of data/observations and is obtained, possibly, by some form of sample averaging). One extreme example of such a complication is the fact that the inverse may not even exist. Thus, when the data that are available for adaptation and redesign are limited, use of inverses of (sample average) estimated high-dimension matrices is not viewed favorably (this issue will be discussed in detail in the next section). In such cases, it is preferable to proceed with alternative methods that *approximate* the optimum solution and, hopefully, avoid implicit or explicit use of inverses. Then, at the adaptive implementation stage, we may utilize estimates of the approximate solutions. Algorithmic designs that aim at approximating the optimum MMSE/MVDR filter include (1) the generalized sidelobe canceler (GSC) and its variations, (2) the auxiliary vector (AV) filter, and (3) the orthogonal multistage filter (also "called nested Wiener filter"). The relative performance of these methods in limited data support environments is examined in Section 4.

3.2. Generalized Sidelobe Canceler (GSC)

For a given (not necessarily normalized) constraint vector $\mathbf{v}_0 \in \mathbb{C}^P$, any "distortionless" linear filter $\mathbf{w} \in \mathbb{C}^P$ that satisfies $\mathbf{w}^H\mathbf{v}_0 = \rho$ can be expressed/decomposed as $\mathbf{w} = (\rho^*/\|\mathbf{v}_0\|^2)\mathbf{v}_0 - \mathbf{u}$ for some $\mathbf{u} \in \mathbb{C}^P$ such that $\mathbf{v}_0^H\mathbf{u} = 0$, as shown in Fig. 2 (this decomposition is an application of the projection theorem in linear algebra). There are two general approaches for the design of the filter part $\mathbf{u}$: (1) eigen-decomposition-based approaches and (2) non-eigendecomposition-based approaches.

Algorithmic eigendecomposition-based designs that focus on the MMSE/MVDR filter part $\mathbf{u}$, which is orthogonal to the constraint vector, or "look" direction $\mathbf{v}_0$, include the Applebaum–Howells arrays, beamspace partially adaptive processors, or generalized sidelobe cancelers

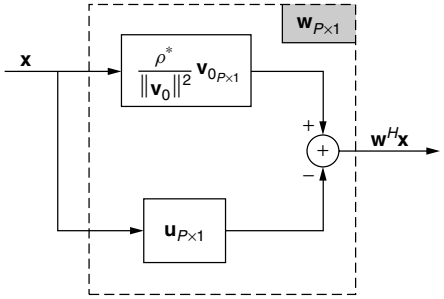


Figure 2. General decomposition of a linear filter $\mathbf{w}$ that satisfies $\mathbf{w}^H\mathbf{v}_0 = \rho$ to two orthogonal components ($\mathbf{u}^H\mathbf{v}_0 = 0$).

(GSCs). More recent developments have been influenced by principal-components analysis (PCA) reduced-rank processing principles. The general goal of these designs is to approximate the MMSE/MVDR filter part $\mathbf{u}$ by utilizing different rank-reducing matrices as explained below [16,17,21–26]. The approximation is of the general form (Fig. 3)

$$\mathbf{u}_{P\times1} \simeq \mathbf{B}_{P\times(P-1)}\mathbf{T}_{(P-1)\times p}\mathbf{w}_{p\times1}^{\mathrm{GSC}} \tag{8}$$

where $\mathbf{B}$ is a matrix that satisfies $\mathbf{B}^H\mathbf{v}_0 = \mathbf{0}_{P-1}$ and is, thus, called "blocking matrix" since it blocks signals in the direction of $\mathbf{v}_0$ ($\mathbf{B}$ is a full column-rank matrix that can be derived by Gram–Schmidt orthogonalization of a $P \times P$ orthogonal projection matrix such as $\mathbf{I} - (\mathbf{v}_0\mathbf{v}_0^H/\|\mathbf{v}_0\|^2)$, where $\mathbf{I}$ is the identity matrix). $\mathbf{T}$ is the rank-reducing matrix with $1 \le p < P-1$ columns that have to be selected and $\mathbf{w}^{\mathrm{GSC}}$ is a vector of weights of the p columns of $\mathbf{T}$ that is designed to minimize the variance at the output of the "overall" filter $\mathbf{w}$, $E\left\{\left|\left(\frac{\rho^*}{|\mathbf{v}_0|^2}\mathbf{v}_0 - \mathbf{u}\right)^H\mathbf{x}\right|^2\right\}$. The solution to the latter optimization problem (assuming that $\mathbf{T}$ is given) is

$$\mathbf{w}^{\mathrm{GSC}} = \frac{\rho^*}{\|\mathbf{v}_0\|^2}[\mathbf{T}^H\mathbf{B}^H\mathbf{R}\mathbf{B}\mathbf{T}]^{-1}\mathbf{T}^H\mathbf{B}^H\mathbf{R}\mathbf{v}_0 \tag{9}$$

We note that the rank-reducing matrix $\mathbf{T}$ "reduces" the dimension of the linear filter (number of parameters to be designed) from P (filter $\mathbf{w}$) to p (filter $\mathbf{w}^{\mathrm{GSC}}$), $1 < p < P-1$. The p columns of the rank-reducing matrix $\mathbf{T}$ can be chosen in various ways. We can choose the p columns to be the eigenvectors that correspond to the P maximum eigenvalues of the *disturbance-only* autocorrelation matrix [27]. This choice is mean-square (MS) optimum under the assumption that the disturbance-only eigenvectors are not rotated by the

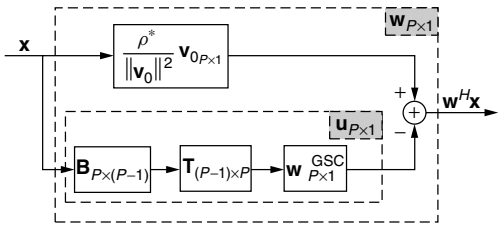


Figure 3. Generalized sidelobe canceler structure.

blocking matrix being used (i.e., when the disturbance subspace is orthogonal to the constraint vector $\mathbf{v}_0$), which is not valid in general. We can address this concern by choosing alternatively the p columns of $\mathbf{T}$ to be the eigenvectors that correspond to the p maximum eigenvalues of the *blocked* data autocorrelation matrix $\mathbf{B}^H\mathbf{R}\mathbf{B}$ [28,29]. If, however, the columns of the rank-reducing matrix $\mathbf{T}$ *have to be* eigenvectors of the blocked-data autocorrelation matrix (there is no documented technical optimality to this approach), then the best way in the minimum output variance sense is to choose the p eigenvectors $\mathbf{q}_i$ of $\mathbf{B}^H\mathbf{R}\mathbf{B}$ is to choose the p eigenvectors $\mathbf{q}_i$ with corresponding eigenvalues λ_i that maximize the ratio $\dfrac{|\mathbf{v}_0^H\mathbf{R}\mathbf{B}\mathbf{q}_i|^2}{\lambda_i}$, $i = 1, \ldots, p$ [30]. This algorithm is also known as "cross-spectral metric" reduced-rank processing [31]. Non-eigendecomposition-based alternatives for the synthesis of $\mathbf{u}$ include the auxiliary vector (AV) filters and the orthogonal multistage filters (also called "nested Wiener filters") [5,32–41].

3.3. Auxiliary Vector (AV) filters

Auxiliary vector (AV) filters are non-eigendecomposition-based filters that approximate the optimum MMSE/MVDR solution [5,32–38]. The AV algorithm is a statistical optimization procedure that generates a sequence of filters (AV filters). Each filter in the sequence has the general structure described in Fig. 2, where the vector $\mathbf{u}$ is approximated by a weighted sum of auxiliary vectors that maintain orthogonality *only* with respect to the distortionless direction $\mathbf{v}_0$ (and they are, in general, *nonorthogonal* to each other). The number of auxiliary vectors used to approximate the filter part $\mathbf{u}$ in Fig. 2 is increasing with the filter index in the sequence. Both the auxiliary vectors and the corresponding weights are subject to design (they are designed according to the maximum cross-correlation and minimum-variance criteria, respectively, as explained in detail below). An important characteristic of the AV algorithm (besides the nonorthogonality of the auxiliary vectors) is that it is a conditional optimization procedure; that is, each filter in the sequence is a function of the previously generated filter. Furthermore, AV filters do not require any explicit or implicit matrix inversion, eigendecomposition, or diagonalization. Finally, under ideal setups (perfect known input autocovariance matrix) the AV filter sequence converges to the MMSE/MVDR optimum solution, [33,34].

A pictorial presentation of generation of the sequence of AV filters is given by Fig. 4a. The sequence is initialized at the appropriately scaled constraint vector $\mathbf{w}_0 = \dfrac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, which is MMSE/MVDR optimum only when the vector inputs are white (i.e., when $\mathbf{R} = \sigma^2\mathbf{I}$, $\sigma > 0$). Next, we incorporate in $\mathbf{w}_0$ an "auxiliary" vector component $\mathbf{g}_1$ that is orthogonal to $\mathbf{v}_0$, and we form $\mathbf{w}_1 = \mathbf{w}_0 - \mu_1\mathbf{g}_1$, where $\mathbf{g}_1 \in \mathbb{C}^P - \{\mathbf{0}\}$, $\mu_1 \in \mathbb{C}$, and $\mathbf{g}_1^H\mathbf{v}_0 = 0$. We assume for a moment that the orthogonal auxiliary vector $\mathbf{g}_1$ is arbitrary but nonzero and fixed, and we concentrate on the selection of the scalar μ_1. The value of μ_1 that minimizes the variance of the output of the filter $\mathbf{w}_1$ can be found by direct differentiation of the variance $E\{|\mathbf{w}_1^H\mathbf{x}|^2\}$ or simply as the value that minimizes the MS error between $\mathbf{w}_0^H\mathbf{x}$ and $\mu_1^*\mathbf{g}_1^H\mathbf{x}$. This leads to $\mu_1 = \mathbf{g}_1^H\mathbf{R}\mathbf{w}_0/\mathbf{g}_1^H\mathbf{R}\mathbf{g}_1$.

Since $\mathbf{g}_1$ is set to be orthogonal to $\mathbf{v}_0$, the expression of μ_1 shows that if the vector $\mathbf{R}\mathbf{w}_0$ happens to be "on $\mathbf{v}_0$" [i.e., if $\mathbf{R}\mathbf{w}_0 = (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$ or equivalently $(\mathbf{I} - \mathbf{v}_0\mathbf{v}_0^H)\mathbf{R}\mathbf{w}_0 = \mathbf{0}$], then $\mu_1 = 0$. Indeed, if $\mathbf{R}\mathbf{w}_0 = (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$, then $\mathbf{w}_0$ is *already* the MMSE/MVDR filter. To avoid this trivial case and continue with our presentation, we assume that $\mathbf{R}\mathbf{w}_0 \neq (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$. By inspection, we also observe that for the MS-optimum value of μ_1 the product $\mu_1\mathbf{g}_1$ is independent of the norm of $\mathbf{g}_1$. Hence, so is $\mathbf{w}_1$. At this point, we set the auxiliary vector $\mathbf{g}_1$ to be a normalized vector that maximizes the magnitude of the cross-correlation between $\mathbf{w}_0^H\mathbf{x}$ and $\mathbf{g}_1^H\mathbf{x}$ (i.e., $\mathbf{g}_1 = \arg\max_{\mathbf{g}} |\mathbf{w}_0^H\mathbf{R}\mathbf{g}|$) subject to the constraint that $\mathbf{g}_1^H\mathbf{v}_0 = 0$ and $\mathbf{g}_1^H\mathbf{g}_1 = 1$. For the sake of mathematical accuracy, we note that both the criterion function $|\mathbf{w}_0^H\mathbf{R}\mathbf{g}|$ to be maximized as well as the orthogonality constraint are phase-invariant. Without loss of generality, to avoid any ambiguity in our presentation and to have a uniquely defined auxiliary vector, we choose the one and only auxiliary vector $\mathbf{g}_1$ that satisfies the maximization problem and places the cross-correlation value on the positive real line ($\mathbf{w}_0^H\mathbf{R}\mathbf{g}_1 > 0$).

The general inductive step is as follows. At step $k+1$, we define the AV filter $\mathbf{w}_{k+1} = \mathbf{w}_k - \mu_{k+1}\mathbf{g}_{k+1}$, where $\mathbf{g}_{k+1}$ and μ_{k+1} are to be *conditionally* optimized given the previously identified AV filter $\mathbf{w}_k$. The auxiliary vector $\mathbf{g}_{k+1}$ is chosen as the vector that maximizes the magnitude of the cross-correlation between the output of the previous filter $\mathbf{w}_k$ and the output of $\mathbf{g}_{k+1}$ (Fig. 4a), again subject to $\mathbf{g}_{k+1}$ being orthonormal to $\mathbf{v}_0$ *only* (we note that the choice of the norm does not affect the solution since $\mu_k\mathbf{g}_k$, $k = 1, 2, \ldots$, is $\mathbf{g}$-norm-invariant; we also emphasize that $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \mathbf{g}_4 \ldots$, *are not* necessarily orthogonal to each other). The value of μ_{k+1} minimizes the output variance of $\mathbf{w}_{k+1}$ given $\mathbf{w}_k$ and $\mathbf{g}_{k+1}$ (or equivalently minimizes the MS error between $\mathbf{w}_k^H\mathbf{x}$ and $\mu_{k+1}^*\mathbf{g}_{k+1}^H\mathbf{x}$). The solution for $\mathbf{g}_{k+1}$ and μ_{k+1} is given below, while the iterative algorithm for the generation of the infinite sequence of AV filters $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$ is presented in Fig. 4b (we note that in Fig. 4b we dropped the unnecessary normalization of $\mathbf{g}_1, \mathbf{g}_2, \ldots$ since $\mu_k\mathbf{g}_k$ is independent of the norm of $\mathbf{g}_k$):

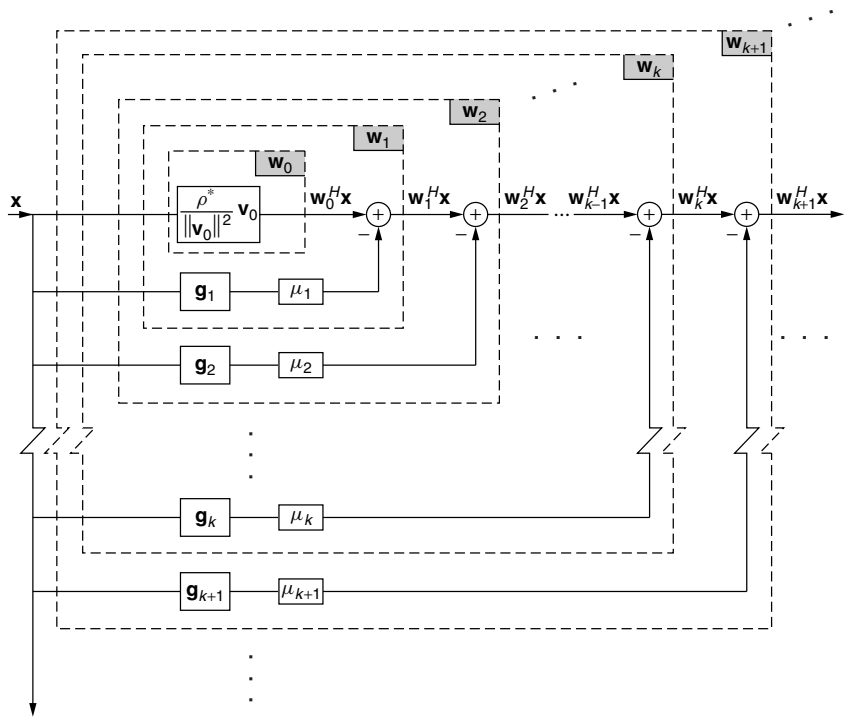
1. The scalar μ_{k+1} that minimizes the variance at the output of $\mathbf{w}_{k+1}$ or equivalently minimizes the MS error between $\mathbf{w}_k^H\mathbf{x}$ and $\mu_{k+1}^*\mathbf{g}_{k+1}^H\mathbf{x}$ is

$$\mu_{k+1} = \frac{\mathbf{g}_{k+1}^H\mathbf{R}\mathbf{w}_k}{\mathbf{g}_{k+1}^H\mathbf{R}\mathbf{g}_{k+1}}, \quad k = 0, 1, 2, \ldots \tag{10}$$

2. Suppose that $\left(\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\mathbf{w}_k \neq \mathbf{0}$ ($\mathbf{w}_k \neq \mathbf{w}_{\text{MVDR}}$). Then, the auxiliary vector

$$\mathbf{g}_{k+1} = \frac{\mathbf{R}\mathbf{w}_k - \dfrac{\mathbf{v}_0^H\mathbf{R}\mathbf{w}_k}{\|\mathbf{v}_0\|^2}\mathbf{v}_0}{\left\|\mathbf{R}\mathbf{w}_k - \dfrac{\mathbf{v}_0^H\mathbf{R}\mathbf{w}_k}{\|\mathbf{v}_0\|^2}\mathbf{v}_0\right\|}, \quad k = 0, 1, 2, \ldots \tag{11}$$

(**a**)



(**b**)

Auxiliary-Vector (AV) algorithm

Input:

Autocovariance matrix $\mathbf{R}$, constraint vector $\mathbf{v}_0$,

desired response $\mathbf{w}^H\mathbf{v}_0 = \rho$.

Initialization:

$$\mathbf{w}_0 := \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0.$$

Iterative computation:

For $k = 1, 2, \ldots$ do

begin

$$\mathbf{g}_k := \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\,\mathbf{w}_{k-1}$$

if $\mathbf{g}_k = 0$ then EXIT

$$\mu_k := \frac{\mathbf{g}_k^H\mathbf{R}\,\mathbf{w}_{k-1}}{\mathbf{g}_k^H\mathbf{R}\,\mathbf{g}_k}$$

$$\mathbf{w}_k := \mathbf{w}_{k-1} - \mu_k\mathbf{g}_k$$

end

Output:

Filter sequence $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$.

Figure 4. (**a**) Block diagram representation and (**b**) algorithmic description/generation of the auxiliary vector (AV) filter sequence $\mathbf{w}_1, \mathbf{w}_2, \ldots$.

maximizes the magnitude of the cross-correlation between $\mathbf{w}_k^H\mathbf{x}$ and $\mathbf{g}_{k+1}^H\mathbf{x}$ (which is equal to $|\mathbf{w}_k^H\mathbf{R}\mathbf{g}_{k+1}|$), subject to the constraints $\mathbf{g}_{k+1}^H\mathbf{v}_0 = 0$ and $\mathbf{g}_{k+1}^H\mathbf{g}_{k+1} = 1$. In addition, $\mathbf{w}_k^H\mathbf{R}\mathbf{g}_{k+1}$ is real positive ($\mathbf{w}_k^H\mathbf{R}\mathbf{g}_{k+1} > 0$).

With respect to the convergence of the filter sequence $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$ to the MVDR filter $\rho^*\dfrac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0}$, we can show that [33]

1. The generated sequence of auxiliary vector weights $\{\mu_k\}, k = 1, 2, \ldots,$ is real-valued, positive, and bounded: $0 < \dfrac{1}{\lambda_{\max}} \le \mu_k \le \dfrac{1}{\lambda_{\min}}$, $k = 1, 2, \ldots$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the corresponding maximum and minimum eigenvalues of $\mathbf{R}$
2. The sequence of auxiliary vectors $\{\mathbf{g}_k\}, k = 1, 2, \ldots,$ converges to the $\mathbf{0}$ vector: $\lim_{k\to\infty}\mathbf{g}_k = \mathbf{0}$

3. The sequence of AV filters $\{\mathbf{w}_k\}, k = 1, 2, \ldots,$ converges to the MVDR filter: $\lim_{k\to\infty} \mathbf{w}_k = \rho^* \dfrac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H \mathbf{R}^{-1}\mathbf{v}_0}$.

3.4. Orthogonal Multistage Filters

An alternative mechanism to approximate the optimum MMSE/MVDR solution can be obtained through the use of the orthogonal "multistage" filter decomposition procedure [39,40] (also called "nested Wiener filter"). It can be shown theoretically that the l-stage filter in [39,40], $\mathbf{w}_{l-\text{stage}}$, $0 \le l \le P-1$, is equivalent to the following structure. First, change the auxiliary vector generation recursion in (11) or Fig. 4b to impose orthogonality with respect not only to the constraint vector $\mathbf{v}_0$ but also to *all previously defined* auxiliary vectors that we denote now as $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{k-1}, k \le P-1$:

$$\mathbf{y}_k = \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2} - \sum_{i=1}^{k-1} \frac{\mathbf{y}_i\mathbf{y}_i^H}{\|\mathbf{y}_i\|^2}\right)\mathbf{R}\mathbf{w}_{k-1} \tag{12}$$

Next, terminate the recursion at $k = l$, $0 \le l \le P-1$, and organize the l (orthogonal to each other and to $\mathbf{v}_0$) vectors $\mathbf{y}_1, \ldots, \mathbf{y}_l$ in the form of a blocking matrix $\mathbf{B}_{P\times l} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_l]$. Then

$$\mathbf{w}_{l-\text{stage}} = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0 - \mathbf{B}_{P\times l}\vec{\boldsymbol{\alpha}}_{l\times 1} \tag{13}$$

where

$$\vec{\boldsymbol{\alpha}} = \frac{\rho^*}{\|\mathbf{v}_0\|^2}[\mathbf{B}^H\mathbf{R}\mathbf{B}]^{-1}\mathbf{B}^H\mathbf{R}\mathbf{v}_0 \tag{14}$$

is the MS *vector*-optimum set of weights of the vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_l$ [5,37]. We note that "vector-optimum" implies that the elements of the vector $\vec{\boldsymbol{\alpha}}$ (weights of the columns of $\mathbf{B}$) are designed/optimized jointly (and *not* in a conditional one-by-one manner). The multistage decomposition algorithm [39,40] is a computationally efficient procedure for the calculation of the weight vector $\vec{\boldsymbol{\alpha}}$ tailored to the particular structure of $\mathbf{B}^H\mathbf{R}\mathbf{B}$ (tridiagonal matrix); the calculation incorporates an implicit matrix inversion operation [in view of (14)]. The same computational savings can be achieved by the general forward calculation algorithm of Liu and Van Veen [42] that returns all intermediate stage filters along the way, up to the stage of interest l.

We conclude this section with a few comments on the relative merits and characteristics of the structures presented so far. From a general input space synthesis/decomposition point of view, the main distinguishing features of the AV algorithm with respect to the multistage algorithm are the *non-orthogonal* AV synthesis approach and the *conditional statistical optimization* procedure. Nonorthogonal synthesis allows the designer to grow an *infinite* sequence of AV filters on a "best effort" basis that takes into account the whole interference space at every step. Conditional optimization results in estimators that do not require any implicit or explicit matrix inversion or decomposition operation and, thus, plays a key role in developing superior adaptive filtering schemes in short-data-record environments, as illustrated in the next section.

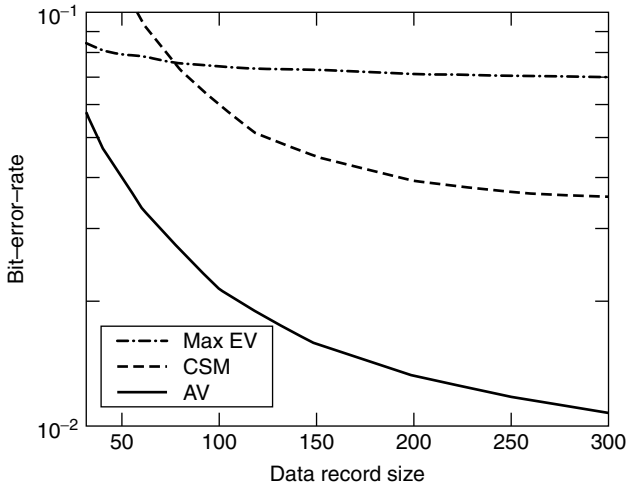


Figure 5. BER as a function of the data record size for the AV, "maximum eigenvector" (MaxEV), and "cross-spectral metric" (CSM) receivers of the same order (3 auxiliary vectors, 3 eigenvectors). *Operational environment*: synchronous DS-CDMA system, user of interest at 12 dB, 12 interferers at 10–14 dB, processing gain $L = 32$, and arbitrary normalized signatures (cross-correlation with the signature of the user of interest ~0.2).

Figure 5 presents an illustrative example of the relative merits of various receiver designs in terms of BER. The example is based on a simple single-path synchronous DSCDMA signal model ($P = L$). The BER performance of the receiver $\mathbf{w}_3$ (that utilizes three auxiliary vectors $\mathbf{g}_1, \mathbf{g}_2$, and $\mathbf{g}_3$) is compared with the BER performance of the "maximum eigenvector" (Max EV) [28,29] receiver and the "cross-spectral-metric" (CSM) [30,31] receiver (both of which use three eigenvectors). As a numerical example that illustrates the convergence of the AV-filter sequence to the ideal MMSE/MVDR solution under perfectly known (ideal) autocorrelation matrix $\mathbf{R}$, in Fig. 6 we plot the squared norm error between the AV filter of the user of interest $\mathbf{w}_k$ and $\mathbf{w}_{\text{MMSE/MVDR}}$ as a function of k (i.e., the number of auxiliary vectors used or equivalently the index of the AV filter in the sequence).

4. ADAPTIVE FILTER ESTIMATION

4.1. Known Channel

4.1.1. SMI, GSC, Auxiliary Vector, and Multistage Estimators. We recall that the MMSE/MVDR filter is a function of the *true* input autocorrelation matrix $\mathbf{R}$ and the *true* constraint vector, $\mathbf{v}_0$. However, in almost every practical adaptive filtering application neither $\mathbf{R}$ nor $\mathbf{v}_0$ is known to the receiver. In this section (4.1) we present various estimates of the optimum MMSE/MVDR filter when $\mathbf{R}$ is unknown and sample average estimated from a data packet (data record) of size J, that is, $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$:

$$\hat{\mathbf{R}}(J) = \frac{1}{J}\sum_{j=0}^{J-1} \mathbf{x}_j\mathbf{x}_j^H \tag{15}$$

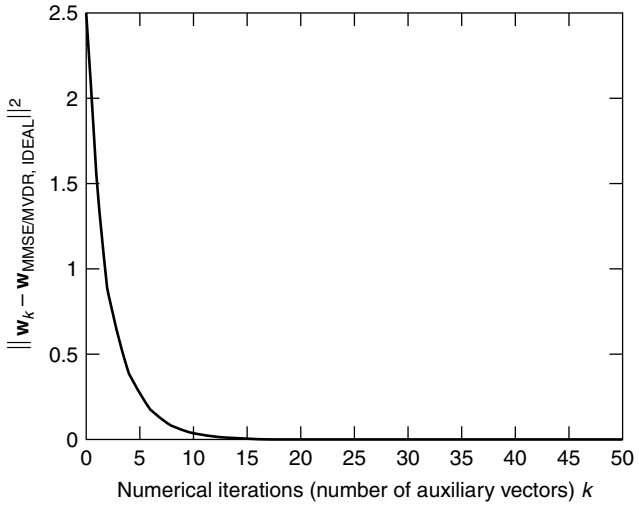


Figure 6. Convergence of the AV filter sequence to the ideal MMSE/MVDR solution as a function of the number of iterations k in Fig. 4b. The sequence of *conditionally optimized* AV filters (that utilize *nonorthogonal* auxiliary vectors) converges to the $\mathbf{w}_{\mathrm{MMSE/MVDR}}$ optimum solution for a perfectly known input autocorrelation matrix $\mathbf{R}$.

Throughout this (4.1) section, $\mathbf{v}_0$ is assumed to be known (since $\mathbf{v}_0$ is a function of the channel parameters we label this section as the "known channel" case); a procedure for the estimation of $\mathbf{v}_0$ from the same data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$ will be presented in the Section 4.2 ("unknown channel").

When $\mathbf{R}$ is unknown, the most widely used MMSE/MVDR filter estimator is obtained from (6) by using the sample average estimate $\hat{\mathbf{R}}(J)$ in place of $\mathbf{R}$. This estimator is known as the sample matrix inversion (SMI) filter. If we choose to work with the approximate solutions presented in Section 3 and utilize the sample average estimate of the autocorrelation matrix $\hat{\mathbf{R}}(J)$ instead of $\mathbf{R}$ in Eqs. (9)–(14), we obtain a GSC, AV, or multistage-type estimator of the MMSE/MVDR solution, respectively. We note that, for Gaussian inputs, $\hat{\mathbf{R}}(J)$ is a maximum-likelihood (ML), consistent, unbiased estimator of $\mathbf{R}$. On the other hand, the inverse of $\hat{\mathbf{R}}(J)$, which is utilized explicitly by the SMI filter and implicitly by both the GSC and the orthogonal multistage decomposition estimator, is not always defined. We can guarantee (with probability 1) that $\hat{\mathbf{R}}(J)$ is invertible only when the number of observations J is greater than or equal to the dimension of the input space (or filter dimension) P and the input distribution belongs to a specific class of multivariate elliptically contoured distributions that includes the Gaussian [43–46]. On the basis of the convergence properties of the AV filter sequence discussed in the previous section we can show that the corresponding sequence of AV filter estimates $\hat{\mathbf{w}}_k(J)$ converges, as $k \to \infty$, to the SMI filter [33]:

$$\hat{\mathbf{w}}_k(J) \xrightarrow[k\to\infty]{} \hat{\mathbf{w}}_\infty(J) = \hat{\mathbf{w}}_{\mathrm{SMI}} = \rho^* \frac{[\hat{\mathbf{R}}(J)]^{-1}\mathbf{v}_0}{\mathbf{v}_0^H[\hat{\mathbf{R}}(J)]^{-1}\mathbf{v}_0} \tag{16}$$

4.1.2. Properties of the Sequence of AV Estimators. The AV filter sequence of estimators begins with $\hat{\mathbf{w}}_0(J) = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, which is a zero-variance, fixed-value, estimator that may be severely biased ($\hat{\mathbf{w}}_0(J) \neq \mathbf{w}_{\mathrm{MMSE/MVDR}}$) unless the input is white (i.e., $\mathbf{R} = \sigma^2\mathbf{I}$, for some $\sigma > 0$). In the latter trivial case, $\hat{\mathbf{w}}_0(J)$ is already the perfect MMSE/MVDR filter. Otherwise, the next filter estimator in the sequence, $\hat{\mathbf{w}}_1(J)$, has a significantly reduced bias due to the optimization procedure employed, at the expense of nonzero estimator (co)variance. As we move up in the sequence of filter estimators $\hat{\mathbf{w}}_k(J)$, $k = 0, 1, 2, \ldots$, the bias decreases rapidly to zero[1] while the variance rises slowly to the SMI [$\hat{\mathbf{w}}_\infty(J)$] levels [cf. (16)]. To quantify these remarks, we plot in Fig. 7 the norm-square bias $\|E\{\hat{\mathbf{w}}_k(J)\} - \mathbf{w}_{\mathrm{MMSE/MVDR}}\|^2$ and the trace of the covariance matrix $E\{[\hat{\mathbf{w}}_k(J) - E\{\hat{\mathbf{w}}_k(J)\}][\hat{\mathbf{w}}_k(J) - E\{\hat{\mathbf{w}}_k(J)\}]^H\}$ as a function of the iteration step (filter index) k, for the same signal model as in Fig. 6 and data packet (data record) size $J = 256$. Bias and covariance trace values are calculated from 100,000 independent filter estimator realizations for each iteration point k; that is, we generate 100,000 independent data packets (J received random vectors per packet). For each packet we evaluate $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \ldots$. Then, we evaluate expectations as sample averages over 100,000 data packets.

Formal, theoretical statistical analysis of the generated estimators $\hat{\mathbf{w}}_k(J), k = 0, 1, 2, \ldots$, is beyond the scope of this presentation. We do note, however, that for multivariate Gaussian input distributions, an analytic expression for the covariance matrix of the SMI estimator $\hat{\mathbf{w}}_\infty(J)$ can be found in [46]

$$E\{[\hat{\mathbf{w}}_\infty(J) - E\{\hat{\mathbf{w}}_\infty(J)\}][\hat{\mathbf{w}}_\infty(J) - E\{\hat{\mathbf{w}}_\infty(J)\}]^H\} = \frac{|\rho|^2}{(\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0)(J-P+1)}\left(\mathbf{R}^{-1} - \frac{\mathbf{R}^{-1}\mathbf{v}_0\mathbf{v}_0^H\mathbf{R}^{-1}}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0}\right) \tag{17}$$

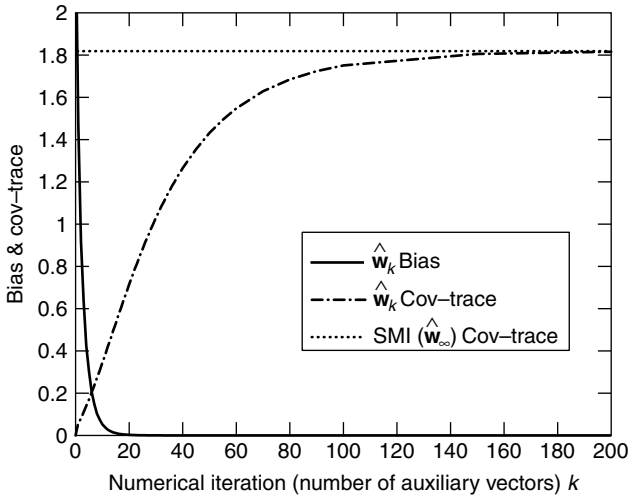


Figure 7. Norm-square bias and covariance trace for the sequence of estimators $\hat{\mathbf{w}}_k(J), k = 0, 1, \ldots$. The signal model is as in Fig. 6; data record size $J = 256$.

[1] The SMI estimator is unbiased for multivariate elliptically contoured input distributions [46,47]: $E\{\hat{\mathbf{w}}_\infty(J)\} = \mathbf{w}_{\mathrm{MMSE/MVDR}} = \rho^*\frac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0}$.

Since under these input distribution conditions $\hat{\mathbf{w}}_\infty(J)$ is unbiased, the trace of the covariance matrix is the MS filter estimation error. It is important to observe that the covariance matrix and, therefore, the MS filter estimation error depend on the data record size J, the filter length P, as well as the specifics of the signal processing problem at hand (actual $\mathbf{R}$ and $\mathbf{v}_0$). It is also important to note that when the input distribution is not Gaussian (e.g., for the CDMA signal model example considered earlier in our discussion, the input is Gaussian-mixture-distributed), then the analytic result in (17) is not directly applicable and can be thought of as only an approximation (a rather close approximation for DSCDMA systems). From the results in Fig. 7 for $J = 256$, we see that the estimators $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \ldots,$ up to about $\hat{\mathbf{w}}_{20}(J)$ are particularly appealing. In contrast, the estimators $\hat{\mathbf{w}}_k(J)$ for $k > 20$ do not justify their increased covariance trace cost since they have almost nothing to offer in terms of further bias reduction.

We emphasize that since the AV filters $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots$ can be considered as approximations of the MMSE/MVDR optimum filter under idealset-ups, the AV-filter estimates $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \hat{\mathbf{w}}_3(J), \ldots$ have been viewed so far not only as estimates of the filters $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots$ but also, and most importantly, as estimates of the MMSE/MVDR optimum filter in (6) and (7). In this context, the mean-square estimation error expression $E\{\|\hat{\mathbf{w}}_k(J) - \mathbf{w}_{\mathrm{MMSE/MVDR}}\|^2\}$ captures the bias/variance balance of the individual members of the estimator sequence $\hat{\mathbf{w}}_k(J)$, $k = 0, 1, 2, \ldots$. In Fig. 8 we plot the MS estimation error as a function of the iteration step k (or filter index) for the same signal model as in Fig. 6, for $J = 256$ [part (a)] and $J = 2048$ [part (b)]. As a reference, we also include the MS-error of the constraint LMS estimator and the RLS estimator.

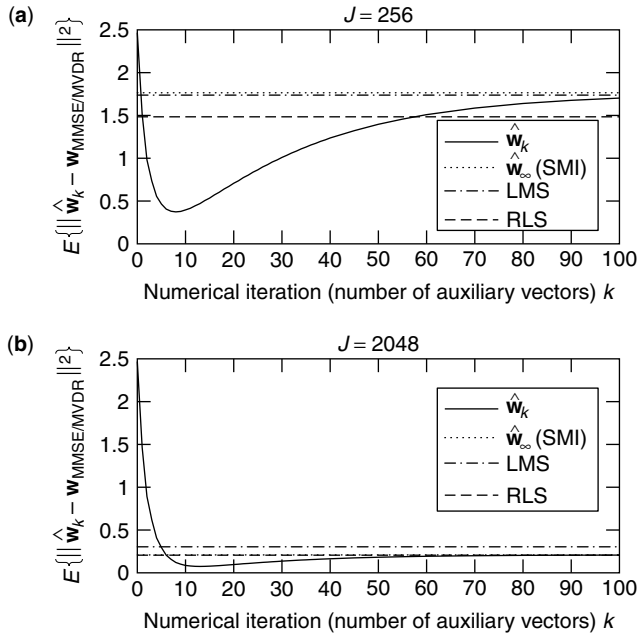


Figure 8. MS estimation error for the sequence of estimators $\hat{\mathbf{w}}_k(J)$, $k = 0, 1, \ldots$: (**a**) data record size $J = 256$; (**b**) $J = 2048$.

The constraint LMS estimator is given by the following recursion:

$$\hat{\mathbf{w}}_{\mathrm{LMS}}(j) = \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\left[\hat{\mathbf{w}}_{\mathrm{LMS}}(j-1) - \mu\mathbf{x}_j\mathbf{x}_j^H\hat{\mathbf{w}}_{\mathrm{LMS}}(j-1)\right] + \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0, \quad j = 1, \ldots, J \tag{18}$$

with $\hat{\mathbf{w}}_{\mathrm{LMS}}(0) = \dfrac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$ and some $\mu > 0$. The recursion of the RLS estimator can be obtained from the SMI formula in (16) by utilizing the following iterative estimation of $\mathbf{R}^{-1}$ that is based on the matrix inversion lemma:

$$\hat{\mathbf{R}}^{-1}(j) = \hat{\mathbf{R}}^{-1}(j-1) - \frac{\hat{\mathbf{R}}^{-1}(j-1)\mathbf{x}_j\mathbf{x}_j^H\hat{\mathbf{R}}^{-1}(j-1)}{1 + \mathbf{x}_j^H\hat{\mathbf{R}}^{-1}(j-1)\mathbf{x}_j}, \quad j = 1, \ldots, J \tag{19}$$

where $\hat{\mathbf{R}}^{-1}(0) = \dfrac{1}{\varepsilon_0}\mathbf{I}$ for some $\varepsilon_0 > 0$. Theoretically, the LMS gain parameter $\mu > 0$ has to be less than $\dfrac{1}{2 \cdot \lambda_{\max}^{\mathrm{blocked}}}$, where $\lambda_{\max}^{rmblocked}$ is the maximum eigenvalue of the "blocked data" autocorrelation matrix $\left(\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\left(\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)$. While this is a theoretical upper bound, practitioners are well aware that empirical, data-dependent "optimization" or "tuning" of the LMS gain $\mu > 0$ or the RLS initialization parameter $\varepsilon_0 > 0$ is necessary to achieve acceptable performance (in our study we set $\mu = \dfrac{1}{200 \cdot \lambda_{\max}^{\mathrm{blocked}}}$ and $\varepsilon_0 = 20$, respectively) [8,9,48–51]. This data-specific tuning frequently results in misleading, overoptimistic conclusions about the short-data-record performance of the LMS/RLS algorithms. In contrast, when the AV filter estimators $\hat{\mathbf{w}}_k(J)$ generated by the algorithm of Fig. 4b are considered, tuning of the real-valued parameters μ and ε_0 is virtually replaced by an integer choice among the first several members of the $\{\hat{\mathbf{w}}_k(J)\}$ sequence. In Fig. 8a, for $J = 256$ all estimators $\hat{\mathbf{w}}_k(J)$ from $k = 2$ up to about $k = 55$ outperform in mean-square error (MSE) or their RLS, LMS, and SMI $[\hat{\mathbf{w}}_\infty(J)]$ counterparts. $\hat{\mathbf{w}}_8(J)$ ($k = 8$ auxiliary vectors) has the least MSE of all (best bias/variance tradeoff). When the data record size is increased to $J = 2048$ (Fig. 8b), we can afford more iterations and $\hat{\mathbf{w}}_{13}(J)$ offers the best bias/variance tradeoff (lowest MSE). All filter estimators $\hat{\mathbf{w}}_k(J)$ for $k > 8$ outperform the LMS/RLS/SMI $[\hat{\mathbf{w}}_\infty(J)]$ estimators. For such large data record sets ($J = 2048$), the RLS and the SMI $[\hat{\mathbf{w}}_\infty(J)]$ MSE are almost identical. Figure 9 offers a three-dimensional plot of the mean-square estimation error as a function of the sample support J used in forming $\hat{\mathbf{w}}_k(J)$ and the number of auxiliary vectors k (or filter index). The dark line that traces the bottom of the MS estimation error surface identifies the best number of auxiliary vectors (or the index of the best filter) for any given data record size J.

4.1.3. How to Choose the Best AV Estimator. We recall that, when the autocovariance matrix is sample-average-estimated, the sequence of AV estimators converges

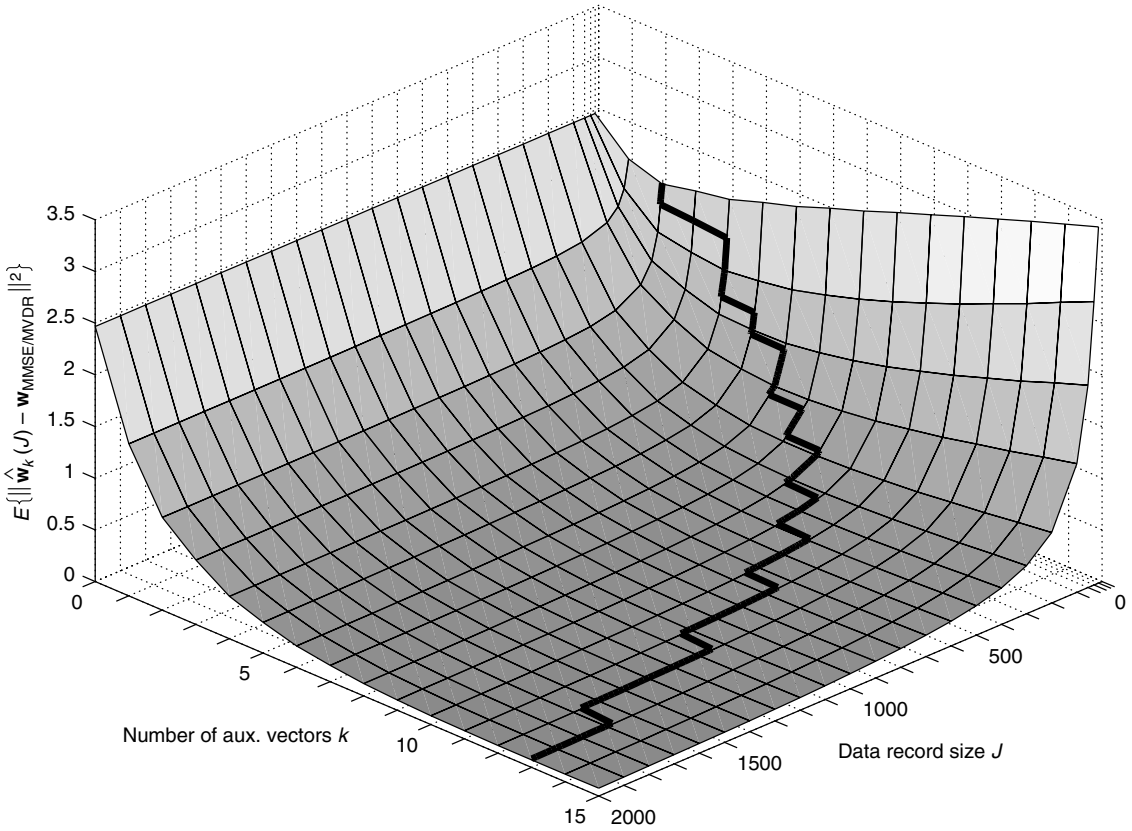


Figure 9. MS estimation error versus number of auxiliary vectors k and sample support J (the signal model is the same as in Fig. 6).

to the SMI filter. Evidently, the early, nonasymptotic elements of the sequence offer favorable bias/variance balance characteristics and outperform in mean-square filter estimation error, as we have seen in Figs. 7–9, the unbiased SMI filter estimator as well as the (constraint) LMS, RLS. We will later see that they also outperform the orthogonal multistage decomposition, and diagonally loaded (DL) SMI filter estimators. In the context of digital wireless communication receivers, superior mean-square filter estimation error translates to superior BER performance under short-data-record receiver adaptation. Selecting the most successful (in some appropriate sense) AV estimator in the sequence for a given data record is a critical problem. Below we present two data-dependent selection criteria [52,53]. The first criterion minimizes the cross-validated sample average variance of the AV filter output and can be applied to general filter estimation problems; the second criterion maximizes the estimated $\mathcal{J}$ divergence of the AV filter output conditional distributions and is tailored to general hypothesis testing (detection) applications.

In particular, the *cross-validated minimum output variance* (CVMOV) rule is motivated by the fact that minimization of the output variance of filters that are constrained to be distortionless in the vector direction of a signal of interest is equivalent to maximization of the output SINR. Cross-validation is a well-known statistical method. In the context of AV filtering, cross-validation is used to select the filter parameter of interest (number of auxiliary vectors k) that minimizes the output variance, which is estimated on the basis of the observations (training data) that have not been used in the process of building the filter itself. A particular case of this general method used in this presentation is the "leave one out" method [54]. The following criterion outlines the CVMOV AV filter selection process.

Criterion 1. For a given data packet (data record) of size J, the cross-validated minimum-output-variance AV filter selection rule chooses the AV filter estimator $\hat{\mathbf{w}}_{k_1}(J)$ that minimizes the cross-validated sample average output variance:

$$k_1 = \arg\min_k \left\{ \sum_{j=1}^{J} \hat{\mathbf{w}}_k(J\backslash j)^H \mathbf{x}_j \mathbf{x}_j^H \hat{\mathbf{w}}_k(J\backslash j) \right\} \tag{20}$$

where $(J\backslash j)$ identifies the AV filter estimator that is evaluated from the available data record after removing the jth sample.

While the CVMOV criterion can be applied to general filter estimation problems, the second criterion, the *maximum $\mathcal{J}$-divergence* criterion, is tailored to applications that can be formulated as binary hypothesis testing problems on AV-filtered data. For any scalar binary hypothesis testing problem, if f_0 and f_1 denote the conditional distributions of the detector input under

hypothesis H_0 and H_1, respectively, then the $\mathcal{J}$-divergence distance between f_0 and f_1 is defined as the sum of the Kullback–Leibler (KL) distances between f_0 and f_1 [55]

$$\mathcal{D}(f_0, f_1) \triangleq \mathcal{KL}(f_1, f_0) + \mathcal{KL}(f_0, f_1) \tag{21}$$

where the KL distance of f_1 from f_0 is defined as $\mathcal{KL}(f_1, f_0) \triangleq \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_0(x)}\, dx$.

The choice of the output $\mathcal{J}$ divergence as one of the underlying rules for the selection of the AV filter is motivated by the fact that the probability of error of the optimum (Bayesian) detector for any scalar binary hypothesis testing problem is lower bounded by

$$P_e \geq \pi_0 \pi_1 \exp\left\{\frac{-\mathcal{D}(f_0, f_1)}{2}\right\} \tag{22}$$

where π_0 and π_1 are the a priori probabilities of H_0 and H_1, respectively. The right-hand side of (22) is a monotonically decreasing function of the $\mathcal{J}$ divergence between the conditional distributions of the detector input. When the conditional distributions under H_0 and H_1 are Gaussian with the same variance, (22) is satisfied with equality. The latter implies that the larger the $\mathcal{J}$ divergence, the smaller the probability of error or, equivalently, the larger the $\mathcal{J}$ divergence, the easier the detection problem. Thus, maximization of the $\mathcal{J}$ divergence implies minimization of the probability of error. Because of the above mentioned properties and their relationship to the probability of error of the optimum detector, $\mathcal{J}$ divergence has been extensively used in the detection literature as a hypothesis discriminant function. In the context of AV filtering, we denote the AV scalar filter output conditional distributions under H_0 and H_1 by $f_{0,k}(\cdot)$ and $f_{1,k}(\cdot)$, respectively, where the index k indicates the dependence of the distributions on the specific AV filter $\hat{\mathbf{w}}_k$ used from the available sequence $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots$. Then, the $\mathcal{J}$ divergence between $f_{0,k}(\cdot)$ and $f_{1,k}(\cdot)$ is also a function of k; for this reason, in the rest of our presentation it will be denoted as $\mathcal{D}(k)$. To the extent that the conditional distributions of the AV filter output under H_0 and H_1 are approximated by Gaussian distributions with opposite means and equal variances (which is a reasonable, in general, assumption), we can show in a straightforward manner that

$$\mathcal{D}(k) \approx \frac{4E^2\{b_0 \mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}]\}}{\mathrm{Var}\{b_0 \mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}]\}} \tag{23}$$

where $\mathrm{Var}(\cdot)$ denotes variance. The following criterion outlines the $\mathcal{J}$-divergence AV-filter selection process.

Criterion 2. For a given data packet (data record) of size J, the $\mathcal{J}$ divergence AV filter selection rule chooses the AV filter estimator $\hat{\mathbf{w}}_{k_2}(J)$ that maximizes the estimated $\mathcal{J}$ divergence $\hat{\mathcal{D}}(k)$ between the AV filter output conditional distributions:

$$k_2 = \arg\max_k \{\hat{\mathcal{D}}(k)\} \tag{24}$$

If we substitute b_0 in (23) by $\hat{b}_0 = \mathrm{sgn}(\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}])$ and evaluate expectations by sample averaging, then we can obtain a blind estimate of the $\mathcal{J}$ divergence:

$$\hat{\mathcal{D}}_B(k) = \frac{4\left[\frac{1}{J}\sum_{j=1}^{J} |\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}_j]|\right]^2}{\frac{1}{J}\sum_{j=1}^{J} |\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}_j]|^2 - \left[\frac{1}{J}\sum_{j=1}^{J} |\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}_j]|\right]^2} \tag{25}$$

where the subscript B identifies the blind version of the $\mathcal{J}$-divergence function. Then, we can evaluate $k_2 = \arg\max_k \{\hat{\mathcal{D}}_B(k)\}$. We recall that in (25) $\mathbf{x}$ denotes the received signal vector of the general form $\mathbf{x} = b_0\sqrt{E_0}\mathbf{v}_0 + \mathbf{z}$ where [cf. (2)] $\mathbf{v}_0 \in \mathbb{C}^P$ is a known deterministic signal vector, $E_0 > 0$ represents the unknown energy scalar, $\mathbf{z} \in \mathbb{C}^P$ is a zero-mean disturbance vector (i.e., it may incorporate ISI, MAI, and additive noise effects), and b_0 is $+1$ or -1 with equal probability. We also recall [cf. (3)] that the decision on $H_0(b_0 = -1)$ or $H_1(b_0 = +1)$ is based on the real part of the AV filter output $\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}]$, where $\hat{\mathbf{w}}_k(J)$ is the AV estimator that utilizes k auxiliary vectors.

4.1.4. Properties of the Multistage and the DL-SMI Estimators. A *finite set* of P filter estimators with varying bias/variance balance can be obtained through the use of the orthogonal "multistage" filter decomposition procedure [40]. In the context of filter estimation from a data record of size J, $\hat{\mathbf{w}}_{0-\mathrm{stage}}(J)$ is the matched filter and $\hat{\mathbf{w}}_{(P-1)-\mathrm{stage}}(J)$ is the SMI estimator. In Fig. 10b we plot the MS estimation error of $\hat{\mathbf{w}}_{l-\mathrm{stage}}(J)$ as a function of l, $0 \leq l \leq P-1 = 31 (J = 60)$. We identify the *best* multistage estimator ($l = 3$ stages), and in Fig. 10c we compare against the AV estimator sequence. We see that all AV estimators $\hat{\mathbf{w}}_k(J)$ from $k = 3$ to 8 outperform in MSE the best multistage estimator ($l = 3$ stages).

An alternative bias/variance trading mechanism through real-valued tuning is the diagonally-loaded (DL) SMI estimator obtained by adding an amount (Δ) to each element of the diagonal of $\hat{\mathbf{R}}(J)$ in the SMI formula (16) [56] [in this way we ensure the invertibility of $\hat{\mathbf{R}}(J)$]

$$\hat{\mathbf{w}}_{\mathrm{DL-SMI}}(\Delta) = \rho^* \frac{[\hat{\mathbf{R}}(J) + \Delta\mathbf{I}]^{-1}\mathbf{v}_0}{\mathbf{v}_0^H[\hat{\mathbf{R}}(J) + \Delta\mathbf{I}]^{-1}\mathbf{v}_0} \tag{26}$$

where $\Delta \geq 0$ is the diagonal loading parameter. We observe that $\hat{\mathbf{w}}_{\mathrm{DL-SMI}}(\Delta = 0)$ is the regular SMI estimator, while $\lim_{\Delta\to\infty} \hat{\mathbf{w}}_{\mathrm{DL-SMI}}(\Delta) = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, which is the properly scaled matched filter. In Fig. 10a we plot the MS estimation error of the DL-SMI estimator as a function of the diagonal loading parameter $\Delta (J = 60)$. We identify the *best possible* diagonal loading value $\Delta \simeq 3.45$ (at significant computational cost), and in Fig. 10c we compare the best DL-SMI estimator against the AV estimator sequence for which *no* diagonal loading is performed. Interestingly, $\hat{\mathbf{w}}_k(J)$ from $k = 4$–7 outperform in MSE the best possible DL-SMI estimator ($\Delta \simeq 3.45$). In Fig. 11 we plot the MS error of the $\Delta = 3.45$ DL-SMI estimator together with the MS error of the *best* multistage and AV estimators over the data support

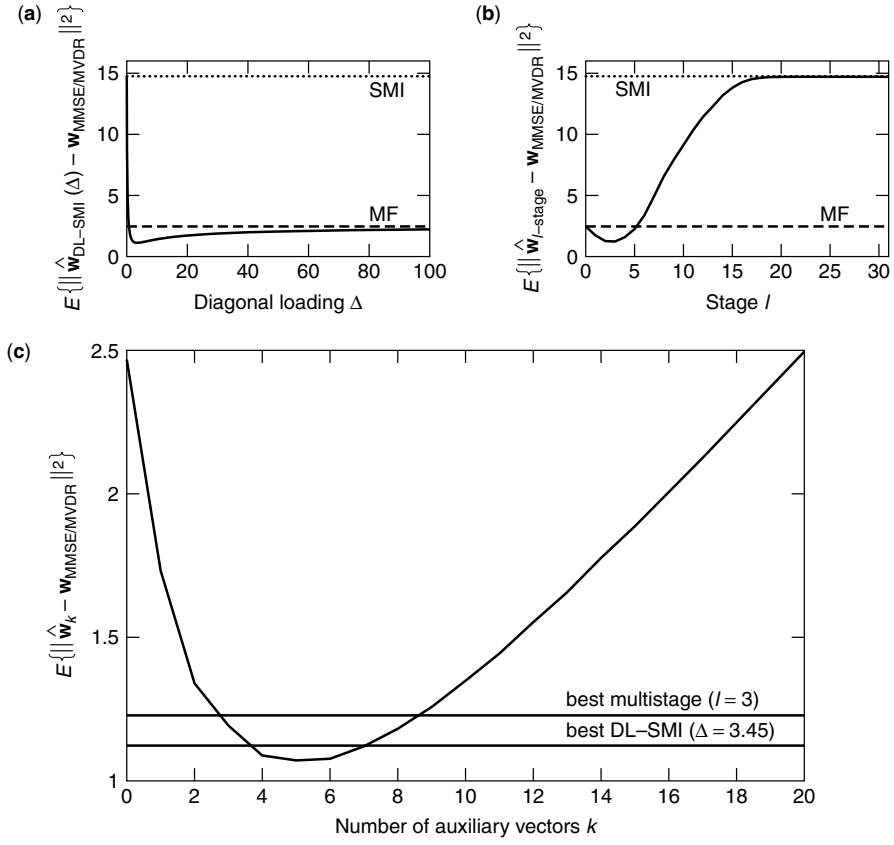


Figure 10. MS estimation error studies for (**a**) diagonally loaded SMI, (**b**) multistage (also called "nested Wiener"), and (**c**) auxiliary vector estimators (the signal model is the same as in Fig. 6 and $J = 60$).

(packet size) range $J = P/2 = 16$ to $J = 3P = 96$. The total computational complexity of the multistage algorithm is of order $O((J + l)P^2)$. The AV algorithm has computational complexity $O((J + k)P^2)$, where k is the desired number of auxiliary vectors $\mathbf{g}_1, \ldots, \mathbf{g}_k$. All intermediate AV filters are returned. The computational complexity of DL-SMI is of order $O((J + P)P^2)$. Estimators of practical interest have $l \ll J$ or $k \ll J$. Therefore, the complexity of all algorithms is dominated by $O(JP^2)$, which is required for the computation of $\hat{\mathbf{R}}(J)$ (the computational complexity of the RLS estimator is, similarly, of the order $O(JP^2)$, the complexity of the GSC estimator is $O(JP^2 + P^3)$, while the complexity of LMS estimator — with no data recycling — is of order $O(JP)$). In terms of performance, the *explicit* or *implicit* matrix inversion that the SMI and the multistage algorithm (at any given stage) performs, respectively, affects adversely their behavior under short-data-record adaptation.

4.1.5. Performance Illustrations. We illustrate the overall short-data-record adaptive filter performance in Figs. 12 and 13 for a multipath fading DSCDMA system that employs antenna array reception. We consider processing gain 31, 20 users, 5 antenna elements, and 3 resolvable multipaths per user with independent zero-mean complex Gaussian fading coefficients of variance 1. The maximum cross-correlation between the assigned user signatures reaches 30%. The total SNR's (over the three paths) of the 19 interferers are set at $\mathrm{SNR}_{2-6} =$ 6 dB, $\mathrm{SNR}_{7-8} = 7$ dB, $\mathrm{SNR}_{9-13} = 8$ dB, $\mathrm{SNR}_{14-15} = 9$ dB, $\mathrm{SNR}_{16-20} = 10$ dB. The spacetime product (filter length) is

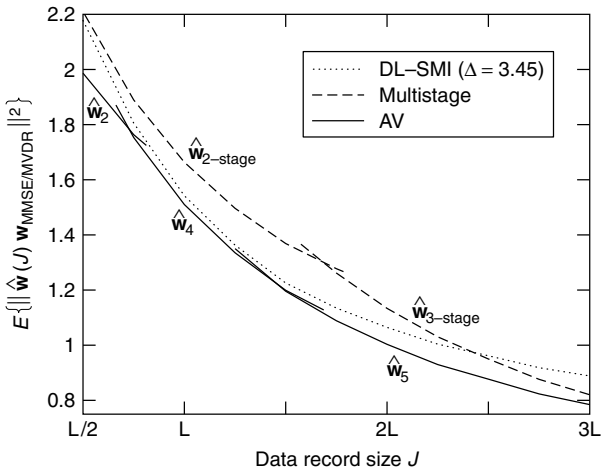


Figure 11. MS estimation error for the *best* multistage and AV estimators over the data support range $J = P/2 = 16$ to $J = 3P = 96$. The MS estimation error of the $\Delta = 3.45$ DL-SMI estimator is also included as a reference (the signal model is the same as in Fig. 6).

$P = (31 + 2)5 = 165$. The experimental results are averages over 1000 runs (100 different channel realizations and 10 independent data record generations per channel). In Fig. 12, we plot the BER[2] of the AV estimators

[2] The BER of each filter under consideration is approximated by $Q(\sqrt{\mathrm{SINR}_{\mathrm{out}}})$, since the computational complexity of the BER

$\hat{\mathbf{w}}_{k_1}(J)$ and $\hat{\mathbf{w}}_{k_2}(J)$ as a function of the SNR of the user of interest for data records of size $J = 230$. We also plot the BER curve of the "genie" assisted BER-optimum filter $\hat{\mathbf{w}}_{k_{\text{opt}}}(J)$ as well as the corresponding curves of the ideal MMSE/MVDR filter $\mathbf{w}_{\text{MMSE/MVDR}}$, the SMI filter estimator $\hat{\mathbf{w}}_{\infty}(J)$, the S-T RAKE matched-filter (MF) $\hat{\mathbf{w}}_0(J)$, and the multistage filter (with the preferred number of stages[3] $l = 7$). We observe that both $\hat{\mathbf{w}}_{k_1}(J)$ and $\hat{\mathbf{w}}_{k_2}(J)$ are very close to the "genie" BER-optimum AV filter estimator choice and outperform significantly the SMI filter estimator, the multistage filter estimator, and the matched filter. We also observe that for moderate to high SNR of the user of interest, the $\mathcal{J}$-divergence selection rule is slightly superior to the CVMOV selection rule. Figure 13 repeats the study of Fig. 12 as a function of the data record size. The SNR of the user of interest is fixed at 8 dB.

Concluding our discussion in this section (4.1), we note that the key for a successful solution (in the sense of superior filter output SINR or BER performance) to the problem of adaptive receiver design under limited data support is to employ receivers with varying bias/variance characteristics and to effectively control these characteristics in a data-driven manner. For this reason, operations and filter design/optimization criteria that suffer from "data starvation" (e.g., implicit or explicit matrix inversion and/or eigendecomposition) should be avoided. From a general input space synthesis/decomposition point of view, the *nonorthogonal* synthesis and the conditional statistical optimization are two principles that allow to grow an *infinite* sequence of AV filters on a "best effort" basis that takes into account the whole interference space at each step. These two features play a key role, leading to superior adaptive filtering performance in short-data-record environments. In particular, under

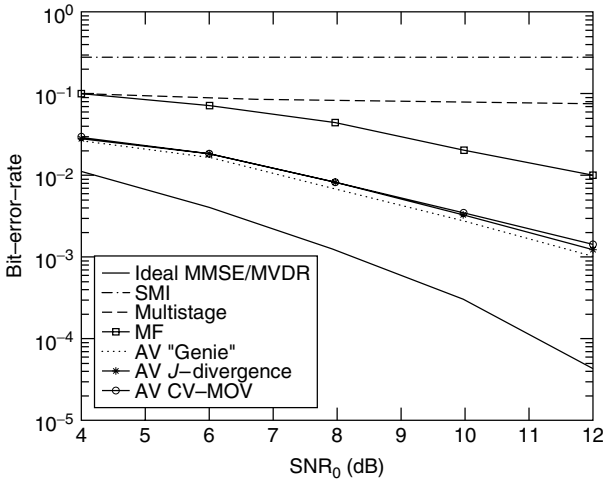


Figure 12. BER versus SNR for the user signal of interest for a multipath fading antenna array received signal model ($L = 31$, $K = 20, M = 5, N = 3$) with $P = 165$ and $J = 230$.

expression for this antenna-array CDMA system prohibits exact analytic evaluation [57].

[3] Honig and Xiao [41] argued that $l = 7$ ($D = 8$ in their notation [41]) stages is "nearly optimal over a wide range of loads and SNRs."

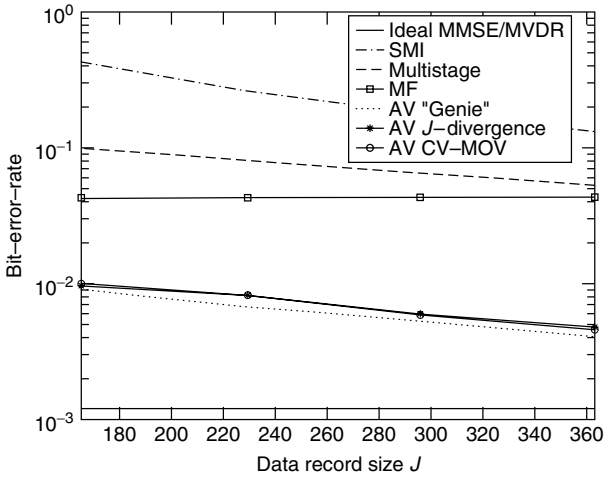


Figure 13. BER versus data record size (the signal model is the same as in Fig. 12, and $\text{SNR}_0 = 8$ dB).

short-data-record adaptation, the early, non-asymptotic elements of the sequence of AV estimators are mildly biased but exhibit much lower variance than other alternatives (for digital communications applications, the latter implies superior BER performance). As the available data record increases we can afford to go higher and higher in the sequence of generated estimators. In the limit, if we are given infinitely many input data, we can go all the way up to the convergence point of the algorithm, which is the ideal MMSE/MVDR receiver. The significant role of conditional statistical optimization in short-data-record adaptive filtering is evident even when *orthogonal* vectors are utilized. For example, it has been seen that the nonconditional (vector-optimum) scheme that utilizes orthogonal vectors with vector-optimum weights (12)–(14) (or, equivalently, the filter obtained by the algorithm of Goldstein et al. [40] that performs implicitly a matrix inversion) exhibits inferior short-data-record performance than does the structure that utilizes orthogonal vectors and conditionally optimized weights [5]. As a few concluding notes, an online version of the AV algorithm has been presented [58,59]. Application of AV filtering to the problem of rapid synchronization and combined demodulation of DSCDMA signals has been considered [60–62]. Detailed results on data record size requirements to achieve a given output SINR (or BER) performance level can be found in earlier studies [63,64].

4.2. Unknown Channel

The second part of this section is devoted to the estimation of the channel-processed constraint vector $\mathbf{v}_0$ from the same data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$. To be consistent with our discussion and illustrative studies presented in the previous sections, we consider the general case of ST DSCDMA signal model described by (1)–(5). We recall that Q, L, N, M, and J denote the number of active users in the system, the processing gain, the number of paths experienced by the transmitted signal of each user, the number of antenna elements, and the data packet (data record) size, respectively. We also recall that $\mathbf{v}_0$ is the ST

RAKE filter of the user of interest (user 0), defined by $\mathbf{v}_0 \triangleq E_{b_0}\{\mathbf{x}b_0\}$, where the statistical expectation operation $E_{b_0}\{\cdot\}$ is taken with respect to the bit of the user of interest b_0 only. Clearly, $\mathbf{v}_0$ consists of shifted versions of the ST matched filter multiplied by the corresponding channel coefficients [cf. (4)]:[4]

$$\mathbf{v}_0 = \sum_{n=0}^{N-1} c_{0,n} \begin{bmatrix} \underbrace{0 \ldots 0}_{n} & \mathbf{d}_0^T & \underbrace{0 \ldots 0}_{N-n-1} \end{bmatrix}^T \odot \mathbf{a}_{0,n} \qquad (27)$$

Hence, $\mathbf{v}_0$ is a function of the binary signature vector (spreading sequence) of the user of interest $\mathbf{d}_0$, the channel coefficients $c_{0,0}, c_{0,1}, \ldots, c_{0,N-1}$, and the corresponding angles of arrival $\theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,N-1}$ [cf. (5)]. While the spreading sequence is assumed to be known to the receiver, the channel coefficients and the angles of arrival are in general unknown.

4.2.1. Subspace Channel and Angle-of-Arrival Estimation. In this section we explain how the channel coefficients $\mathbf{c}_0 \triangleq [c_{0,0}, c_{0,1}, \ldots, c_{0,N-1}]^T$ and the angles of arrival $\boldsymbol{\theta}_0 \triangleq [\theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,N-1}]^T$ for the user of interest, user 0, can be estimated by subspace-based techniques from the ST data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$ [65,66]. We note that while adaptive subspace (eigendecomposition)-type MMSE/MVDR filtering is not a favorable approach under limited data support (the resulting estimates exhibit high variance), subspace-type channel estimation techniques do not suffer from "data starvation" as illustrated later.

Let the binary data of each user be organized in identically structured packets of J bits. The channel estimation procedure that we employ utilizes J_p pilot bits (bits that are known to the receiver). Thus, the qth user data packet, $\{b_q(0), b_q(1), \ldots, b_q(J-1)\}, q = 0, 1, \ldots, Q-1$, contains $J - J_p$ information bits and J_p pilot bits. The J_p known bits will be utilized later for the supervised recovery of the phase of the subspace channel estimates since blind second-order channel estimation methods return phase-ambiguous estimates. An example of the data packet structure is shown in Fig. 14, where the J_p pilot bits appear as a *midamble* in the transmitted packet [67]. Without loss of generality, we assume that each user transmits one data packet per slot and the slot duration is T_s seconds. Therefore, the data packet size J is the number of information bits transmitted by each user in one time slot, $T_s = JT$, where T is the duration of each information bit transmission.

The rank r_s of the *signal subspace* of the received data vectors $\mathbf{x}$ can be controlled by considering one-sided or two-sided truncations of $\mathbf{x}$ (the latter eliminates ISI). The possible values of r_s, depending on the data format of choice, are as follows:

1. *No truncation*: data dimension $= M(L+N-1)$, $2Q+1 \le r_s \le 3Q$.
2. *One-sided truncation*: data dimension $= ML$, $2Q \le r_s \le 3Q-1$.
3. *Two-sided truncation*: data dimension $= M(L-N+1)$, $Q \le r_s \le 2Q-1$.

To have a guaranteed minimum rank of the *noise subspace* of $M(L-N+1)-(2Q-1)$, we choose to truncate $\mathbf{x}$ from both sides (case (3)) as shown in Fig. 15, and we form the "truncated" received vector $\mathbf{x}^{\text{tr}}$ of length $M(L-N+1)$ as follows:

$$\mathbf{x}^{\text{tr}} = \begin{bmatrix} \mathbf{x}((N-1)T_c) \\ \mathbf{x}(NT_c) \\ \vdots \\ \mathbf{x}((L-1)T_c) \end{bmatrix}$$

Then, with respect to the jth information bit of user 0, $\mathbf{x}_j^{\text{tr}}$ can be expressed as

$$\mathbf{x}_j^{\text{tr}} = b_0(j)\frac{\sqrt{E_0}}{L}\mathbf{A}_0\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{c}_0 + \text{MAI}_j + \mathbf{n}_j \qquad (29)$$

where MAI_j accounts comprehensively for multiple-access interference of rank $r_s - 1$, $\mathbf{B}(\boldsymbol{\theta}_0)$ is a block diagonal matrix of the form $\mathbf{B}(\boldsymbol{\theta}_0) \triangleq \text{diag}(\mathbf{a}_{0,0}, \mathbf{a}_{0,1}, \ldots, \mathbf{a}_{0,N-1})$, and $\mathbf{A}_0 = \mathbf{A}_0^s \odot \mathbf{I}_M$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix, and

$$\mathbf{A}_0^s = \begin{bmatrix} d_0[N-1] & d_0[N-2] & \ldots & d_0[0] \\ d_0[N] & d_0[N-1] & \ldots & d_0[1] \\ \vdots & \vdots & & \vdots \\ d_0[L-1] & d_0[L-2] & \ldots & d_0[L-N] \end{bmatrix} \qquad (30)$$

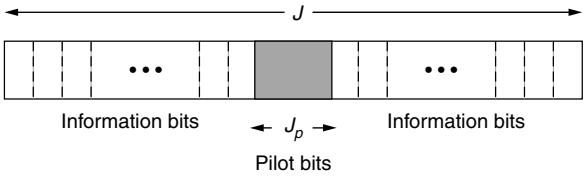


Figure 14. Data packet structure of total length J bits that contains a midamble of J_p pilot bits.

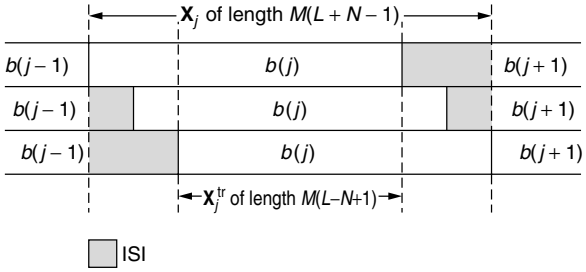


Figure 15. Data collection and ISI trimming.

[4] For the sake of mathematical accuracy

$$\mathbf{v}_0 = \sqrt{\frac{E_0}{L}} \sum_{n=0}^{N-1} c_{0,n} \begin{bmatrix} \underbrace{0 \ldots 0}_{n} & \mathbf{d}_0^T & \underbrace{0 \ldots 0}_{N-n-1} \end{bmatrix}^T \odot \mathbf{a}_{0,n}$$

The positive multiplier $\sqrt{\frac{E_0}{L}}$ is dropped in (27) as inconsequential.

Let $\mathbf{R}_{\text{tr}} = E\{\mathbf{x}^{\text{tr}}\mathbf{x}^{\text{tr}H}\}$ be the autocorrelation matrix of $\mathbf{x}^{\text{tr}}$. We form a sample-average estimate

$$\hat{\mathbf{R}}_{\text{tr}} = \frac{1}{J}\sum_{j=0}^{J-1}\mathbf{x}_j^{\text{tr}}\mathbf{x}_j^{\text{tr}H} \tag{31}$$

based on the truncated J available input vectors $\mathbf{x}_j^{\text{tr}}, j = 0, 1, \ldots, J-1$. If $\hat{\mathbf{R}}_{\text{tr}} = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^H$ represents the eigendecomposition of $\hat{\mathbf{R}}_{\text{tr}}$, where the columns of $\hat{\mathbf{Q}}$ are the eigenvectors of $\hat{\mathbf{R}}_{\text{tr}}$ and $\hat{\mathbf{\Lambda}}$ is a diagonal matrix consisting of the eigenvalues of $\hat{\mathbf{R}}_{\text{tr}}$, then we use the eigenvectors that correspond to the $M(L-N+1)-(2Q-1)$ smallest eigenvalues to define our estimated noise subspace. Let the matrix $\hat{\mathbf{U}}_n$ of size $[M(L-N+1)] \times [M(L-N+1)-(2Q-1)]$ consist of these "noise eigenvectors." We estimate $\mathbf{c}_0$ and $\boldsymbol{\theta}_0$ indirectly through an estimate of the $MN \times 1$ vector

$$\mathbf{h}_0 \triangleq \mathbf{B}(\boldsymbol{\theta}_0)\mathbf{c}_0 \tag{32}$$

We estimate $\mathbf{h}_0$ as the vector that minimizes the norm of the projection of the signal of the user of interest, user 0, $\mathbf{A}_0\mathbf{h}_0$, onto the estimated noise subspace $\hat{\mathbf{U}}_n$:

$$\hat{\mathbf{h}}_0 = \arg\min_{\mathbf{h}_0} \|(\mathbf{A}_0\mathbf{h}_0)^H\hat{\mathbf{U}}_n\| \quad \text{subject to} \quad \|\hat{\mathbf{h}}_0\| = 1 \tag{33}$$

The solution to this constrained minimization problem is the eigenvector that corresponds to the minimum eigenvalue of $\mathbf{A}_0^H\hat{\mathbf{U}}_n\hat{\mathbf{U}}_n^H\mathbf{A}_0$. After obtaining $\hat{\mathbf{h}}_0$, we may extract the desired vectors $\hat{\mathbf{c}}_0$ and $\hat{\boldsymbol{\theta}}_0$ by applying least-squares (LS) fitting to $\hat{\mathbf{h}}_0$. Then, the estimate $\hat{\mathbf{v}}_0$ is completely defined by (27).

Since the channel estimation method described above is based on a blind second-order criterion, the phase information is absorbed, which means that the estimate $\hat{\mathbf{v}}_0$ is phase-ambiguous. Inherently, adaptive filter estimators that utilize a phase-ambiguous estimate of $\mathbf{v}_0$ are also phase-ambiguous. Next, we consider the recovery (correction) of the phase of linear filters when the vector $\mathbf{v}_0$ is known within a phase ambiguity.

4.2.2. Phase Recovery. Without loss of generality, let $\tilde{\mathbf{v}}_0$ denote a phase ambiguous version of $\mathbf{v}_0$

$$\tilde{\mathbf{v}}_0 e^{j\psi} = \mathbf{v}_0 \tag{34}$$

where ψ is the unknown phase. We consider the class of linear filters $\mathbf{w} \in \mathbb{C}^{M(L+N-1)}$ that are functions of the ST RAKE vector $\mathbf{v}_0$ and share the following property:

$$\mathbf{w}(\mathbf{v}_0) = \mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi} \tag{35}$$

Such filters include (1) the ST RAKE filter itself, $\mathbf{v}_0$, (2) the ST MMSE/MVDR filter of (6), and (3) the auxiliary vector sequence of ST filters $\{\mathbf{w}_k\}$.

As seen in (35), for this class of filters the phase ambiguity of $\tilde{\mathbf{v}}_0$ leads to a phase ambiguous linear filter $\mathbf{w}(\tilde{\mathbf{v}}_0)$. Phase ambiguity in digital communications can be catastrophic since it may result in receivers that exhibit BER equal to 50%. Given $\tilde{\mathbf{v}}_0$, we attempt to correct the phase of $\mathbf{w}(\tilde{\mathbf{v}}_0)$ as follows. As a selection criterion for the phase correction parameter ψ we consider the minimization of the mean-square error (MSE) between

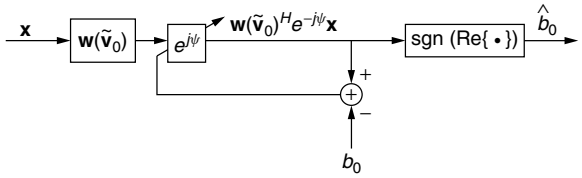


Figure 16. Supervised (pilot-assisted) phase correction for the spacetime linear filter $\mathbf{w}(\tilde{\mathbf{v}}_0)$.

the output of the phase corrected filter $[\mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi}]^H\mathbf{x}$ and the desired information bit b_0 (Fig. 16):

$$\hat{\psi} = \arg\min_{\psi} E\{|[\mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi}]^H\mathbf{x} - b_0|^2\}, \quad \psi \in [-\pi, \pi) \tag{36}$$

The optimum phase correction value according to this criterion is given by

$$\hat{\psi} = \text{angle}\{\mathbf{w}(\tilde{\mathbf{v}}_0)^H E\{\mathbf{x}b_0\}\} \tag{37}$$

Essentially, (37) suggests to project the phase ambiguous $\mathbf{w}(\tilde{\mathbf{v}}_0)$ filter onto the ideal ST RAKE filter $\mathbf{v}_0 = E\{\mathbf{x}b_0\}$. However, $E\{\mathbf{x}b_0\}$ is certainly not known. Since we have assumed that a pilot information bit sequence of length J_p is included in each packet, the expectation $E\{\mathbf{x}b_0\}$ can be sample-average-estimated by $\frac{1}{J_p}\sum_{j=1}^{J_p}\mathbf{x}_j b_0(j)$, where $b_0(j), j = 1, 2, \ldots, J_p$, is the jth *pilot* information bit and $\mathbf{x}_j$ is the corresponding input data vector. Then, the phase-corrected adaptive filter estimate is given by

$$\mathbf{w}(\hat{\mathbf{v}}_0, \hat{\mathbf{R}})e^{j\hat{\psi}}, \quad \hat{\psi} = \text{angle}\left\{\mathbf{w}(\hat{\mathbf{v}}_0, \hat{\mathbf{R}})^H\left[\sum_{j=1}^{J_p}\mathbf{x}_j b_0(j)\right]\right\} \tag{38}$$

Since J represents the packet size of the DSCDMA system and J_p is the number of midamble pilot information bits per packet, then the ratio J_p/J quantifies the wasted bandwidth due to the use of the pilot bit sequence. Ideally, J_p/J is to be kept small. As we will see in the next section, a few pilot bits (on the order of 5 bits) are sufficient for effective recovery of the filter phase. As a numerical example, when the packet size is set at $J = 256$ and $J_p = 5$ is chosen, then $J_p/J \simeq 2\%$ only.

5. PACKET ERROR RATE, CAPACITY, AND THROUGHPUT ANALYSIS

5.1. Packet Error Rate

So far, we have concentrated on the design/estimation of the receiver filter $\mathbf{w}$ in (3). The filter estimate is generated adaptively on an individual packet-by-packet basis. All J received vectors of the packet are utilized for the design of $\mathbf{w}$ (estimation of $\mathbf{R}$, $\mathbf{v}_0$, and the number of auxiliary vectors k_1 or k_2), while J_p of them are also used for supervised phase correction (estimation of ψ). Then, the $J - J_p$ information bits of user 0 associated with the remaining $J - J_p$ received vectors are detected by (3).

The effectiveness of the filter is characterized statistically by the probability distribution of the number of bit errors in a packet:

$$p(i) \triangleq \Pr\{i \text{ bit errors in the packet}\}, i = 0, 1, \ldots, J - J_p. \tag{39}$$

Without loss of generality, we define the bit error rate (BER) as the probability of erroneous detection of $b_0(0)$ (the first bit of user 0 in the packet):

$$\begin{aligned}\text{BER} &\triangleq \Pr\{\hat{b}_0(0) \neq b_0(0)\} \\ &= \sum_{i=0}^{J-J_p} \Pr\{\hat{b}_0(0) \neq b_0(0) \mid i \text{ bit errors in the packet}\} p(i) \\ &= \sum_{i=0}^{J-J_p} \frac{i}{J-J_p} p(i) = \frac{1}{J-J_p} \sum_{i=0}^{J-J_p} i p(i)\end{aligned} \tag{40}$$

It is interesting to note at this point that if the filter $\mathbf{w}$ were independent of the information bit stream $b_0(0), b_0(1), \ldots$, then the BER could have been expressed analytically as a weighted sum of the value of the error function $Q(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_x^\infty \mathrm{e}^{t^2/2}\, dt$ evaluated at 2^{K-1} different points (interfering bit combinations) weighted by the probability of each point $2^{-(K-1)}$. However, this independence assumption does not hold true in our system since the data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$, which includes the information bits $b_0(0), b_0(1), \ldots, b_0(J-1)$ to be detected, is directly utilized for the calculation of $\mathbf{w}$. Therefore, performance cannot be evaluated using the analytic BER expression; instead, we can only rely on (40).

A data packet is received successfully within a single time slot if the number of errors in the detection of the $J - J_p$ information bits is less than or equal to the maximum number of (correctable) bit errors allowed by the forward error correction (FEC) module. If no FEC is present, then the packet is successfully received when all $J - J_p$ information bits are detected correctly. The packet error rate (PER) is defined as the probability of receiving an uncorrectable packet within a single time slot and is given by

$$\begin{aligned}\text{PER}(h) &\triangleq \Pr\{\text{more than } h \text{ bit errors in the packet}\} \\ &= \sum_{i=h+1}^{J-J_p} p(i) = 1 - \sum_{i=0}^{h} p(i)\end{aligned} \tag{41}$$

where h is the maximum number of correctable bit errors per packet. By setting $h = 0$ we obtain the PER of a system without FEC:

$$\text{PER}(0) = \sum_{i=1}^{J-J_p} p(i) = 1 - p(0) \tag{42}$$

To examine the PER performance of the general DSCDMA system described in (1)–(5) equipped with the packet-rate adaptive ST AV filter receiver, we proceed with an illustration. We consider a Q-user system with Gold signatures of length $L = 31$. We fix the packet size at $J = 256$ bits and use $J_p = 5$ of them as pilot midamble bits. Each user signal experiences $N = 3$ independent Rayleigh fading paths with equal average received energy per path and independent angles of arrival uniformly distributed in $\left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right)$. We consider averages over 20,000 independently drawn multipath Rayleigh fading ST channels. The receiver antenna array consists of $M = 4$ elements. With these numbers, the multipath extended ST product (or, equivalently, the length of the adaptive filter) is $M(L + N - 1) = 132$. The total received predetection SNRs of each user, namely, the sum of the received SNRs over all paths defined as $2E_q \displaystyle\sum_{n=0}^{N-1} E\{|c_{q,n}|^2\}/N_0$, $q = 0, 1, \ldots Q-1$, is set to 11 dB (we recall that $\dfrac{N_0}{2}$ is the AWGN power spectral density assumed to be identical for every spatial channel/antenna element).

In Fig. 17a, we plot the PER as a function of the number of active users Q using (41) for various receivers: (1) ST RAKE, (2) SMI, (3) auxiliary vector, (4) and the orthogonal multistage decomposition filter (also known as "nested Wiener filter") with the preferred number of stages $l = 7$. We also add to this study the multistage filter with $l = 1$ stages, which we found empirically to be the best number of stages for this specific problem. No FEC is assumed ($h = 0$). In a rather more interesting study, Fig. 17b shows the PER under $h = 4$ FEC.

5.2. Capacity

System BER/PER performance improvements due to the use of the AV receiver allow us to accommodate more users for a set quality-of-service (QoS) PER constraint (or reduce the transmitting power of the handset or increase the range/coverage of the base-station transceiver for a preset maximum number of active users). The QoS constraint is based strictly on the specific user application requirements and determines the *user capacity* of the system. Let us express the PER as a function of both the FEC capabilities h and the number of active users Q, $\text{PER}(h, Q)$. Then, we define the user capacity $C(h)$ as the maximum number of users under which the QoS constraint is not violated:

$$C(h) \triangleq \max\{Q: \text{PER}(h, Q) \leq \lambda_{\text{QoS}}\} \tag{43}$$

where λ_{QoS} is the QoS constraint threshold.

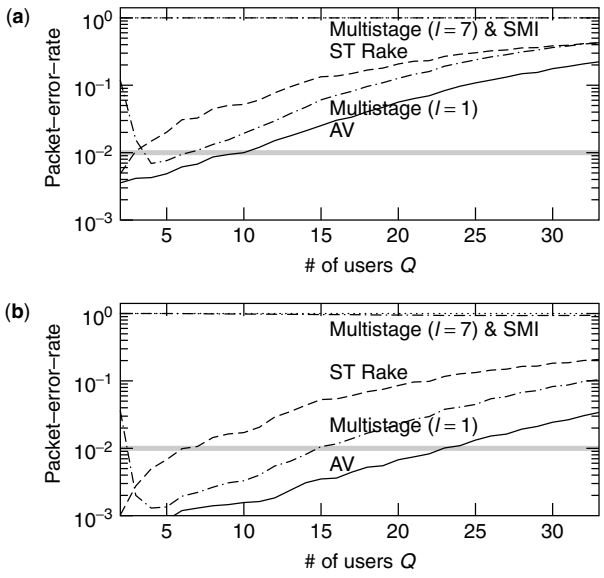


Figure 17. PER versus number of users Q for a system with (**a**) no FEC and (**b**) 4-bit FEC.

We return to Fig. 17 to examine the user capacity of the system for the adaptive receivers under consideration. We set the QoS constraint threshold to $\lambda_{\text{QoS}} = 10^{-2}$. From Fig. 17a we conclude that in the absence of FEC the user capacity of the AV system is 10 and the capacity of the RAKE system is 2. Significant capacity improvement is achieved with 4-bit FEC (Fig. 17b). The AV system supports 23 users, while the RAKE scheme supports only 6 users. Neither the $l = 7$ multistage nor the SMI receiver can meet the QoS constraint for any number of users $Q \geq 1$, with or without FEC. The best multistage ($l = 1$) receiver can support 4–6 users when no FEC is considered and 3–14 users with 4-bit FEC. However, it cannot meet the QoS requirement when $Q \leq 2$. We conclude that we have two viable solutions: the AV and plain ST RAKE receiver systems. The AV system allows up to $\frac{23}{31} \simeq 74\%$ loading for this Gold-coded system with processing gain $L = 31$ and 4-bit FEC (with $M = 4$ antenna elements and multipath fading reception with 11 dB total predetection SNR per user).

5.3. Throughput

If a packet is received with only correctable errors, a positive acknowledgment (ACK) is sent back to the user over a different downlink channel (FDD). Once an uncorrectable error is detected in the packet, a negative acknowledgment (NAK) is sent back to the mobile which then retransmits after waiting for a random number of time slots. If packet arrival is modeled as Poisson-distributed, then the probability of the arrival of Q new messages during a time-slot interval T_s is given by

$$A(Q) = \frac{{G_N}^Q e^{-G_N}}{Q!}, \qquad Q \geq 1 \tag{44}$$

where $G_N \triangleq (\lambda T_s / L)$ is the normalized offered traffic load (we recall that L is the system processing gain) and λ is the packet arrival rate. *Packet throughput* is defined as a measure of the ratio of the average number of successful transmissions to the number of transmission attempts made in a particular time slot. Let $\text{PSR}(h, Q)$ be the packet success rate (that is the probability that a packet is received successfully within a single time slot) with h-bit FEC in the presence of Q users. Then, $\text{PSR}(h, Q) = 1 - \text{PER}(h, Q)$ and the *normalized* packet throughput of the system S_N under slotted ALOHA accessing can be expressed as

$$\begin{aligned} S_N(h) &= \frac{1}{L} \sum_{Q=1}^{C(h)} Q A(Q) \text{PSR}(h, Q) \\ &= \frac{1}{L} \sum_{Q=1}^{C(h)} Q A(Q)(1 - \text{PER}(h, Q)) \end{aligned} \tag{45}$$

In Fig. 18a we plot the normalized packet throughput S_N versus the offered traffic G_N without FEC ($h = 0$). As a reference, we add the familiar packet throughput

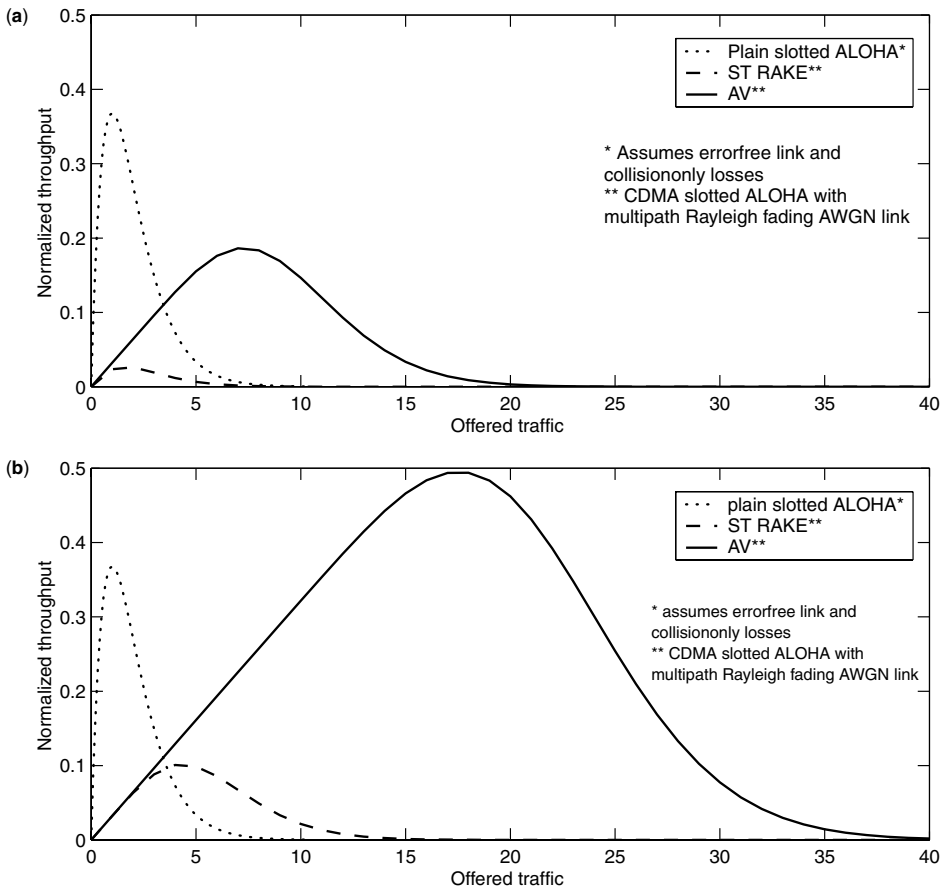


Figure 18. Normalized throughput versus offered traffic for a system with (**a**) no FEC and (**b**) 4-bit FEC.

curve for a plain slotted ALOHA system that assumes an *ideal error-free* link and *collision-only* packet losses. As one might expect, the CDMA slotted ALOHA (RAKE or AV) system with a multipath fading AWGN link and no FEC exhibits a lower maximum achievable throughput value than the plain ALOHA system with an error-free link. However, enhancing the system with FEC (Fig. 18b) gives rise to a significant improvement of the AV system whose throughput performance at higher traffic loads now overshoots the ideal-link plain slotted ALOHA which allows only one user per slot for successful transmission. We observe that with the adaptive AV filter receiver we can achieve up to 0.4938 throughput when we have FEC capability of $h = 4$ bits.

6. CONCLUDING REMARKS

Wireless cellular and personal communication service (PCS) networks have experienced significant growth driven by a strong market interest for highly mobile, widely accessible, two-way voice and data communications. Current research efforts are focusing on system improvements to meet future demand and quality of service requirements. User capacity increase may be sought in the form of a synergy of effective multiple accessing schemes and advanced receiver technology (e.g., code-division multiple access with adaptive antenna arrays). Manageable complexity (hardware and software) at the physical layer can be achieved by means of linear equalizers as opposed to more complex structures. Improved receiver output SINR and BER performance may be sought in the form of intelligent modulation techniques as well as intelligent signal processing at the receiver end of the communications link. However, realistically, receiver output SINR and BER improvements in rapidly changing channel environments can be achieved only by means of adaptive short-data-record-optimized receiver designs (as opposed to designs based on ideal asymptotic optimization solutions).

This article focused on linear MMSE/MVDR adaptive filtering with applications to adaptive receiver designs for mobile communications. We presented three alternative methods that approximate the optimum solution under perfectly known input statistics (input autocorrelation matrix and input/desired-output cross-correlation vector): (1) The generalized sidelobe canceler, (2) the auxiliary vector filters, and (3) the multistage filters. When the input statistics are unknown and estimated, these three approximate solutions provide estimates of the optimum solution with varying performance levels (output SINR and BER). When estimation is based on a short data record, that is, when system adaptation and redesign has to be performed with limited data support (which is the case for most systems of practical interest), then the performance differences become even more pronounced. In mobile packet data communications, for example, the size of the data record that is available for receiver adaptation and redesign is limited by the coherence time of the communication link and may be of the order of 300 data symbols or less in practical situations. In this context, for a given system transmission bit rate, the packet size may be designed to be sufficiently small to conform with the coherence time of the link. Then, packet-rate adaptive receiver designs may be pursued.

A viable solution for adaptive MMSE/MVDR system designs under limited data support is provided by the auxiliary vector (AV) algorithm. AV estimators exhibit varying bias/covariance characteristics — the bias of the generated estimator sequence decreases rapidly to zero while the estimator covariance trace rises slowly from zero (for the initial, fixed-valued, matched-filter estimator) to the asymptotic covariance trace of the SMI filter. Sequences of practical estimators that offer such control over favorable bias/covariance balance points are always a prime objective in the estimation theory literature. Indeed, under quasistatic fading over the duration of a packet and packet-rate adaptation, members of the generated sequence of AV estimators outperform in MS estimation error LMS/RLS-type, SMI and diagonally loaded SMI, and orthogonal multistage decomposition filter estimators. In addition, the troublesome, data-dependent tuning of the real-valued LMS learning gain parameter, the RLS initialization parameter or the SMI diagonal loading parameter, is replaced by an integer choice among the first several members of the estimator sequence. In that respect, we presented two data-driven criteria for the selection of the best AV filter estimator in the sequence.

As a representative application throughout this article we considered a wireless multiuser multipath fading AWGN link with direct-sequence spread-spectrum signaling and slotted ALOHA accessing. We developed a complete adaptive antenna-array CDMA linear filter receiver design that adapts itself and detects the transmitted information bits on an individual packet-by-packet basis. The receiver incorporated seamlessly packet-rate blind subspace-based spacetime channel estimation and supervised recovery of the spacetime channel phase through the use of a few packet midamble pilot bits. Illustrative examples showed that very limited midamble pilot signaling (on the order of $\frac{5}{256} \simeq 2\%$) can be sufficient for phase recovery and effective adaptive receiver design. Therefore, differential modulation to overcome the phase ambiguity problem is not absolutely necessary.

Bit error rate, packet error rate, and user capacity studies and comparisons were also included. Through the development of the probability mass distribution of the bit errors in a packet, we can translate these findings to packet throughput results. Interestingly, the adaptive AV receiver designed for a CDMA system in multipath Rayleigh fading and AWGN that assumed a modest 11 dB total received predetection SNR per user, four antenna elements, and 4-bit FEC, offers a $\frac{0.4938 - 0.3679}{0.3679} 100\% \simeq 34\%$ improvement in terms of normalized maximum packet throughput over plain (non-CDMA) slotted ALOHA with an *error-free* link. In this context, the packet-rate adaptive receiver design using the AV filtering principles coupled with FEC techniques seems to provide a viable solution in improving the performance of DS-CDMA mobile communication links.

BIOGRAPHY

Stella N. Batalama received the Diploma degree in computer engineering and science from the University of Patras, Greece in 1989 and the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in 1994.

From 1989 to 1990 she was with the Computer Technology Institute, Patras, Greece. From 1990 to 1994 she was a Research Assistant in the Communication Systems Laboratory, Department of Electrical Engineering, University of Virginia. In 1995 she joined the Department of Electrical Engineering, State University of New York at Buffalo, where she is presently an Associate Professor. During the summers of 1997–2002 she was Visiting Faculty in the U.S. Air Force Research Laboratory, Rome, New York. Her research interests include small sample support adaptive filtering and receiver design, adaptive multiuser detection, robust spread-spectrum communications, supervised and unsupervised optimization, and distributed detection.

Dr. Batalama is currently an associate editor for the *IEEE Transactions on Communications* and the *IEEE Communications Letters*.

BIBLIOGRAPHY

1. T. K. Liu and J. A. Silvester, Joint admission congestion control for wireless CDMA systems supporting integrated services, *IEEE J. Select. Areas Commun.* **16**: 845–857 (Aug. 1998).
2. H. Bischl and E. Lutz, Packet error rate in the non-interleaved Rayleigh channel, *IEEE Trans. Commun.* **43**: 1375–1382 (April 1995).
3. R. D. J. van Nee, R. N. van Wolfswinkel, and R. Prasad, Slotted ALOHA and code-division multiple-access techniques for land-mobile satellite personal communications, *IEEE J. Select. Areas Commun.* **13**: 382–388 (Feb. 1995).
4. X. Wu and A. Haimovich, Space-time processing for CDMA communications, *Proc. Conf. Information Science and Systems*, Baltimore, March 1995, pp. 371–376.
5. D. A. Pados and S. N. Batalama, Joint space-time auxiliary-vector filtering for DS/CDMA systems with antenna arrays, *IEEE Trans. Commun.* **47**: 1406–1415 (Sept. 1999).
6. I. S. Reed, J. D. Mallet, and L. E. Brennan, Rapid convergence rate in adaptive arrays, *IEEE Trans. Aerospace Electron. Syst.* **10**: 853–863 (Nov. 1974).
7. B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, Adaptive antenna systems, *Proc. IEEE* **55**: 2143–2158 (Dec. 1967).
8. R. L. Plackett, Some theorems in least squares, *Biometrika* **37**: 149 (1950).
9. R. A. Wiggins and E. A. Robinson, Recursive solution to the multichannel filtering problem, *J. Geophys. Res.* **70**: 1885–1891 (1965).
10. J. S. Thompson, P. M. Grant, and B. Mulgrew, Smart antenna arrays for CDMA systems, *IEEE Personal Commun.* 16–25 (Oct. 1996).
11. L. E. Brennan and I. S. Reed, Theory of adaptive radar, *IEEE Trans. Aerospace Electron. Syst.* **9**: 237–252 (March 1973).
12. E. J. Kelly, An adaptive detection algorithm, *IEEE Trans. Aerospace Electron. Syst.* **22**: 115–127 (March 1986).
13. L. C. Godara, Applications of antenna arrays to mobile communications, Part I: Performance improvement, feasibility, and system considerations, *IEEE Proc.* **85**: 1031–1060 (July 1997).
14. J. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
15. E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, UMTS/IMT-2000 based on wideband CDMA, *IEEE Commun. Mag.* **36**: 70–80 (Sept. 1998).
16. S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1991.
17. V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
18. J. Capon, High-resolution frequency-wavenumber spectrum analysis, *Proc. IEEE* **57**: 1408–1418 (Aug. 1969).
19. N. L. Owsley, A recent trend in adaptive spatial processing for sensor arrays: Constraint adaptation, J. W. R. Griffiths et al., eds., *Signal Processing*, Academic Press, New York, 1973, pp. 591–604.
20. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1990.
21. S. P. Applebaum and D. J. Chapman, Adaptive arrays with main beam constraints, *IEEE Trans. Antennas Propag.* **24**: 650–662 (Sept. 1976).
22. P. W. Howells, Explorations in fixed and adaptive resolution at GE and SURC, *IEEE Trans. Antennas Propag.* **24**: 575–584 (Sept. 1976).
23. B. D. Van Veen and R. A. Roberts, Partially adaptive beamformer design via output power minimization, *IEEE Trans. Acoust. Speech, Signal Process.* **35**: 1524–1532 (Nov. 1987).
24. L. J. Griffiths and C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. Antennas Propag.* **30**: 27–34 (Jan. 1982).
25. P. Strobach, Low-rank adaptive filters, *IEEE Trans. Signal Process.* **44**(12): 2932–2947 (Dec. 1996).
26. P. A. Thompson, An adaptive spectral analysis technique for unbiased frequency estimation in the presence of white noise, *Proc. 13th Asilomar Conf. Circ. Systems Computers*, Nov. 1980, pp. 529–533.
27. N. L. Owsley, in S. Haykin, ed., *Array Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.
28. B. D. Van Veen, Eigenstructure based partially adaptive array design, *IEEE Trans. Antennas Propag.* **36**: 357–362 (March 1988).
29. A. M. Haimovich and Y. Bar-Ness, An eigenanalysis interference canceler, *IEEE Trans. Signal Process.* **39**: 76–84 (Jan. 1991).
30. K. A. Byerly and R. A. Roberts, Output power based partially adaptive array design, *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 1989, pp. 576–580.
31. J. S. Goldstein and I. S. Reed, Reduced-rank adaptive filtering, *IEEE Trans. Signal Process.* **45**: 492–496 (Feb. 1997).
32. S. N. Batalama, M. J. Medley, and D. A. Pados, Robust adaptive recovery of spread-spectrum signals with short data records, *IEEE Trans. Commun.* **48**: 1725–1731 (Oct. 2000).

33. D. A. Pados and G. N. Karystinos, An iterative algorithm for the computation of the MVDR filter, *IEEE Trans. Signal Process.* **49**: 290–300 (Feb. 2001).

34. D. A. Pados and G. N. Karystinos, Short-data-record estimators of the MMSE/MVDR filter, *Proc. ICASSP 2000*, Istanbul, Turkey, June 2000, Vol. 1, pp. 384–387.

35. D. A. Pados and S. N. Batalama, Low-complexity blind detection of DS/CDMA signals: Auxiliary-vector receivers, *IEEE Trans. Commun.* **45**: 586–1594 (Dec. 1997).

36. A. Kansal, S. N. Batalama and D. A. Pados, Adaptive maximum SINR rake filtering for DS-CDMA multipath fading channels, *IEEE J. Select. Areas Commun.* 1965–1973 (Dec. 1998).

37. D. A. Pados and S. N. Batalama, Joint space-time auxiliary-vector filtering for antenna array DS/CDMA systems, *Proc. 1998 Conf. Information Science and Systems*, Princeton, NJ, March 1998, Vol. 2, pp. 1007–1013.

38. D. A. Pados, T. Tsao, J. H. Michels, and M. C. Wicks, Joint domain space-time adaptive processing with small training data sets, *Proc. IEEE Radar Conf.*, Dallas, TX, May 1998, pp. 99–104.

39. J. S. Goldstein, I. S. Reed, P. A. Zulch, and W. L. Melvin, A multistage STAP CFAR detection technique, *Proc. IEEE Radar Conf.*, Dallas, TX, May 1998, pp. 111–116.

40. J. S. Goldstein, I. S. Reed, and L. L. Scharf, A multistage representation of the Wiener filter based on orthogonal projections, *IEEE Trans. Inform. Theory* **44**: 2943–2959 (Nov. 1998).

41. M. L. Honig and W. Xiao, Performance of reduced-rank linear interference suppression, *IEEE Trans. Inform. Theory* **47**: 1928–1946 (July 2001).

42. T.-C. Liu and B. Van Veen, A modular structure for implementation of linearly constrained minimum variance beamformers, *IEEE Trans. Signal Process.* **39**: 2343–2346 (Oct. 1991).

43. N. E. Nahi, *Estimation Theory and Applications*, R. E. Krieger, Huntington, NY, 1976.

44. C. D. Richmond, Derived PDF of maximum likelihood signal estimator which employs an estimated noise covariance, *IEEE Trans. Signal Process.* **44**: 305–315 (Feb. 1996).

45. R. L. Dykstra, Establishing the positive definiteness of the sample covariance matrix, *Ann. Math. Stat.* **41**(6): 2153–2154 (1970).

46. C. D. Richmond, PDF's, confidence regions, and relevant statistics for a class of sample covariance-based array processors, *IEEE Trans. Signal Process.* **44**: 1779–1793 (July 1996).

47. A. O. Steinhardt, The PDF of adaptive beamforming weights, *IEEE Trans. Signal Process.* **39**: 1232–1235 (May 1991).

48. B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, Adaptive antenna systems, *Proc. IEEE* **55**: 2143–2158 (Dec. 1967).

49. L. C. Godara and A. Cantoni, Analysis of constrained LMS algorithm with application to adaptive beamforming using perturbation sequences, *IEEE Trans. Antennas Propag.* **34**: 368–379 (March 1986).

50. V. Solo, The limiting behavior of LMS, *IEEE Trans. Acoust. Speech, Signal Process.* **37**: 1909–1922 (Dec. 1989).

51. J. M. Cioffi and T. Kailath, Fast recursive-least-squares transversal filters for adaptive filtering, *IEEE Trans. Acoust. Speech, Signal Process.* **32**: 304–337 (April 1984).

52. H. Qian and S. N. Batalama, Data-record-based criteria for the selection of an auxiliary-vector estimator of the MVDR filter, *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Oct. 2000, pp. 802–807.

53. H. Qian and S. N. Batalama, Data-record-based criteria for the selection of an auxiliary-vector estimator of the MMSE/MVDR filter, *IEEE Trans. Commun.* (in press).

54. C. R. Rao, *Handbook of Statistics 9*. New York, NY: Elsevier, 1993.

55. D. Kazakos and P. Papantoni-Kazakos, *Detection and Estimation*, Computer Science Press, New York, 1990.

56. B. D. Carlson, Covariance matrix estimation errors and diagonal loading in adaptive arrays, *IEEE Trans. Aerospace and Electron. Syst.* **24**: 397–401 (July 1988).

57. H. V. Poor and S. Verdú, Probability of error in MMSE multiuser detection, *IEEE Trans. Inform. Theory* **43**: 858–871 (May 1997).

58. I. N. Psaromiligkos and S. N. Batalama, Interference-plus-noise covariance matrix estimation for adaptive space-time processing of DS/CDMA signals, *Proc. IEEE VTC 2000—Vehicular Technology Conf.*, Boston, Sept. 2000, Vol. 5, pp. 2197–2204.

59. I. N. Psaromiligkos and S. N. Batalama, Recursive AV and MVDR filter estimation for maximum SINR adaptive space-time processing, *IEEE Trans. Commun.* (in press).

60. I. N. Psaromiligkos and S. N. Batalama, Blind self-synchronized demodulation of DS-CDMA communications, *Proc. IEEE ICC 2000—Int. Conf. Communications*, New Orleans, LA, June 2000, pp. 2557–2560.

61. I. N. Psaromiligkos, M. J. Medley, and S. N. Batalama, Rapid synchronization and combined demodulation for DS/CDMA communications. Part I: Algorithmic developments, *IEEE Trans. Commun.* (in press).

62. I. N. Psaromiligkos and S. N. Batalama, Rapid synchronization and combined demodulation for DS/CDMA communications. Part II: Finite data-record-size performance analysis, *IEEE Trans. Commun.* (in press).

63. S. N. Batalama and I. N. Psaromiligkos, Data record size requirements of MVDR-optimized adaptive antenna arrays, *Proc. IEEE ICASSP 2000—Int. Conf. Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000, Vol. V, pp. 3069–3072.

64. I. N. Psaromiligkos and S. N. Batalama, Data record size requirements for adaptive space-time DS/CDMA signal detection and direction-of-arrival estimation, *IEEE Trans. Commun.* (in press).

65. S. Gopalan, G. N. Karystinos, and D. A. Pados, Capacity, throughput, and delay of slotted ALOHA DS-CDMA links with adaptive space-time auxiliary-vector receivers, *IEEE Trans. Wireless Commun.* (in press).

66. S. E. Bensley and B. Aazhang, Subspace-based channel estimation for code division multiple access communication systems, *IEEE Trans. Commun.* **44**: 1009–1020 (Aug. 1996).

67. P. Chaudhury, W. Mohr, and S. Onoe, The 3GPP proposal for IMT-2000, *IEEE Commun. Mag.* **37**: 72–81 (Dec. 1999).

PACKET-SWITCHED NETWORKS

DIMITRIOS STILIADIS
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

Packet-switched networks are becoming the dominant method of communication, replacing earlier schemes based on the telephone-type circuit-switched networks. With the increasing popularity of the World Wide Web, electronic mail (email), and multimedia applications, the traditional role of a data network as a means of transmitting data between computers is expanding. The same integrated network is now used by applications such as teleconferencing, distance education, real-time video and voice, email, Facsimile (fax), and distributed systems. At the same time, link speeds are experiencing dramatic increases, and the number of users is growing exponentially.

Telephone networks are based on the idea of *circuit switching* (Fig. 1). The dominant application supported by the circuit-switching architecture is voice, and it is assumed that the bandwidth required between two users is determined by the bandwidth needed for transmitting good-quality analog voice, or 8 kHz. The network establishes a dedicated bidirectional path between end nodes (i.e., telephones), and while the call is active, end users can continuously use this path to transmit information. The network is responsible for allocating enough resources throughout the communication path so that data can be transmitted as a continuous flow. Once the resources are allocated, they cannot be reused by another user, until the call is complete. This allocation is usually done by using either time-division multiplexing (TDM) or frequency-division multiplexing (FDM).

One of the most important properties of circuit-switched networks is that the sum of the capacities required by all the active communication paths cannot exceed the capacity of the link. The network must perform *admission control* and cannot accept new connections once the sum of the capacities needed by the active connections reaches a maximum threshold. When the network cannot admit any more connections, new requests by end users receive a "busy signal" similar to the one encountered in phone networks (or the *call is blocked*). Because of the requirement for admission control, establishing a connection and allocating resources (i.e., initiating a new phone call) is a relatively expensive operation that is tackled by a set of distributed computers.

The ideas of voice switching can be extended to the context of data communications by assuming that the two end nodes are computers exchanging data. The network offers a dedicated communication path between the two computers, and once a path is established, it may remain active for a long time period. However, nodes do not constantly transmit data during this period, but they may remain idle for some long intervals of time. Consider, for example, the case where one of the nodes is a desktop computer and the other is a file server. The computer will issue a request to the file server for some data, and once the data are received, the computer will begin to process them. When processing is complete, it will write the data back to the file server. While the computer is processing the data, no information is exchanged on the communication path.

Several data applications may exhibit a similar *bursty* behavior, where they use the bandwidth for some intervals of time and remain idle for others. File access and the World Wide Web (WWW) are prominent examples of such applications. Furthermore, the notion of burstiness can be extended to real-time applications such as voice or video. During a voice call, communicating parties do not always talk, but might have long periods of silence, during which a voice signal corresponding to silence does not need to be transmitted over the network.

If multiple users reserve bandwidth that is needed only during short intervals of time, valuable resources are wasted. In the example of Fig. 1, let us assume that only one session can be active at any time at the central link. Assume that user A places a phone call to user B, and they start talking. Although all the bandwidth of the critical link is reserved for this communication, user A only talks half of the time, and thus the link is 50% utilized. Now let us assume that user C tries to place a phone call to user D. The network cannot accept this call, since the bandwidth

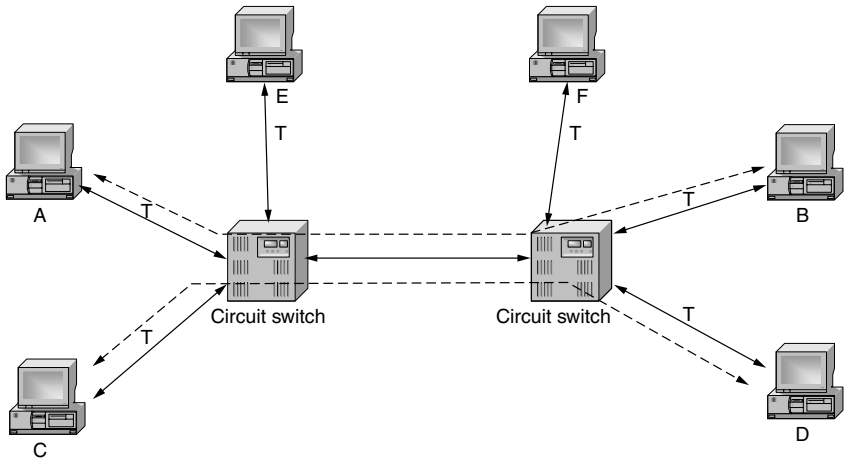


Figure 1. Circuit-switched network.

of the critical link is already allocated to the other phone call. Therefore, when an dedicated path is set up between bursty applications, the bandwidth might not be properly utilized.

Now assume that we could have a network where the path is set up when the link is actually needed and the resources are released when no information is transmitted. In this case, user A would use the critical link while talking to user B. The rest of the time the critical link would be free and user C could also transmit information. Both calls would take place at the same time, and resources would be shared in a more efficient manner.

Unfortunately, this is not easy to achieve in a circuit-switched network, since establishing a communication path requires the configuration of several nodes in the network. However, it is exactly this observation that led to the idea of packet-switched networks, also referred to as *store-and-forward networks*.

2. PACKET-SWITCHED NETWORKS

Communication between two users can be thought of as an exchange of *messages*. For example, when two people talk to each other, they form sentences, and each sentence can be considered as a message. When an end user browses the Web, he/she reads several pages of information, and each page can be considered as a message. A message can have variable sizes and the size of any given message is not bounded. For this reason, the concept of a *packet* is introduced. A packet can be part of a message or it can encapsulate several messages. However, depending on the technology, the packet size is either constant or bounded. For example, packets in a local network based on Ethernet technology are no more than 1500 bytes [1].

When two users communicate, they form a series of messages (and thus packets) and transmit them to one another. Referring to the network of Fig. 2, when user A talks, a packet is created and it is transmitted over the access link. The packet arrives at the intermediate node (referred to as *packet switch* or *router*), and assuming that the link is free, it is transmitted to user B. At a later time a packet from node C is transmitted to the intermediate node and through the critical link to user D. Information from the two users is *multiplexed* on the link on a packet-by-packet basis.

A problem arises when packets from both users arrive at a packet switch at the same time, but only one of the packets can be transmitted at the core link. The packet switch must store in some local memory the packet received from one of the two users (let us say user A) and transmit the packet from the other user. When the transmission is complete, it will retrieve the packet that originated from user A and transmit this packet over the link. It is exactly this concept, that is the foundation of *packet-switched* or *store-and-forward* networks.

Although packet networks may operate efficiently in most cases, *congestion* may degrade their performance during some time periods. Congestion occurs when during a period of time, the bandwidth needed for transmitting all arriving packets exceeds the capacity of the outgoing link. When this happens, the memory in the packet switches may overflow and messages might get lost. In order for packet switched networks to avoid or prevent congestion, they must support a range of mechanisms that control the end user behavior.

3. STATISTICAL MULTIPLEXING

Let us consider again the network of Fig. 2 and assume that the capacity of all the links between users and network nodes is equal to 2 packets per second, and thus, the time to transmit a packet is equal to half a second. Let us also assume that users transmit information half of the time. For example, user A transmits 10 packets and then waits for 5 s before transmitting more packets. If both users transmit their packets at exactly the same time, some of these packets must be buffered at the packet switch.

Let us now assume that the capacity of the core link is equal to the sum of the capacities of the input links, or 4 packets per second. Note, that this is also equal to the capacity needed from a circuit-switched network if both calls are active at the same time. In this scenario, both packets can be transmitted on the link without any delay (Fig. 3). A packet is transmitted over this link within 0.25 s and the link remains idle for half of the time. The maximum queueing delay of any packet is bounded by 0.25 s.

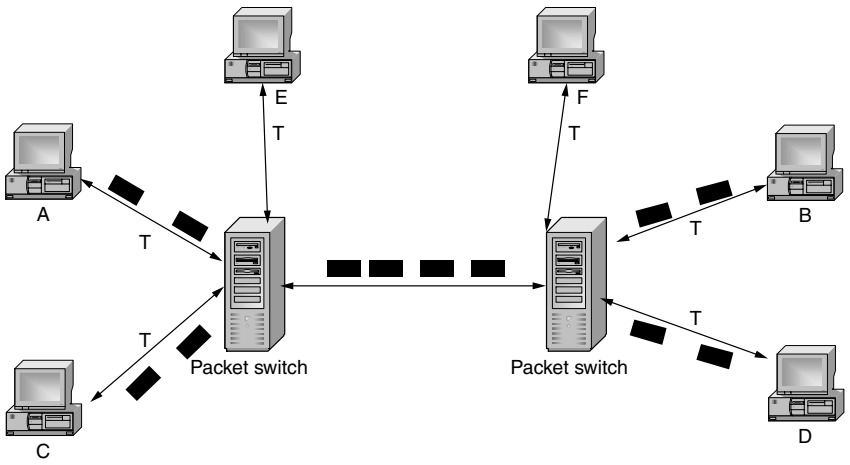


Figure 2. Packet-switched network; information transmitted in terms of packets.

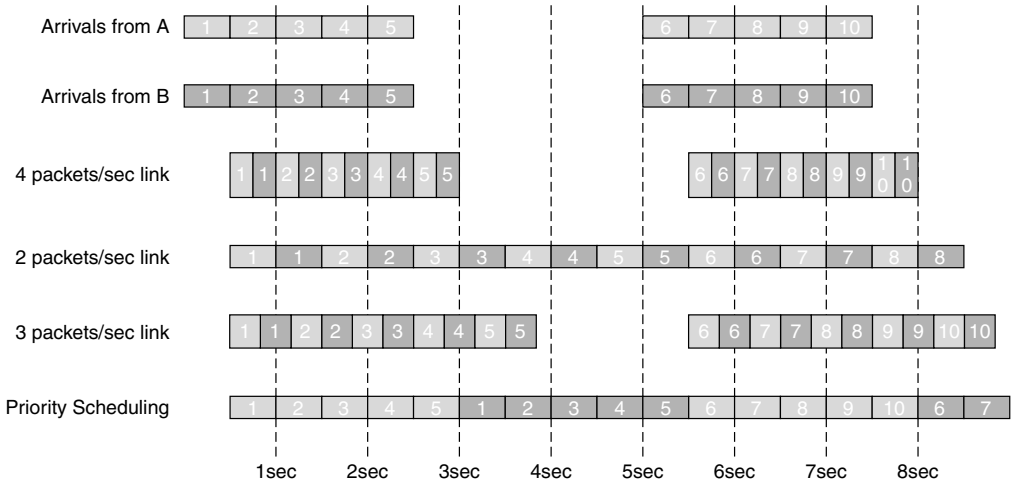


Figure 3. Sequence of packet arrivals and transmissions by a packet switch for different link capacities and scheduling disciplines.

If we set the capacity of the core link to 2 packets per second, some packets are delayed by as much as 2 s (or 4 packet transmission times) and the packet switch must maintain a buffer of at least 4 packets. However, the utilization of the core link is 100%. If we set the capacity of the core link equal to 3 packets per second, no packet will see a delay longer than 0.33 s and the utilization of the core link is 66%.

We can observe in the example above that by modifying the capacity of the core link, we can trade off link utilization for delays and buffers in the core switches. This is exactly the concept of statistical multiplexing. The ratio derived as the sum of the incoming link capacities divided by the outgoing link capacity is referred to the *statistical multiplexing gain* [2]. For example, when we set the core link capacity equal to the access link capacity, the statistical multiplexing gain is $\frac{4}{2}$ or 2.

The concept of statistical multiplexing can be viewed from a different perspective. If we assume that the buffers and link capacities at the core node are fixed (and thus, the maximum delay we can afford is fixed), we need to determine whether traffic from a particular source will not encounter excessive losses. This leads to the *effective bandwidth theory* [3]. If some statistical model for the arrival traffic is available, we can determine the loss probability of different sources.

3.1. Timescales

The concept of statistical multiplexing is not unique to data networks, but it was actually developed within the context of circuit switching. The main difference between the two approaches is in terms of times of interest or, *timescales*. In circuit switching it is assumed that end nodes use the network mostly for phone-calls. The main concept behind engineering voice networks is that not all users will initiate a call at the same time. Thus, a large number of users can share the same resources on a *call-by-call* basis. Statistical models are used to describe how often users actually initiate a call (*call arrival rate*) and how long is the call duration (*call holding time*). Based on these parameters, one can estimate the link capacities required to reduce the probability that the network will reject a call because of lack of resources. This was the original concept of statistical multiplexing of resources among a large number of users.

When we move to packet-switched networks, multiplexing is done on a *packet-by-packet* basis as opposed to a call-by-call basis. The lengths of times (timescales) of interest are much shorter. The reason for our interest in these shorter timescales is derived from the nature of data applications. Users would like to have a constant high-speed connection to the network and transmit very fast only for short intervals of time. Consider the operation of the Web (WWW). Users access Webpages, process the received data (i.e., read the context), and make decisions about accessing more data (i.e., follow links). The response time of the network is critical, and the duration of the connection to the Internet might be very long.

During the late 1990s, when many users were connecting to the Internet using modems and their telephone lines, the phone network was constantly overloaded. The traffic models used to engineer the network assumed that the call holding time is around 3 min. These models were failing to capture the actual behavior of end users however, since an increasing number of users kept their phone lines busy for hours when connected to the Internet. Although the network would reject calls, the bandwidth of the lines was not fully utilized.

3.2. Traffic Scheduling

Since packet-switched networks are used for a variety of end-user applications, it is reasonable to expect that different applications might have different requirements from the network. For example, email messages might be stored in the routers for several seconds without any problem. On the other hand, messages that carry voice must be delivered immediately since they are part of an interactive communication. Similarly, some applications might be able to afford information loss, whereas for other applications the network must support mechanisms that will guarantee that all information is delivered without packet losses.

Let us consider the network of Fig. 2, and let us assume that the core link capacity is 2 packets per second. In the previous case (Fig. 3), we assumed that packets are served in an first come–first served order (FCFS or FIFO) (where packets are transmitted in the same order as they were received). If traffic from user A is real-time traffic and requires minimum possible delays, whereas traffic from user C is email traffic, we can modify the way packets are selected for transmission. When packets from both users are buffered in the switch, it will always transmit packets from node A first (*static priority* order). The sequence of departures is also shown in Fig. 3. Real-time packets see no delay, whereas email packets see a maximum delay of 2 s. Thus, the method used for selecting packets for transmission (or *traffic scheduling discipline*) can determine the maximum and average queueing delays of different users. Notice however, that if the scheduler always transmits packets when packets are available in its queues, the average delay of *all* packets does not depend on the traffic scheduling discipline. Interested readers are referred to Zhang [4] for an overview of various traffic scheduling mechanisms.

4. CONNECTION-ORIENTED VERSUS CONNECTIONLESS NETWORKS

There is one main taxonomy of packet-switched networks, which is based on the method used to decide how packets are forwarded (or *routed*), and how resources are allocated [5].

In the previous examples we concentrated on a switch where traffic from all input links is multiplexed on a specific output link. This type of a switch is also known as a *multiplexer*. In the more general case, however, a packet switch might receive traffic from several interfaces and forward the packets to different interfaces. A large network will consist of multiple packet switches as shown in Fig. 4. When a packet arrives in such a switch a decision must be made as to which is the outgoing interface.

There are two main philosophies or methods used to resolve this question:

1. *Connection-Oriented* or *Virtual Circuit Switched Networks*. In these networks a virtual circuit is established between the source and destination nodes before any communication starts. The establishment of the virtual circuit does not necessarily lead to an explicit resource allocation as in the case of circuit-switched networks. However, some state information is associated with each switch that determines how all packets that belong to this connection must be forwarded. When a packet is created, a unique connection identifier is attached to it. All intermediate nodes will use this connection identifier to determine the output interface for this packet. The path that all packets of the connection follow is determined a priori during the connection setup phase. Technologies such as *asynchronous transfer mode* (ATM) [6] and *multiprotocol label switching* (MPLS) [7] are based on this principle.
2. *Connectionless Networks*. In these networks no explicit path is established a priori. Every end node in the network is associated with some address, and every switch or router has a "view" of the topology of the network. In other words, a *forwarding table* is stored in every switch with an entry corresponding to every other node. This table determines on a packet-by-packet basis, the interface that must be

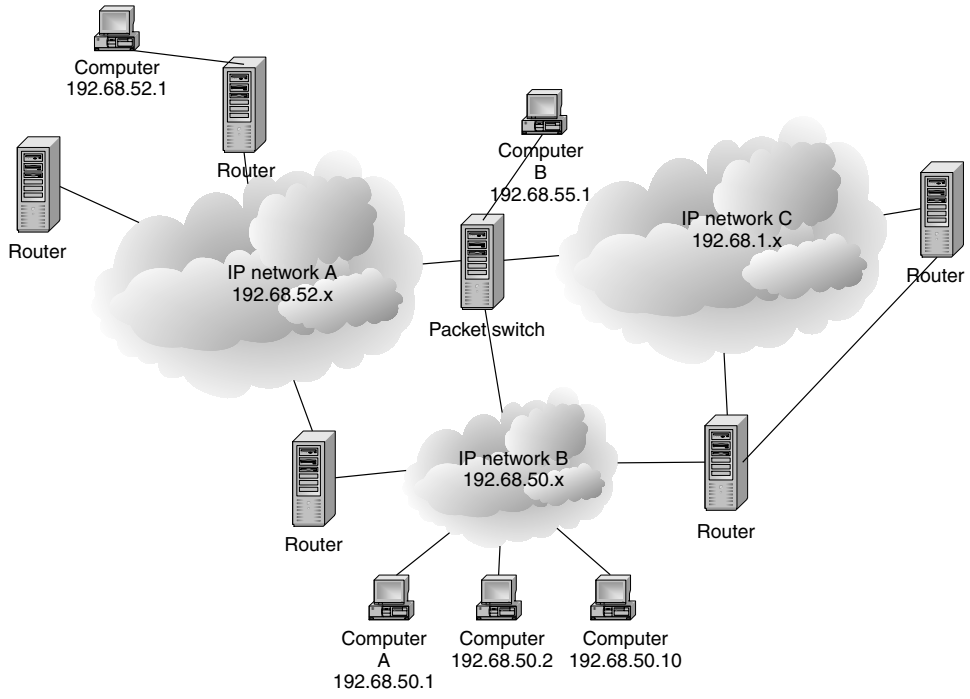


Figure 4. Global Internet architecture as an interconnection of multiple networks.

used in order to reach specific end nodes. Packets carry a *destination address* that identifies the end nodes, and switches use only this information to decide how to forward them. The network supports the mechanisms that allow individual switches to discover the topology of the network and create this forwarding table. Networks based on the *Internet Protocol* use this approach.

A connectionless network operates very much like regular mail and the post office. Senders address letters to specific recipients and the post office uses the address to forward the letters from one city to another and finally to the recipient. The originating post office cares only about the destination town or state. Local post offices make decisions about street addresses and apartment numbers. Similarly, in a packet network, routers in the middle of the network need to know only the router that is closest to the destination user. Only the final (or *edge*) router needs to know about specific users.

An important feature of connectionless networks is that packets between two end nodes can be routed through multiple paths of the network at the same time. For example, subsequent packets from workstation A to workstation B in Fig. 4 might take different paths. It is exactly this feature of connectionless networks, however, that makes them extremely reliable. If one of the nodes fails, packets can be still forwarded to their destination through a different path. This is known as the *shelf-healing* property of IP networks. On the other hand, in a connection-oriented network if a node or a link fails, a completely new path must be established from scratch. Since establishing a path might take a long period of time, traffic may be interrupted during this period and packets may get lost. For this reason connection oriented networks support techniques where multiple paths are reserved a priori.

Note that, depending on the path delays of a connectionless network, it is possible for packets that use different paths arrive to their destination out of order (i.e., if packet p1 is transmitted before packet p2 from node (workstation) A, it is not guaranteed to arrive at node B before packet p2). For this reason, the network must support mechanisms that will reorder packets at the end node.

In addition to these differences, which are related mainly to how the forwarding decisions are made, there is another fundamental distinction between the two philosophies. In connection-oriented networks, that are very similar to traditional phone networks, the end user must notify the network in advance of the bandwidth and the type of service he/she wants to use. The network uses this information to *admit* the user, avoid congestion, and offer different services. In connectionless networks the user transmits the packet to the network and hopes that enough capacity is available for the packet to be delivered to its destination. The network does not promise to any node that packets will be delivered after a predetermined delay, but it does its best (*best-effort networks*). The user is responsible for adapting his/her bandwidth requirements based on feedback received by the network, in order to prevent congestion.

5. PROTOCOLS AND LAYERING

During the design of networks (both packet-switched and circuit-switched), it became apparent that a large number of technologies and architectures must work together. For example, packet networks might work on top of Ethernet, and Ethernet is defined to work either over copper using electrical signals or over fiber using optical signals. The Ethernet network can also interchange information with a wireless network or a network using strictly optical signals [like a SONET (synchronous optical network)].

In order to allow a variety of transmission media to interwork with a variety of networking architectures and physical links, the notion of protocol layering was introduced [8]. First, a *protocol* can be considered as a set of rules, messages, and behaviors, that allows two end nodes to communicate to each other. Networks are built using a layered architecture of protocols.

To understand the concept of layering, let us consider the case of the voice network. End users know only about telephone numbers. When they want to communicate with other users, they dial a number on their phone, which can be considered as a *module*. The telephone will interact with the telephone network, that is another module, to establish a connection and the end users can begin talking. The user does not need to know how the telephone works, and the telephone does not to know how exactly the telephone network establishes a connection. However, the user expects a behavior from the phone, and the phone expects a behavior from the network.

In other words, each module in the network follows a set of rules to communicate with another module and expects a prespecified behavior. In the case of packet-networks, there are different modules in the network architecture that decide how packets are formed and how delivery is guaranteed. These modules expect a specific behavior from modules downstream, and finally from the links that are used to interconnect network nodes.

Figure 5 illustrates the most commonly accepted layer architecture as defined by the reference model of the Open Systems Interconnection (OSI) model developed by the International Standards Organization (ISO). Most networks follow this model. A description of the different layers follows.

5.1. Physical Layer

The physical layer defines the electrical, optical, electromagnetic or other properties of actual physical links. Examples are electrical signal voltages, signal frequencies, coding schemes, clock recovery and distribution schemes, and error correction schemes. Layers above the physical layer do not need to worry about such details, and they expect usually a low loss transmission of information among nodes.

5.2. Data-Link Layer

The data-link layer provides a variety of mechanisms that offer a logical communication path between two nodes. Mechanisms defined are packet formats, error checking, and retransmission mechanisms. A commonly used data-link layer is encountered in Ethernet networks. The basic

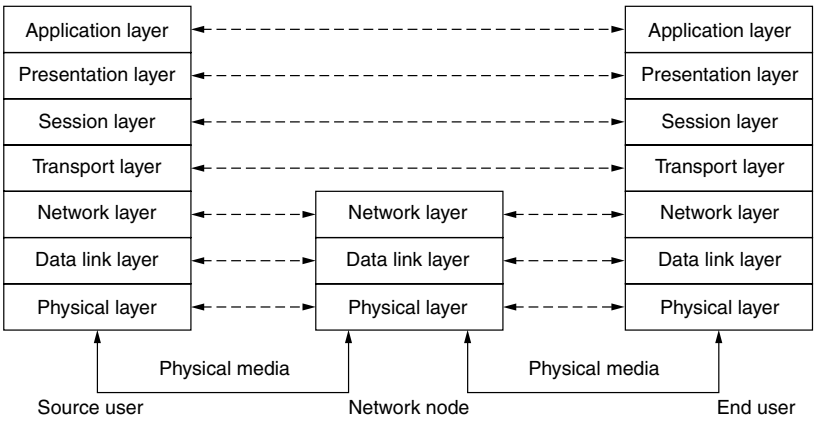


Figure 5. Network protocol layers.

principle of Ethernet (or any *multiple-access network* in general) is that several nodes are connected through the same physical media. Each node broadcasts information on the media, and only the destination node actually uses this information. The data-link layer will designate the source and destination nodes, and it will resolve contentions that appear when multiple nodes try to use the media at exactly the same time. The operation of this layer might involve retransmission of packets if communication fails. The data-link layer is limited to communication between two neighboring network nodes only.

5.3. Network Layer

The main function of the network layer is to determine where and how packets must be forwarded. When a packet arrives to this layer, the network layer uses local information to determine whether this packet is destined for some local process or it must be transmitted to another node. For packets arriving from higher layers, the network layer will format the packet in such a way that it will be recognized by other nodes in the network. The network layer does not care for the method used to transmit a packet to a neighboring node, and it assumes that the data-link and physical layers will provide this communication path. The *Internet Protocol* (IP) is an example of a network-layer protocol.

In some networks, the network layer might also assist with flow control. It might provide mechanisms that will avoid or reduce the probability of congestion. For example, if a downstream node is congested, information might be provided to the upstream nodes to reduce their transmission rate. The nodes might pick a different path to transmit traffic or they might propagate this information all the way back to the end user.

5.4. Transport Layer

The transport layer is responsible for establishing a logical connection between two end nodes. Connections might have properties such as bidirectional or unidirectional, or error-free. The transport layer will receive information from higher layers, packetize it, and pass it to the network layer. The transport layer might also verify that information is delivered correctly to the other application, by adding error checking mechanisms and defining the end-to-end protocol for retransmissions that will fix errors. In connectionless networks the transport layer will also guarantee in-order delivery of packets. The Transport Control Protocol (TCP) and User Datagram Protocol (UDP) encountered in the Internet are examples of transport-layer protocols.

5.5. Session Layer

The session layer is the user–network interface. It translates a user request to a request from the network. Users do not need to know how the network establishes connections and how it guarantees packet delivery. The session layer maps user requirements to network functions. If the application needs an error-free communication, the session layer will interface to a transport protocol like TCP. If the application needs an one-way transmission of information, where error-free delivery is not required, the session layer will interface to a transport protocol like UDP.

The session layer might also assist with resolving names to addresses that are recognized by the network. For example, end users do not really know about numeric addresses used by the transport and network layers (like IP addresses), but mostly remember mnemonics (like World Wide Web addresses).

5.6. Presentation Layer

The presentation layer might apply specific transformations on the data when they are delivered across the network. For example, data might be encrypted to prevent any other users from receiving the information. Sometimes data might be compressed in order to provide a speedier delivery. The presentation layer will also determine issues like ordering of bytes in big-endian or little-endian formats.

5.7. Application Layer

The application layer consists of end-to-end applications that use the network as a means of providing a service. An example of an application layer protocol is HTTP, which is the protocol used between a Web browser and a Web server.

5.8. Assigning Tasks to Layers

Although in the previous sections we gave some definitions of layers, depending on the network technology some of the functions can be moved to different layers. Flow control

is a clear example of such a function. In the case of ATM networks, flow control or congestion control is a part of the network layer definition, and *all* nodes in the network are responsible for assisting in this task. On the other hand, in IP networks, congestion control is a transport-layer function, and intermediate nodes do not interfere. The basic concept of TCP is that end nodes will detect packet losses, and will use these losses as an indication of congestion (i.e., several packets were queued on the same node and some packets had to be dropped). When congestion is detected the rate of data transmission of the end nodes is decreased. One can easily notice that if the data-link and physical layers do not guarantee error-free transmission, when a packet is lost because of link quality, it will be misinterpreted by TCP as an indication of congestion. This can lead to performance problems that have been the focus of several studies [9].

6. INTERNET AND TCP/IP

In this section we briefly discuss the principles of the most popular packet-switched network, the Internet. The basic architecture follows the principles of connectionless networks as described earlier [10] (Fig. 4). Data are forwarded in terms of packets of variable size, and the maximum packet size is 64K bytes (64 kilobytes). A header is associated with every packet that includes among other fields the addresses for the source and destination nodes of the packet. Routers use this header to determine how packets must be forwarded.

One can imagine the Internet as being composed of a large number of independently controlled networks without centralized control. Each network consists of three basic components:

1. A global assignment of addresses to nodes
2. Routers that forward packets to different interfaces and run the protocols that allow them to decide how to forward packets (*routing protocols*)
3. End nodes that use the TCP/IP protocol stack

6.1. Addressing

The Internet architecture provides a global means for assigning addresses to end nodes or routers, and once an address is assigned to a node, it is unique across the Internet. Whenever another user wants to communicate with this node, it must forward packets toward this address. IP version 4[1] addresses are 32 bits or 4 bytes. A decimal notation of four numbers separated by a dot is used to represent these addresses. For example 192.68.52.1 is an IP address that maps to the 32-bit number *C*0443401 (Fig. 6).

In order to minimize administrative overheads a hierarchical method of address assignment, was developed. Several end nodes that can be accessed through similar paths are assigned addresses within a specific range. Take for example, the networks in Fig. 4. Networks A and B

[1] We will refer to IP version 4 address within this text. There are newer versions of the IP protocol (IPv6) that use a slightly different address space.

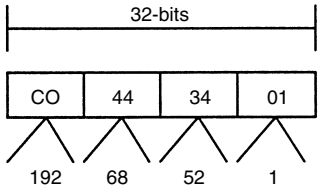


Figure 6. Representation of IP addresses.

are two global networks that can span the same country or continent. There are only two routers connecting networks A and B, and the IP addresses assigned to nodes on network A can be all aggregated together in a single number 192.68.52.x. This means that every node in network A has an address between 192.68.52.0 and 192.68.52.255. Similarly, IP addresses within network B are aggregated as 192.68.50.x. When any node in network B wants to communicate with a node in network A, the packets must go through one of the paths that interconnect networks A and B.

Note, that even though two nodes might be located in the same area they might connect to the Internet using a different network. Their IP address is determined by the network they connect to and not by their physical location. This is contrary to the method used in the telephone network, where address assignment is based on geographic boundaries (i.e., all telephone numbers within the same area have the same area code).

The IP addresses in these examples can be thought of as consisting of two components: (1) the *network number*, which is the 24 most significant bits, and identifies the network where a host resides; and (2) the *host or interface address*, which is the 8 less significant bits, which identify individual hosts within the network. The maximum number of hosts in a given network depends on the number of bits that are used to identify the network.

Originally, the IP address space was partitioned into four *static* classes of addresses (classes A–D) with different number of bits assigned to the network number. For example, a class A address uses 8 bits for the network number and 24 bits for hosts. Only a small number of networks (127) can have a class A address, and they were assigned to large organizations. The problem with this approach is that if an organization does not need all these addresses, a significant portion of the address space is wasted. For example, if a network is assigned a class A address, it can have up to 2^{24} interface addresses. If the network needs only a quarter of them, the rest will remain unused.

As the popularity of the Internet increased, it became apparent that the address space must be utilized properly, and that wasting address space resources can limit the scalability of the network. The new version of IP (version 6) is designed to address this problem, and it introduces 128-bit addresses. In the meantime, however, the concept of classless interdomain routing (CIDR) was introduced to allow a more efficient use of the address space [11]. Each network is now identified by two numbers: (1) an IP address and (2) a mask that determines the number of most significant bits of the address that identify the network. For example, a network might be identified with an address of 192.128.0.0/9, which means that the 9 most

significant bits correspond to the network number, and the 23 less significant bits identify hosts within this network. This method allows a variable number of hosts in any given network, and the address space is utilized more efficiently.

6.2. Routing

The second component is routing. This is a set of protocols run by the IP routers that allow them to construct a local view of the topology of the network. The routers must know which interface to use in order to reach a specific IP address in the network. However, routers do not need to maintain a distinct entry for each individual address. Because of the aggregation method described earlier, routers in network A can maintain a single entry for all nodes in network B, since they are all reachable through the same path [12]. This is actually the main benefit of the hierarchical address assignment. Routers at the core of the network must only maintain state information that determines how different networks are reached. Routers at the edge of the network must maintain state information on individual hosts.

In order for the routers to discover the network topology, they exchange information. Such information might include the list of neighboring routers or information learned from other routers in the network. For example, router R1 knows that one of its interfaces is connected to network A. It will thus notify the routers or end nodes in network B that if they want to reach network A, they must forward packets toward R1.

Once the network topology is discovered, most routing protocols assume that the network is represented by a graph, and routing is the task of finding the shortest path between any two nodes in this graph. Other criteria, such as finding the least-congested path, might be used in order to optimize network performance. Routing protocols are distributed among the various routers, and there is no centralized place that decides how packets must be forwarded. Each router makes independent decisions, and it is crucial that the routing algorithms lead to a stable network configuration and do not create loops. Assume two routers A and B and a destination C. Let us assume that router A views that in order for a packet to reach destination C, it must send the packet to node B. If router B has also an entry that says that if a packet is destined to destination C, it must be forwarded to node A, then any packet destined to node C will be constantly transmitted between nodes A and B.

As we mentioned earlier, the Internet can be considered as an interconnection of various networks. Each of these individually managed networks is called an *autonomous system* (AS). Routers that belong to the same AS use an *interior gateway protocol* (IGP) to exchange information, and forwarding is mostly based on shortest-path criteria. Routing between ASes uses an *exterior gateway protocol* (EGP), that is based mainly on policies determined by contracts between network operators.

6.3. TCP/IP Stack

The third component is that all end nodes must follow the TCP/IP protocol stack. This means that they must encapsulate packets using the IP headers and use the correct destination IP addresses. Most IP traffic uses one of two basic transport protocols. When a reliable end-to-end communication is required without any packet losses, the TCP protocol is used. TCP allows end users to open a logical "connection" to another end user. Once the connection is established, TCP enables the transmission of data between the end users and guarantees reliable delivery. For example, if a packet is lost in the network, TCP will take care of retransmissions.

TCP is also an important part of the congestion control mechanisms of IP networks. The basic principle is that end users must adapt their transmission rate to the bandwidth that is available from the network. Users start with a low transmission rate and increase the rate periodically. When losses are detected, they are interpreted as an indication of congestion and the rate is dropped. This process is constantly repeated, and allows TCP to calibrate the transmission rate to the available resources. One can consider this as a *closed-loop* flow control protocol, since information from the network is used to adjust the rate of the transmitter [13].

When nonreliable communication is sufficient, the User Datagram Protocol (UDP) can be used. UDP establishes an simplex connection between two users and allows delivery of data without flow control or error recovery. UDP is especially useful in real-time streaming applications, like video or voice, where small data losses are acceptable.

6.4. Domain Name Servers

There is an additional layer, handled by the *domain name servers*, that maps mnemonic addresses to IP addresses and can be regarded as a directory service. Most users are aware of World Wide Web (WWW) addresses such as *www.wiley.com.* These addresses must be translated to actual IP addresses before a communication can start. End nodes first send a request to a DNS asking for the IP address of the mnemonic address of the destination. The DNS will translate the WWW address to an IP address. When the address is resolved, the end node can use the IP protocol to communicate with the desirable destination. Note that the DNS address itself is statically configured by the end user.

7. HISTORY OF PACKET-SWITCHED NETWORKS

The first theoretical work in packet switching appeared in L. Kleinrock's Ph.D. thesis in 1962, and it is still considered as forming the theoretical foundation [14]. Around the same time Baran invented the fundamental concepts behind store-and-forward switching [15]. Among others, Baran's work developed the concepts of packets or messages, adaptive routing based on failures, and decoupling between logical and physical addresses. Similar concepts were developed independently in the Cyclades packet-switched network in France [16].

The telecommunications industry did not pay much attention to these concepts until the U.S. Department of Defense Advanced Research Project Agency (DARPA) sponsored a research program for the development of ARPAnet. The ARPAnet was mainly developed in order to provide a reliable and low cost communication network

among timesharing systems scattered throughout the country. This can be considered as the first wide-area packet-switched network, that was demonstrated by 1969. By 1972 the ARPAnet had 4 hosts, expanding to 23 by 1973.

From that point on the expansion of the Internet has been exponential, and it entered our everyday lives with the development of the World Wide Web in the early 1990s. Currently millions of hosts are interconnected in the Internet through a maze of networks without any centralized control, and data traffic is increasing exponentially.

BIOGRAPHY

Dimitrios Stiliadis received his Ph.D and M.S. degrees in computer engineering from the University of California at Santa Cruz, in 1996 and 1994, respectively. Prior to that he received the Diploma in computer engineering from the University of Patras, Greece, in 1992. Since 1996, he has been with the High-Speed Networks Research Department of Bell Laboratories, where he is currently a distinguished member of technical staff. During these years he has been leading the architecture of several generations of packet switching equipment. His recent research has been in issues related to traffic management, switch scheduling, and applications of optical technologies to packet networks. He is a corecipient of the 1998 IEEE Fred W. Ellersik Award.

BIBLIOGRAPHY

1. W. Stallings, *Handbook of Computer-Communications Standards*, Vol. 2, *Local Network Standards*, Macmillan, 1987.
2. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.
3. R. Guerin, H. Ahmadi, and M. Nagshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Select. Areas Commun.* **9**(7): 968–981 (Sept. 1991).
4. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *Proc. IEEE* **83**(10): 1374–1396 (Oct. 1995).
5. S. Keshav, *An Engineering Approach to Computer Networking*, Addison-Wesley, 1997.
6. D. E. McDysan and D. L. Spohn, *ATM Theory and Applications*, McGraw-Hill, 1998.
7. B. Davie and Y. Rekhter, *MPLS: Technology and Applications*, Morgan Kaufmann Publishers, 2000.
8. H. Zimmerman, OSI reference model — the ISO model of architecture for open systems interconnection, *IEEE Trans. Commun.* **28**(4): 425–432 (April 1980).
9. H. Balakrishnan, V. N. Padmanabhan, S. Sheshan, and R. Katz, Comparison of mechanisms for improving TCP performance over wireless links, *Proc. ACM SIGCOMM '96*, Sept. 1996.
10. W. R. Stevens, *TCP/IP Illustrated Volume 1, 2, 3*, Addison-Wesley, 2000.
11. V. Fuller et al., *Classless Inter-Domain Routing*, RFC 1519, *ftp://ds.internic.net/rfc/rcf1519.txt*, June 1993.
12. R. Perlman, *Interconnections: Bridges, Routers, Switches, and Internetworking Protocols*, Addison-Wesley, 2000.
13. V. Jacobson, Congestion avoidance and control, *Proc. ACM SIGCOMM 88*, Aug. 1998, pp. 314–329.
14. L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964.
15. P. Baran, On distributed communications networks, *IEEE Trans. Commun. Syst.* (March 1964).
16. B. M. Leiner et al., *A Brief History of the Internet*, Internet Society *http://www.isoc.org/internet/history/brief.shtml*.

PAGING AND REGISTRATION IN MOBILE NETWORKS

CHRISTOPHER ROSE
Rutgers WINLAB
Piscataway, New Jersey

1. INTRODUCTION

A communications network routes messages from senders to recipients. In *fixed* networks where units are not mobile, the terminal (such as a telephone handset, computer, video display, or a host of other possible devices) resides at a fixed physical location which rarely changes. In contrast, a *mobile communication network* routes messages between senders and receivers who may often change location. The seemingly simple addition of terminal mobility to the networking problem complicates it in both obvious and subtle ways.

We will start by exploring conceptually simple methods for accommodating mobility and then examine the deeper implications of these methods on network organization and design. However, rather than plunging directly into what could easily become an opaque technicalese treatise, it is easiest to use an analogy that we will expand as needed. So, consider a postal address, 536 West 145th Street, Apartment 21 in New York City — the author's childhood residence. Any messages for the author would be delivered to this address via a mail carrier or through a specific set of copper wires running from a central office to this address. The "address" of the telephone was (212) AU1-4676, which could be reached from anywhere in the "developed" world at that time. This basic fixed scenario was the dominant communications network structure — static identifiers corresponding to fixed physical locations for the terminal equipment, and tacitly, for the reams of equipment necessary to carry traffic between arbitrary addresses. For the postal service this would include postoffices and the city streets along which mail carriers traveled. For telephone service this would mean central switching offices and the cables that threaded their way under and around the city between central offices and homes.

Of course, people move about in their day to day lives, and the fixed scenario could only awkwardly accommodate mobility. A mail carrier arriving before August 6, 1970 would have been able to properly deliver mail to the author. After that date, without a *forwarding address*, the mail would remain undelivered or marked "Return to sender." Likewise, a telephone call that arrived while the

author was in school across the street at PS 186 would be missed, or the caller would have to be given a new number corresponding to the new physical location. At the time there was no (inexpensive) method to guarantee delivery of messages to a recipient in motion.

This simple analogy illustrates that having fixed terminals associated with fixed physical locations makes routing information through any sort of network a relatively straightforward task in principle. Certainly there are details such as communications link congestion/availability that the service provider must handle, but the basic notion of network topology stasis remains intact.

Now, suppose that units require messages to be delivered wherever they happen to be. The network needs to know exactly where the unit is — or in the parlance of the mobility management field — the unit's *point of network attachment*. There are two ways the network can ascertain the unit's location: (1) the network can search for the unit in likely places (paging) or (2) the unit can tell the network where he/she is (registration) — and we have thus provided an en passant definition of paging and registration as basic building blocks (or *primitives*) for handling mobility in communications networks.

The basic ideas of paging and registration lead to a variety of techniques that we will explore more carefully in Section 2. However, before proceeding we pause here to note that "units" need not be rigidly defined. In the Internet age where computer programs such as *Web crawlers* might autonomously search the Web for information, a *unit* could in principle be a *program*. Conceptually, paging and registration for such programs is not a big leap; however, from the network management standpoint one must consider that such programs could change their physical location *much* more rapidly than any "physically realized" unit such as a person or a piece of hardware and this could lead to unique stresses on the network.

Finally, suppose that in addition to keeping track of mobile units, the actual network topology were labile. By analogy, imagine the perplexity of a mail carrier who found not only that had the author moved from 536 West 145th Street but also that the connecting avenues and streets had changed relative locations as well! The analog to this somewhat nightmarish scenario might be the rule rather than the exception in some types of ad hoc network architectures where the communication infrastructure is *composed* of mobile nodes communicating wirelessly as opposed to the usual sets of fixed location equipment such as cables, microwave towers, and switching centers.

Interestingly (and perhaps obviously), all these scenarios can be handled by suitably abstracted versions of the basic paging and registration paradigm. We therefore carefully describe paging and registration along with their associated costs for conventional mobile networks in following sections and later apply the concepts to the more exotic scenarios which will almost certainly arise in the future.

2. PAGING AND REGISTRATION

The key idea behind paging and registration is that routing messages to a unit is a cooperative affair where the network makes an effort to find the unit through paging and the unit, who needs to be found, registers its location with the system. The optimization problem arises since both paging and registration require some sort of communication, and communication bandwidth is a valuable commodity — thus, there are *costs* associated with paging and registration.

Specifically, it is obvious that a unit could always be immediately located by the system if that unit always registered each change in location, and in response the system updated its global routing databases. However, the network cost of such updates, especially for units whose point of network attachment changed often could be prohibitive [1–6]. Thus, mobile networks trade unit localization delay against the cost of perfect location information available through constant registration. In other words, some uncertainty in the actual unit point of attachment is tolerated and is resolved by seeking the unit through paging — which requires time and some cost in communications bandwidth. In addition, the unit concurrently offers periodic updates on location within cost constraints, or alternatively, the system actively seeks a unit when paging cost exceeds some threshold. Finally, if a unit never receives calls, then there is little need to ever register location since the system will never need to find that unit. All these issues combined form the paging/registration problem.

2.1. Paging in a Typical Cellular System

Consider a cellular telephone system where units can reside in one of N possible locations as in Fig. 1. These locations might be associated with cellular base stations or be *location areas* composed of many cells associated with a mobile telephone switching office (MTSO). If the unit location is uncertain, then the system must find the unit when an incoming call arrives. To do so, in a wireless system, paging messages are sent to possible unit locations. The paging message uniquely specifies the desired mobile terminal and thus requires passage of information, which further implies use of bandwidth — the most precious resource in a wireless system. The system could page a unit in all possible locations simultaneously and could in principle find a unit almost immediately.

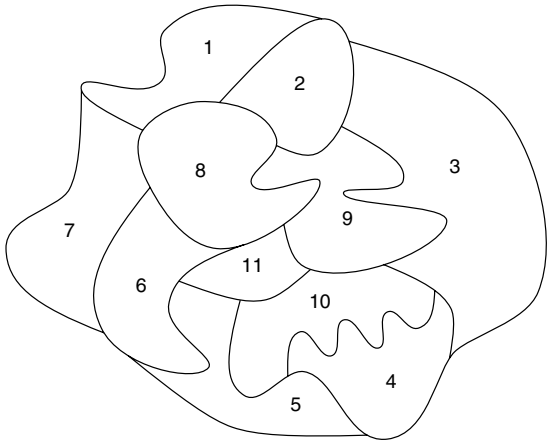


Figure 1. Possible locations associated with a mobile unit.

However, paging a single unit in all locations precludes paging more than one unit simultaneously, and in systems where signaling channels and traffic channels are shared, consumes resources everywhere, which can actually increase the average delay in finding a unit [7,8] and/or degrade quality of service (QoS).

Thus, if some delay can be tolerated in finding a unit, sequential paging of a unit starting with the most likely location will minimize the expected amount of paging bandwidth used. If some delay bound must be met, then groups of locations can be paged simultaneously and in this way various points on a paging delay/cost performance curve can be achieved. Regardless, in all such paging problems, the key result is that locations should be searched in order of the *a priori* probability of unit residence [9].

For such *unit-centric* approaches to mobility management, the unit location probability distribution plays a key role in minimizing the system cost of paging; the problem of assembling and maintaining such information has been carefully examined [10,11]. Specifically, it was shown how the Lempel–Ziv empirical sequence coding method could be applied to constructing and maintaining compact mobility profiles for different users.[1] In addition, the tradeoff between paging group size and paging delay for multiple unit systems has been considered [7,8].

Of course, some practical issues must be mentioned.[2] There is the possibility that a unit, owing to the propagation environment, will not receive a page. Thus, paging messages may be repeated up to some system-specified number of retries before a failure is declared at a given location. This somewhat complicates the analysis, but the basic premise (search most likely first) still holds. It should also be noted that most current systems do not bother to maintain a dynamic register of likely unit locations, even though it could significantly reduce paging channel use [12] since the complexity of implementation is thought to exceed the benefit.

2.2. Registration

Classically, registration strategies have used the previously mentioned concept of a *location area* — groups of (usually) contiguous locations. These areas can be global in the sense that they are identical for all mobile units, or personal in that each unit has its own location area. The two concepts are illustrated in Figs. 2 and 3.

Under the classical scenario, a service area is partitioned into groups of locations. Incoming calls result in page requests at all locations in the appropriate location area. When a mobile unit crosses a group boundary, a registration is mandated. Location area boundaries are chosen based on incoming call rates, aggregate mobility patterns, and the relative cost of paging and registration.

[1] The papers by Bhattacharya and Das [10,11] received top prize at ACM MobiCom'99 and are required reading for anyone who studies mobility management

[2] Details about paging message structure and protocols for current systems can be found in (for example) the paper by Goodman et al. [7].

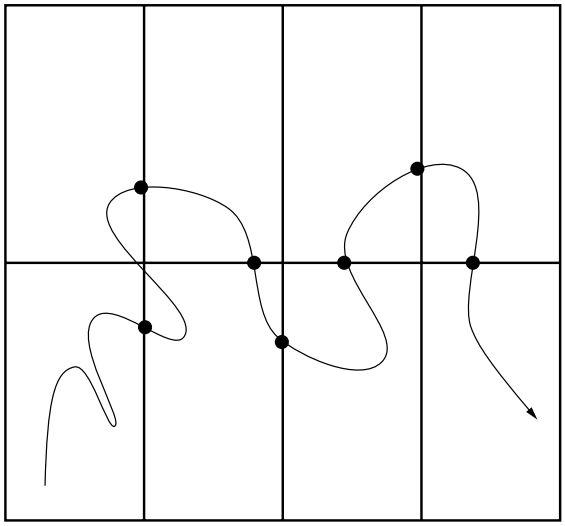


Figure 2. Illustration of classical registration strategy with geographically fixed location areas. Registration occurs at location area boundary crossings (denoted by solid dots). The arrival of an incoming call triggers polling requests over all locations contained in the current location area.

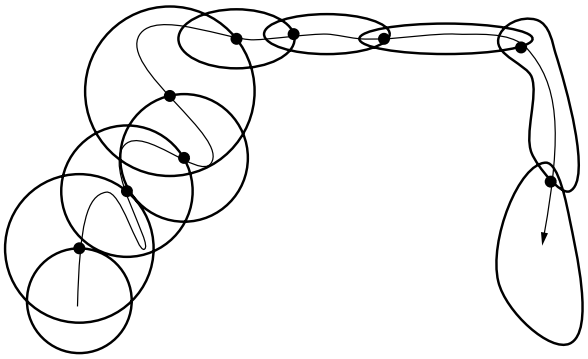


Figure 3. Illustration of personal location areas, recalculated after every contact with system. The location area shape may change with time depending on the mobility characteristics of the unit. For example, the mobile unit in this example might first be driving around a city and then enter a high-speed thoroughfare. The location areas change shape to accommodate the changed mobility. Registration points are denoted by solid dots.

Unfortunately the simple classical strategy suffers from spurious registrations since mobile units that dwell near boundaries can "pingpong" between location areas. Furthermore, since mobile units may have different mobility characteristics, designing location areas for the aggregate can be inefficient.

A key element of the registration problem is the rate at which calls arrive to the mobile unit. If calls never arrive, then the unit need never register since paging cost is always zero. Likewise, if calls arrive regularly then registration might not often be necessary because the network has reasonably accurate location information. Thus, the procedure used depends both upon the mobility process and the incoming call rate as well as the relative cost of registration.

Using these ideas, the concept of personal location areas was proposed [13,14] where each mobile unit contact with

the system results in the calculation of a new location area, centered around the current location. Since the boundaries are reset after each registration, this method does not suffer from spurious location updates, and since location areas can be designed for each unit, the problem of designing for the aggregate does not exist.

However, as with the classical scheme, all locations in a location area are paged in response to an incoming call even if the call arrives shortly after a previous call has terminated. Furthermore, since the boundaries of these location areas are fixed, high-velocity mobile units must register more frequently or require larger location areas, both of which tend to increase cost.

2.3. Paging/Registration Using Optimal Paging and the Mobility Index

Suppose a unit follows a brownian motion pattern[3] in which case the cost of paging is proportional to $(\mathcal{D}t)^{n/2}$, where n is the number of degrees of freedom in unit motion, τ is the time since the unit location was known exactly, and $\mathcal{D}$ is a constant [15]. It is useful to further define a *mobility index* as $\rho = \mathcal{D}/\lambda$, where λ is the incoming call rate. The inverse of ρ is known as the *call to mobility ratio* [16,17]. The paging cost can then be written as $(\rho t)^{n/2}$, where the new variable t, the time since last sighting, is measured in units of average call interarrival times $1/\lambda$. The utility of the mobility index lies in its explicit inclusion of the frequency with which a unit is contacted. For example, if the unit is often paged, then its whereabouts are better known to the system, thereby decreasing the unit mobility index.

These basic notions have been applied to registration strategies based on time and/or place [18]. Other work has also considered versions of this basic problem, for example, the paper by Bar-Noy et al. [19]. It is assumed that the mobile unit knows the paging method employed by the network and thereby knows the cost (or expected cost) to be incurred at the current point in (time, place). Assuming Poisson arrivals of paging requests makes for a relatively straightforward optimization.

When only time information is used by the mobile unit, a deadline registration policy is appropriate. Thus, the mobile unit should register at time τ^* after the time of last contact with the network, where τ^* depends on the mobility index and the relative costs of registration and paging. Intervening call initiations or received calls reset the timer. With both time and place information, however, optimality is an open question.

Application of the timer-only registration method leads to an improvement over simpler place-based methods [13], especially at higher mobile unit velocities since the optimal paging algorithm is affected only by location uncertainty and not directly by the unit velocity. Suboptimal registration methods based on time *and* place have been explored [20,21], and similar to an optimal method based on distance from last sighting derived for the random walk with drift velocity zero [22], threshold rules for registration were obtained. It was found that the (time, place)-based method performed only slightly better than the timer-alone method over a range of conditions. The primary improvement afforded by spatial information was a reduction in cost variance.

[3] Brownian motion is the simplest and most often used mobility model. In certain ways it constitutes a worst-case mobility scenario since for finite location variance (a surrogate for average energy expended in moving an object), the location uncertainty (entropy) is maximum.

3. PAGING, REGISTRATION, AND INFORMATION THEORY ABSTRACTIONS

3.1. Entropy and Paging Cost

Since paging amounts to issuing queries about unit location and these queries require signaling messages, two obvious questions arise:

How much information does the network need to resolve a mobile unit location at a given point in time?

What is the relationship of this information to the amount of signaling required?

The obvious simple answer to the first question is *the entropy of the location distribution*. Likewise, considering that locations vary with time, we see that the average information *rate* necessary to completely specify unit location for all time is the entropy rate of the motion process [23,24]. However, the structure of allowable queries about location profoundly affects the relationship between information content and the necessary signaling in the paging problem.

For example, consider the game of "Guess my number," where a number n is chosen between 1 and J according to some probability distribution p_j [25,26]. Using the standard information-theoretic formulation, the minimum average number of yes/no questions necessary to identify n is the entropy of the probability distribution. However, in a mobile communications system, each polling event at a given location requires signaling in the network, and possibly use of a radio channel as well in the case of a wireless system. Thus, from a signaling standpoint, the appropriate queries are of the form "Is your *number* (*location*) *k*?" This leads to problems with relying solely on entropy as a measure of location uncertainty. Specifically, one can easily describe distributions whose entropies exist and are finite, but for whom the paging cost is infinite [27].

3.2. Beyond Paging and Registration

We have so far considered only the effort associated with finding a given mobile unit through paging and registration; that is, the unit location was not known exactly. Here we take a slightly different view and assume that the location of every unit is known somewhere. Now we ask how much signaling effort is necessary to efficiently disseminate the location information over the network.

Consider then the *universal phone number* (UPN), where a single number is used to identify each mobile unit and route the call appropriately. At the heart of the UPN problem is a question similar to that which arose in the paging/registration context: *Who needs to know?* Thus, the frequency with which a given UPN routing entry is used is the primary index of the importance of maintaining its accuracy.

Let us assume that the answer to this question is given by the rate, r_{ij}, at which unit i calls unit j. This concept of calling rate, coupled to a view of mobility as a set of time-varying location distributions, allows a simple lower bound on the amount of information that must be disseminated over the network.

Specifically, let a random variable τ_{ij} describe a renewal process for the time between calls from unit i to unit j. We assume that τ_{ij} has a density function $f_{\tau_{ij}}(t)$ with mean $1/r_{ij}$. At the initiation of a call at time t, from the perspective of unit i, the location uncertainty of unit j is given by the location probability distribution $p_j(t, t_0, x_j(t_0))$, where $x(t_0)$ was the position of unit j at time t_0.[4] In information-theoretic terms, however, the entropy of this distribution represents the average amount of information required to exactly specify the location of unit j to unit i at time t.

Define $H_{ij}(t)$ with $t \geq t_0$ as the entropy in bits of the location distribution $p_j(t, t_0, x_j(t_0))$. The average number of bits needed by unit i to specify the location of unit j just before the next call is

$$\overline{H}_{ij} = \int_0^\infty f_{\tau_{ij}}(t') H_{ij}(t')\, dt' \tag{1}$$

and by renewal theory [28], the *absolute minimum* average number of bits per second needed by unit i to determine the location of unit j is simply $\overline{H}_{ij}/E[\tau_{ij}] = r_{ij}\overline{H}_{ij}$.

The minimum necessary aggregate network signaling load can then be determined by considering the average number of hops [29]; that is, location information from unit j must be routed to unit i and must traverse some number of nodes. As a simple example, assume M mobile units. With uniformly distributed mobile unit locations, we define γ as the mean number of hops from any node to the rest of the network. Therefore, the aggregate signaling rate associated with location information dissemination is

$$\mathcal{R} = \gamma \sum_{ij} r_{ij}\overline{H}_{ij} \tag{2}$$

The generality of this approach allows connection rates between *any* two entities to be defined; thus, not only "mobile units" but possibly routing tables as well — harkening back to the perplexed postman from the introduction who must travel streets that rearrange themselves from day to day. Possibly, this basic method can be extended to derive lower bounds for signaling costs associated with any mobility management scheme on any given network.

4. MOBILITY MANAGEMENT AND THE FUTURE

When considering mobility in present-day communications systems, what usually comes to mind is a person, conveyance, or mobile computer. However, in future networks, programs as well as physical objects might also be mobile. For example, suppose that you are an investment banker and seek financial market information distributed over the network. Specifically, you might wish to exploit small price differences over numerous "cybermarkets" — analogous to but more extensive than current arbitrage practices.

If the processing and communication capacity of the machines and network were infinite or there were no competing units, it would be a relatively simple task to spawn search processes on all machines in the network and have them report back information in some fashion. However, capacity constraints allow only a small number of programs to be launched into the network and run on other machines. For efficient search or to execute trades, the programs should be able to communicate and modify their behavior as information is gathered. For effective search, the programs should be able to relocate themselves to appropriate databases or markets. Such migrant programs charged with ferreting out information are generically called *mobile agents* [30–33].

The need for communication between agents implies a need for registration and paging since the agents are effectively mobile units to which calls may be routed. In motion processes that involve movement of mass, there are inherent constraints on motion between physical locations. Mobile agents, however, suffer few such constraints since their motion processes are influenced primarily by the location of contacts relevant to the search. Thus, one instant an agent could be active at a host in Los Angeles, and when done relocate to a suggested host in Madras.

For a simple random walk in n dimensions, paging cost is proportional to $(\rho t)^{n/2}$, where, once again, ρ is the mobility index and t is the time elapsed since last contact with the system (measured in units of intercall arrival). The intrinsic lack of constraints on both mobility index (how fast and far) and motion dimensionality (how many choices) for mobile agents suggests that groups of intercommunicating agents moving rapidly over a network of effective dimension $n \geq 3$ could severely stress signaling resources. Under this not-too-futuristic scenario, efficient mobility management could easily become the principal issue in network design.

5. CONCLUSIONS

The concept of mobility management based on time-varying mobile unit location probability distributions was introduced. Mobile units could be cellular telephone units, mobile computers or even mobile computer programs such as mobile agents. Regardless, the problem of finding mobile units boils down to the intuitively pleasing problem of searching for units in the most likely places as characterized by a unit location probability distribution.

Since location uncertainty is at the heart of the mobility management problem and information theory is the lingua franca of uncertainty, it is tempting to apply information theory to the paging problem, but sometimes unilluminating since even if the entropy of the location distribution is finite, it could still require infinite paging effort to find the unit on average. Thus, it is safest to rely directly on the ordered location probability function (ordered from most likely to least likely locations).

[4] Paging and registration are done by the *system* as opposed to the caller. Thus, the only a priori information available to the caller is the location distribution $p_j(t, t_0, x_j(t_0))$, which is independent of the registration/paging method used to track unit j.

We outlined various simple paging/registration procedures based on location probability distributions and compared them to more classical methods. A byproduct of this study was the definition of a *mobility index*, which is a useful reification of mobile unit location uncertainty and its growth as a function of time since the last contact. We then showed how paging/registration cost varies as a function of the mobility index.

We also showed how this cost varies with the dimensionality of the motion process which led to a consideration of non-classical mobility: groups of mobile programs ranging over a network in a coordinated way seeking out information. One could imagine agents that roam the network for information [33] or to buy and sell goods [31]. The results suggest that severe stress could be placed on a communication network by widespread use of such programs.

In the networks context, we then suggested ways in which time-varying location probability distributions for mobile units might be used to underbound the amount of signaling traffic necessary for distributing location information. In this case, an information theoretic approach *is* helpful.

BIOGRAPHY

Dr. Christopher Rose received the B.S. (1979), M.S. (1981), and Ph.D. (1985) degrees all from the Massachusetts Institute of Technology in Cambridge, Massachusetts, Dr. Rose joined AT&T Bell Laboratories in Holmdel, New Jersey, as a member of the Network Systems Research Department in 1985 and in 1990 moved to Rutgers University, where he is currently an Associate Professor of Electrical and Computer Engineering and Associate Director of the Wireless Networks Laboratory. He is Editor for the *Wireless Networks* (ACM), *Computer Networks* (Elsevier), and *Transaction on Vehicular Technology* (IEEE) journals and has served on many conference technical program committees. Dr. Rose was Technical Program Co-Chair for MobiCom'97 and Co-Chair of the WINLAB Focus'98 on the U-NII, the WINLAB Berkeley Focus'99 on Radio Networks for Everything and the Berkeley WINLAB Focus 2000 on Picoradio Networks. Dr. Rose, a past member of the ACM SIGMobile Executive Committee, is currently a member of the ACM MobiCom Steering Committee and has also served as General Chair of ACM SIGMobile MobiCom 2001 (Rome, July 2001). In December 1999 he served on an international panel to evaluate engineering teaching and research in Portugal.

His current technical interests include mobility management, short-range high-speed wireless (Infostations), and interference avoidance methods for unlicensed band networks.

BIBLIOGRAPHY

1. K. Meier-Hellstern, E. Alonso, and D. O'Neill, The use of SS7 and GSM to support high density personal communications, in J. M. Holtzman and D. J. Goodman, eds., *Wireless Communications: Future Directions*, Kluwer Academic, 1993.
2. M. J. Beller, E. H. Lipper, and M. P. Rumsewicz, Switching system impacts of PCS traffic, *2nd Int. Conf. Universal Personal Communications*, Ottawa, Canada, Oct. 1993.
3. G. Columbo, L. DeMartino, C. Eynard, and L. Gabrielli, Mobility load control in future personal communications networks, *2nd Int. Conf. Universal Personal Communications*, Ottawa, Canada, Oct. 1993.
4. C. N. Lo, S. Mohan, and R. S. Wolff, An estimate of network database transaction volumes to support voice and data personal communications services, *Proc. 8th ITC Specialist Seminar on Universal Personal Communications*, Santa Margherita, Italy, Oct. 1992.
5. E. H. Lipper and M. P. Rumsewicz, Teletraffic considerations for widespread deployment of PCS, *IEEE Networks* (*Special Issue on Nomadic Personal Communications*) (Sept./Oct. 1994).
6. E. H. Lipper, Switching system performance problems for universal personal communications, *Computer Networks and ISDN Systems*, 1995 (P. Enslow, Editor-in-Chief).
7. D. Goodman, P. Krishnan, and B. Sugla, Minimizing queueing delays and number of messages in mobile phone location, *ACM-Mobile Networks Appli.* (*MONET*) **1**(1): 39–48 (1996).
8. C. Rose and R. Yates, Ensemble polling strategies for increased paging capacity in mobile communications networks, *ACM Wireless Networks* **3**(2): 159–167 (1997).
9. C. Rose and R. Yates, Minimizing the average cost of paging under delay constraints, *ACM Wireless Networks* **1**(2): 211–219 (1995).
10. A. Bhattacharya and S. K. Das, LeZi-update: An information-theoretic approach to track mobile users in PCS networks, *Proc. ACM Mobicom'99*, Seattle, Aug. 1999.
11. A. Bhattacharya and S. K. Das, LeZi-update: An information-theoretic approach for personal mobility tracking in PCS networks, *ACM/Kluwer J. Wireless Networks* **8**(2): 121–135 (March 2002).
12. C. U. Saraydar and C. Rose, Minimizing the paging channel bandwidth for cellular traffic, *ICUPC'96*, Boston, Oct. 1996, pp. 941–945.
13. H. Xie, S. Tabbane, and D. J. Goodman, Dynamic location area management and performance analysis, *Proc. IEEE Vehicular Technology Conf. VTC'93*, Secaucus, NJ, May 1993.
14. S. Tabbane, An alternative strategy for location tracking, *IEEE J. Select. Areas Commun.* **13**(5): 880–892 (June 1995).
15. C. Rose, Minimizing the average cost of paging and registration: A timer-based method, *ACM Wireless Networks* **2**(2): 109–116 (June 1996).
16. R. Jain, Y.-B. Lin, C. Lo, and S. Mohan, A caching strategy to reduce network impacts of PCS, *IEEE J. Select. Areas Commun.* **12**(8): 1434–1444 (Oct. 1994).
17. R. Yates, C. Rose, S. Rajagopalan, and B. Badrinath, Analysis of a mobile-assisted adaptive location management strategy, *ACM Mobile Networks Appli.* (*MONET*) **1**(2): 105–112 (1996).
18. C. Rose, State-based paging/registration: A greedy approach, *IEEE Trans. Vehic. Technolo.* **48**(1): 166–173 (Jan. 1999).
19. A. Bar-Noy, I. Kessler, and M. Sidi, To update or not to update? *ACM Wireless Networks* **1**(2): 175–186 (1995).
20. C. Rose, *State-based paging/registration: A greedy technique*. Winlab-TR 92, Rutgers Univ., Dec. 1994.
21. C. Rose, A greedy method of state-based registration, *IEEE Int. Conf. Communications ICC'96*, Dallas, TX, June 1996.

22. U. Madhow, M. L. Honig, and K. Steiglitz, Optimization of wireless resources for personal communications mobility tracking, *IEEE Trans. Network.* **3**(6): 698–707 (Dec. 1995).

23. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

24. R. E. Blahut, *Data Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1988.

25. J. L. Massey, Guessing and entropy, *IEEE Int. Symp. Information Theory*, Trondheim, Norway, 1994, p. 204.

26. E Arikan, An inequality on guessing and its application to sequential decoding, *IEEE Trans. Inform. Theory* **42**(1): 99–105 (Jan. 1996).

27. C. Rose and R. Yates, Location uncertainty in mobile networks: A theoretical framework, *IEEE Commun. Mag.* **35**(2): 94–101 (Feb. 1997).

28. S. M. Ross, *Stochastic Processes*, Wiley, New York, 1983.

29. C. Rose, Mean internodal distance in multihop store & forward networks, *IEEE Trans. Commun.* **40**(8): 1310–1318 (1992).

30. P. Maes, Agents that reduce work and information overload, *Commun. ACM* **37**(7): 31–40 (1994).

31. A. Chavez and P. Maes, Kasbah: An agent marketplace for buying and selling goods, *1st Int. Conf. Practical Application of Intelligent Agents and Multi-Agent Technology*, London, 1996.

32. P. Maes, Intelligent programs, *Sci. Am.* **273**(3): 84–88 (Sept. 1995).

33. H. Lieberman, Letizia: An agent that assists Web browsing, *1995 Int. Joint Conf. Artificial Intelligence*, Montreal, CA, 1995.

PARABOLIC ANTENNAS

ALESSANDRO ORFEI
CNR, Istituto di
Radioastronomia
Bologna, Italy

1. INTRODUCTION

The aim of this article is to introduce the important tutorial matters related to the specific field of parabolic antennas and to give an overview as complete as possible on the key parameters and factors to understand the operation of this widely used tool to transmit and receive radiowaves. Parabolic antennas can be designed in various ranges of radiofrequencies, spanning from ~100 MHz to 100 GHz, and a many applications take advantage of this well-established and versatile technology. Whether very small parabolas are used for commercial links or large reflectors are needed to detect faint signals coming from the sky, the fundamentals are the same; the particular application will address which aspects among others have to be taken into account.

2. ANTENNA PARAMETERS AND CHARACTERISTICS

There are many parameters characterizing the operation and performance of parabolic antennas. Their importance in designing a system is dependent on the application; for instance, large antennas have to be carefully designed with respect to structural and mechanical constraints, because induced deformations, arising from gravity, wind, and temperature, dictate most of the performances. On the other hand, simple small parabolas used in receiving satellite TV signals need reasonable design, but the cost and ease of effective mass production are of primary importance.

2.1. Geometry of Used Configurations, Primary Focus-Fed Parabola

The simplest way to collect electromagnetic energy is to use a paraboloidal mirror only. Exploiting the geometric definition, a plane wave incoming at different points of the parabolic surface will be focused at the focus (called *primary focus*) of the parabola. If the phase center of a *feed* (or *illuminator*) coincides with the focus, the energy will be collected by the waveguide and then transformed in an electric signal (Fig. 1). A paraboloidal surface is obtained simply by rotating a parabola about its *focal axis*, that is the axis joining the *vertex* of the paraboloid and the primary focus. The distance between these two points is called *focal length*.

2.2. Geometry of Used Configurations, Cassegrain/ Gregorian System

Collection of electromagnetic energy could also be achieved by adding a portion of a hyperbolic shape as a second mirror. Exploiting the geometrical definition of hyperbola the energy will be focused at one of the hyperbole foci if the other one coincides with the paraboloidal primary focus. In this case the first one is called a *secondary focus* (F2) of the paraboloid and the phase center of the feed is placed to coincide with it. Again, the hyperboloid is obtained by revolution of an hyperbola about its axis. This kind of configuration is called *Cassegrain* system (Fig. 2a). The goal is also achieved if the second mirror is a portion of an ellipsoid (Fig. 2b). In this case the configuration is called a *Gregorian* system. In both systems the *secondary mirror* is also called a *subreflector*.

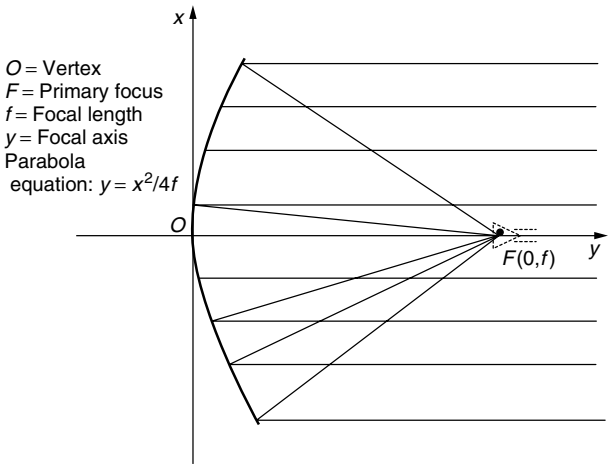


Figure 1. Primary focus-fed system.

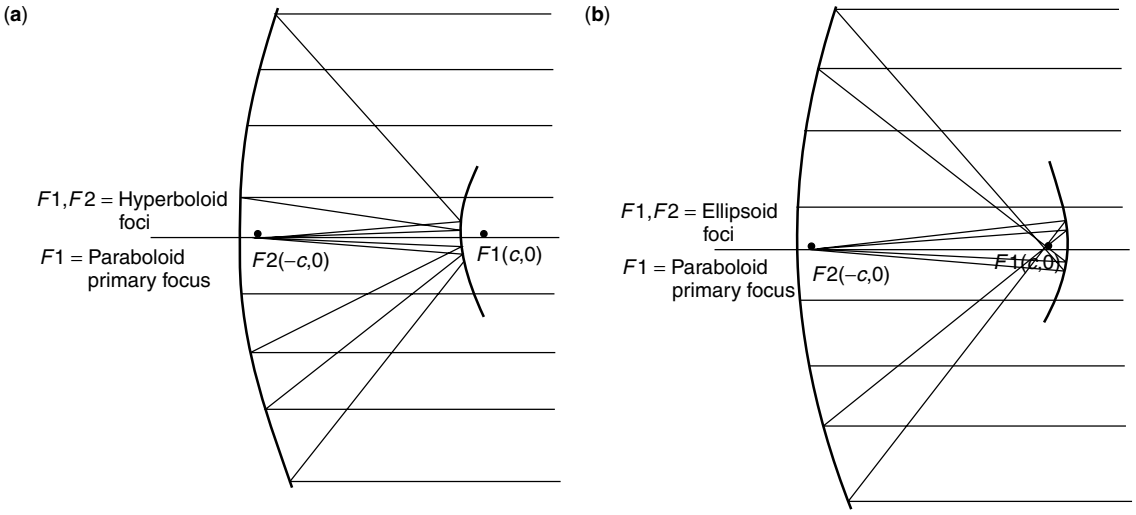


Figure 2. (**a**) Cassegrain system; (**b**) Gregorian system.

In applications where very high antenna efficiency and very particular antenna illumination are necessary, both mirrors have a shape that slightly deviates from a perfect parabolic/hyperbolic or parabolic/elliptic pair. In this case the configuration is called a *shaped* system. A shaped form of the illumination function can have a dip in the central portion of the mirror because that surface is obscured by the primary focus arrangement and so it is not useful in picking up a signal. The central dip also improves the return loss of the system, avoiding a standing wave.

2.3. Geometry of Used Configurations, Offset System

The *offset* system can be built starting with one of the three classic configurations shown in Figs. 1 and 2. Let's imagine that we remove part of the paraboloidal surface, keeping only the subsurface that collects the electromagnetic energy that doesn't interfere with either the feed or the secondary mirror (Fig. 3). This particular and very difficult-to-design configuration avoids the efficiency loss for the blockage effect due to the obstruction of the feed or secondary mirror. An impressive realization of this configuration is the GBT (Green Bank Telescope, Green Bank WV), a radiotelescope with the primary dish of 100 m in diameter.

2.4. Geometric Parameters

Referring to Fig. 4, seven parameters are used to describe and design the optics of the parabolic system:

D = diameter of the primary mirror
d = diameter of the secondary mirror
f = focal length
Φ_p = primary mirror edge half-angle
Φ_s = secondary mirror edge half-angle
L = secondary mirror depth
F = secondary mirror focal length

For describing the Cassegrain/Gregorian system, only four of them have to be fixed/the other ones are dependent by the following three equations:

$$\tan\frac{\Phi p}{2} = \pm\frac{D}{4f}\begin{cases}+\text{Cassegrain;}\\ -\text{Gregorian}\end{cases} \tag{1}$$

$$\frac{1}{\tan\Phi p}+\frac{1}{\tan\Phi s}=\frac{2F}{d} \tag{2}$$

$$1-\frac{\sin\left(\frac{\Phi p-\Phi s}{2}\right)}{\sin\left(\frac{\Phi p+\Phi s}{2}\right)}=\frac{2L}{F} \tag{3}$$

One of the most representative parameter for the design is the ratio f/D, because many of the electromagnetic characteristics have mathematical dependence on it.

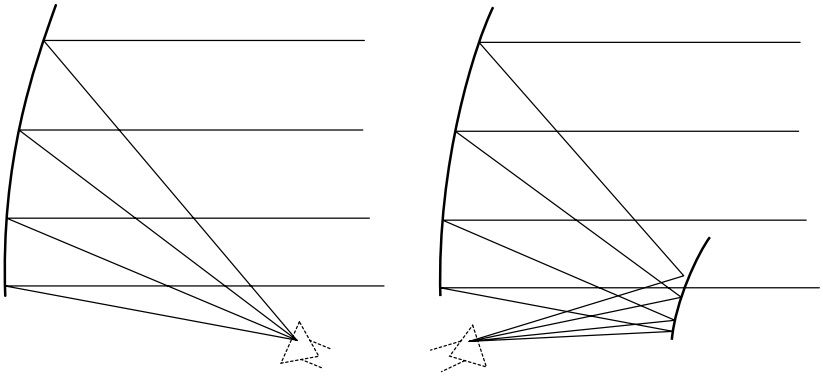


Figure 3. Offset configurations.

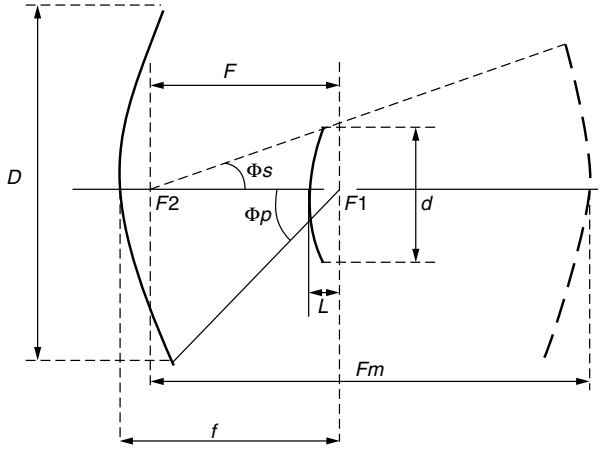


Figure 4. Antenna parameters.

Figure 4 also shows the concept of an *equivalent parabola*, which is useful in introducing the term known as *magnification* M, defined as the ratio between the focal length of the equivalent parabola and the focal length of the real system

$$M = \pm\frac{Fm}{f} = \frac{\tan\dfrac{\Phi p}{2}}{\tan\dfrac{\Phi s}{2}} (+\text{Cassegrain}, -\text{Gregorian}) \qquad (4)$$

The equivalent parabola has the same diameter and feed of the real system but is a paraboloid of different curvature. Its property is to focalize the incoming rays at the same point of the real dual-mirror system. This concept is useful in appreciating that a dual-mirror system gives a longer focal length, avoiding any lengthening of the mechanical structure in front of the primary mirror.

2.5. Electromagnetic Characteristics of Parabolic Antennas

2.5.1. Pattern. The combination of each geometry previously described and the illuminator gives the properties by means of which the antenna can irradiate or receive an electromagnetic signal. The first characteristic to be considered is the *pattern*. The antenna pattern relates the spatial distribution of the transmitted or received power [power pattern, $P(\theta, \phi)$]. Similarly, the pattern gives the value of the electric field at every point in space [field pattern, $E(\theta, \phi)$]. Usually the patterns are functions of spherical coordinates and often are normalized values with respect to the maximum of the function. The simplest pattern refers to an isotropic antenna, namely, an antenna that radiates the same amount of power in all directions. However, an isotropic antenna is inappropriate in cases where the antenna must pick up signals from a specific direction at a time while avoiding spurious signals coming from the ground (increasing *antenna noise temperature*) or unwanted other transmitters (interference). If the useful signal is to be transmitted to or received from different directions, the antenna can be pointed in the desired direction. In this way the antenna is directive and the *antenna gain* will be much higher than an isotropic one, but only at specific desired directions in space. Therefore, from Fig. 5a–c, many electromagnetic parameters can be defined.

The range of angles of maximum propagation are referred as the *main lobe* and numerically are within the half-power beamwidth (HPBW), namely, all the directions between the maximum power received (or transmitted) and its half-power. HPBW is often called the antenna *beam* and the following equation holds:

$$\text{HPBW} = k_i \frac{\lambda}{D} \qquad (5)$$

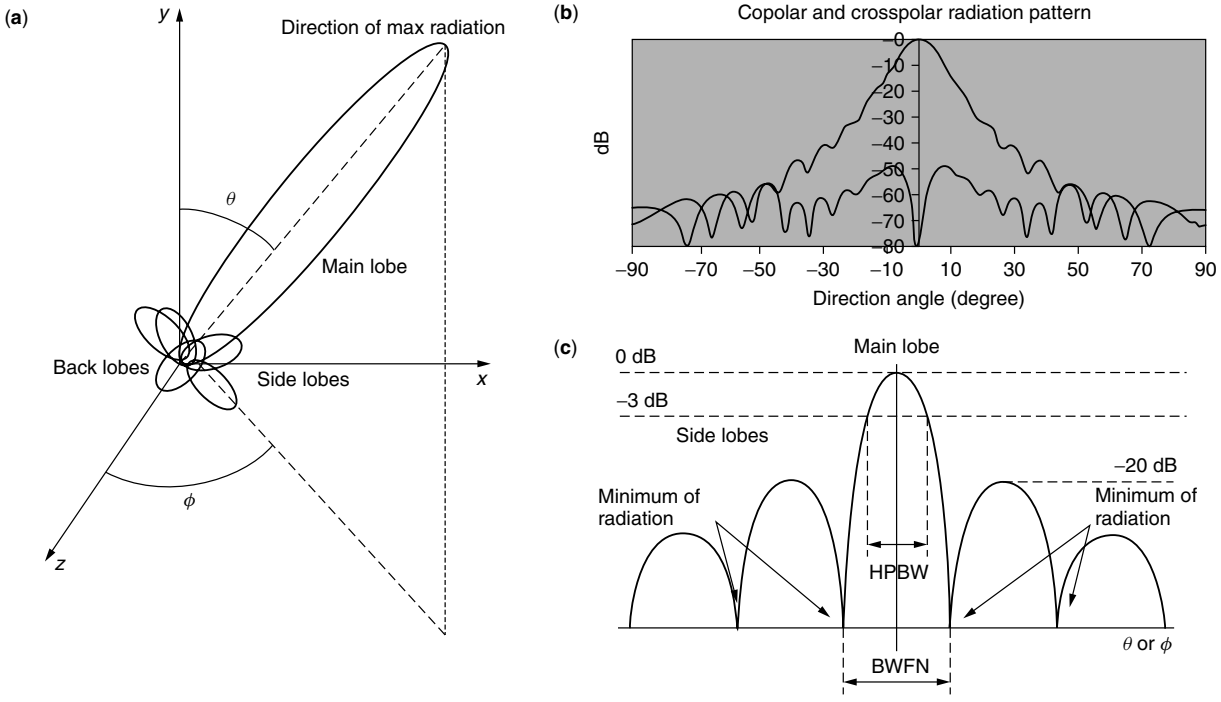


Figure 5. Antenna pattern.

where λ is the operating wavelength, D is the diameter of the antenna aperture, and k_i is a factor, always near unity, depending on which kind of function the feed illuminates in the mirror of the antenna. Often HPBW is also considered a measure of the antenna *resolution*; that is, its capability to discriminate two closely spaced objects in space. To clarify this statement, in Fig. 5c, note that the distance between the direction of max pick up and the first minimum is half of beamwidth between first nulls (BWFN). This means that two objects spaced at these locations are distinguishable because from one point the pickup is maximum while the other one is null. It could be shown that HPBW is approximately equal to BWFN/2.

Figure 5c represents a cut in either the θ or Φ plane (see Fig. 5a) and shows how low the near sidelobes are with respect to the main lobe. This is indicated as -20 dB. This illustration could be viewed as a zoom of the more complete polar pattern in Fig. 5b. In this figure the relative amount of power radiated (or received) as a function of angle is shown (polar or *copolar pattern*) together with the radiation in a perpendicular direction where, ideally, the radiation should be zero (*cross-polar pattern*).

2.5.2. Gain and Aperture Efficiency. It can be shown that the antenna gain has a direct relation to the collecting area of the antenna:

$$G = \eta \frac{4\pi}{\lambda^2} Ag \tag{6}$$

Although G is a dimensionless quantity, often given in decibels by simply taking $10 \log_{10}(G)$, there are application fields, such as radio astronomy, in which the antenna gain assumes the meaning of how much the antenna temperature is increased when the antenna surface receives a given amount of power density per unit bandwidth. In this case the dimension of gain is kelvins over jansky, where jansky is a unit defined as

$$1 \text{ Jy} = 10^{-26} \frac{W}{\text{m}^2 \cdot \text{Hz}} \tag{7}$$

and the antenna gain can be calculated in the following way:

$$G_R = 10^{-26} \eta \frac{Ag}{2k} \tag{8}$$

where k is the Boltzmann constant. The two ways to express the antenna gain are equivalent; what simply changes is the measurement unit. If G_R is known, we can calculate η and use Eq. (6) to get G; conversely, if G is available, we can calculate η and then get G_R from (8).

Ag is the geometric area of the aperture of the parabolic antenna. *Aperture* is the cross section subtended by the dish, and for parabolic antennas it is a circle with diameter D. λ is the wavelength at which the gain is to be calculated or measured, and η is called the aperture efficiency. The *aperture efficiency* is a number less than one and acts in reducing the real area that is effective in collecting the electromagnetic energy coming into the aperture. η originates from many causes, each of them described by an appropriate efficiency parameter, and in general depends on a lot of variables, including the frequency, direction of pointing, structural deformations of the antenna due to gravity, temperature and wind, and type of function by which the feed illuminates the dish. Trying to clarify as simply as possible, we could start by stating (9), which relates the causes that most affect the efficiency:

$$\eta = \eta_b \eta_x \eta_{\text{ph}} \eta_{\text{sp}} \eta_{\text{diff}} \eta_{\text{ill}} \eta_{\text{surf}} \eta_{f\text{loss}} \eta_{\text{vswr}} \eta_{s\text{loss}} \tag{9}$$

η_b is the *blocking efficiency*. It comes from the obstruction of the feed or subreflector, and the supporting legs raise at the incoming electromagnetic energy. Practical values span from 0.85 to 0.95, except for the antenna offset solution. A rule of thumb to get the order of magnitude of η_b is to compute the ratio of the total blocked area A with the area of the antenna aperture Ag, then

$$\eta_b = \left(1 - \frac{A}{Ag}\right)^2 \approx 1 - 2\frac{A}{Ag} \tag{10}$$

η_x is the *cross-polarization efficiency*. It arises when the polarization of the incoming wave doesn't match the polarization of the antenna. The extreme example should be a dipole sensitive at the horizontal polarization: η_x is zero for vertical polarized waves, so they cannot be detected. Parabolic antennas have a high degree of symmetry, so they are sensitive to both linear and circular polarizations and η_x has very high values (0.99 or better), particularly if the feed is circular and designed so that the amplitude and phase patterns are equal in the two orthogonal planes of maximum radiation (E and H planes). The situation worsens if the antenna components are not perfectly aligned to each other or the feed is not symmetric. In the case that the antenna is an offset system, a very careful design must be developed if a low cross-polarization level is needed.

η_{ph} is the *phase efficiency*. If all rays (see Figs. 1–3) don't arrive in phase (e.g., because the mirror deforms under structural loads or because the feed is not able to perfectly illuminate in phase all directions), a small amount of power can be lost. This term is generally negligible for classic parabolic antennas, 0.99 or so, but it could be very low for shaped antennas when used with primary focus receivers, because the equalizing effect of the subreflector is absent. However, in this case, the effect is very frequency-dependent, putting a limit on the highest usable frequency of receivers placed at the primary focus.

η_{sp} is the *spillover efficiency*. When a feed illuminates the primary or secondary mirror of the antenna, the energy at the edges cannot go sharply to zero. Thus, a fraction of the total illumination energy will be lost: the ratio between this fraction and the total energy is called "spillover efficiency". This term is generally the result of a compromise, as the illumination is far from uniform, the higher is η_{sp} but the lower is the effective area. Normally antennas have an illumination function different from uniform, so a certain amount of tapering is used. The *taper* is how much the function is lower at the edge with respect to its maximum value. This is one of the design parameters for the feed and is evaluated at the angle Φp or Φs depending on which geometry is chosen, and usually the tapering is higher in the primary focus configuration than the secondary. The reason is that a primary mounting of the feed "sees" the ground at angles greater than Φp,

so it picks up an unwanted noise temperature at about 300 K that must be attenuated. Instead, by illuminating the subreflector, the feed "sees" the atmosphere at angles greater than Φs, that it is cooler than the ground.

η_{diff} originates from *diffraction* due to edge effects for both primary and secondary mirrors.

η_{ill} is the *illumination efficiency* and accounts for the fact that the illumination function is not uniform over the aperture, so it is a measure of the reduction of gain due to tapering.

η_{surf} is the *surface efficiency*. The rays colliding on the antenna surface find a nonideal shape. The subreflector and primary mirror surfaces have roughness. Furthermore, large-diameter antennas have a primary mirror consisting of a lot of aluminum panels drawn close, which also means that their relative alignment is a concern. Sometimes larger subreflectors are made by panels as well. Antennas suffer important deformation due to gravitational and wind effects as their dimensions increase; also temperature-induced deformation of the surfaces must be taken into account. The net result is that incident rays are reflected by nonperfect surfaces, resulting in phase errors that reduce the gain. Manufacturing errors of the panels and surface, temperature, and wind effects on the mirrors and on the antenna supporting structure, and also gravitational deformations, are treated like random errors. Therefore the parameter indicating the departure from the ideal shape is the RMS (root-mean-squared) value of the real surface with respect to the ideal one. Because of the many the causes, the RMS values must be combined to obtain the *total surface* RMS σ. Usually σ is the RSS (root sum squared) of the RMS values:

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2} \tag{11}$$

The surface efficiency follows a Gaussian function depending on the σ/λ ratio:

$$\eta_{\text{surf}} = e^{-[4\pi(\sigma/\lambda)]^2} \tag{12}$$

$\eta_{f\text{loss}}$ is the *feed loss efficiency*. When the electromagnetic energy passes through the feed, the waveguide attenuates a small amount of power, so the power for illuminating the aperture is slightly less than the supplied one. The insertion losses are generally low — let's say in the range 0.1–0.4 dB, which means that $\eta_{f\text{loss}} = 0.91 \div 0.98$.

η_{vswr} is the feed *return loss efficiency*. Together with the insertion loss mentioned previously, the feed loses power because a part goes back due to a nonperfect impedance matching. Return losses less than -15 dB are easily obtained, so $\eta_{\text{vswr}} \geq 0.97$.

$\eta_{s\text{loss}}$ is the *surface loss efficiency*. The surface of the mirrors are conducting electric currents; thus ohmic losses due to material resistivity arise. The effect is low, $\eta_{s\text{loss}} = 0.99$.

To conclude this section, it is worthwhile to mention other possible causes that reduce the antenna gain. Up to now a perfect geometric alignment of the three antenna components — the primary mirror, the secondary mirror, and the feed — were assumed. In the case that either the feed or the subreflector are not well positioned on the focus in the direction of its axes, a *defocusing* effect occurs. Defocusing exists in a movable antenna also if it were properly aligned. In fact, the alignment holds for a single position of the antenna, because the antenna deformation, as the elevation changes, will move the focus so that the feed or subreflector will be defocused. If a proper tracking of the focus movement is necessary to recover that amount of loss, a mechanical facility must be added in the antenna design in order to move the feed or subreflector.

Another useful concept is the *field of view* (*FOV*). It is related to a displacement of the feed with respect to the focus position outside the focal axis. If this is the case, a reduction of gain, together with an increase of sidelobe level and cross-polar pattern, will be experienced. The field of view could be defined as the space region where the feed can be displaced losing no more than a fixed amount of gain. Of course, FOV is not an absolute parameter, but it depends on the amount of gain loss that is considered acceptable. Generally it has an angular dimension, but it could also be expressed as HPBW times or as a multiple of wavelength. The FOV concept suggests that in the real world the focus of an antenna must not be viewed as a point just outside of which the system doesn't completely work. Instead a "focal surface" exists where the performance of the antenna worsens with the distance that the feed is placed with respect to the focus. The FoV can be exploited to use the antenna with more than one frequency by placing receivers working at different wavelengths, or using feed arrays (many identical feeds working at the same frequency).

2.6. Structural and Mechanical Aspects of Parabolic Antennas

Small parabolic and stationary dishes don't give particular structural and mechanical problems. They are mounted on a lattice or on a mast, pointed in a fixed direction, and their performances are not affected by temperature, wind stress, or operating environment. On the other hand, some applications call for the antenna to be protected from all these causes, and thus it is completely enclosed in a *radome*, a microwave transparent dielectric housing protecting the antenna from adverse environmental conditions.

Between these two extremes a lot of applications use parabolic antennas without radomes, that experience all weather conditions and, because of their dimension and weight, also gravitational deformations both for the dish and the supporting framework. The general characteristics of a movable and large reflector antenna are described in the following text. The following structural and mechanical elements are shown in Fig. 6.

- A *concrete foundation* with pillars, which is embedded some meters in the ground.
- A *track* over which the *wheels* move. This is a very common solution to allow the rotation of the antenna in the azimuth direction.
- The *alidade*, which is the supporting structure of the antenna.
- The *elevation wheel*, which allows the antenna to rotate in the elevation direction.
- The *backup structure*, which supports the primary mirror.

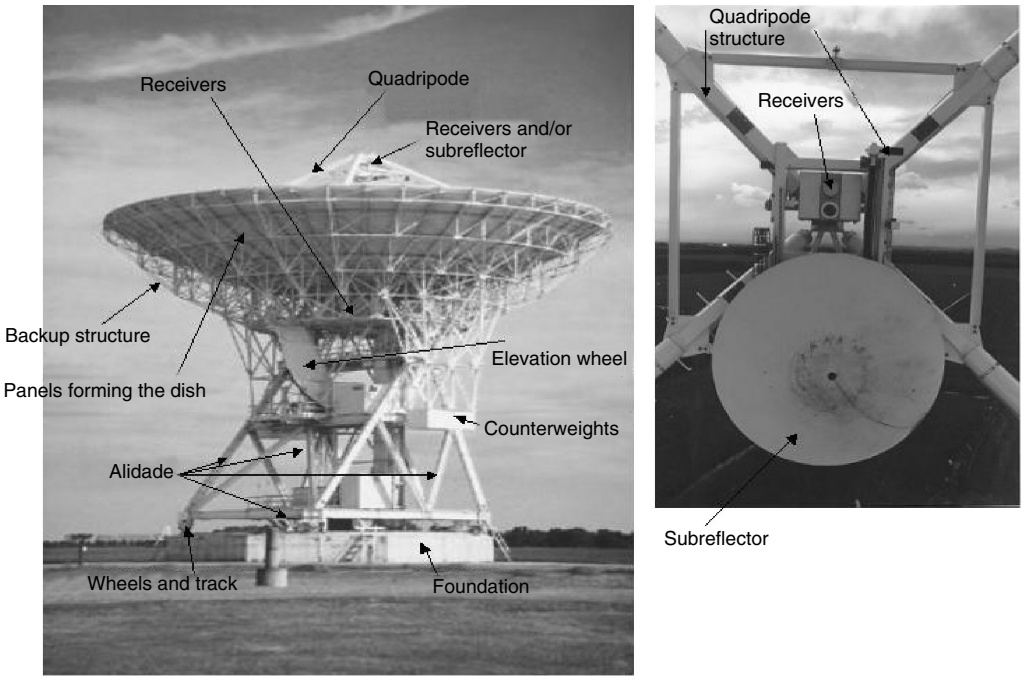


Figure 6. Elements of a parabolic antenna (Medicina observatory radiotelescope, Bologna, Italy).

The *quadripode*, usually three or four legs, which supports the secondary mirror and primary focus receivers.

The *panels*, which form the primary mirror of the antenna.

The right side of Fig. 6, a zoom of the primary mirror location, shows a very particular arrangement of the mechanical coexistence of the subreflector and the primary focus receivers. This allows a fast change among receivers placed at both the primary and secondary focus.

The wheels and track are made of iron alloy, and this is interfaced with the concrete foundation by means of either a suitable grout solution or, better, iron plates over which the track is bolted and grout under the plates. The plates are bolted in the concrete by anchor bolts. The alidade, elevation wheel, and backup and quadripode structure consist of large steel beams, trusses, frames, and brackets. The panels are made of aluminum, covered by a highly reflective white paint, the dimension and quantity of which are determined by the diameter of the dish; to give an idea, a mirror with a diameter around 30 m needs 200 or more panels, with an area of 3–5 m^2 each. The specifications regarding the panels are given in terms of manufacturing error and deformation under gravity, temperature, and wind. All of these are RMS values with respect to the ideal parabolic contour and they contribute, together with the RMS value of the subreflector surface, the RMS deformation of the structure and the mirrors alignment error, to the total surface accuracy σ [see formula (11)]. σ is said to determine the minimum operating wavelength of the antenna by means of the conventional relation

$$\sigma = \frac{\lambda \min}{20} \tag{13}$$

The subreflector can be made of fiberglass or, if the dimension is too large, by aluminum panels. Usually, the dimension for the subreflector is around $D/10$, where D is the diameter of the primary dish. It results as a compromise by taking the blockage at minimum values together with an efficient illumination of the paraboloid. The shaping of the antenna may instead require a subreflector diameter higher than $D/10$.

Sun exposure heats in a nonuniform way all the elements of the structure, both deforming the shape of the mirror and changing the direction of pointing. Further, in these heavy antennas (a 30-m antenna can weigh 200 tons or more), gravity takes a great role in deforming the backup structure. The departure from a true parabolic shape changes with respect to the elevation. To overcome the gravitational effect, some antennas adopt a structural design called *homology*; the mirror maintains a parabolic shape, changing the focal length only (i.e., the mirror opens or closes, maintaining the symmetry). By focusing the subreflector, the focal length can be tracked for each elevation. This design results in a much heavier antenna and significantly increases the cost.

2.7. Pointing

2.7.1. Overview. This subject is rarely taken into account in most applications and books, but it is worthwhile to mention both for completeness and for those applications that need a precise tracking of a target.

In Section 2.5 the term *beam* and the acronym HPBW were introduced, indicating a measure of the angular size where most of the radiation is contained. If the antenna has to point at a target and, above all, has to track it, the response of the antenna servosystem to the target coordinates must be within the pointing performance, to maintain the target inside the antenna beam. It is easily

recognized here that a pointing error can be viewed as a loss of power. It acts like the antenna efficiency terms, multiplying by a factor of <1 the amount of power received (transmitted) from (to) the target. To give a quantitative example, suppose that the antenna main lobe is Gaussian so it can be expressed in terms of HPBW by the following equation (see also Fig. 7):

$$\eta_{\text{point}} = e^{-(1.665*\theta/\text{HPBW})^2} \tag{14}$$

η_{point} is unity for perfect pointing, but rapidly decreases if errors occur. If the pointing error $\theta = \text{HPBW}/2$, half of the power is lost and an error of equal to one-fifth of the beam is enough to loose 10% of power.

Pointing errors can be divided into two classes: systematic and nonsystematic. *Systematic* errors are repeatable errors, and a mathematical model can be predicted. Gravitational deformations induce a pointing error. The erection of the antenna leaves unavoidable errors such as a slight nonorthogonality between the azimuth and elevation axis, nonperfect horizontal azimuth plane, and a mechanical axis of the mirror different from the electromagnetic direction of maximum pickup. All these factors induce systematic pointing errors that can be measured or derived from the best-fitting technique.

Nonsystematic errors are random errors and are due to temperature effects on all the elements of the antenna structure, for example unevenly expanding alidade trusses, and wind forces acting so that average wind and its gusts slightly move the antenna. These are not predictable and environment-dependent, in the sense that the pointing accuracy is a function of the amount of wind, absolute temperature, and its drift. Generally the antenna is said to work in three possible conditions — precision, normal, and extreme operation, indicating worsening of performance as wind and temperature effects increase.

2.7.2. Beam Deviation Factors. In Section 2.5 the case of a displaced feed was reported, listing the effects on the antenna pattern. A feed displacement gives pointing displacement as well. Generally speaking, the movements of all elements forming the antenna geometry, primary or secondary focus feed, subreflector, or primary mirror, cause pointing errors. These movements can be the translation or rotation of the element that originates a misalignment angle α, causing a pointing angle error β: the ratio between these angles is called the *beam deviation factor* (BDF), which is a function of the antenna parameters, namely, focal length f, diameter D, magnification M, and secondary mirror focal length F and for most cases its value ranges from 0.7 to 0.9. In the following, a survey of possible situations is presented,

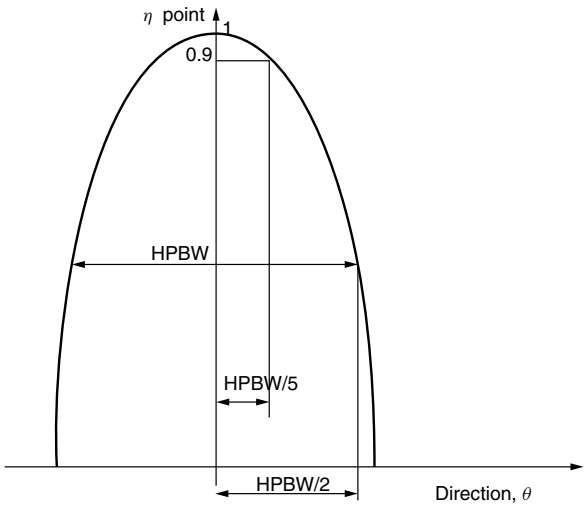


Figure 7. Pointing error.

- Primary feed lateral displacement (Fig. 8a):

$$\text{BDF} = \frac{\beta}{\alpha_p} = \frac{\beta}{\tan^{-1} d/f} = \frac{\sin^{-1}\left[\dfrac{d*\left(1+k_i(D/4f)^2\right)}{f*\left(1+(D/4f)^2\right)}\right]}{\tan^{-1} d/f} \tag{15}$$

k_i is the same factor appearing in Eq. (5). In the case $d \ll f$, so that $\tan^{-1}$ and $\sin^{-1}$ are equal to their argument, a more practical and usual relation can be used:

$$\text{BDF} = \frac{1 + k_i * (D/4f)^2}{1 + (D/4f)^2} \tag{16}$$

- Secondary feed lateral displacement (Fig. 8b):

$$\text{BDF} = \frac{\beta}{\tan^{-1}\left(\dfrac{d}{M*F}\right)} \tag{17}$$

- Subreflector rotation (Fig. 8b):

$$\text{BDF} = \frac{\beta}{\tan^{-1}\left(\dfrac{F}{f}*\dfrac{2\alpha_s}{M+1}\right)} \tag{18}$$

- Subreflector displacement (Fig. 8b):

$$\text{BDF} = \frac{\beta}{\tan^{-1}\left(\dfrac{h}{f}*\dfrac{M-1}{M}\right)} \tag{19}$$

The situation is such that by moving an element to one side, the beam is deviated to the opposite side (as shown in Fig. 8a), so to recover the right pointing, the antenna must be moved according to the element. BDF values allow us to understand the pointing error sensitivity of the antenna with respect to misalignments or effects due to the stiffness of the antenna elements.

2.8. Antenna Noise Temperature

Regardless of whether the antenna is receiving, or transmitting the wanted signal, a certain amount of noise power is picked up by its pattern. The amount of noise power is expressed as an equivalent resistor at temperature T_a, called the *antenna noise temperature*, which, when matched at the antenna receiver input in place of the antenna, gives the same amount of noise power. It is recalled here that the relation between power and temperature is

$$P = kT_aB \tag{20}$$

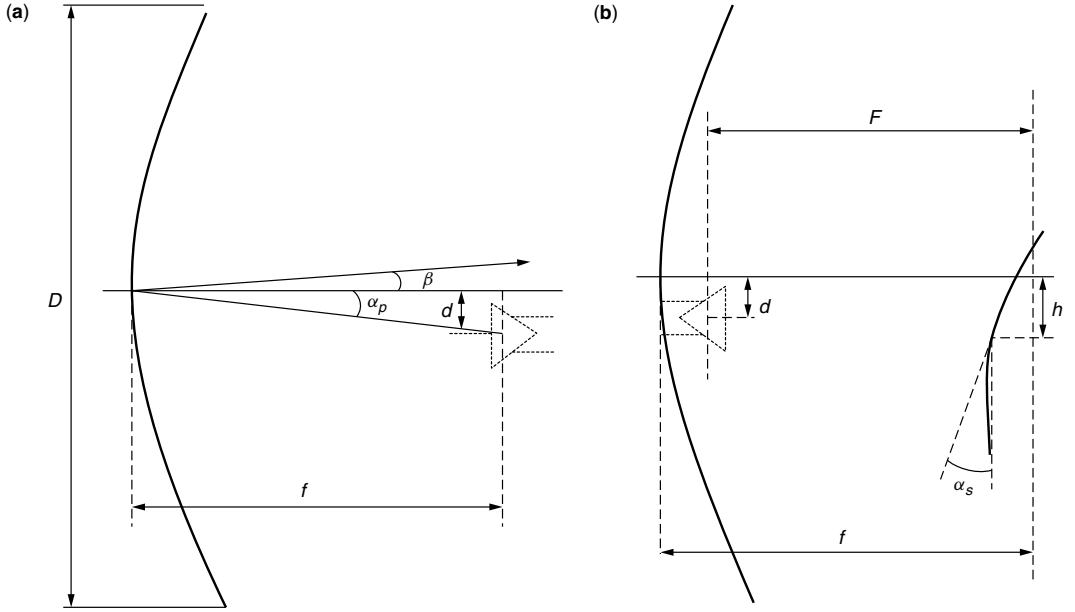


Figure 8. Beam deviation factors: (**a)** primary and (**b)** secondary feed lateral displacement, subreflector rotation and lateral displacement.

where k is the Boltzmann constant, B is the bandwidth, and T_a is the equivalent temperature.

The noise power comes from different origins:

- Noise power picked up from ground (conventionally at a physical temperature equal to 290 K) by backlobes or by antenna spillover.
- Noise power coming from cosmic sources, from the sky or planets and received by either the mainlobe or sidelobes. At least the antenna receives the big bang remnant, 2.73 K, which is uniformly distributed over all the sky and not dependent on frequency. Galaxies or other sky objects emit signals in the form of noise. The same could be said for planets and the sun of our solar system. The amount of noise power picked up depends strongly on frequency, on the object, and on how the antenna lobes point toward the object.
- Noise power from the atmosphere. The atmosphere is an absorbing medium, so it acts like an attenuator producing noise. The atmosphere noise varies with frequency and with the elevation angle at which the antenna is pointing. Low elevation angles give a higher amount of noise because the radio path is longer than that at high elevation.

2.9. G/T

By adding the antenna noise temperature to the receiver noise temperature, the so called *system temperature* T is obtained; that is, the overall noise of the complete antenna receiving system proportional to the inferior limit on the detection of signals. In any case, the detection of signals also depends on the antenna gain; the larger the reflector, the higher is the amount of power received, but this also holds for antenna noise. The ratio between T and the antenna gain G, the term G/T, gives a figure on the capability of the antenna to detect small signals. G/T can also be used to compare different antenna receiving systems; if a larger antenna has a very noisy receiver, its performance can be worse than a smaller antenna having a receiver, at the state of the art, with very low noise.

3. APPLICATIONS

Parabolic antennas have many applications in various fields requiring very different performance. Depending on the application, the design will encompass all or part of the characteristics described in Section 2.

3.1. Radio Astronomy

This relatively young science uses many large reflector antennas located all over the world to receive signals from many different natural sky objects such as galaxies, quasars, stars, and planets. The nature of the received signals is generally noise or plenty of spectral lines. If an object has to be resolved, the interferometry technique is used. This means that many parabolic antennas are used in observing the same object at the same time. In that case the value D of the interferometer is the distance among the antennas that can span from some kilometers to thousands of kilometers. This last case is called a *very-long-baseline-interferometry* (VLBI) technique and allows us to reach angular resolution as low as milliarcseconds at microwave frequencies [by applying Eq. (5)].

The parabolic antennas used in radio astronomy are of all types described in Sections 2.1 to 2.3, and operative frequencies span from about 300 MHz to 30 GHz for centimetric wave antennas to hundreds of gigahertz for millimeter and submillimeter antennas.

This field of application often needs a very careful design for all the antenna characteristics described in Section 2.

3.2. Microwave Relay Link

A microwave relay link is intended as a link that transmits and receives the radio signal (e.g., the broadcasting of

analog and/or digital signals) between parabolic antennas many tens of kilometers in distance and on a line-of-sight path. The diameters of the antennas used are a few meters, and the operative frequency range is in the microwave region from about 2 to 20 GHz. The Friis formula addresses the design of the link:

$$\mathrm{Pr} = \frac{\mathrm{Pt} * \mathrm{Gt} * \mathrm{Gr}}{(4 * \pi * r)^2} * \lambda^2 \tag{21}$$

P_r = received power
P_t = transmitted power
G_t, G_r = gain of transmitting and receiving antenna
r = distance of the link
λ = operating wavelength in free space of the transmitted and received signal

3.3. Satellite Communication

Most of the earth stations in a satellite communication system are parabolic antennas of all types described in Section 2. Both single feed and feed array configurations are used. In this last case, displaced location of the feeds with respect to the reflector are used to get different pointing directions in order to receive signals from different satellites. The used frequencies span from a few gigahertz to over 12 GHz. The diameter of the antennas can range well over 10 m.

3.4. Remote Sensing

Parabolic antennas can be used in the microwave range for radiometric measurements of the atmosphere parameters and earth and sea surface characteristics. Also in this case, scanning beam techniques and multifrequency measurements call for use of a feed array.

BIOGRAPHY

Alessandro Orfei received his degree in 1983 from Bologna University, Department of Electronic Engineering. He worked for three years at the G. Marconi Foundation Laboratories in the field of the fiber optics, and then for three years in a private company as a design engineer in the field of telecommunication. Since 1989, he has been a researcher at the Istituto di Radioastronomia–Consiglio Nazionale delle Ricerche (Italy). He is in charge of all work concerning the VLBI (very-long-baseline interferometry) 32m antenna at the Medicina Radio Observatory (Bologna, Italy).

PARTIAL-RESPONSE SIGNALS FOR COMMUNICATIONS

Apostolos Rizos
AWARE, Inc.
Bedford, Massachusetts

1. INTRODUCTION

Partial-response signals are those where intentional controlled intersymbol interference (ISI) is introduced between successive symbols, either for channel-matching or bandwidth-efficiency purposes. In contrast to a partial response signal, a *full-response signal* does not introduce any intentional ISI between successive symbols.

To explain these definitions in more detail, let us remember the basic representation of a linear digital modulation technique [1]

$$s(t) = \sum_{n=0}^{\infty} I_n g_T(t - nT) \tag{1}$$

A binary sequence d_n with values $\{0, 1\}$ is mapped to an information-bearing sequence I_n, which is typically an amplitude value that will scale the transmitting filter output $g_T(t)$ (the subscript T refers to the transmitter and not the symbol rate $1/T$). For binary PAM (pulse amplitude modulation), I_n will take values $A, -A$, so that a binary 1 will be transmitted using pulse $A \cdot g_T(t)$ and a binary 0 will be transmitted using the inverse pulse $-A \cdot g_T(t)$.

Note that for higher-order modulations (e.g., 4-PAM), k bits of the binary sequence are mapped into a 2^k-level information symbol I_n. The information sequence I_n may also be complex-valued, which corresponds to a QAM (quadrature amplitude modulation) scheme that needs to be modulated onto a carrier; the real part of the information signal will modulate the cosine of the carrier, while the imaginary part will modulate the sine (which is orthogonal to the cosine, and hence, can be distinguished from it). Without loss of generality, we will assume the information sequence to take values $\{1, -1\}$, and the cascade of the information sequence and the transmitting filter $g_T(t)$ to give a total energy $\mathcal{E}_b$ for each transmitted bit.

Assuming a linear transmission channel with impulse response $c(t)$, and a receiving filter with impulse response $g_R(t)$, then the output of the receiver filter will be

$$r(t) = \sum_{n=0}^{\infty} I_n x(t - nT) + v(t) \tag{2}$$

where $v(t)$ is the noise from the channel and $x(t)$ is the combined effect of $g_T(t), c(t), g_R(t)$. Its frequency representation (Fourier transform) will be the product of the Fourier transforms of its components

$$X(f) = G_T(f)C(f)G_R(f) \tag{3}$$

By sampling the output of the receiving filter every T seconds, we obtain the decision variables that are used for the estimation of the information sequence

$$r_m = x(0)I_m + \sum_{n=0, n\neq m}^{\infty} I_n x(mT - nT) + v(mt) \tag{4}$$

We see that the mth received sample r_m depends on the mth information symbol I_m, but also [depending on the values of $x(nT), n \neq 0$] on the adjacent symbols $I_n, n \neq m$. This is called *intersymbol interference* (ISI). A full-response

signal is designed in such a way that the combination of transmit and receiver filters will give no ISI

$$x(n) = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \tag{5}$$

while a partial-response signal is designed in such a way that the combination of transmit and receive filters will give controlled ISI, $x(n) \neq 0$ (but equal to predetermined values) for more than one sample n. The controlled ISI amount in a PR system is chosen in such a way so as to satisfy the system requirements, such as small bandwidth or no DC spectral component.

An example of a full-response signal is shown in Fig. 1. The basic transmitting pulse is a rectangular pulse $p(t)$ of duration T, and we assume that both the channel and the receiver filter introduce no change to the signal: $C(f) = G_R(f) = 1, \forall f$. We see that each symbol does not interfere with adjoining symbols.

There is a problem with a signal such as this rectangular pulse; it is not bandwidth-limited; that is, its Fourier transform $P(f)$ is nonzero for a nonbounded frequency range. Nonband-limited signals pose problems in communications for two reasons. First, the characteristics of the transmission medium (and the transmitter and receiver components) usually result in some form of nonideal frequency response and bandwidth limitation for the signal — the signal has to be of limited bandwidth to be able to be transmitted without significant frequency content loss or distortion from the channel. The second reason is that in order to accommodate many channels in large-capacity trunks, most telecommunication standards impose some form of frequency-division multiplexing (FDM), whereas each individual channel has to satisfy a bandwidth constraint in order to be multiplexed and not interfere with adjacent channels.

For the remainder of this discussion we will assume that the channel characteristics are such that it is an ideal channel of bandwidth W, with

$$C(f) = \begin{cases} 1, & |f| \leq W \\ 0, & |f| > W \end{cases} \tag{6}$$

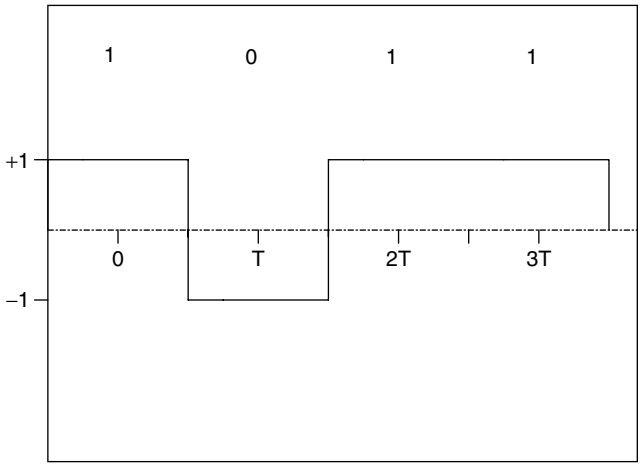


Figure 1. An example of a full-response signal.

Dealing with a channel that, besides the bandwidth limitations, also introduces amplitude or phase distortion to the transmitted signal is a subject of channel equalization.

Hence, the problem of designing full-response signals [i.e., signals that satisfy Eq. (5)] that also have a limited bandwidth W, having $G_T(f)G_R(f) = 0$ for $|f| \geq W$, arises. The pioneering work on this was done by Nyquist [2]. He established the condition (*Nyquist condition*) that $X(f)$ has to satisfy in order to have zero ISI in the received samples.

The major result of the Nyquist condition is that in order to transmit without ISI a signal with symbol rate $1/T$, there is a minimum bandwidth requirement of $W \geq 1/2T$. The resulting maximum symbol rate of $1/T = 2W$ is called *Nyquist rate*. It is interesting to note the duality with sampling theory, where it is known (again by Nyquist) that in order to uniquely sample a bandlimited signal, one has to use a sampling rate (frequency) of at least $2W$; a higher rate results in correlation (i.e., ISI, from a communications point of view) between successive samples.

A family of pulses that satisfy the Nyquist criterion for zero ISI is the one with a *raised cosine spectrum*. The bandwidth occupancy B of this pulse is determined by the rolloff factor α, which takes values $0 \leq \alpha \leq 1$, giving

$$B = \left(\frac{1}{2T}\right) \cdot (1 + \alpha) \tag{7}$$

An example of two pulses with raised-cosine spectrum, with rolloff factors $\alpha = 0, 0.5$ is shown in Fig. 2. We notice that the pulses have zero value at multiple integers of the symbol interval $t = nT, n \neq 0$, which is the non-ISI criterion of Eq. (5). Thus, superimposing shifted versions (by nT) of these pulses leads to a combined signal where only one constituent pulse contributes in the value of the signal at a specific sampling instant $t = mT$.

The limiting case of pulses with the raised-cosine spectrum is when the rolloff factor is $\alpha = 0$. Then the spectrum has an ideal rectangular characteristic with the smallest possible bandwidth $B = 1/2T$ for no ISI. The resulting pulse is the sinc( ) function

$$x(t) = \frac{\sin(\pi t/T)}{\pi t/T} = \text{sinc}\,\frac{t}{T} \tag{8}$$

However, filters that have such a sharp frequency response are practically nonrealizable. For the filter to be causal (i.e., to have its impulse response to the right of the $t = 0$ axis in Fig. 2a), the truncation of its impulse response must be of a reasonable length and a plus a delay function is required. Since the tails of the sinc( ) function decay quite slowly (proportionally to $1/t$) the truncation length has to be extremely large, to avoid the significant loss of the signal characteristics, and this makes the filter realization very difficult. Another detrimental effect of the heavy tails of the sinc( ) pulse is that a mistiming error (sampling at a time slightly off the multiples of T) results in a nonconverging (due to the $1/t$ decay) series of ISI components.

Hence, usually for a realizable full-response system a raised-cosine spectrum characteristic with $\alpha >$

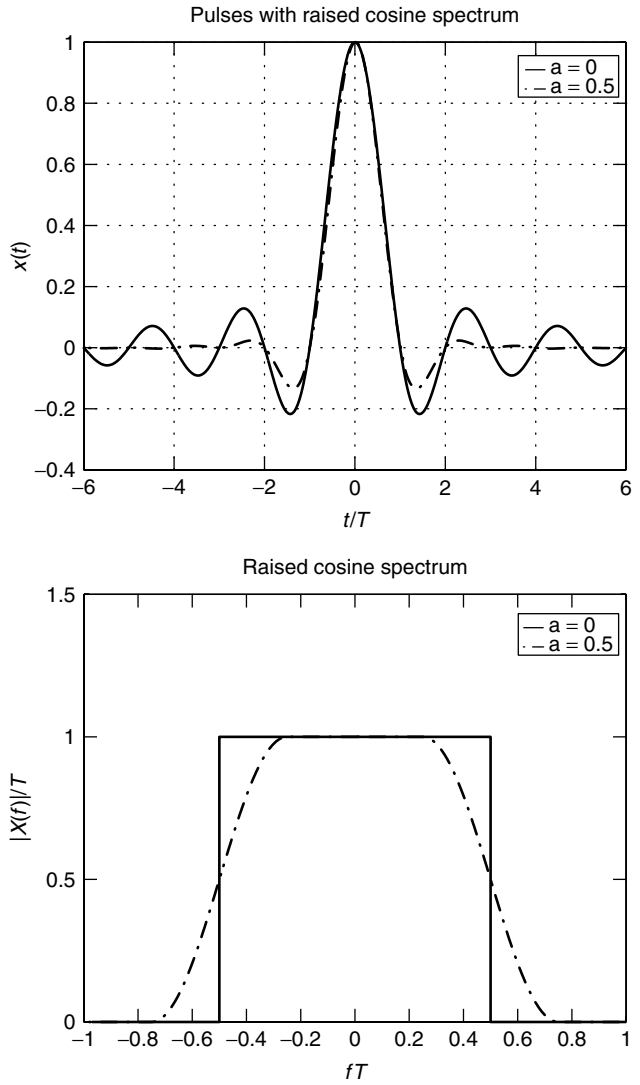


Figure 2. Pulses with raised-cosine spectrum: (**a**) $x(t)$ in the time domain; (**b**) frequency response $X(f)$.

0 is chosen. This leads to a filter with a more realistic truncation length requirement. However, this also results to a signal bandwidth occupancy that is larger than the optimum $(1/2T)$ one, typically by 15–25%.

2. SIMPLE PARTIAL-RESPONSE SIGNALS

In the previous section we noticed that the Nyquist limit on the transmission rate over band-limited channels $1/T = 2W$ is practically nonrealizable with full-response signaling. However, if one relaxes the zero-ISI condition and allows for a controlled amount of ISI, then one can obtain realizable filters that have the Nyquist bandwidth $W = 1/2T$. This was first observed by Lender [3] and later extended by Kretzmer [4], Kobayashi [5], and Pasupathy [6].

Let's examine the simplest case of a partial-response signal, one that has a composite response $x(n)$ [as given by Eq. (3)] with values[1]

$$x(n) = \begin{cases} 1, & n = 0 \\ 1, & n = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

It is easily shown that for $W = 1/2T$ the actual corresponding pulse is

$$x(t) = \text{sinc}\left(\frac{t}{T}\right) + \text{sinc}\left(\frac{t}{T} - \frac{1}{T}\right) \tag{10}$$

with frequency response

$$X(f) = 2Te^{-j\pi fT}\cos(\pi fT),\ |f| < \frac{1}{2T}(0 \text{ otherwise}) \tag{11}$$

This is the first partial-response pulse that was examined and is called a *duobinary* pulse. We notice that it is equivalent to a digital FIR filter with coefficients [1 1] followed by a filter with an ideal rectangular frequency response [to which $\text{sinc}(t/T)$ corresponds]. We mentioned above that the ideal rectangular filter is not practically realizable on its own; however, the pulse given by Eq. (9) and shown in Fig. 3a is much more easily realizable since its tails decay rapidly and its frequency response (shown in Fig. 3b) decays smoothly toward zero.

Another simple partial-response scheme is the *modified duobinary pulse*, which is characterized by the following composite response:

$$x(n) = \begin{cases} 1, & n = -1 \\ -1, & n = 1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

The modified duobinary pulse and its spectrum are given by the following relationships and are shown in Fig. 4; also

$$x(t) = \text{sinc}\left(\frac{t+T}{T}\right) - \text{sinc}\left(\frac{t-T}{T}\right) \tag{13}$$

$$X(f) = j \cdot 2T\sin(\pi f 2T),\ |f| < \frac{1}{2T}(0 \text{ otherwise}) \tag{14}$$

We notice that the modified duobinary spectrum has a null at DC ($f = 0$), which makes it suitable for channels that don't pass DC (e.g., circuits with transformer couplings) or for SSB modulation. Currently, the most significant application of the modified duobinary pulse is in magnetic recording systems. From Fig. 4a, we notice that the modified duobinary pulse is similar to the read-back signal response to a pulse in a magnetic recording system. On the basis of this observation, one can shape with minimal equalization the combined magnetic channel into a modified duobinary system response, and use maximum-likelihood decoding (explained in the next subsection) to estimate the data sequence. The use of these ideas in magnetic recording systems [7] (labeled as PRML, from

[1] One should notice that the given sample values result in a higher energy per bit with respect to a zero-ISI system with $x(0) = 1$, and this should be taken into account when one calculates the SNR of the signal and its bit error probability.

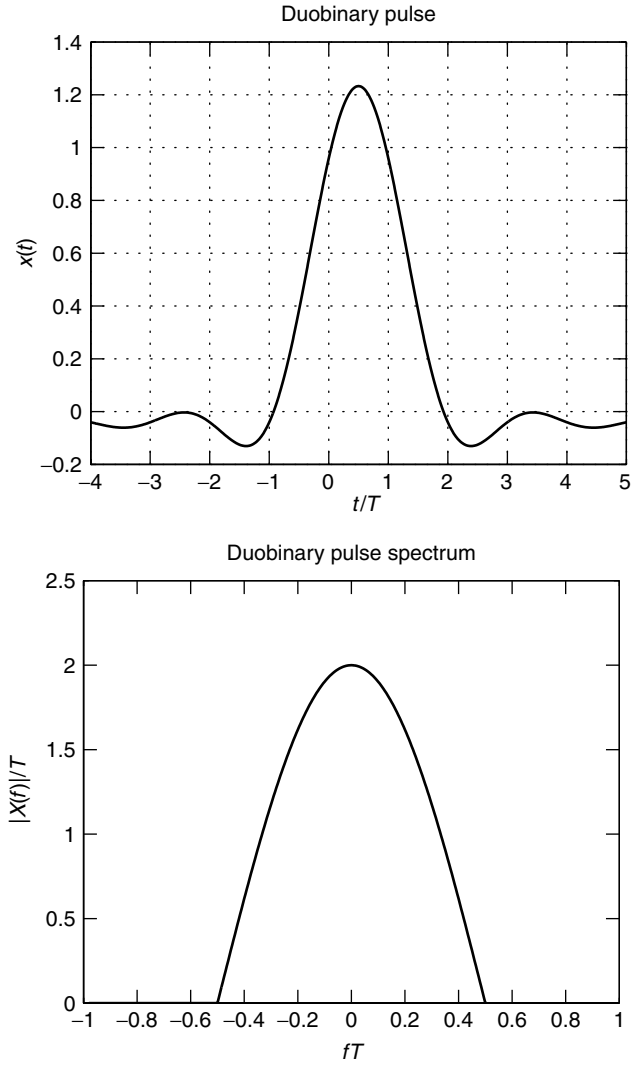


Figure 3. Duobinary pulse: (*a*) $x(t)$ in the time domain; (*b*) frequency response $X(f)$.

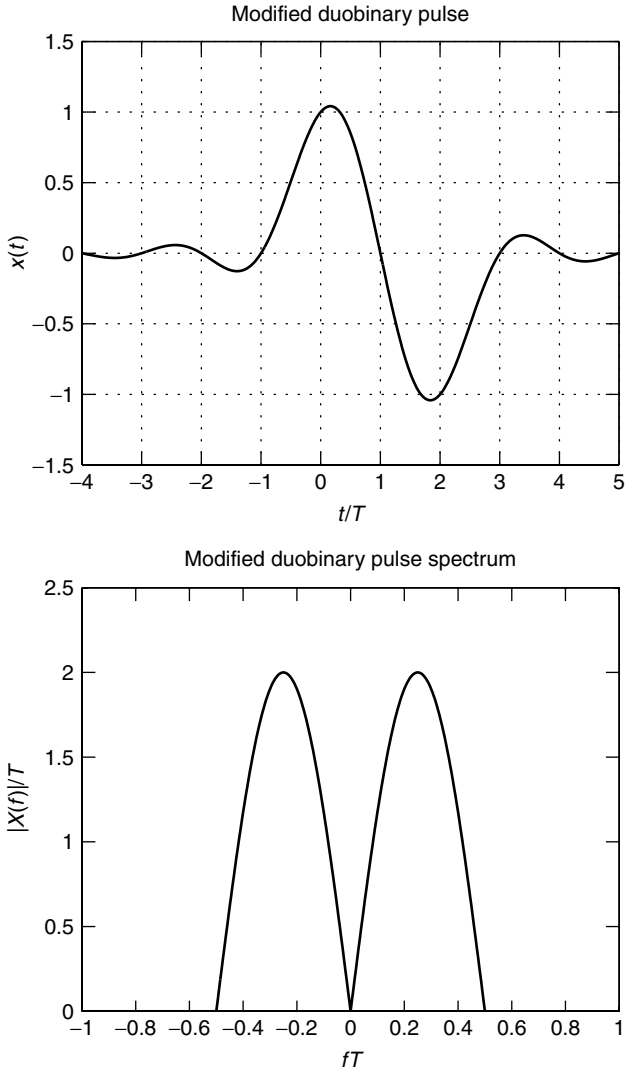


Figure 4. Modified duobinary pulse: (**a**) $x(t)$ in the time domain; (**b**) frequency response $X(f)$.

partial-response maximum likelihood) was one of the main reasons in the impressive increase of recording density, and hence capacity, of the hard disks in the last decade.

2.1. Detection of Partial-Response Signals

Let's examine the detection methods that may be employed for the simplest duobinary partial response signal. From Eqs. (4) and (9) we see that the output of the sampler, which is used as the decision variable, is

$$r_m = I_m + I_{m-1} + v_m$$

If I_{m-1} has already been detected with value I'_{m-1}, then its effect on r_m may be eliminated through subtraction, and hence I_m may be estimated from

$$I'_m = \text{sgn}(r_m - I'_{m-1})$$

where sgn() denotes the sign function, since in a binary scheme the detection rule for a signal with possible values $\{1, -1\}$ is $x > 0?1 : -1$, for zero-mean Gaussian noise. The detected value I'_m can then be used for the detection of I_{m+1} and so on. However, the use of this method can lead to serious error propagation: an error (because of the noise) in the estimate of a particular symbol will adversely affect the estimate of the next symbol and so on.

A technique that may be used to eliminate the dependence on the previous symbol estimate is *precoding*. The binary source sequence d_n is transformed to a new binary sequence p_n through the precoding operation

$$p_n = d_n \ominus p_{n-1} \tag{15}$$

where $\ominus$ denotes modulo-2 subtraction (equivalent to modulo-2 addition). Then the precoded sequence p_n is mapped into the information sequence, $I_n = 2p_n - 1$, that is used for transmission using the duobinary pulse. Using the relationship between d_n, p_n, I_n, and r_n, one can show

Table 1. Example of Precoding for Duobinary Signal

Binary sequence d_n		0	1	1	1	0	0	1	0
Precoded sequence p_n	1	1	0	1	0	0	0	1	1
Information sequence I_n	1	1	−1	1	−1	−1	−1	1	1
Duobinary signal r_n		2	0	0	0	−2	−2	0	2
Estimate d'_n		0	1	1	1	0	0	1	0

that the estimate of d'_m may be obtained as

$$d'_m = \frac{r_m}{2} - 1(\text{modulo } 2) \tag{16}$$

directly from r_m, without any use of the previous symbols. The (noiseless) example presented in Table 1 demonstrates this detection operation. We notice that, according to Eq. (16), values of $r_m = \pm 2$ (which occur with probability $\frac{1}{4}$ each) map to a binary 0 in the original binary source sequence, and the value $r_m = 0$ (which occurs with probability $\frac{1}{2}$) maps to a binary 1. In the case where noise affects the sample r_m the corresponding decision regions will be $|r_m| \geq 1$ and $-1 < r_m < 1$ respectively.

In a similar way, the modified duobinary signal can be detected on a symbol-by-symbol basis, without any dependence on the previous decisions, and hence without any error propagation, by using the precoding operation of

$$p_n = d_n \oplus p_{n-2}$$

The abovementioned precoding techniques may be extended directly from a binary to an M-ary ($M = 2^k$) modulation scheme, by using modulo-M instead of modulo-2 operations.

It can be shown [8] that the performance of a duobinary or modified duobinary scheme that employs a symbol-by-symbol detector is approximately 2.1 dB worse than the performance of a full-response scheme with the same constellation size. However, the symbol-by-symbol detector does not exploit a significant property of the partial-response signal: its memory. As an example, we notice that two consecutive received samples of a valid duobinary signal cannot have the values $2, -2$ or $-2, 2$. However, the symbol-by-symbol detector will not factor this in and will decode the preceding received samples according to the rule of Eq. (16), which will lead to at least one binary error.

There is an estimation method that does factor in the memory present in communication signals and gives the sequence of symbols with the minimum probability of error. This is the maximum-likelihood sequence estimator (MLSE), which can be implemented efficiently using the Viterbi algorithm [9]. This algorithm was first proposed for the decoding of convolutional codes, where the output depends on a finite number of previous inputs, and its use was later extended for systems with finite ISI, either unintentional (channel distortion) or intentional (partial response signals). It uses the squared distance[2] between the received samples and the possible valid sequences, and declares as detected sequence the one having the smallest total squared distance from the received one.

It can be shown that a duobinary or modified duobinary signal with MLSE detection at the receiver has the same performance as a full-response signal. So the MLSE detector does not have the 2.1-dB performance degradation that characterizes the symbol-by-symbol detector. As a side note, we should mention that there were research efforts after the introduction of partial response signaling to evaluate the free Euclidean distance of a *faster-than-Nyquist* scheme [10]. This name has been used for transmission schemes that employ a sinc( ) pulse with zero crossings at $1/T$ (and a bandwidth of $W = 1/2T$), but which are spaced at $T' < T$ (i.e., have a higher baud rate than does the pulse designed for zero ISI). This closer spacing of the sinc pulses leads to ISI, since the sample points for successive symbols no longer correspond to the zero crossing of the pulse. However, it was proved that the minimum Euclidean distance of such a scheme is the same as that of a full-response (no-ISI) system, as long as $T' > 0.802 \cdot T$. Since the use of the ideal rectangular spectrum is not practically realizable and the receiver will have to deal with an infinite series of ISI terms in order to estimate the received sequence, these ideas were just a form of mathematical exercise. However, the underlying principle of the existence of PR schemes that have bandwidth less than the Nyquist rate with a free distance that is the same as a full-response system led to additional research for bandwidth-efficient implementable PR techniques.

3. PARTIAL RESPONSE FOR BANDWIDTH EFFICIENCY

The duobinary and modified duobinary systems that we outlined above were designed with the goal of achieving the Nyquist rate of $W = 1/2T$, with practically implementable filters. Still, one can achieve even better bandwidth efficiency by using a partial response scheme. Since partial response can be viewed as just one form of filtering, one may use a longer filter $x(n)$ to achieve better bandwidth characteristics.

The disadvantage of this approach is the increased intersymbol interference between symbols, due to the longer $x(n)$ that spans $L + 1$ successive samples, where by L we denote the memory of the partial-response scheme.

As noted in a previous section, for received sequences coming from a channel/system with finite memory, the MLSE detector is the optimum one. We noted that the major determining factor for its performance is the minimum Euclidean distance $d_{\min}$ between any two valid sequences.

It can be shown [8] that any partial response system has minimum distance $d_{\min}$ less or equal to a full-response system of the same constellation size and transmitted energy level. Hence, the goal of the partial response scheme is to lose as little performance as possible, while maximizing the bandwidth efficiency. Said and Anderson [8] offered an optimization framework for finding the best partial-response scheme for a certain memory length L; given a constraint on the bandwidth B,

[2] For the case of linear modulation in AWGN the likelihood of a received sample r depends on its squared distance $|r - s|^2$ from the nominal value s to be received.

the partial-response scheme $x(n)$ with the best (largest) $d_{\min}$ would be found; or, given a certain $d_{\min}$ (i.e., a given performance level) the partial-response scheme $x(n)$ with the best bandwidth characteristics B would be found.

Besides the potential loss in $d_{\min}$, there is a second disadvantage associated with such a partial-response scheme, namely, the complexity of its receiver. An MLSE, implemented through a Viterbi algorithm (VA), has complexity quite bigger than the slice-and-quantize operation of a full-response receiver for an M-ary PAM. The MLSE complexity grows proportionally to the number of states in the finite memory system to which the PR scheme is equivalent. The number of states is equal to 2^{ML}, where L is the memory of the PR scheme and M is the constellation size. So for large constellation sizes (where the advantages of a smaller-bandwidth scheme are more easily exhibited), or for longer filters $x(n)$ (which give better bandwidth efficiency), the complexity issue makes implementation problematic. To overcome this disadvantage, reduced-complexity schemes, such as *reduced-state sequence estimation* (RSSE) [11,12], were proposed. These schemes were originally proposed for the traditional channel ISI problem, of which partial response is a special case. It was shown that with careful design and state reduction choices these schemes can offer performance very close to MLSE with significant computational cost savings.

Using these ideas, practical PR schemes were found [13], with bandwidth occupancy $B \approx 0.35 \cdot 1/T$ (roughly half of the bandwidth of a realizable full-response system), which exhibits a performance penalty of $<2-3$ dB compared to an equivalent (same constellation size) FR scheme. However, because of its much better bandwidth characteristics, the symbol (baud) rate of the PR scheme can be increased w.r.t. (with respect to) to the FR scheme, while still satisfying spectral occupancy constraints. If the bandwidth of the full-response scheme is $B_{\mathrm{FR}} = 0.7 \cdot 1/T_{\mathrm{FR}} = 1.4 \cdot 1/2T_{\mathrm{FR}}$ (raised cosine with 40% rolloff), and the bandwidth of the PR scheme is $B_{\mathrm{FR}} = 0.35 \cdot 1/T_{\mathrm{PR}}$, then the symbol rate of the PR scheme can be increased to twice that of the full response scheme: $1/T_{\mathrm{PR}} = 2 \cdot 1/T_{\mathrm{FR}}$. This will allow the use of PR scheme with a constellation size of $M^{1/2}$, if the full-response scheme is M-ary, which for moderate to large M more than compensates — due to the increased distance, and hence noise immunity, between energy levels — for the performance penalty paid for the ISI between symbols.

4. LINE CODING: MODULATION SIGNALS WITH MEMORY

A more generalized form of partial-response signaling is employed in modulation systems that are used mainly for high-speed baseband transmission, and their purpose is to shape the spectrum of the transmitted signal to match the spectral characteristics of the channel. The way to shape the spectrum is typically through the introduction of restrictions, namely, memory, on the generator pulse. These techniques are usually called *transmission line coding* or *modulation coding*. We should make the distinction here between the above type of coding, where correlation between transmitted symbols is introduced, and the more usual notion of coding that involves an increase in the bit rate between the source data sequence and the line sequence. In the latter case, the code typically encodes k source bits into n transmitted bits (code rate $R = k/n < 1$), and the receiver use this redundancy to detect and correct errors. These are called *channel coding* or *error correction* techniques, and they are different from modulation coding techniques.

Focusing on transmission line codes, we depict in Fig. 5, some of the most commonly employed ones.

The first, and probably simplest, modulation scheme for baseband transmission is the *non-return-to-zero* (NRZ) method. In NRZ a binary $d_k = 1$ is mapped into a square waveform of amplitude +A, and a binary $d_k = 0$ is mapped into a square waveform of amplitude −A. Actually, this scheme is a memoryless system, equivalent to binary PAM transmission. It does not pose any demodulation problems, but it is not suitable for spectrum shaping because of its lack of correlation between symbols. Furthermore, it does not offer another desirable property for baseband transmission — it does not guarantee a minimum rate for pulse transitions, since a sequence of multiple 1s (or 0s) leads to a constant voltage signal throughout its duration. Pulse transitions are necessary for deriving timing and synchronization from the received signal at the receiver.

Numerous slight variations of the NRZ scheme exist. The system that is described above is also called by some authors a *(bi)polar NRZ scheme*, to distinguish it from a unipolar NRZ scheme, where the two amplitude levels are $[+A, 0]$ instead of $[+A, -A]$ of bipolar NRZ. The unipolar NRZ scheme has a more pronounced DC component than the bipolar one, and is used less in practise.

Another memoryless system is the *return-to-zero* (RZ) method. In this system, the basic square pulse (of amplitude $[+A, -A]$, for binary $d_k = 1$, 0 respectively) returns to 0 for the second half of the symbol duration, so it has a 50% duty cycle. The return to 0 guarantees voltage transitions at a rate of $2/T$, thus eliminating

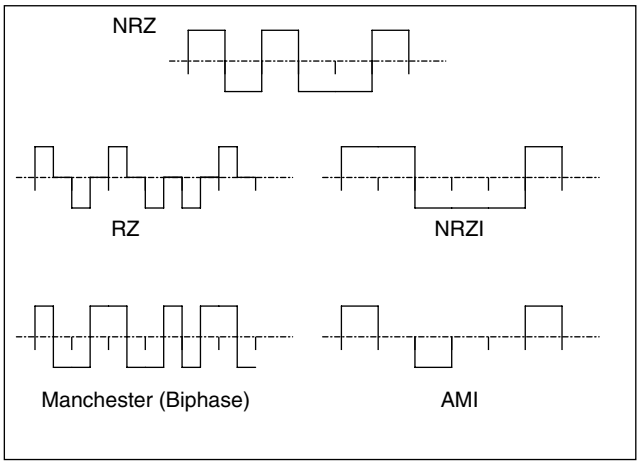


Figure 5. Examples of line coding techniques: NRZ, RZ, Manchester, NRZI, AMI (all mapping the binary information sequence [1, 0, 1, 0, 0, 1]).

the timing/synchronization problems that an NRZ scheme might have. However, its narrower pulses occupy a larger bandwidth, and this is a serious disadvantage for an RZ scheme.

Similar to the RZ scheme is biphase (or Manchester) encoding, which employs a square wave of period T that transitions from $+$A to $-$A (instead of to 0, like the RZ scheme) at the middle of the symbol interval. The abovementioned pulse is used for a binary $d_k = 1$, while its antipodal pulse is used for 0. Although relatively bandwidth-inefficient, the Manchester scheme is used in some practical systems, including some Ethernet and TTY applications, because of its lack of DC, and timing information provision.

A line code that employs memory in its waveform generation is the *non-return-to-zero, inverse* (NRZI) scheme. This employs the same basic pulse as the NRZ scheme, but with the transitions from one amplitude level to the other happening when the information bit is $d_k = 1$, and the previous symbol level is retained when $d_k = 0$. NRZI can be directly generated from NRZ, if the actual data sequence is passed through a precoder of the form $p_k = p_{k-1} \oplus d_k$, and then p_k is used to generate the transmitted waveform through NRZ. NRZI is used in magnetic recording systems, since it has good bandwidth characteristics and no DC component. However, it presents the usual problem of no guaranteed timing information, if many consecutive 0's appear in the data sequence.

Finally, another line coding technique with memory is *alternate mark inversion* (AMI). An AMI waveform employs zero voltage/amplitude for a binary 0, while it employs a square pulse of alternating amplitude for binary 1s, so if the previous 1 bit were sent using $-$A voltage, the next 1 bit would be sent using a pulse with $+$A voltage. AMI has good bandwidth characteristics and no DC component, and its memory offers some elementary error detection capabilities, since two consecutive pulses of the same sign (with any number of zero-voltage symbols in between them) mean that an error has occurred. Although AMI might present the receiver with timing/synchronization problems when multiple consecutive 0s exist in the transmitted sequence, it is a widely employed scheme, especially in T1/T3 trunklines.

The problem of the transmitted waveform being constant for many symbols, if multiple consecutive 0s are being transmitted, that practical NRZI and AMI schemes face is addressed in two ways.

1. The first one is to introduce codes that operate on the binary source sequence and generate a channel sequence with a restriction in the maximum number of successive zeros. Typically this technique is used in conjunction with NRZI in magnetic recording systems. These codes usually impose an additional restriction in the minimum number of 0s between two 1s in a sequence, and this is used to increase the distance (and, hence, reduce interference) between pulse transitions. These codes are known as *runlength-limited codes* (RLL codes), and have a code rate of less than 1; that is, they introduce redundancy. A good tutorial paper on them has been written by Immink [22].
2. The second technique to address the consecutive 0s problem is used in conjunction with AMI. We have seen that AMI offers some basic error detection, since two successive (ignoring intermediate 0s) pulses of the same polarity do not correspond to a valid input sequence. The idea for the second technique is to replace a string of consecutive zeros with a string that violates the AMI principle (and contains pulse transitions). The receiver will detect the "error" condition, and, since the received string has a predetermined pattern, it will substitute the original string of all zeros in its place. This way both sufficient timing information is available, and the original data are not lost.

The class of these codes is usually denoted as *binary bipolar with n zero substitution* (BnZS). A common one is B8ZS: a pattern of 8 zeros is replaced by a string where the bits 4 and 7 violate the bipolar principle. Let's take as an example the string $+(00000000)0-\cdots$, where $+$ stands for amplitude A, $-$ stands for amplitude $-$A (both correspond to 1s in the information sequence), and 0 is zero amplitude (0s in the information sequence). This string is a valid bipolar sequence, since the two consecutive nonzero pulses have opposite polarity. The B8ZS code replaces the constant string of 8 zeros (inside parentheses) by the string with the two bipolar violations (BPV) resulting in the transmitted string of $+(000+-0-+)0-\cdots$. The receiver, on detecting the two BPVs at the specified locations, replaces the 8 bits with all zeros, and the original sequence is restored. If the most recent pulse sign before the 8-zero string is negative, then the replacement string will be $(000-+0+-)$ in order to give the two BPVs at bits 4 and 7. The B8ZS scheme is employed by many commercial carriers. A similar, but simpler, scheme is the B6ZS code, where 6 consecutive 0s are replaced by a string with bipolar violation at the second and fifth bits (i.e., the B8ZS scheme, without the two first initial zero bits).

Another two members of the BnZS family are the B3ZS and B4ZS codes, which are very similar. Let's examine B4ZS [which is widely used, and is also called *high-density bipolar 3* (HDB3)]. Here, a string of four zeros is replaced by the string B00V or 000V, where B stands for a pulse that satisfies the bipolar rule, while V is a pulse violating the rule. The choice between these two strings is made such that the number of pulses satisfying the bipolar rule between violations is odd. As an example, suppose that we have the string $+(0000)0-0+\cdots$, and the number of valid pulses after the last (not shown) violation is even. Then, the marked four 0s will be replaced by B00V, because the first bit (valid pulse) will make for an odd number of valid pulses between the last violation and the current one. Hence, the transmitted string will be $+(-00-)0+0-\cdots$. Note that, contrary to B8ZS, the B4ZS code does affect the polarity of the subsequent nonzero pulses. The B3ZS scheme is the same as B4ZS, but with three consecutive zeros replaced by either B0V or 00V patterns (again, with an odd number of valid B pulses between successive V violations).

5. NOTES ON THE LITERATURE

Partial-response systems were first proposed by Lender [3], and later extended by Kretzmer [4], Pasupathy [6], and others. Faster-than-Nyquist signaling was investigated by Mazo [15], Foschini [10], and Hajela [16].

The combination of partial response with TCM encoding was investigated by Wolf and Ungerboeck [17], Ketchum [18], and Forney and Calderbank [19].

The use of partial-response shaping in magnetic recording systems is covered in various papers. As an example we mention the papers by Cideciyan et al. [7], and Tyner and Proakis [20], where additional references in the evolution of PRML systems can be found. Typically nowadays, the target response in a PRML system is a *generalized* partial-response scheme, which is the product of the modified duobinary with a system of the form $(1+D)^n$. A tutorial paper on the general issue of coding for magnetic recording channels was written by Immink et al. [21].

A paper that surveys the use of modulation coding in copper wire subscriber lines was published by Lechleider [22], while a good reference for BnZS and other currently employed line coding techniques is the one by Bellamy [23].

The correlation between a PR system and precoding techniques to combat ISI in a decision feedback equalization scheme can be found in the paper by Forney and Eyuboglu [24] and references cited therein. A paper that covers the state of partial response signaling is the one by Said and Anderson [8], while a typical implementation of a PR system for a copper cable channel is given in [25].

BIOGRAPHY

Apostolos D. Rizos received the B.S. degree in physics from the University of Athens, Athens, Greece, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Northeastern University, Boston. Since 1999, he has been with AWARE Inc., Bedford, Massachusetts, where he is a Senior DSP Engineer, working on algorithms for DSL modems. Before that, he was a technical consultant with Delphi Communication Systems, Maynard, Massachusetts, working on underwater acoustic modems. His interests lie in modulation and coding for communication systems, and computationally efficient algorithm implementations.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
2. H. Nyquist, Certain topics in telegraph transmission theory, *AIEE Trans.* **47**: 617–644 (1928).
3. A. Lender, The duobinary technique for high-speed data transmission, *IEEE Trans Commun. Electron.* **82**: 214–218 (May 1963).
4. E. R. Kretzmer, Generalization of a technique for binary data communication, *IEEE Trans. Commun. Technol.* **14**: 67–68 (Feb. 1966).
5. H. Kobayashi, Correlative level coding and maximum likelihood decoding, *IEEE Trans. Inform. Theory* **17**: 586–594 (Sept. 1971).
6. S. Pasupathy, Correlative coding: A bandwidth-efficient signaling scheme, *IEEE Commun. Soc. Mag.* **15**: 4–11 (July 1977).
7. R. D. Cideciyan et al., A PRML system for digital magnetic recording, *IEEE J. Select. Areas Commun.* **10**: 38–55 (Jan. 1992).
8. A. Said and J. B. Anderson, Bandwidth-efficient coded modulation with optimized linear partial-response signals, *IEEE Trans. Inform. Theory* **44**: 701–713 (March 1998).
9. G. D. Forney Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 363–378 (May 1972).
10. G. D. Foschini, Contrasting performance of faster binary signaling with QAM, *AT&T Bell Labs Tech. J.* **63**: 1419–1445 (Oct. 1984).
11. V. Eyuboglu and S. Qureshi, Reduced-state sequence estimation with set partitioning and decision-feedback, *IEEE Trans. Commun.* **36**: 13–20 (Jan. 1988).
12. P. Chevillat and E. Eleftheriou, Decoding of trellis-encoded signals in the presence of intersymbol interference and noise, *IEEE Trans. Commun.* **37**: 669–676 (July 1989).
13. A. D. Rizos and J. G. Proakis, Reduced-complexity sequence detection approaches for PR-shaped, coded linear modulations, *Proc. IEEE GLOBECOM'97*, Phoenix, AZ Nov. 1997.
14. K. A. S. Immink, Runlength-limited codes, *Proc. IEEE* **78**: 1745–1759 (Nov. 1990).
15. J. E. Majo and H. J. Landau, On the minimum distance problem for faster-than-Nyquist signaling, *IEEE Trans. Inform. Theory* **34**: 1420–1427 (Nov. 1988).
16. D. Hajela, On computing the minimum distance for faster than Nyquist signaling, *IEEE Trans. Inform. Theory* **36**: 289–295 (March 1990).
17. J. Wolf and G. Ungerboeck, Trellis coding for partial-response channels, *IEEE Trans. Commun.* **34**: 765–772 (Aug. 1986).
18. J. Ketchum, Performance of trellis-codes for M-ary partial-response, *Proc. IEEE GLOBECOM'87*, 1987, pp. 1720–1724.
19. G. D. Forney Jr. and A. R. Calderbank, Coset codes for partial response channels; or, coset codes with spectral nulls, *IEEE Trans. Inform. Theory* **35**: 925–943 (Sept. 1989).
20. D. J. Tyner and J. G. Proakis, Partial response equalizer performance in digital magnetic recording channels, *IEEE Trans. Magn.* **29**: 4194–4208 (Nov. 1993).
21. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorder, *IEEE Trans. Inform. Theory* **44**: 2260–2299 (Oct. 1998).
22. J. W. Lechleider, Line codes for digital subscriber lines, *IEEE Commun. Mag.* **27**: 25–32 (Sept. 1989).
23. J. C. Bellamy, *Digital Telephony*, 3rd ed., Wiley, New York, 2000.
24. G. D. Forney Jr. and V. Eyuboglu, Combined equalization and coding using precoding, *IEEE Commun. Mag.* **29**: 25–34 (Dec. 1991).
25. G. Cherubini, S. Olcer, and G. Ungerboeck, A quaternary partial-response class-IV transceiver for 125 Mbit/s data transmission over unshielded twisted-pair cables: principles of operation and VLSI realization, *IEEE J. Select. Areas Commun.* **13**: 1656–1669 (Dec. 1995).

PATH LOSS PREDICTION MODELS IN CELLULAR COMMUNICATION CHANNELS

H. L. BERTONI
Polytechnic University
Brooklyn, New York

1. INTRODUCTION

Modern wireless services involve two-way transmission of individual signals, rather than one-way broadcast of the same signal to many listeners. In order to accommodate many users for such services in an allocated frequency band, the concept of spatial frequency reuse was developed. The simplest implementation of frequency reuse involves spatial separation of the simultaneous use of the same frequency channel linking subscribers with their nearest access point (or base station). By keeping the interference just low enough to achieve a desired quality of service, the subscribers can be accommodated with a minimum of physical infrastructure. In other words, using the smallest distance possible between cochannel cells allows for the most channels per base station. The conflicting requirements of achieving radio coverage, while maintaining the desired limit to interference, places a premium on accurate methods for prediction of the received radio signal strength and other channel characteristics. A survey of many of the characteristics can be found in Refs. 1 and 2.

One approach to understanding the channel characteristics is to make measurements over a wide range of system parameters, such as frequency and bandwidth, antenna height, and distance between antennas [3–5]. The measurement results can be reduced to best-fit formulas that give the system parameter dependence in a way that is simple to use in prediction software [6–8]. Because most subscribers live in cities, it is important to measure these quantities in different building environments. An alternative approach is to use theoretical methods for predicting radiowave propagation; the most common are ray optics and the uniform theory of diffraction (UTD). In this article we discuss only the theoretical methods for predicting the path loss (or path gain), and compare them with measurement-based models. The prediction of other channel characteristics can be found elsewhere [1,9].

Path gain PG is defined as the ratio of the received signal strength to the total radiated power. The commonly used term "path loss" refers to the reciprocal of path gain, and is usually expressed in decibels. As an example, for antennas located in free space the path gain is given by

$$PG_0 = \left(\frac{\lambda}{4\pi R}\right)^2 g_1 g_2 \tag{1}$$

where R is the separation between antennas, λ is the wavelength, and g_1 and g_2 are the directive antenna gains. It is often convenient to consider the path gain between idealized isotropic antennas for which $g_1, g_2 = 1$. Provided that the carrier frequency is the same, reciprocity of Maxwell's equations implies that the path gain is the same, no matter which end of the wireless link is the transmitter and which is the receiver.

2. RAY CONCEPTS FOR UNDERSTANDING AND PREDICTING PROPAGATION

Modern wireless systems use UHF (300 MHz–3 GHz) and microwave radiolinks to connect base station antennas and subscribers located between the buildings, or even inside buildings. Thus the buildings have a major influence on the received signal. Since the wavelength λ at these frequencies is small compared to the building size, it is appropriate to think of the radiowaves as traveling along rays radiated in all directions by the source. The rays travel along straight lines until they encounter the buildings or ground where they are reflected according to the laws of geometric optics (GO). A ray incident on a building edge or corner creates a family of diffracted rays propagating away from the edge in a cone whose half-angle is equal to the angle between the incident ray and the edge, as described by the uniform theory of diffraction (UTD) [10]. The influence of each interaction on the ray fields is contained in a reflection or diffraction coefficient, together with an algebraic dependence on the length of the ray segments that conserves energy within a tube of neighboring rays. Subsequent ray encounters with buildings act in cascade. Thus, rays connecting base station antenna and subscriber may be multiply reflected by the building walls, and/or diffracted at the corners and rooftops, as suggested in Fig. 1 for propagation from the base station antenna to the subscriber.

The ray segments reaching the subscriber come from all directions in the horizontal plane, and a wedge of angles in the vertical direction. Reciprocity of Maxwell's equations implies that the same ray paths, with arrows reversed, apply for propagation from the subscriber to the base station antenna. For an elevated base station, as shown in Fig. 1, the arriving rays come from limited wedge of angles. However, base station antennas located below the rooftops will receive rays coming from all directions in the horizontal plane.

Each ray carries a copy of the transmitted symbol bit $p(t)$ that is delayed by the path length L_j divided by the speed of light c, and has complex amplitude A_j. The received voltage at location x along a street is the sum of

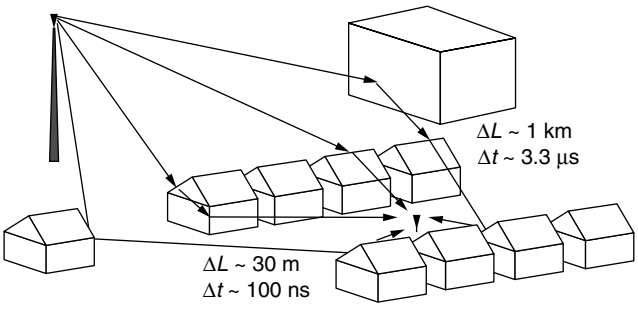


Figure 1. Multiple ray paths by which signals propagate from an elevated base station to a subscriber at street level. (©2002 by H. L. Bertoni.)

the ray contributions

$$V(x)e^{j\omega t} = \sum_j A_j p(f - L_j/c)e^{j\omega(t-L_j/c)} \qquad (2)$$

where c is the speed of light and $\omega = 2\pi f$. Because the differences in path length ΔL of the various rays reflected from objects near to the mobile are on the order of the street width, which is on the order of 30 m, the time difference for this cluster of rays is on the order of 100 ns. In addition, rays from the base station may be reflected from large structures at a greater distance before arriving at the building in the vicinity of the subscriber. Such paths may have delays on the order of 1 μs, followed by additional delays due to scattering in the vicinity of the subscriber, which results in another cluster of rays arriving at the subscriber. Figure 2 is an example of the time-delay profile of the signal received during wideband pulsed measurements made in Paris at 890 MHz with an omnidirectional receiving antenna and a system time resolution of 0.1 μs [11]. Echos with significant amplitude are received with time delays up to about 5 μs.

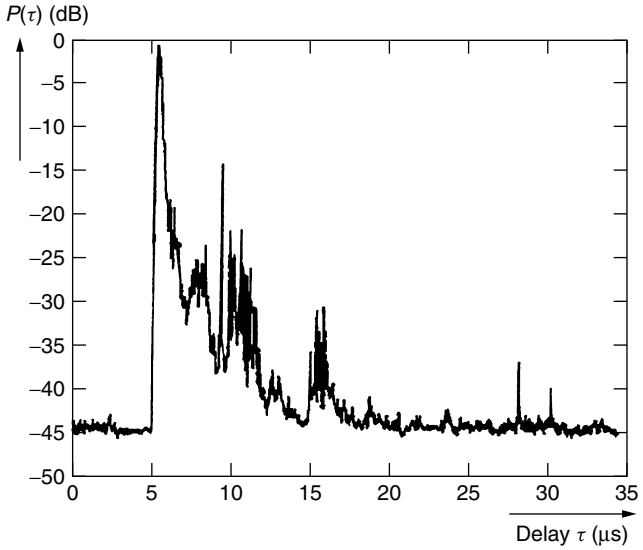


Figure 2. Time-delay profile of a pulsed signal in the 890-MHz band measured in Paris showing significant arrivals at up to 5 μs. (From Ref. 11, with permission.)

For systems whose bandwidth BW is small enough so that individual ray signals overlap in time ($1/BW \gg \Delta L/c$), all terms in (2) can be approximated using $p(t - L_j/c) \approx p(t - R/c)$, where R is the distance from the base station. In this case the rays arriving in all directions at the subscriber add coherently and generate a standing wave or interference pattern in space. The rapidly varying curve in Fig. 3 is a plot of the total received power as a function of the position x of a vehicle that is non-line-of-sight (NLoS) of the base station [12]. The amplitude is seen to undergo variations of up to 20 dB over distances on the order of one half the wavelength λ, which at 910 MHz is about $\frac{1}{3}$ m. In a moving vehicle, this spatial variation is perceived as a rapid time variation, which has led to the term "fast fading." Taking a sliding average of the received power over a distance of about 10–20λ smooths out the rapid fluctuation, as shown in Fig. 3, and is known as the small-area average.

The small average in Fig. 3 is seen to vary by about ±6 dB from the overall average, and variation has a scale length of 5–10 m. This variation is often referred to as "slow fading" or "shadow loss" since it results from shadowing by buildings, trees, and other objects. Finally, for outdoor or indoor links, the signal amplitude shows a systematic dependence on the distance or range R between base station and subscriber. Typically, measurements of the small-area average are plotted in dB against $\log R$, and a linear regression line is fit to the measurements. The regression fit gives the range dependence, as discussed in a later section, while the deviation from the regression line is interpreted as the shadow fading.

The sliding average of power is proportional to the spatial average $\langle |V(x)|^2 \rangle$ of the voltage in Eq. (2). Since the ray amplitudes A_j are slowly varying functions of distance x, and since the pulse duration ($1/BW$) is typically long compared to the time difference over the averaging interval of about $20\lambda/c = 20/f$, the sliding average is therefore

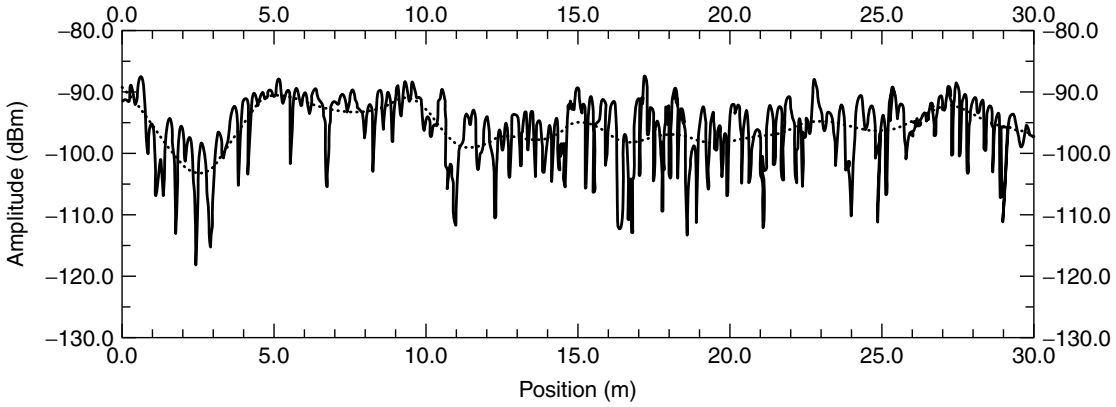


Figure 3. Narrowband [continuous-wave (CW)] measured signal variation, and the sliding average, as a function of distance along a street for non-LOS conditions. (From Ref. 12, with permission.)

given by

$$\langle |V(x)|^2 \rangle = \sum_{j,k} A_j A_k^* p(t - L_j/c) p^*(t - L_k/c) \langle e^{-j\omega(L_j - L_k)/c} \rangle$$
$$\approx \sum_j |A_j|^2 |p(t - L_j/c)|^2 \quad (3)$$

The final approximation in this equation is obtained by recognizing that the phase differences $\omega(L_j - L_k)/c$ for $j \neq k$ go through 2π variations due to changes in the pathlengths L_j with distance x along the street. As a result, the spatial average of the exponential vanishes for $j \neq k$. Thus the spatial average power is equal to the sum of the ray powers. For each symbol bit, the total received energy is found by integrating (3) over time, and is seen to be proportional to $\sum |A_j|^2$. Thus the total bit energy is proportional to the sum of the ray powers. Again for narrow band systems $|p(t - L_j/c)|^2 \approx |p(t - R/c)|^2$ and the time variation can be taken outside of the summation in (3).

3. TWO-RAY MODEL FOR FLAT EARTH

The simplest propagation environment occurs when there is only flat earth between the base station and the subscriber. In this case the received signal can be computed from the two-ray model consisting of a direct ray and a ground-reflected ray, as shown in Fig. 4. Because the two ray paths are of nearly equal length, it is necessary to add the ray fields coherently, and not simply add the ray powers [13]. For isotropic antennas the path gain of the two-ray model is given by [13]

$$PG = \left(\frac{\lambda}{4\pi}\right)^2 \left| \frac{e^{-jkr_1}}{r_1} + \Gamma(\theta) \frac{e^{-jkr_2}}{r_2} \right|^2 \quad (4)$$

where r_1 is the direct distance from the transmitter to the receiver, r_2 is the distance through reflection point. The Fresnel reflection coefficient $\Gamma(\theta)$ depends on the angle of incidence θ and the polarization, and is given by

$$\Gamma(\theta) = \frac{\cos\theta - a\sqrt{\varepsilon_r - \sin^2(\theta)}}{\cos\theta + a\sqrt{\varepsilon_r - \sin^2(\theta)}} \quad (5)$$

Here $a = 1/\varepsilon_r$ for vertical polarization and $a = 1$ for horizontal polarization, where ε_r is the relative dielectric constant of the ground. For average ground, the relative dielectric constant is $\varepsilon_r = 15 - i60\sigma\lambda$, and the conductivity σ is around 0.005 mho/m [13]. As the distance between the transmitter and receiver increases, the angle θ approaches 90°, the reflection coefficient Γ approaches -1 and r_2 approaches r_1.

Measured path loss between vertically polarized dipoles and between vertically polarized bicones is shown in Fig. 5 when the antennas are located along a flat road whose only features are low vegetation and wooden telephone poles [13]. For comparison, the dashed curve in Fig. 5 is a plot of the signal predicted by (4) for vertical polarization. For small horizontal separation $R < 10$ m, the antenna patterns have an influence on the measurements. However, for $R > 10$ m, only the antenna gains are important and they result in a vertical offset of the curves (the signal for dipoles is few decibels greater than for isotropic antennas, and for bicones it is a few decibels smaller).

Using a logarithmic scale for the horizontal separation R, as in Fig. 5, the received power is seen to vary about straight lines having two distinct slopes separated by a breakpoint R_B. Before the R_B, the radio signal oscillates severely as a result of alternating regions of destructive and constructive combination of the two rays, while after the R_B it decreases more rapidly with distance. The breakpoint lies near the last peak in the two-ray model, at a distance $R_B = 4h_1h_2/\lambda$, where h_1, h_2 are the antenna heights [1]. Well beyond the breakpoint distance the path gain of (4) reduces to

$$PG = \frac{h_1^2 h_2^2}{4R^4} \quad (6)$$

so that the received signal decreases more rapidly than the $1/R^2$ dependence of free space.

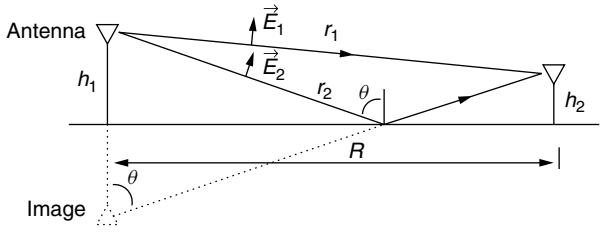


Figure 4. Propagation over flat earth as described by the two ray model. The ground reflected ray appears to come from the image of the source. For large R, the fields of both rays are nearly parallel.

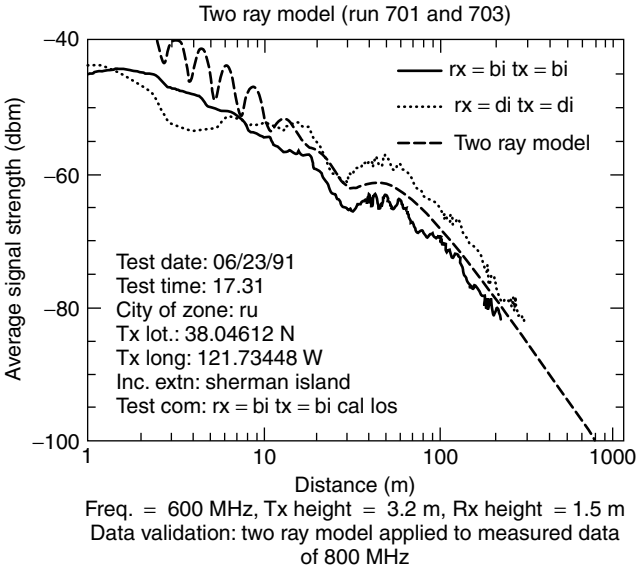


Figure 5. Measured and predicted variations of the received signal for propagation over flat earth for antennas of height 3.2 and 1.6 m at 800 MHz. (From Ref. 13, with permission.)

The two-ray model also serves as a basis for understanding the received signal when the two antennas are located within a line of sight along a street in an urban environment. In this case the direct and ground reflected rays give the dominant contributions, while additional contributions come from rays that are reflected by the buildings lining the streets. Building-reflected rays result in additional rapid variations about the simple two-ray model, but do not change the overall variation. Accounting for a single reflection in the building walls and ground reflection leads to the six-ray model, which has been used to obtain the plot of Fig. 6. Similar results are obtained from measurements in urban environments [1,14].

4. PROPAGATION OVER BUILDINGS IN RESIDENTIAL ENVIRONMENTS

In residential sections of cities, and in suburban regions, the buildings are of more or less uniform height, and the propagation may take place past many rows of buildings between the base station and subscriber. For base station antennas near to or above the rooftops, the radiowaves to a subscriber will propagate primarily over the rooftops, except for subscribers on the few streets aligned with the base station. Signals propagating through the buildings are highly attenuated by the exterior and interior walls. Except in the distant suburbs, the gaps between buildings are small and are seldom aligned with the base station, or aligned from row to row.

To predict the range dependence of the spatial average path gain for macrocells, the individual buildings in a row are replaced by a continuous smooth obstacle, as seen in the end view in Fig. 7. All rows are assumed to have the same height, and each row of buildings is separated by the same distance d. Using this model of the buildings, the mean path gain PG is the product of three factors [1]:

$$PG = (PG_0)(Q^2)(PG_2) \tag{7}$$

Here, PG_0 is the free-space path gain given by Eq. (1). The term Q in (7) is the reduction of the fields arriving at the buildings near to the mobile as a result of diffraction past the previous rows of buildings. The term PG_2 represents the signal resulting from diffraction from the rooftops down to ground level. These two terms are discussed in more detail below.

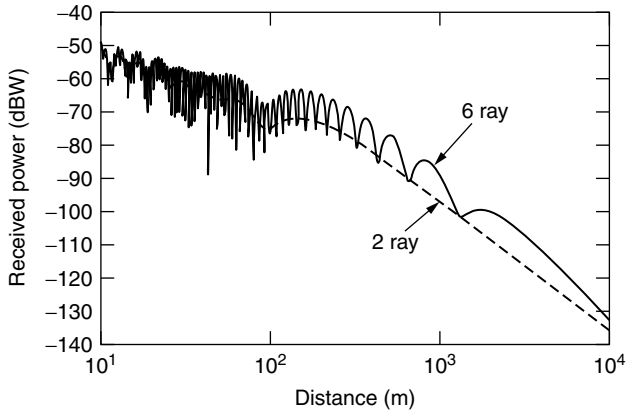


Figure 6. Comparison of the predictions of the two-ray model and the six ray model accounting for reflections form the buildings lining a street, as well as reflections from the ground. (© 1999 by H. L. Bertoni.)

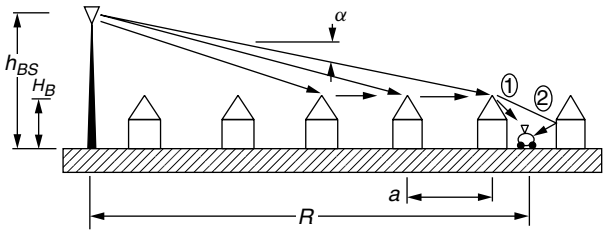


Figure 7. Side view of propagation over the rooftops to the last building before the subscriber and subsequent diffraction down to street level.

4.1. Q: Reduction of the Rooftop Fields

Except close to the base station, the horizontal distance from the base station to the buildings around the subscriber is large compared to the elevation of the base station antenna above the average building height. As a result, the glancing angle α shown in Fig. 7 is given by

$$\alpha = \tan^{-1}\left(\frac{h_{BS} - H_B}{R}\right) \approx \frac{h_{BS} - H_B}{R} \tag{8}$$

where h_{BS} is the base station height and H_B is the average building height. Provided that α is small, the radiowave propagating from the base station to the rooftops near the subscriber will undergo a cascade of multiple diffraction events at all of the previous rows of buildings. The mathematical treatment of multiple diffraction can be found in Ref. 1, while the simpler results are presented here.

A simple case occurs when the rooftops are all of the same height, the rows of buildings are all separated by the same distance d, and the base station antenna is at the same height as the rooftops. In this case the additional loss of the field reaching the Mth row of buildings from the base station is $Q = 1/M$ [1]. Because $M \approx R/d$, when (1) and $Q = 1/M$ are substituted into (7), the path gain is found to vary with distance as $PG \propto 1/R^4$, which is like the dependence for large separations of antennas above a flat earth. However, the proportionality constant for PG, and its dependence on frequency and antenna height will be different in the two cases.

When the base station antenna is well above the rooftops, as is the case for macrocellular applications, relatively simple expressions can again be found for Q. In this case the number of rows of buildings crossed by the radiowave is large and the glancing angle α is small. For example, in older cities the row separation is $d = 50$ m, so that 40 rows are crossed when the signal propagates to a distance of $R = 2$ km. In this case the reduction of the rooftop fields due to diffraction by previous rows of buildings can be expressed in terms of the parameter g_p, which is defined by

$$g_p = \alpha\sqrt{\frac{d}{\lambda}} \tag{9}$$

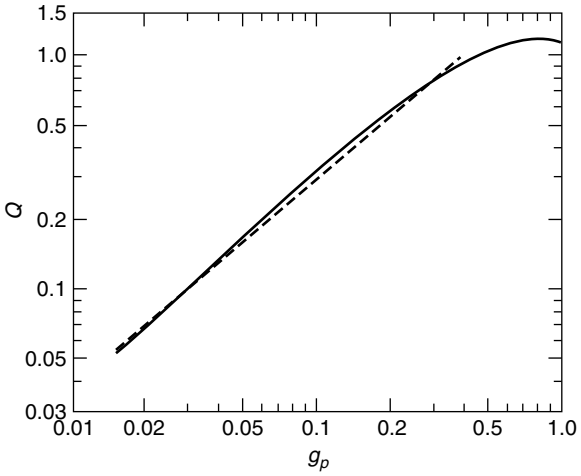


Figure 8. Dependence of the reduction of the rooftop fields Q on the dimensionless parameter g_p. The dashed line gives the simple approximation. (From Ref. 15, with permission.)

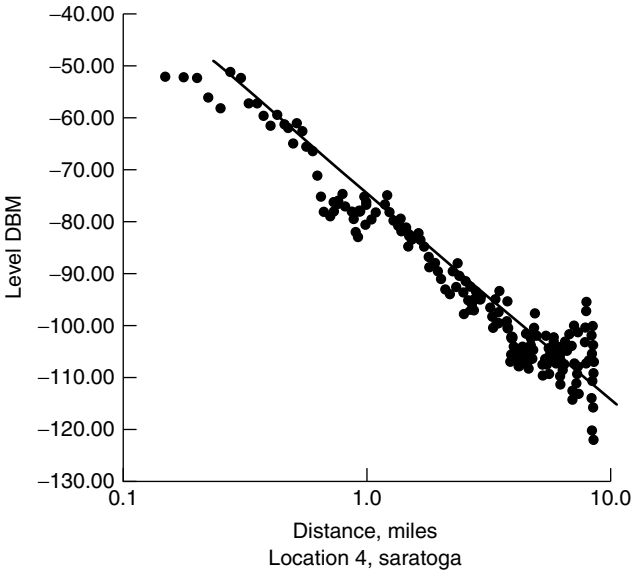


Figure 9. Measured small-area average power in dBm (dots) plotted versus R on a logarithmic scale. The solid line represents the theoretical prediction. (From Ref. 15, with permission.)

for propagation perpendicular to the rows of buildings. The variation of $Q(g_p)$ with g_p is plotted in Fig. 8 [15], and can be expressed in terms of a third order polynomial.

For large angles α, such as on satellite links, $g_p > 1$ and $Q(g_p) \approx 1$ so that only the last row of buildings before the mobile affects the received signal. A simple approximation to $Q(g_p)$ is given by the straight line shown dashed in Fig. 8. This approximation is given by

$$Q(g_p) = 2.35\, g_p^{0.9} \tag{10}$$

and is accurate to within 0.8 dB over the range $0.01 < g_p < 0.4$ [1].

4.2. PG_2: Diffraction from Rooftop to Ground Level

Many diffraction paths exist whereby the waves above the buildings reach ground level. In Fig. 7 the two rays giving the major contribution are as shown. The first of these is diffracted from the rooftop of the building nearest the mobile in the direction of the base station, while the second is reflected from the face of the building across the street. The field resulting from diffraction at the building edge is in the form of a cylindrical wave, with the edge acting as an equivalent line source. Because of the rapid spatial variation resulting from the interference of the two waves, the spatial average power will be the sum of the individual ray powers. With the foregoing assumptions, the spatial average received power is given by

$$PG_2 = \left[\frac{1}{\rho_1}|D(\theta_1)|^2 + |\Gamma|^2\frac{1}{\rho_2}|D(\theta_2)|^2\right] \tag{11}$$

where Γ is the reflection coefficient of the building opposite to the mobile, $k = 2\pi/\lambda$ and $D(\theta_i)$ is the diffraction coefficient.

For a receiver in the middle of the street, the distances ρ_1 and ρ_2 in from the diffracting edge (Fig. 9) are given by

$$\begin{aligned}\rho_1 &= \sqrt{(H_B - h_m)^2 + x^2}\\ \rho_2 &= \sqrt{(H_B - h_m)^2 + (2d - w - x)^2}\end{aligned} \tag{12}$$

while the angles θ_i for $i = 1, 2$ are

$$\theta_i = \arcsin\frac{(H_B - h_m)}{\rho_i} \tag{13}$$

In these expressions, H_B is the building height, h_m is the mobile height, w is the front-to-back dimension of the building, and x is the distance of the receiver from the diffracting edge of the building just before the mobile.

The diffraction coefficient $D(\theta_i)$ in Eq. (11) depends on the boundary condition at the rooftop edge of the building, which is rarely known. However, for diffraction angles θ_i away from 90°, the diffraction coefficient is not very sensitive to the boundary conditions. Thus we may use the diffraction coefficient for an absorbing edge, which for propagation perpendicular to the rows of buildings is [1]

$$D(\theta_i) = \frac{1}{\sqrt{2\pi k}}\left(\frac{1}{\theta_i - \alpha} + \frac{1}{2\pi + \theta_i - \alpha}\right) \approx \frac{1}{\sqrt{2\pi k}}\frac{1}{\theta_i} \tag{14}$$

We can simplify (11) by accounting for Γ, which for common building materials is $\Gamma \approx 0.3$. The value of Γ compensates for the differences in sizes of $D(\theta_1)/\sqrt{\rho_1}$ and $D(\theta_2)/\sqrt{\rho_2}$, so that the second term is close to the first term. The near equality of the two terms is seen from the deep fades observed in the fast fading pattern. Thus, the path gain for diffraction down to street level can be rewritten as

$$PG_2 \approx \frac{2}{\rho_1}[D(\theta_1)]^2 \approx \frac{1}{\pi k}\frac{1}{\rho_1\theta_1^2} \approx \frac{1}{\pi k}\frac{d/2}{(H_B - h_m)^2} \tag{15}$$

The variation of (11) with x has been validated by measurements in Japan [16] and England [17].

4.3. Path Gain for Macrocells

Macrocells in cities have $h_{BS} - H_B \sim 10$ m and 1 km $< R <$ 10 km. Since $d \approx 50$ m, the value of g_p falls in the range

0.015–0.15 at 900 MHz and 0.021–0.21 at 1800 MHz, so that we may use expression (12) for $Q(g_p)$. Combining expressions (1), (8), (9), (12), and (15) into (7), the path gain for isotropic antennas can be expressed in decibels as

$$PG_{dB} = 10\log\left(\frac{\lambda}{4\pi R}\right)^2 + 10\log\left[(2.35)^2\left(\frac{h_{BS}-H_B}{R}\right)^{1.8} \times \left(\frac{d}{\lambda}\right)^{0.9}\right] + 10\log\left[\frac{d/(2\pi k)}{(H_B-h_m)^2}\right] \quad (16)$$

Substituting $k = 2\pi/\lambda$ and $\lambda = c/f$, combining the various constant terms in (14) and expressing the frequency f_M in megahertz, the range R_k in kilometers, the path gain, can be written as

$$PG_{dB} = -92.5 - 21\log f_M + 10\log\left[\frac{d^{1.9}(h_{BS}-H_B)^{1.8}}{(H_B-h_m)^2}\right] - 38\log R_k \quad (17)$$

It is seen from (15) that the R dependence of Q combines with the free-space path to give the overall range dependence of 38 log R_k, corresponding to a range index $n = 3.8$ that is close to values measured in North American cities [4]. As a result of the near cancellation of the frequency dependence in Q^2 and PG_2, the path gain is seen to vary inversely with frequency to the 2.1 power, which is nearly that of the free-space path gain.

The predictions given by (17) for the received signal are shown in Fig. 9 superimposed on the small area average received power (dots) measured in Philadelphia [4]. The horizontal range is plotted on a logarithmic scale, for which (17) plots as a straight line. Excellent agreement is seen with the slope index of propagation and the average signal level. The deviations of the small area averages in Fig. 9 from the straight line correspond to the shadow fading. This deviation can be modeled in terms of the differences in building height along the rows; gaps between buildings, including street intersections; and the presence of trees [1,17].

4.4. Measurement-Based Models

In designing the original CMR systems, extensive measurements of the small-area average power versus R were made by various groups around the world. Okumura et al. [3] made an extensive set of measurements in and around Tokyo. Their work examined the effects of base station antenna height, frequency, building environment, terrain roughness, and other factors on the range dependence of the received signal, which was presented as curves of median received field strength (proportional to voltage) versus R for various parameters. Subsequently, Hata [4] fitted curves with simple formulas based on the slope intercept form $L = -10\log A + 10n\log R$ for the path loss L in decibels between isotropic antennas. Recall that the path loss in decibels is the negative of PG_{dB}. Hata's formulas were made to fit the measurements over the range of parameters: $150 \le f_M \le 1{,}500$ MHz, $1 \le R_k \le 20$ km, $30 \le h_{BS} \le 200$ m, and $1 \le h_m \le 10$ m. Over this range the result for urban areas is

$$L = 69.55 + 26.16\log f_M - 13.82\log h_{BS} - a(h_m) + (44.9 - 6.55\log h_{BS})\log R_k \quad (18)$$

The term $a(h_m)$ gives the dependence of path loss on subscriber antenna height, and is defined such that $a(1.5) = 0$.

To compare the predictions of (17) with (18), assume that $f_M = 1000$ MHz, $H_B = 10$ m (3 stories), $h_m = 1.5$ m, and $d = 50$ m. If we further assume that $h_{BS} = 20$ m, which is somewhat below the range of the Hatta model but consistent with practice, then the path loss obtained from (17) and (18) is

$$\text{Theory:}\quad L = 123.8 + 38\log R_k$$
$$\text{Hatta:}\quad L = 130.9 + 36.4\log R_k \quad (19)$$

The close agreement of the theory and measurements shown in this equation is a further substantiation of diffraction as a key process in the propagation from an elevated base station to subscribers at street level.

4.5. Range Dependence for Microcells

Microcellular systems make use of base station antennas located at about the height of three-story buildings, or on lampposts, to cover cells of radius 1 km or less. In a high-rise building environment, this placement is well below the building so that propagation is around the buildings rather than over them. For residential areas, this base station antenna height will be near to, or below the rooftops. In both environments, the location of the base station antenna relative to the buildings needs to be taken into accounted. Over microcells, the street grid likely to be rectangular, and line-of-sight (LoS) streets are a more significant fraction of all the streets in a cell, calling for their separate treatment. Measurement models appropriate to such antenna placement have been proposed for high-rise and residential environments [7,18]. Theoretical models have also been evaluated for predicting path loss, as discussed, for example, in Ref. 1.

5. 3D RAYS FOR SITE-SPECIFIC PREDICTIONS

Computer codes have been written to compute the ray paths and the ray fields working from databases of buildings and ground elevation. An example of a building database is shown in Fig. 10, which represents a simplified view of the high rise section of Rosslyn, Virginia [19]. In creating a building database there are tradeoffs that limit useful fidelity. The cost of creating the database increases with the detail included, as does the running time of the ray code. However, accuracy of the ray predictions does not continue to increase as more detail is added. The Rosslyn database of Fig. 4 shows the shape of the major geometric components of the buildings, but omits many smaller features, such as windows, balconies and decorative masonry.

Because almost all buildings have vertical sides, building databases are usually constructed with this assumption in order to reduce computation time. It is common to take the roofs to be flat, and some computer codes assume the ground to be flat. The most inexpensive databases describe each building using a polygon to represent its footprint, and a single building height. The more elaborate database of Fig. 10 stacks individual building elements, each of which has a polygonal base, while the height at each corner of the polygon can be different to accommodate slanting roofs.

In practice it is difficult to create a building database with position accuracy better than 0.5 m, or to include the architectural detail that can introduce phase shift in the reflection coefficients. For these reasons it is not possible to accurately compute the phases of the individual rays that would be needed to predict the spatial interference pattern for narrowband signals. Although the exact fading pattern cannot be predicted, its statistical properties can. The most important parameter is the small-area average received power, which corresponds to the sliding average in Fig. 2. This average, which is found by spatially averaging the power or field magnitude squared, can be computed by adding ray powers, as discussed in the text following Eq. (3).

The primary problem when using GO and UTD is to find the rays connecting the transmitter and the receiver. Even in simple environments there are a very large, and possibly infinite, number of such rays to be found. Geometric optical (GO) rays that undergo only reflections at the building walls and the ground are found using either the image method or the shooting–bouncing ray approach, which is most commonly employed for outdoor environments. It has been found that five to seven reflections must be accounted for to ensure accurate predictions [20].

In the shooting–bouncing ray (SBR) approach, rays starts from the transmitter in all directions over a sphere at incremental angular separations. For each ray from the transmitter, the first intersection with a building surface is computed. Using the laws of geometric reflection, the reflected ray is traced to the point of intersection with the first building surface, and so on. Because the rays have finite angular separation, there is a vanishing probability that a ray will pass through a predefined receiver point. To overcome this problem, the receiver is given a finite cross section whose diameter is equal to the product of the angular separation and the total pathlength. This procedure replaces the actual ray to the receiver by a single neighboring ray that undergoes reflections at the same building surfaces.

Diffracted rays are found by tracing the GO rays to the edge, and then treating the diffracting edge as a secondary source of rays that are traced using GO. An example of rays diffracted at a corner of one building and at the roof of another is shown in Fig. 11. The ray incident at one point along the edge excites diffracted rays that leave the edge in a Keller cone whose half-angle is equal to the angle between the incident ray and the edge [10]. Since the cone angle will vary along the edge, the edge is divided into small segments and a secondary ray trace is carried out from each segment. Moreover, each ray incident on an edge segment will, in general, have a different cone angle, hence requiring a separate ray trace.

Because each edge initiates a series of ray traces for each ray family that illuminates the edges, the three-dimensional (3D) ray trace is very time consuming, and can only accommodate rays that undergo no more than two diffraction events. In order to speed the code and to account for more diffraction events at the rooftops, the vertical plane launch (VPL) approximation has been proposed [19,21]. This approximation is involves replacing the Keller cone for diffraction at horizontal edges by the vertical planes. For the common case when the horizontal displacement of the rays is larger than the vertical displacement, the distortion of the ray path will be small with this approximation. When viewed from above, the rays diffracted in the forward direction lie in the plane of incidence, while rays diffracted backwards lie in the plane of reflection.

The VPL approximation allows the ray paths to be constructed by first carrying out a 2D ray trace in

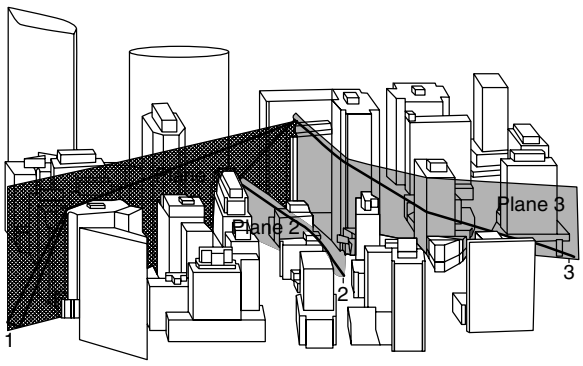


Figure 10. Building database of Rosslyn, Virginia, showing the vertical planes launched from a base station atop a building. The ray paths in the planes are those found from the VPL approximation. (From Ref. 19, with permission.)

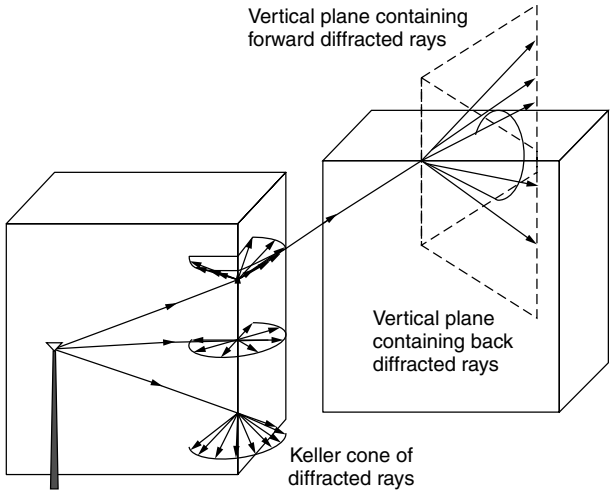


Figure 11. Rays diffracted at an edge lie in the Keller cone, whose half-angle is equal to that between the incident ray and the edge. In the VPL approximation, the cone for diffraction at horizontal edges is unrolled into the vertical planes. (© 2002 by H. L. Bertoni.)

the horizontal plane, followed by a simpler analytic determination of the ray paths in the vertical dimension. The procedure, in effect, traces vertical planes launched by the transmitter as they pass over and are reflected from building surfaces, as suggested in Fig. 10. These planes may also diffract at vertical building corners. After unfolding the vertical plane for each ray, the path in the vertical plane is determined by accounting for possible reflection or diffraction at horizontal edges. Because the shooting and bouncing ray method needs be used in only two dimensions, much less computer time is required. Moreover, by using analytic methods in the vertical plane, many more diffraction events at horizontal edges can be accommodated. An example of the VPL prediction of the spatial average received power in Rosslyn is shown in Fig. 12 [22]. The transmitter is located atop the building shown in Fig. 10, while the receiving locations are spaced approximately 5 m apart along a 2-km drive path on six different streets. For comparison with these predictions, measurements were made as the receiving van was driven along all the streets. The predictions are generally in good agreement with the measurements, except for 1360 receiver numbers and higher, which are on a street at the edge of the database. When compared to measurements, the error of ray predictions typically has an average of 1–2 dB and a standard deviation of 6–8 dB.

6. SUMMARY

Measurements and theory give complementary ways of predicting path loss and other statistical properties of the radio channel. Given the highly variable nature of the path loss, theoretical predictions are in good agreement with measurements. When validated against a set of measurements, the theoretical models allow for the variation of system parameters, such as frequency and antenna height, and building geometry. Thus a good theoretical model increases the value of a set of measurements by carrying it to a much larger range of parameters and building environments. Although we have discussed only a few aspects of path loss prediction in this article, theoretical models can be used to predict other channel characteristics. For example, computation of the ray fields via GO and UTD directly gives the directions of departure and arrival at both ends of the link, time delay of the ray, and the contribution of the ray fields to the received voltage. By making predictions for many mobile locations, the ray codes can be used for Monte Carlo simulation of statistical channel parameters, such as delay spread or angle spread.

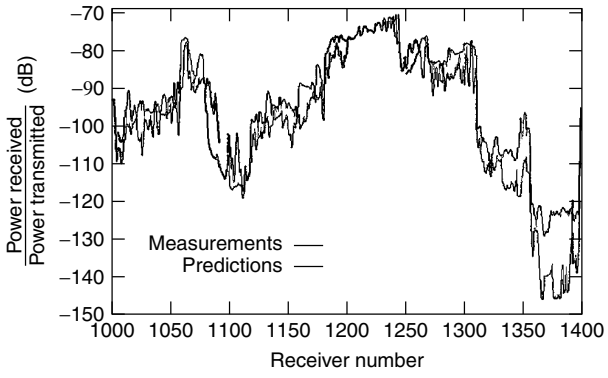


Figure 12. Comparison of measurements and predictions of the small-area path gain from a rooftop base station and mobiles at approximately 5-m intervals along six different streets in Rosslyn for a frequency of 900 MHz.

BIOGRAPHY

Henry L. Bertoni is on the faculty of Polytechnic University in Brooklyn, serving as head of the ECE Department (1990–95, 2001–present), and as vice provost of graduate studies (1995–96). His research has dealt with theoretical aspects of wave phenomena in electromagnetics, ultrasonics, acoustics, and optics. He has authored or coauthored over 80 journal papers and nine book chapters on these topics. Four journal articles have received best paper awards. His current research deals with characterizing the radio channel for modern wireless application, and the theoretical prediction of these characteristics. He and his students were the first to explain the physical mechanisms underlying characteristics observed in the measurements of cellular mobile radio signals. Much of this work is described in his recent book *Radio Propagation for Modern Wireless Systems*, Prentice Hall PTR, 2000. Dr. Bertoni is a fellow of the IEEE. He was the first chairman of the Technical Committee on Personal Communications of the IEEE Communications Society, and was IEEE representative to, and chairman of, the Hoover Medal Board of Award. He is a member of the International Scientific Radio Union and the Radio Club of America. From 1998 to 2001 he was a distinguished lecturer of the IEEE Antennas and Propagation Society.

BIBLIOGRAPHY

1. H. L. Bertoni, *Radio Propagation for Modern Wireless Applications*, Prentice-Hall PTR, Englewood Cliffs, NJ, 2000.
2. H. L. Bertoni, Radio channel characteristics observed for cellular and microcellular links: A tutorial review, *J. Commun. Networks* **1**: 249–265 (1999).
3. Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, Field strength and its variability in VHF and UHF land-mobile radio service, *Rev. Electric. Commun. Lab.* **16**: 825–873 (1968).
4. G. D. Ott and A. Plitkins, Urban path-loss characteristics at 820 MHz, *IEEE Trans. Vehic. Technol.* **VT-27**: 189–197 (1978).
5. H. H. Xia et al., Microcellular propagation characteristics for personal communications in urban and suburban environments, *IEEE Trans. Vehic. Technol.* **43**: 743–752 (1994).
6. M. Hata, Empirical formula for propagation loss in land mobile radio service, *IEEE Trans. Vehic. Technol.* **29**: 317–325 (1980).
7. D. Har, H. H. Xia, and H. L. Bertoni, Path loss prediction model for microcells, *IEEE Trans. Vehic. Technol.* **48**: 1453–1462 (1999).
8. L. J. Greenstein, V. Erceg, Y. S. Yeh, and M. V. Clark, A new path-gain/delay-spread propagation model for digital cellular channels, *IEEE Trans. Vehic. Technol.* **46**: 477–485 (1997).

9. C. Cheon, G. Liang, and H. L. Bertoni, Simulating radio channel statistics for different building environments, *IEEE J. Select. Areas Commun.* **19**: 2191–2200 (2001).

10. D. A. McNamara, C. W. I. Pistorius, and J. A. G. Malherbe, *Introduction to the Uniform Geometrical Theory of Diffraction*, Archtech House, Norwood, MA, 1990.

11. J. Fuhl, J.-P. Rossi, and E. Bonek, High-resolution 3-D direction-of-arrival determination for urban mobile radio, *IEEE Trans. Antennas and Propag.* **45**: 672–682 (1997).

12. M. Lecours, I. Y. Chouinard, G. Y. Delisle, and J. Roy, Statistical modeling of the received signal envelope in a mobile radio channel, *IEEE Trans. Vehic. Technol.* **37**: 204–212 (1988).

13. H. H. Xia et al., Radio propagation characteristics for line-of-sight microcellular and personal communications, *IEEE Trans. Antennas Propag.* **41**: 1439–1447 (1993).

14. A. J. Rustako, Jr., N. Amitay, g. J. Owens, and R. S. Romano, Radio propagation at microwave frequencies for line-of-sight microcellular mobile and personal communications, *IEEE Trans. Vehic. Technol.* **40**: 203–210 (1991).

15. J. Walfish and H. L. Bertoni, A theoretical model of UHF propagation in urban environments, *IEEE Trans. Antennas Propag.* **36**(10): 1788–1796 (1988).

16. F. Ikegami, S. Yoshida, T. Takeuchi, and M. Umehira, Propagation factors controlling mean field strength on urban streets, *IEEE Trans. Antennas Propag.* **32**: 822–829 (1984).

17. L. R. Maciel and H. L. Bertoni, Theoretical prediction of slow fading statistics in urban environments, *Proc. IEEE ICUPC'92 Conf.*, 1992, pp. 1–4.

18. E. Damosso, ed., *COST Action 231: Digital Mobile Radio; towards Future Generation Systems*, European Commission, Directorate G, Brussels, 1999.

19. G. Liang and H. L. Bertoni, A new approach to 3-D ray tracing for propagation prediction in cities, *IEEE Trans. Antennas Propag.* **46**: 853–863 (1998).

20. G. E. Athanasiadou, A. R. Nix, and J. P. McGeehan, Microcellular ray tracing propagation model and evaluation of its narrow-band and wide-band predictions, *IEEE J. Select. Areas Commun.* **18**: 322–335 (2000).

21. J. P. Rossi et al., A ray-launching method for radio-mobile propagation in urban area, *Digest of IEEE APS Symp.* London, Ontario, Canada, June 1991, pp. 1540–1543.

22. G. Liang, private communication, Jan. 2002.

PEAK-TO-AVERAGE POWER RATIO OF ORTHOGONAL FREQUENCY-DIVISION MULTIPLEXING

CHINTHA TELLAMBURA
Monash University
Clayton, Victoria, Australia

MATTHEW G. PARKER
University of Bergen
Bergen, Norway

1. INTRODUCTION

OFDM techniques have been proposed for digital TV broadcasting and high-speed wireless networks over multipath channels [1]. OFDM is commonly implemented using discrete Fourier transform (DFT) techniques and has been adopted, or is being investigated, for wireless LANs, wireless ATM, digital audio broadcasting [2], terrestrial digital videobroad-casting [3], and the broadband wireless local loop. OFDM offers many advantages such as resistance to multipath and excellent performance under noisy conditions.

With OFDM, the single carrier wave is replaced by simultaneous transmission of the signal on multiple, equally spaced subcarriers [4]. A baseband version of OFDM is called *discrete multitone transmission* (DMT). OFDM systems require the calculation of discrete Fourier transforms (DFTs). It is the availability of technology that allows for the implementation of fast transform (FFT) algorithms on integrated circuits at a reasonable price, that has made OFDM and DMT the modulation method of choice for the commercial applications given above [5].

Unfortunately, there are a number of difficulties with implementing OFDM and DMT:

- When the sinusoidal signals of the N subcarriers add constructively, the peak power can be N times the mean power; that is, the peak-to-average power ratio (PAR) of the transmitted signal can as large as N.
- Radiofrequency amplifiers are used to achieve the linearity over the entire signal. This causes high battery demand in mobile/wireless applications.
- The allocation of the radio spectrum for radio LANs limits the isotropically radiated peak envelope power. If the output peak is clipped, this generates out of band radiation due to intermodulation distortion as well as in-band distortion.

The first difficulty listed above is in fact the cause of the second and third. These problems limit the usefulness of OFDM for some applications.

2. PEAK-TO-AVERAGE POWER RATIO OF OFDM

The complex baseband OFDM signal may be represented as

$$s(t) = \frac{1}{\sqrt{N}} \sum_{n=-\infty}^{\infty} \sum_{k=0}^{N-1} a_{k,n} e^{j2\pi k \Delta f t} g[t - k(T + T_g)] \qquad (1)$$

where $j^2 = -1$, N is the total number of subcarriers, and $a_{k,n}$ is the data symbol for the k subcarrier and the nth OFDM symbol (i.e., N subcarrier OFDM system transmits a block of N data symbols per OFDM symbol). The frequency separation between any two adjacent subcarriers is $\Delta f = 1/T$. The unit rectangular pulse $g(t)$ is of duration $T + T_g$, where T_g is known as the "guard interval."

Because there is no overlap between different OFDM symbols, for the PAR problem it is sufficient to consider a single OFDM symbol ($n = 0$). In practice, filtering can cause some degree of intersymbol interference, which will be neglected here. The guard interval is used to repeat parts of each OFDM signal, but has no effect on the PAR. Therefore we may set $T_g = 0$. Since only $n = 0$ is sufficient

for the problem at hand, $a_{k,n}$ may be replaced by a_k. Thus, for PAR considerations, the complex baseband signal may be represented as

$$s(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k e^{j2\pi k \Delta f t}, \quad 0 \le t < T \tag{2}$$

Note that all the subcarriers are mutually orthogonal. Each modulated symbol a_k is chosen from the set $F_q = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ of q distinct elements. The set F_q is called the *signal constellation* of the q-ary modulation scheme. While several modulation schemes are in use it should be noted that the statistical distribution of the PAR is largely independent of the signal constellation. In most applications, one uses phase shift keying (PSK) signaling in which

$$F_q = \{1, \zeta, \zeta^2, \ldots, \zeta^{q-1}\} \tag{3}$$

where $\zeta = e^{j2\pi/M}$. For example, for binary PSK (BPSK) $F_q = \{1, -1\}$ and for quaternary PSK (QPSK), $F_q = \{1, j, -1, -j\}$. Another popular modulation technique is quadrature amplitude modulation (QAM), for which

$$F_q = \{m + jn\} \tag{4}$$

where m and n are selected integers.

Note that for any PSK constellation F_q for any $u \in F_q$, $|u|^2 = 1$, and this condition does not hold for QAM constellations. In general, all elements in F_q occur with equal probability $1/q$.

We shall write an ordered N-tuple $\mathbf{a} = (a_0, a_1, \ldots, a_{N-1})$ and let $(F_q)^N$ denote the set of all ordered N-tuples where each $a_k \in F_q$. We shall refer to any member of the set $(F_q)^N$ as a *data frame*. Note that each a_k carries $\log_2 q$ data bits. Normal values are $q = 2, 4, 8$, and so on. Each OFDM symbol thus carries $N \log_2 q$ data bits. The instantaneous envelope *power* of the signal is the real-valued function $P_a(t) = |s(t)|^2$. We define the peak-to-average power ratio (PAR) as

$$\mathrm{PAR}(\mathbf{a}) = \frac{\max_t P_a(t)}{E(P_a(t))} \tag{5}$$

where $E(\cdot)$ denotes the time average. For a PSK constellation, this value is unity. Strictly speaking, this definition should be called the peak-to mean envelope power ratio (PMEPR), because $s(t)$ is the envelope but not the transmitted signal itself. As such, this is also called the *baseband* PAR. The actual transmitted signal is modeled as

$$S(t) = \Re(s(t)e^{j2\pi f_c t}) \tag{6}$$

where f_c is the carrier frequency and $\Re(z)$ denotes the real part of z. The definition of PAR would now be

$$\mathrm{PAR}(\mathbf{a}) = \frac{\max_t |\Re(s(t)e^{j2\pi f_c t})|^2}{E(|S(t)|^2)} \tag{7}$$

which is also known as the *passband* PAR. It is often easier to work with definition (5) rather than (7). Further, if f_c is large (i.e., $f_c \gg N/T$), which is the case in practice, this is approximately 3 dB higher than the baseband PAR. This difference is more or less fixed. Consequently, we will be using the baseband PAR without any loss of generality or applicability.

The PAR is a function of the data frame, and recall that there are q^N distinct data frames. For any input data frame, we have

$$1 < \mathrm{PAR}(\mathbf{a}) \le N \tag{8}$$

For example, for $N = 256$ the PAR can be as high as 24 dB $[10 \log_{10}(256)]$. Fortunately, very high PAR values are very rare. For example, with BPSK, only four sequences $0000\ldots, 1111\ldots, 0101\ldots$, and $1010\ldots$ achieve $\mathrm{PAR}(\mathbf{a}) = N$. For randomly distributed data, the probability of an occurrence of this is $4/2^N = 2^{2-N}$. This probability is negligible when N is large - as is the case in practice.

The PAR of a sequence is closely related to its out-of-phase aperiodic autocorrelation (APA) values. The APA coefficients of $\mathbf{a}$ are

$$\rho(k) = \sum_{n=0}^{N-1-k} a_{n+k} a_n^* \quad \text{for} \quad k = 0, \ldots, N-1 \tag{9}$$

The PAR is bounded as

$$\mathrm{PAR}(\mathbf{a}) \le 1 + \frac{2}{N} \sum_{k=1}^{N-1} |\rho(k)| \tag{10}$$

This shows that binary or polyphase sequences with low out-of-phase APA values [i.e., small $\rho(k)$ for $k \ge 1$] can be used to construct low PAR signals. Conversely, Schroeder, [6] notes that sequences that have low PAR also have low APA values. The problem of constructing sequences with low APA values (i.e., similar to an impulse function) is a longstanding problem. The general problem of finding sequences that minimize the PAR seems just as difficult.

Example 1. Consider $\mathbf{a} = (1, 1, 1, -1, 1)$, which is a Barker sequence. Its APA is $\{5, 0, 1, 0, 1\}$. Hence applying (10) gives $\mathrm{PAR}(\mathbf{a}) \le 1 + \frac{4}{5}$; thus, the PAR is less than 2.55 dB. By computing (5), the PAR is exactly 2.55 dB. In this case, the upper bound coincides with the exact.

Example 2. Using (10), we can immediately device a simple PAR reduction code. For an information sequence $p_0 = (m_0, m_1, \ldots, m_{n-1})$, where $m_k \in \{1, -1\}$, the encoder output is given by $p_e = (m_0, \ldots, m_{n-1}, -m_{n-2}, m_{n-3}, -m_{n-4}, \ldots, -m_0)$. This code rate is $n/(2n-1)$ and length $N = 2n - 1$. For k odd, $\rho(k)$ of p_e is zero. For example, when $n = 3$, then $\rho(1) = \rho(3) = 0$ and $|\rho(2)| \le 3$ and $|\rho(4)| = 1$. Thus our bound gives PAR ≤ 2.6, almost a 3-dB reduction. For large N, the coding rate is almost $\frac{1}{2}$. The sum of $|\rho(k)|$ for k even is bounded by $(N-1)^2/4$. Thus, the peak is bounded as $\mathrm{PAR} \le 1 + (N-1)^2/(2N)$, a 3 dB reduction.

3. STATISTICAL PROPERTIES OF PAR

3.1. CCDF of the PAR

Since the input data are randomly distributed in many applications (if not, they can be made so by the use of a

suitable scrambling operation), PAR(**a**) itself is a random variable. The complementary cumulative distribution function (CCDF), the probability that the PAR of an OFDM symbol exceeds a certain threshold, is useful for many purposes. The CCDF is defined as

$$F(\zeta) = \Pr(\text{PAR}(\mathbf{a}) \geq \zeta) \tag{11}$$

The CCDF is shown for $N = 32$, 64, 128, and 256 in Fig. 1. For 256 subcarriers, the maximum PAR is 24 dB. To reach this, however, all the subcarriers need to be in phase at some instant in time, and therefore produce an amplitude peak equal to the sum of the amplitudes of the individual subcarriers. This occurs with extremely low probability for large N. For example, the PAR exceeds 12.5 dB for 1 in 10^5 of all the possible transmitted OFDM symbols.

3.2. Gaussian Approach

An exact expression for $F(\zeta)$ is not yet known because computing PAR(**a**) requires the time instances for which the derivative of the envelope power $P_a(t)$ equals zero. This is a root-finding problem for a nonlinear equation and as such there are no analytic formulas for the roots. However, as a way around this difficulty, we make the following assumptions:

A1. $s(t)$ is a complex Gaussian process.
A2. N samples of $s(t)$ given by $s_n = s(nT/N)$ for $n = 0, 1, \ldots, N-1$ are independent and identically distributed complex Gaussian random variables.
A3. The maximum of $|s_n|^2$ is equal to $\max_t = P_a(t)$.

A1 becomes quite accurate as N increases. The independence assumption in A2 is never exact because we know that the N samples must satisfy Parseval's theorem. A3 is never exactly true. Nevertheless, we define random variables (RVs):

$$Y_n = |s_n|^2 \quad n = 0, \ldots, N-1 \tag{12}$$

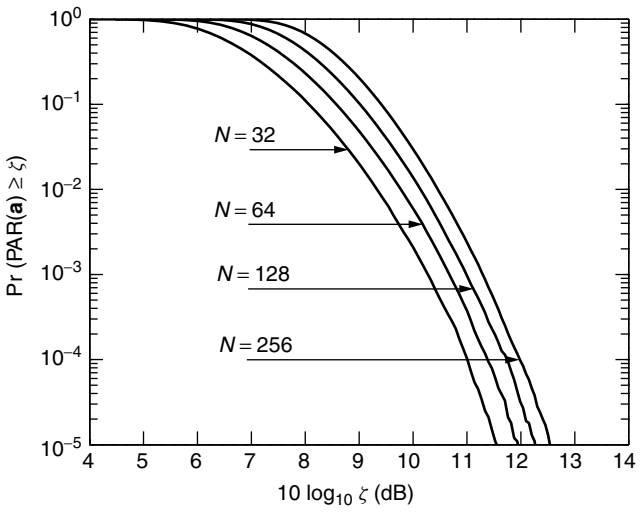


Figure 1. The CCDF for N QPSK subcarriers.

Since the real and imaginary parts of s_n are independent (provided no oversampling is used), with mean zero and the same variance, Y_n approaches a chi-squared distribution with two degrees of freedom. The Y_n values hence are i.i.d. and exponential RVs. Their cumulative density function (CDF) is given by

$$F(y) = \Pr(Y_n \leq y) = 1 - e^{-y} \tag{13}$$

The statistical properties of the maximum of Y_n can be readily derived. We see that the CDF of the maximum is given by

$$\begin{aligned} F_{\max}(y) &= \Pr\{Y_{\max} \leq y\} \\ &= \Pr\{\text{all } Y_n \leq y\} \\ &= (1 - e^{-y})^N \end{aligned} \tag{14}$$

The CDF of the PAR is then obtained as

$$F(\zeta) = 1 - (1 - e^{-\zeta})^N \tag{15}$$

Despite the three assumptions that may not always hold, this result is useful for quick analysis of the statistics of PAR and for determining the achievable PAR reduction for some schemes.

3.3. Asymptotic Results

The statistical behavior of PAR(**a**) for large N is important, and in some practical applications, N can be as large as 2048 or more. We can show that PAR(**a**) grows as $\ln N$; that is, for a randomly picked data sequence, PAR(**a**) is unlikely to be significantly less than $\ln N$. From Eq. (14), we get

$$\begin{aligned} \Pr\{\text{PAR}(\mathbf{x}) \leq \alpha \ln N + h\} &= (1 - e^{-(\alpha \ln N + h)})^N \\ &\simeq e^{-N^{(1-\alpha)}e^{-h}} \text{ for } N \to \infty \end{aligned} \tag{16}$$

where $\alpha \geq 1$. This follows readily from fact that $\lim_{n\to\infty}(1 - \theta/n)^n = e^{-\theta}$. If $\alpha = 1$ and $h = -h$ in (16), we see that

$$\Pr\{\text{PAR}(\mathbf{x}) \leq \ln N - h\} \simeq e^{-e^h} \tag{17}$$

This is the formal mathematical equivalent of our statement above. This also elicits information about clipping and coding for PAR reduction. *Clipping* is a method used to deal with high peak amplitude excursions at the transmitter output. This is necessary because the D/A converter has a limited resolution (i.e., the number of bits) and the power amplifier cannot be linear over an amplitude range that includes the peak amplitudes. A clip occurs when the signal amplitude exceeds a predefined threshold, and hence clipping is described as follows:

$$s_c(t) = \begin{cases} s(t) & \text{if } |s(t)| \leq s_{\text{clip}} \\ s_{\text{clip}} e^{j\angle s(t)} & \text{if } |s(t)| > s_{\text{clip}} \end{cases} \tag{18}$$

where $s_{\text{clip}} > 0$ is the clipping threshold. The probability of clipping is the number of clips per unit time. Of course, each clip introduces symbol errors and out-of-band noise.

Equation (17) shows that for a normal OFDM system, if the clipping threshold is set below $\ln N$, then the clipping probability will be unity for large N. Likewise, (16) suggests if the clipping threshold is set above αN, then the clipping probability can be arbitrarily reduced. Here the right-hand side (RHS) of (16) explicitly shows that the decay rate depends on both α and h.

Additionally, coding seems unnecessary for PAR reduction of $\alpha \ln N$ or higher as clipping will not occur very often. In contrast, keeping the PAR at a level significantly below $\ln N$ using clipping will introduce significant distortion. At this point coding becomes interesting. Consider coding to reduce the PAR to h below $\ln N$, where $|h|$ is small compared to $\ln N$. Assume a binary modulation [i.e., $x_k \in (+1, -1)$]. From (17), the achievable coding rate is

$$R(h) = \frac{\log_2(2^N e^{-e^h})}{N} = 1 - \frac{e^h}{N \log_2 e} \tag{19}$$

As h increases, the achievable coding rate tends to zero. Similarly, if h is negative, the achievable coding rate tends to one, suggesting that the PAR can be limited to a level above $\ln N$ with very little redundancy. Moreover, the required amount of redundancy decays exponentially with the difference between the target PAR and $\ln N$. It appears that any family of good codes (i.e., of nonvanishing coding rate) must have a PAR bound of around $\ln N$.

3.3.1. Code Rate. Consider QPSK modulated N subcarriers, which can accommodate $2N$ information bits at most. Suppose that we need a code rate $R = 1 - K/(2N)$ to limit the PAR to ζ. Thus we have

$$\Pr(\text{PAR} \le \zeta) = \frac{2^{2N-K}}{2^{2N}} \tag{20}$$

which can be rearranged as

$$R = 1 + \frac{1}{2N} \log_2 \Pr(\text{PAR} \le \zeta) \tag{21}$$

This probability term can be estimated by simulation. Figure 2 shows the required code rate to limit the PAR. For $N = 128$, to reduce PAR to 7 dB from 21 dB for the uncoded case, the required code rate is 0.98. This suggests that the PAR can be much reduced by a small amount of redundancy. Unfortunately, it has thus far not been possible to discover such a code.

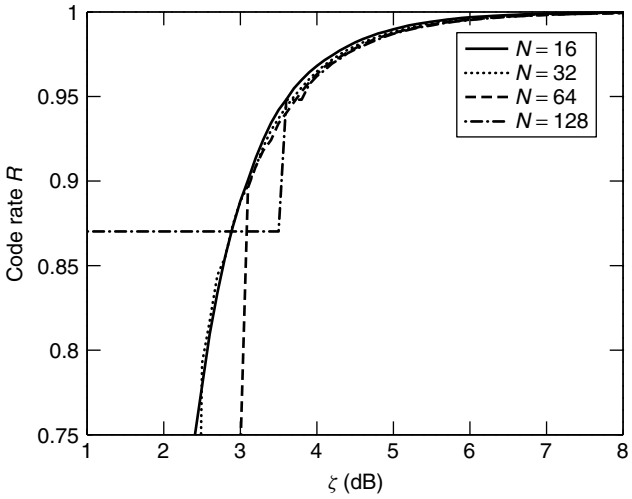


Figure 2. Code rate for limiting the PAR.

4. COMPUTATIONAL METHODS FOR PAR

4.1. Discrete-Time PAR

In order to compute PAR($\mathbf{a}$) exactly, we need all the roots of the equation

$$\frac{dP_{\mathbf{a}}(t)}{dt} = 0 \tag{22}$$

which is difficult to solve, especially for higher-order modulation formats. Most PAR reduction techniques are concerned with reducing PAR($\mathbf{a}$) [Eq. (5)]. However, since most systems employ discrete-time signals, the maximum amplitude of LN samples of $s(t)$ is used to approximate it, where L is the oversampling factor. The sampling can be implemented by an inverse discrete Fourier transform (IDFT). Hence consider the IDFT of length LN of $\mathbf{a}$ expressed as

$$A_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k e^{i2\pi nk/LN} \quad n = 0, 1, \ldots, LN-1 \tag{23}$$

It is seen that A_n are the samples of the waveform [Eq. (2)]. The discrete-time PAR is thus defined as

$$\text{PAR}(\mathbf{a})_{\text{dis}} = \max_{0 \le n < LN} |A_n|^2 \tag{24}$$

$L > 1$ corresponds to oversampling. Of course, if $L \gg 1$, the discrete-time PAR should approach the (continuous-time) PAR. It is therefore clear that

$$\text{PAR}(\mathbf{a})_{\text{dis}} \le \text{PAR}(\mathbf{a}) \tag{25}$$

Moreover [7] has shown that $\text{PAR}(\mathbf{a}) \le 2\text{PAR}(\mathbf{a})_{\text{dis}}$ if $L = \lceil 2\pi \rceil$. In practice, samples of (2) are generated by means of an inverse fast Fourier transform (IFFT), and fed to a digital-to-analog converter followed by an antialiasing lowpass filter. Quite often an oversampling factor of 4 is sufficiently accurate. For BPSK modulated subcarriers, this fact can be verified because the continuous-time PAR can be computed exactly. The most common method to compute PAR($\mathbf{a}$) is to use the discrete-time PAR. Note that we will simply use the term PAR, except when we are concerned about the difference between the discrete- and continuous-time values.

4.2. Using the Infinity Norm

This method was suggested by Van Eetvelt. The peak of a continuous function $s(t)$ is given by the L_∞ norm defined as

$$\max |s(t)| = L_\infty(s(t)) = \lim_{n \to \infty} \left[\frac{1}{T} \int_0^T |s(t)|^n dt \right]^{1/n} \tag{26}$$

To compute the L_∞ norm, we can use the result that for increasing p the L_p norm is nondecreasing. So in practice, taking a sufficiently large power allows the peak value to be approximated as closely as required. However, as computing this integral is not that easy, the use of the discrete-time PAR is much more convenient.

4.3. Computation of Continuous-Time PAR: BPSK case

To compute the continuous-time PAR, the roots of the derivative of the envelope power function (EPF) are required. At first, finding the required roots appears very difficult, since this derivative consists of sinusoidal functions. As such, the problem suggests a general root finding algorithm for nonlinear functions. Fortunately, this difficulty can be avoided for the BPSK case. Using an inverse cosine based transformation, the EPF can be converted to a sum of Chebysev polynomials. Moreover, the required roots are now trapped within the interval from 0 to 1. So the original root finding problem is reduced to a root finding problem for a polynomial. Reliable algorithms for finding all roots of a polynomial (a polynomial of order n will have n roots) are well known. Consequently, using this approach, the absolute peak of the EPF can be evaluated exactly. In this case, we have $a_k \in \{1, -1\}$. It is easy to show that [8,9]

$$P_a(t) = \sum_{k=0}^{N-1} \beta_k \cos(2\pi k t) \tag{27}$$

where

$$\beta_k = \begin{cases} 1 & k = 0 \\ \dfrac{2}{N} \displaystyle\sum_{n=0}^{N-1-k} a_n a_{n+k} & k = 1, 2, \ldots, N-1 \end{cases} \tag{28}$$

To compute PAR($\mathbf{a}$) exactly, the roots of $\dfrac{dP_a(t)}{dt} = 0$ are needed. Let us define

$$Q_a(t) = P_a\left[\frac{\cos^{-1} t}{2\pi}\right] = \sum_{k=0}^{N-1} \beta_k T_k(t) \tag{29}$$

where $T_k(t) = \cos(k \cos^{-1} t)$ is the kth order Chebysev polynomial (Ref. 10, p. 1054). Note that $T_0(t) = 1, T_1(t) = t, T_2(t) = 2t^2 - 1$ and so on (explicit expressions for the coefficients of $T_n(t)$ for any N are available). Since $Q_{\mathbf{a}}(t)$ is a polynomial of degree $(N-1)$, its first derivative is a polynomial of degree $(N-2)$. Recall that a polynomial of degree n will have n roots. These roots can be real or complex and many algorithms exist with which one can find all the roots of a polynomial. For example, a companion matrix can be constructed whose eigenvalues are the desired roots. Since

$$\frac{dP_{\mathbf{a}}(t)}{dt} = \frac{dQ_{\mathbf{a}}(\cos(2\pi t))}{dt} = -2\pi \sin(2\pi t) \frac{dQ_{\mathbf{a}}[\cos(2\pi t)]}{d\cos(2\pi t)} \tag{30}$$

the derivative vanishes both at $t = 0, \frac{1}{2}$ and at the transforms of the real roots of $\dfrac{dQ_a(t)}{dt}$ that lie between -1 and $+1$. Let those roots be $\xi_1, \xi_2, \ldots, \xi_M$ where $M \le N-2$. Define the set

$$\Lambda = \left\{0, \frac{1}{2}, \frac{\cos^{-1}\xi_1}{2\pi}, \ldots, \frac{\cos^{-1}\xi_M}{2\pi}\right\} \tag{31}$$

It is clear that all the required roots of the derivative of $P_a(t)$ are in this set. Note that $P_{\mathbf{a}}(t)$ is a periodic function and only the roots between 0 to 1 need to be considered. Therefore, the continuous-time PAR is obtained by

$$\text{PAR}(\mathbf{a}) = \max_{t \in \Lambda} P_{\mathbf{a}}(t) \tag{32}$$

5. PAR REDUCTION METHODS

We next describe some of the techniques that have been proposed to reduce the effects of high PAR.

5.1. Multiple Signal Generation

The basic idea behind this approach is to generate multiple, independent OFDM symbols to represent an input data frame and select the minimum PAR symbol for transmission. There are several techniques based on this idea and these primarily differ in the way they generate the multiple symbols. Another issue with this approach is the need for side information to tell the receiver which one of the signals has been used. Suppose that $M \ge 1$ independent OFDM symbols are generated for an information sequence. The CCDF of the minimum of these is given by

$$\Pr(\text{PAR}_{\min} \ge \zeta) = [1 - (1 - e^{-\zeta})^N]^M \tag{33}$$

Figure 3 shows this CCDF for $M = 1, 2, \ldots, 16$ and $N = 128$. $M = 1$ is the ordinary OFDM. Even for $M = 4$, a PAR reduction of about 4 dB occurs at a probability of 10^{-6}.

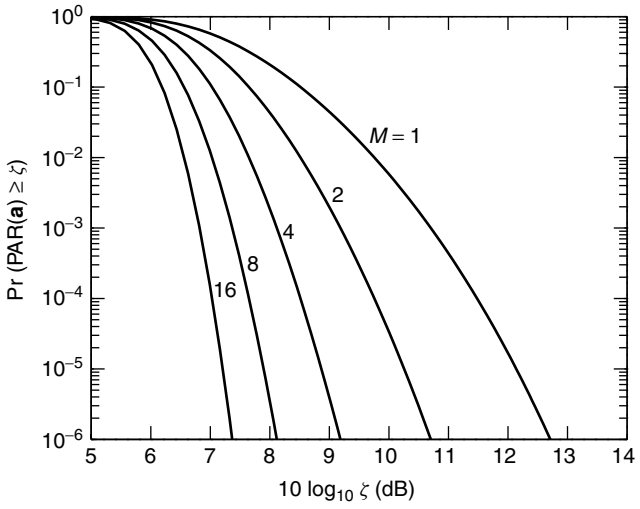


Figure 3. CCDF for the lowest PAR out of M statistically independent signals.

5.1.1. Partial Transmit Sequences. We shall write the input data block as a vector, $\mathbf{X} = [X_0, \ldots, X_{N-1}]^T$. For the PTS approach, the input data vector $\mathbf{X}$ is partitioned into disjoint subblocks, as $\{\mathbf{X}_m | m = 1, 2, \ldots, M\}$, and these are combined to minimize the PAR. While several subblock partitioning schemes do exist, we assume the simplest scheme for which the subblocks consist of a contiguous set of subcarriers and are of equal size. Now, suppose that for $m = 1, \ldots, M$, $\mathbf{A}_m = [A_{m1}, A_{m2}, \ldots, A_{m,LN}]^T$ is the zero-padded IFFT of $\mathbf{X}_m$. These are the partial transmit sequences. The objective is thus to combine these with the aim of minimizing the PAR. The signal samples at the output of the PTS combiner can be written as

$$\mathbf{S} = \begin{bmatrix} A_{11} & A_{21} & \ldots & A_{M1} \\ A_{12} & A_{22} & \ldots & A_{M2} \\ \ldots & \ldots & \ldots & \ldots \\ A_{1LN} & A_{2LN} & \ldots & A_{MLN} \end{bmatrix} \begin{bmatrix} e^{j\phi_1} \\ e^{j\phi_2} \\ \vdots \\ e^{j\phi_M} \end{bmatrix} \tag{34}$$

where $\mathbf{S} = [S_1(\Phi), \ldots, S_{LN}(\Phi)]^T$ contains the optimized signal samples. We shall write the phase factors as a vector, $\Phi = [\phi_1, \phi_2, \ldots, \phi_M]^T$. The phase factors $\{\phi_k\}$ are chosen to minimize the peak of the signal samples $|S_k(\Phi)|$. So the minimum PAR is related to the problem

$$\begin{aligned} &\text{Minimize} \quad \max_{0<k\le LN} |S_k(\Phi)| \\ &\text{subject to} \quad 0 \le \phi_m < 2\pi, m = 1, \ldots, M \end{aligned} \tag{35}$$

Suppose $\hat{\phi}_m$ to be the global optimal solution to this problem. Unfortunately, there appears to be no simple way to obtain $\hat{\phi}_m$ analytically. For coherent demodulation, it is necessary to send $\hat{\phi}_m$ to the receiver as side information. When $\hat{\phi}_m$ is a continuous value, an infinite number of bits will be required as side information. The solution to this problem is to limit $\hat{\phi}_m$ to a level from a finite number of predetermined levels (quantization). For differential demodulation, it is not necessary to send $\hat{\phi}_m$ to the receiver, but $M - 1$ subcarriers at the subblock boundaries have to be set aside as reference carriers.

The phase factors are restricted to a finite set of values and hence (35) is approximated by the problem

$$\begin{aligned} &\text{Minimize} \quad \max_{0<k\le LN} |S_k(\Phi)| \\ &\text{subject to} \quad \phi_m \in \left\{ \frac{2\pi l}{W} | l = 0, \ldots, W-1 \right\} \end{aligned} \tag{36}$$

If the number of rotation angles W is "sufficiently" large, the solution of (36) will approach that of (35). Furthermore, ϕ_1 can be fixed without any performance loss. Now, there are only $M - 1$ free variables to be optimized and hence W^{M-1} distinct phase vectors, Φ_i, need to be tested. As such, (36) is solved using W^{M-1} iterations; the ith iteration involves computing LN signal samples, each of which is denoted by $S_k(\Phi_i)$, using (34) and choosing the maximum $|S_k(\Phi_i)|$ value. At the end of each iteration, the phase vector is retained if the current value of max $|S_k(\Phi_i)|$ is less than the previous maximum. The phase vector that is retained after all the iterations are completed will be an approximation to the global optimal solution of (35).

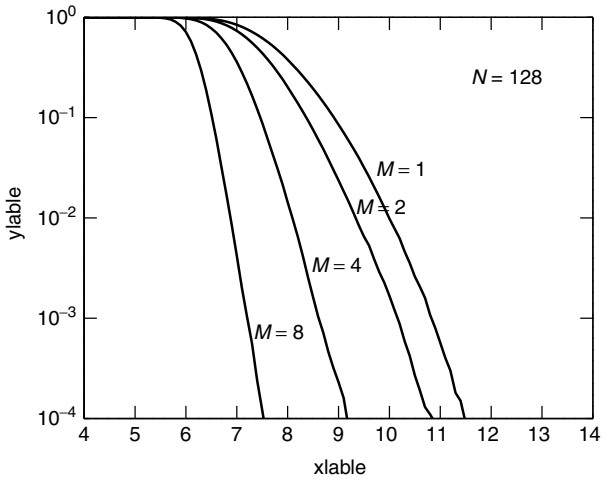


Figure 4. PTS performance for several M for $N = 128$ QPSK subcarriers.

In most reported studies, $W = 2$. In some cases, the use of more rotation angles ($W > 2$) has been found to yield diminishing returns. Figure 4 shows the PAR distribution for this method for varying M, with normal OFDM being $M = 1$. For $M = 8$, the PAR can be reduced by about 4 dB at a probability of 10^{-6}.

5.1.2. Selected Mapping (SLM) Approach. For a given M-PSK sequence, one generates M independent M-PSK sequences by multiplying by M fixed vectors and choosing the sequence with lowest PAR for transmission. This method is simple and very impressive. However, it needs M FFTs to select the best sequence among L. Suppose that we have M fixed phase vectors $\underline{P}_k = (p_k^0, p_k^1, \ldots, p_k^{N-1})$ for $k = 1, \ldots, M$, where $p_k^n \in \{0, 1, \ldots, M-1\}$ for $\forall n, k$. Without loss of generality, $p_1^n = 0$ $\forall n$. For an input data sequence $\mathbf{a}$, we generate M independent' sequences

$$A_k = \mathbf{a} \oplus \underline{P}_k \quad k = 1, \ldots, M \tag{37}$$

where $\underline{u} \oplus \underline{v}$ is the componentwise modulo M addition of $\underline{u}$ and $\underline{v}$. Originally, it was suggested to use the following selection function [9,11]: transmit A_l for $1 \le l \le L$ if

$$\text{SF} = \begin{cases} \text{PAR}(A_l) & \text{Baüml} \\ |W_H - N|^2 + |R_1|^2 & \text{Van Eetvelt} \end{cases} \tag{38}$$

is minimized. Here

$$R_1 = \sum_{k=1}^{N-1} A_{l:k} A_{l:k+1}^*$$

and W_H is the binary Hamming weight of the length N binary sequence. The performance of Baüml's SF is quite good, but requires multiple FFTs per input data frame.

5.1.3. Interleaving Approach. In this approach $K - 1$ interleavers are used at the transmitter [12]. These interleavers produce $K - 1$ permuted frames of the input

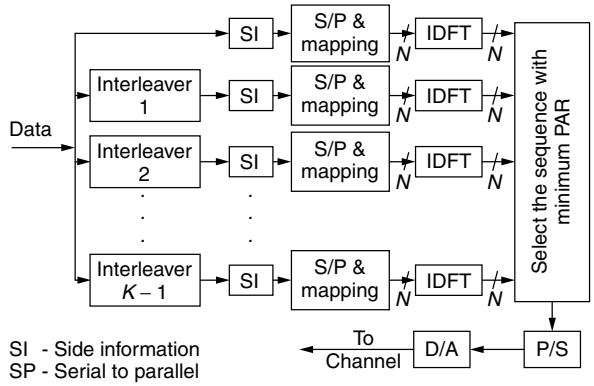


Figure 5. System model.

data before mapping into QPSK symbols. The 4 times oversampled IDFT of each frame (including the uncoded frame) is used to compute its PAR. The minimum PAR frame of all the K frames is selected for transmission. The identity of the corresponding interleaver is also sent to the receiver as side information. Figure 5 describes an OFDM transmitter with interleavers to reduce the PAR. The PAR reduction achievable with this method is similar to that of the PTS method.

5.2. Coding Techniques

5.2.1. Constructing Sequence Families with Low PAR and High Distance. Some definitions: A M-ary code $\mathcal{C}$ is a given set of sequences of symbols where each symbol is chosen from a set $F_M = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ of M distinct elements. The set F_M is often taken to be the set $Z_M = \{0, 1, 2, \ldots, M-1\}$, with $M = 2^h$ for positive integer h. We will denote a codeword as an N-tuple $(b_0, \ldots, b_{N-1})$ by $\mathbf{b}$. The Hamming distance between two sequences or codewords is defined as

$$d_H(\mathbf{a}, \mathbf{b}) = \sum_{n=0}^{N-1} \delta(a_n - b_n)$$

where $\delta(\cdot)$ is the Krönecker delta function. A critical parameter of a code $\mathcal{C}$ is the minimum Hamming distance, or just minimum distance, which measures how good it is at error-correcting. This is defined to be the smallest of the distances between any two distinct codewords:

$$d_{\min} = \min\{d_H(\mathbf{a}, \mathbf{b}) | \mathbf{a}, \mathbf{b} \in \mathcal{C}, \mathbf{a} \neq \mathbf{b}\}$$

The PAR of a code $\mathcal{C}$ is defined as

$$\mathrm{PAR}_{\max} = \max\{\mathrm{PAR}(\mathbf{a}) | \forall \mathbf{a} \in \mathcal{C}\}$$

Finally, the code rate of a code $\mathcal{C}$ is defined as

$$R = \frac{\log_2 \#(\mathcal{C})}{N \log_2 M}$$

where $\#(S)$ denotes the number of elements of set S. The following notation will be used where necessary. An $[n, k, d_{\min}, \eta]$ code is a code of length n, containing k information symbols, with minimum distance $d_{\min}$ and $\mathrm{PAR}_{\max} \eta$.

5.2.2. Golay Complementary Sequences and Reed–Muller Codes. One of the earliest low-PAR code constructions was that proposed by Jones et al. [13], with parameters [4,3,2,1.75]. This was based on a table of low-PAR codewords. Several authors noted that low-PAR codes can be constructed using complementary sequences. A pair of sequences $\mathbf{a}$ and $\mathbf{b}$ is complementary if

$$\rho_{\mathbf{a}}(k) + \rho_{\mathbf{b}}(k) = 2\delta(k) \qquad k = 0, 1, \ldots, N-1$$

Taking the Fourier transform of this, we have

$$P_{\mathbf{a}}(t) + P_{\mathbf{b}}(t) = 2.$$

Hence it is clear the PAR of $\mathbf{a}$ and $\mathbf{b}$ must be less than or equal to 2. Recently, Davis and Jedwab made a significant breakthrough by identifying the relationship between complementary sequences and Reed–Muller codes [14]. Further important results have been found [7,15]. The rth-order Reed–Muller code $\mathrm{RM}(r, m)$ has length $n = 2^m$, minimum Hamming distance $d = 2^{m-r}$, and the number of information bits $k = \sum_{i=0}^{r} \binom{m}{i}$. For example, RM(1,5) is (32,6,16), a low rate linear code.

5.2.3. Constructing Single Sequences with Low PAR. Although OFDM requires the transmission of a large family of sequences with low PAR, it is helpful to also consider the special case where the family size is 1. These single sequences can form the kernel of larger families of sequences with low PAR. The periodic autocorrelation (PA) of a length N sequence, $\mathbf{a}$, is given by

$$\rho_p(k) = \sum_{n=0}^{N-1} a_{n+k} a_n, \quad \text{for } k = 0, 1, \ldots, N-1$$

where all indices are taken, mod N. We can upper bound the PAR of the N-point DFT of $\mathbf{a}$ using,

$$\mathrm{PAR}_p(\mathbf{a}) \leq 1 + \frac{2}{N} \sum_{k=1}^{(N-1)/2} |\rho_p(k)|, \; N \text{ odd} \tag{39}$$

and a similar expression for N even.

Similarly, the negaperiodic autocorrelation (NA) of a length N sequence, $\mathbf{a}$, is given by

$$\rho_n(k) = \sum_{n=0}^{N-1} (-1)^{\left\lfloor \frac{n+k}{N} \right\rfloor} a_{n+k} a_n, \quad \text{for } k = 0, 1, \ldots, N-1$$

where all indices are taken, mod N. We can upper bound the PAR of the N-point negaperiodic DFT of $\mathbf{a}$ using

$$\mathrm{PAR}_n(\mathbf{a}) \leq 1 + \frac{2}{N} \sum_{k=1}^{(N-1)/2} |\rho_n(k)|, \; N \text{ odd} \tag{40}$$

and a similar expression for N even.

Figure 6 shows the continuous power spectrum of an m-sequence, where the N periodic DFT points are bounded by the PA, and interleaved with the N negaperiodic DFT points that are bounded by the NA. There are numerous constructions for sequences with low PA in the literature (e.g., m sequences, trace sequences, Legendre sequences,

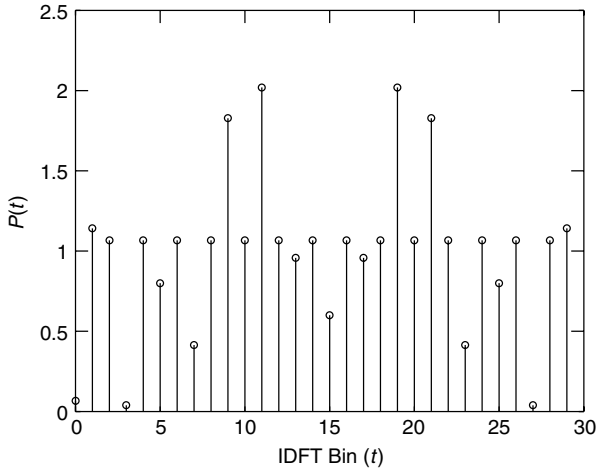


Figure 6. 30-point IDFT power spectrum for length $N = 15$ binary m-seq: 001001101011110.

and cyclotomic constructions). For instance, a binary m sequence guarantees $\rho_p(k) = -1, \forall k, k \neq 0$. Substituting back into (39) gives an upper bound on PAR_p (**a**) of $\dfrac{2N-1}{N}$, which is not particularly tight as the true PAR_p of a binary m sequence is $\dfrac{N+1}{N}$. Sequences with low PA are often proposed for code-division multiple access, but the spectral power peak of a sequence in between N IDFT points can rise considerably. Figure 6 illustrates this fact for a length 15 m sequence which has a low N-point IDFT, but not such a low $2N$-point IDFT. The periodic and negaperiodic N-point IDFTs are bins 0, 2, 4, ..., 26, 28, and bins 1, 3, ..., 27, 29, respectively, of Fig. 6.

Unlike the periodic case, the NA of a sequence has not been studied in such great detail. This is partly because there is a certain overlap between construction techniques for sequences with low PA and low NA, respectively. For instance, for BPSK, if **a** has odd length and low PA, then $\mathbf{a} \oplus 010101\cdots$ has equally low NA. However, the constructions are distinct for the even-length case. The APA and associated PAR upper bound were given in Eqs. (9) and (10). The APA can be viewed as the sum of PA and NA as follows:

$$\rho(k) = \frac{\rho_p(k) + \rho_n(k)}{2}, 0 \leq k < N,$$

$$\rho(k) = \frac{\rho_p(N-k) - \rho_n(N-k)}{2}, -N < k < 0$$

It follows that if a sequence has low PA *and* low NA, then it has low APA and a low upper bound on PAR. Finding constructions that have both low PA and low NA is a difficult problem.

Sequences with low APA are often parameterized by their merit factor (MF), where

$$\mathrm{MF}(\mathbf{a}) = \frac{N^2}{2\sum_{k=1}^{N-1} |\rho(k)|^2} \tag{41}$$

and high MF implies low APA. Finding sequences with highest MF is closely related to the APA problem, and has been studied by a few authors [16,17]. The BPSK sequence with highest known MF is of length 13 and has an MF of 14.08. This sequence has $\max_{k>0} |\rho(k)| = \max_{k>0} |\rho_p(k)| = \max_{k>0} |\rho_n(k)| = 1$. In words, the optimum periodic and negaperiodic properties of **a** guarantee its aperiodic optimality. Although BPSK sequences have been found with MF $\simeq$ 9.0 up to length 117, these sequences are the result of optimized computer search. *Very few* infinite constructions for high MF sequences are known. The best-known are the offset Legendre, twin prime, and (modified) Jacobi constructions. These constructions generate sequences with optimally low PA, moderately low NA, and with MF $\to$ 6.0 as $N \to \infty$, and this is the highest asymptote known for BPSK. A more recent Legendre-type infinite construction has generated sequences with optimally low NA, moderately low PA, and again with MF $\to$ 6.0 as $N \to \infty$. In contrast, both the m sequence and length 2^m complementary sequence have an asymptotic MF of 3.0 [17]. Finally we note that if an infinite construction could be found such that $|\rho_p(k)| \leq c_p, |\rho_n(k)| \leq c_n$, where c_p, c_n are constants, and $k > 0$, then, as $N \to \infty$, the construction would have asymptotically infinite MF. The existence of such a sequence construction is extremely unlikely.

5.2.4. Golay–Davis–Jedwab Codes. Many of the early code proposals were comprehensively generalized by Davis and Jedwab when they proposed an infinite family of binary Golay complementary sequences with parameters $[2^m, \log_2(m!) + m, 2^{m-2}, 2.0]$, and defined them as certain Reed–Muller (RM) RM(2, m) cosets of RM(1, m). We call this family *DJ*, where *DJ* comprises codewords, $c(\mathbf{x})$, with algebraic normal form

$$c(\mathbf{x}) = \sum_{i=0}^{m-2} x_{\pi(i)} x_{\pi(i+1)} \oplus \sum_{i=0}^{m-1} g_i x_i \oplus h\mathbf{1}$$

where $g_i, h \in (0, 1)$, $\mathbf{1}$ is the all-ones vector, π is a permutation of the integers $\{0, 1, \ldots, m-1\}$ and the x_i are binary variables representing length 2^m binary sequences such that element t of x_i is $\left\lfloor \dfrac{t}{2^i} \right\rfloor$, mod 2. The DJ construction was also generalized to higher alphabets and to PARs that are a multiple of 2. The construction is optimal for low N. For instance, for $m = 3$ and binary sequences we can construct an optimal $[8, \log_2(48), 2, 2.0]$ DJ code. There are only 16 more sequences with PAR $\leq$ 2.0, which are not included in the DJ set, and the inclusion of any of these sequences would reduce d from 2 to 1. Unfortunately the rate, k/N, of the DJ construction vanishes rapidly for $N > 32$. Therefore the DJ construction is practically useful only in the context of OFDM for systems requiring no more than 32 subcarriers. It remains an open problem to discover low PAR error-correcting code constructions for $N > 32$ with an acceptable rate. Unfortunately, many OFDM systems require anything from 8 to 8192 subcarriers.

5.2.5. Rudin–Shapiro Recursion and Its Generalizations. The DJ construction can be viewed as Rudin–Shapiro recursion [18]. Let length N sequences, $\mathbf{a_i}$ and $\mathbf{b_1}$ satisfy $P_{\mathbf{a}}(t) + P_{\mathbf{b}}(t) \leq Q$ so that both **a** and **b** have PAR $\leq Q$. Then

it can be shown that sequence concatenations $\mathbf{a_{i+1}} = \mathbf{a_i} \mid \mathbf{b_i}$ and $\mathbf{b_{i+1}} = \mathbf{a_i}|\overline{\mathbf{b_i}}$ both have PAR $\leq Q$, where "|" means concatenation, and "$\overline{*}$" means negation of every component of *. When $Q = 2$, the recursive definition, along with swapping the places of $\mathbf{a}$ and $\mathbf{b}$, and varying the position of the negation, gives the complete DJ binary codeset. This is the case when $\mathbf{a_0}, \mathbf{b_0} \in \{1, -1\}$. More generally, $\mathbf{a_0}$ and $\mathbf{b_0}$ can be any length N_1 pair of sequences, where we wish Q to be as close to 2.0 as possible. Then we can construct a Rudin–Shapiro-type code from this 'seed pair' with parameters $[N_1 2^m, \log_2(m!) + m, d_s 2^{m-2}, Q]$, where $d_s = \min(d_H(\mathbf{a_0}, \mathbf{b_0}), d_H(\mathbf{a_0}, \overline{\mathbf{b_0}}))$.

For N_1 small these codes are roughly the same size as the DJ code with marginally worse PAR. Moreover one can form the union of codes constructed this way, where the distance of the combined codeset is dependent on the set of seed pairs, and the PAR is governed by the constituent code with highest Q. However, as Rudin–Shapiro recursion generates a quadratic extension of an arbitrary-degree seed, the rate of the construction still vanishes as N increases, as quadratics constitute a vanishingly small part of the whole space of 2^N sequences. We can devise higher-degree constructions by viewing the parameters of the Rudin–Shapiro construction as arising from orthogonality of the matrix $\mathbf{R} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, where $\mathbf{v_{i+1}} = \begin{pmatrix} \mathbf{a_{i+1}} \\ \mathbf{b_{i+1}} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \mathbf{a_i} \\ \mathbf{b_i} \end{pmatrix}$, where addition is replaced by concatenation (more generally *tensor sum*). Let $[N_i, k_i, d_i, \eta_i]$ be the code parameters at iteration i of the recursion. Let r and c be the number of rows and columns of the associated matrix, $\mathbf{R}$, respectively, and δ be the number of rows and columns of the associated matrix, $\mathbf{R}$, respectively, and δ be the Hamming distance between rows of $\mathbf{R}$. Then, using matrix $\mathbf{R}$ along with permutation and negation to define the recursion, we get

$$\begin{aligned} N_{i+1} &= cN_i & k_{i+1} &= c\tbinom{r}{c}k_i \\ d_{i+1} &= \min(\delta(N_i - d_i) + (c-\delta)d_i, \delta d_i \\ &\quad + (c-\delta)(N - d_i)) & \eta_{i+1} &= f(\eta_i, i) \end{aligned}$$

where $N_0 = T, k_0 = r, d_0 = d_T, \eta_0 = \eta_T$, and where f is some function determined by the closeness to orthogonality of the matrix $\mathbf{R}$. If $\mathbf{R}$ is orthogonal then $f(\eta_i, i) = \eta_i$. The rate can be improved (as is the case with the DJ code) at the price of distance by including all tensor permutations of the construction. It is possible to increase the rate still further, at the price of distance, by including further permutations of v_i in between iterations. Thus there are a vast number of currently unexplored recursive constructions based on orthogonal and near-orthogonal matrices of various dimensions. Near-orthogonal constructions are of particular interest as they provide a moderate increase in PAR while also providing a large number of rows to maintain rate as N increases.

5.3. Other Techniques

The tone reservation approach has been proposed for the reduction of PAR of OFDM signals by Tellado and Cioffi [19]. In this method, both the transmitter and receiver agree on reserving a small subset of tones for generating PAR reduction signals. The transmitter does not send data on these reserved tones. The complex baseband signal may now be represented as

$$s(t) = \sum_{k \in I_{\text{info}}}^{N-1} c_k e^{j2\pi k \Delta f t} + \sum_{k \in I_{\text{tones}}}^{N-1} b_k e^{j2\pi k \Delta f t}, 0 \leq t \leq T \qquad (42)$$

where I_{info} and I_{tones} are two disjoint tone-index sets such that

$$I_{\text{info}} \cup I_{\text{tones}} = \{0, 1, \ldots, N-1\}$$

The values $b_k, k \in I_{\text{tones}}$, will be called PAR reduction tones (PRTs). The redundancy for this method is

$$R = \frac{|I_{\text{tones}}|}{N} \qquad (43)$$

Parallel combinatory OFDM [20] reduces the PAR without reducing the bandwidth efficiency and without increasing the bit error probability. An OFDM system with N subcarriers using q-PSK can transmit q^N different OFDM symbols. PC-OFDM is based on expanding the q-PSK signal constellation with one extra, zero-amplitude, point. With this expanded signal constellation, the number of different OFDM symbols increases to $(q+1)^N$. A subset of these modified OFDM symbols may be chosen with lower PAR. The authors show that this method will have at least the same bandwidth efficiency, and lower bit error probability, when compared to the original OFDM system.

Another PAR reduction method, and the simplest, is to deliberately clip the OFDM signal before amplification. In particular, since the large peaks occur with very low probability, clipping could be an effective technique for PAR reduction. However, clipping is a nonlinear process and may cause significant in-band distortion, which increases the bit error rate, and out-of-band noise, which reduces the spectral efficiency. Filtering after clipping can reduce the spectral splatter but may also cause some peak regrowth. Peak regrowth can however be reduced by oversampling the OFDM signal and clipping [21,22].

Henkel and Wagner [23] develop a trellis shaping technique for PAR reduction. In this method, a valid code sequence of a convolutional code is added (modulo 2) to a data sequence. The code sequence is chosen to reduce the PAR. If H is the parity-check matrix of the convolutional code, then for a valid code sequence y we have $yH^T = 0$. This property can be used to eliminate the added code sequence at the receiver side. However, it is necessary to precode the information sequence with the left inverse $(H^T)^{-1}$.

BIOGRAPHIES

C. Tellambura received his B.Sc. degree with honors from the University of Moratuwa, Sri Lanka, in 1986, his M.Sc. in electronics from the King's College, UK, in 1988, and his Ph.D. in electrical engineering from the University of Victoria, Canada, in 1993. He was a postdoctoral research fellow with the University of Victoria and the University of Bradford. Currently, he is a senior lecturer at Monash University, Australia. He is an editor for the *IEEE Transactions on Communications* and the *IEEE Journal on Selected Areas in Communications* (Wireless

Communications Series). His research interests include coding, communications theory, modulation, equalization, and wireless communications.

Matthew G. Parker received a B.Sc. in electrical and electronic engineering in 1982 from University of Manchester Institute of Science and Technology, U.K. and, in 1995, a Ph.D. in residue and polynomial residue number systems from University of Huddersfield, U.K. From 1995 to 1998 he was a postdoctoral researcher in the Telecommunications Research Group at the University of Bradford, U.K., researching into coding for peak factor reduction in OFDM systems. Since 1998 he has been working as a postdoctoral researcher with the Coding and Cryptography Group at the University of Bergen, Norway. He has published on residue number systems, number-theoretic transforms, complementary sequences, sequence design, quantum entanglement, coding for peak power reduction, factor graphs, linear cryptanalysis, and VLSI implementation of Modular arithmetic.

BIBLIOGRAPHY

1. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* 5–14 (1990).
2. P. Shelswell, The COFDM modulation system: the heart of digital audio broadcasting, *Electron. Commun. Eng. J.* 127–136 (June 1995).
3. G. K. H. Sari and I. Jeanclaude, Transmission techniques for digital terrestrial TV broadcasting, *IEEE Commun. Mag.* **33**: 100–109 (Feb. 1995).
4. S. B. Weinstein and P. M. Ebert, Data transmission by frequency division multiplexing using the discrete Fourier transform, *IEEE Trans. Commun.* **19**: 628–634 (Oct. 1971).
5. C. Reiners and H. Rohling, *Multicarrier transmission technique in cellular mobile communications systems*, *Proc. IEEE Vehicular Technology Conf.* IEEE, 1994, pp. 1645–1649.
6. M. R. Schroeder, Synthesis of low-peak-factor signals and binary sequences with low autocorrelation, *IEEE Trans. Inform. Theory* **IT-13**: 85–89 (1970).
7. K. G. Paterson, Generalized Reed-Muller codes and power control in OFDM modulation, *IEEE Trans. Inform. Theory* **46**: 104–120 (Jan. 2000).
8. C. Tellambura, Upper bound on the peak factor of N-multiple carriers, *IEE Electron. Lett.* **33**: 1608–1609 (Sept. 1997).
9. P. Van Eetvelt, G. Wade, and M. Tomlinson, Peak to average power reduction for OFDM schemes by selective scrambling, *IEE Electron. Lett.* **32**: 1963–1964 (Oct. 1996).
10. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed., Academic Press, 1994.
11. R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping, *IEE Electron. Lett.* **32**: 2056–2057 (Oct. 1996).
12. A. D. S. Jayalath and C. Tellambura, Reducing the peak-to-average power ratio of an OFDM signal through bit or symbol interleaving, *IEE Electron. Lett.* **36**: 1161–1163 (June 2000).
13. A. E. Jones, T. A. Wilkinson, and S. K. Barton, Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes, *IEE Electron. Lett.* **30**: 2098–2099 (1994).
14. J. A. Davis and J. Jedwab, Peak-to-mean power control in OFDM, Golay complementary sequences, and Reed-Muller codes, *IEEE Trans. Inform. Theory* **45**: 2397–2417 (Nov. 1999).
15. K. G. Paterson and V. Tarokh, On the existence and construction of good codes with low peak-to-average power ratio, *IEEE Trans. Inform. Theory* **46**: 1974–1987 (Sept. 2000).
16. M. J. E. Golay, A new search for skewsymmetric binary sequences with optimal merit factors, *IEEE Trans. Inform. Theory* **36**: 1163–1166 (Sept. 1990).
17. T. Høholdt, *Difference Sets, Sequences and Their Correlation Properties*, vol. 542 of *Series C: Mathematical and Physical Sciences*, Kluwer, 1999.
18. T. Høholdt, H. E. Jensen, and J. Justesen, Aperiodic correlations and the merit factor of a class of binary sequences, *IEEE Trans. Inform. Theory* **31**: 549–552 (July 1985).
19. J. Tellado, *Multicarrier Modulation with Low PAR*, Kluwer, 2000.
20. P. K. Frenger and N. A. B. Sevensson, Parallel combinatory OFDM signalling, *IEEE Trans. Commun.* **47**: 558–567 (April 1999).
21. X. Li and L. J. Cimini, Jr., Effects of clipping and filtering on the performance of OFDM, *IEEE Commun. Lett.* **2**(5): 131–133 (1998).
22. H. Ochiai and H. Imai, Performance of the deliberate clipping with adaptive symbol selection for strictly band-limited OFDM systems, *IEEE J. Select. Areas Commun.* **18**: 2270–2277 (Nov. 2000).
23. W. Henkel and B. Wagner, Another application of trellis shaping: PAR reduction for DMT (OFDM), *IEEE Trans. Commun.* **48**: 1471–1476 (Sept. 2000).

PERMUTATION CODES

EMANUELE VITERBO
Politechnico di Torino
Torino (Turin), Italy

1. INTRODUCTION

Permutation codes were proposed by David Slepian in 1965 [12]. In the quest for efficient codes for the band-limited Gaussian channel, permutation codes are among the first attempts to solve the problem taking into account both coding gains and decoding efficiency. Permutation codes are multidimensional spherical signal constellations with the desirable property that they possess a very simple maximum-likelihood (ML) decoding algorithm. Slepian used the term *permutation modulation* for his permutation codes.

A *variant I* permutation modulation is the set of codewords obtained by taking all permutations of an initial vector in the n-dimensional Euclidean space. A *variant II* permutation modulation is the set of codewords obtained by taking all permutations and sign changes of the components of an initial vector in the n-dimensional Euclidean space. Trivial examples of variant I and variant II modulations are orthogonal and biorthogonal codes,

respectively. Also pulse code modulation (PCM), pulse position modulation (PPM), and simplex codes can be viewed as permutation modulations.

Good permutation modulations may be designed by appropriately selecting the initial vector. Permutation modulations may be very efficiently decoded by essentially applying a sorting algorithm to the received signal vector. Permutation modulations are very special cases of group codes for the band-limited channel proposed by Slepian [10,13].

Karlof later proposed the term *permutation code* for a generalization of permutation modulations [6]. In particular, he studied the group codes obtained from subgroups of the symmetric group (the group of permutations of n objects). The resulting spherical codes may be seen as particular subsets of the corresponding permutation modulation. We can think of permutation modulation as a "full rate" permutation code. In this case the initial vector selection problem and the decoding algorithm are much harder [6–8].

Here, we will focus on permutation codes according to Slepian's definition. This article is organized as follows. The next section introduces the notation and gives detailed definition of permutation modulation. Performance in terms of error probability and the ML decoding algorithm are also presented. Section 3 gives some examples of permutation codes and discusses some application issues.

2. THEORY

Digital transmission over the band-limited additive white Gaussian noise (AWGN) channel is commonly modeled in the n-dimensional Euclidean space $\mathbf{R}^n$ as $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{y}$ is the received signal vector, $\mathbf{x}$ is the transmitted signal vector (or codeword) taken from a finite signal constellation (or codebook) $\mathcal{S}$, and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is a real Gaussian random vector with i.i.d. (independent, identically distributed) components. The space dimension n is related to the time–bandwidth product through the sampling theorem, namely, if T is the signal duration and W the occupied bandwidth, then $n = 2WT$.

Letting $M = |\mathcal{S}|$ be the number of points in the constellation, we define the spectral efficiency as

$$\frac{R}{W} = \frac{2}{n}\log_2 M \text{ bps/Hz} \tag{1}$$

Let $r = \log_2 M$. If we are interested in transmitting binary information, we simply label $2^{[r]}$ codewords by distinct binary vectors of length $[r]$ and disregard the remaining $M - 2^{[r]}$ codewords.[1]

The average signal power of $\mathcal{S} = \{\mathbf{x}_i\}_0^{M-1}$ is given by

$$\mathcal{P} = \frac{1}{nM}\sum_{i=0}^{M-1}\|\mathbf{x}_i\|^2 \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^n$.

The maximum-likelihood (ML) receiver gives an estimate $\hat{\mathbf{x}}$ of the transmitted codeword $\mathbf{x}$ according to the minimum distance criterion

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}_i \in \mathcal{S}} \|\mathbf{y} - \mathbf{x}_i\|^2 \tag{3}$$

[1] Where $[x]$ denotes the greatest integer smaller than x.

We note that the complexity of the ML receiver depends greatly on the structure of the code $\mathcal{S}$. In the worst case a total of M Euclidean distances must be computed. For large values of M this may be impractical; hence it is common to trade some of the performance for a reduced decoding complexity. Many classical forward error-correcting (FEC) codes have been selected for applications because they have simple decoding algorithms.

The average codeword error probability with ML detection is given by

$$P(e) = \frac{1}{M}\sum_{i=0}^{M-1} P(e \mid \mathbf{x}_i) = \frac{1}{M}\sum_{i=0}^{M-1}\int_{\overline{\mathcal{R}_i}} \frac{e^{-\|\mathbf{x}-\mathbf{x}_i\|^2/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}}\,d\mathbf{x} \tag{4}$$

where $\overline{\mathcal{R}_i} = \mathbf{R}^n \backslash \mathcal{R}_i$ is the complement of the decision region corresponding to the codeword $\mathbf{x}_i$, defined as

$$\mathcal{R}_i = \{\mathbf{z} \in \mathbf{R}^n : \|\mathbf{z} - \mathbf{x}_i\| < \|\mathbf{z} - \mathbf{x}_j\|, \quad \forall j \neq i\} \tag{5}$$

These regions are also known as *minimum-distance* or ML regions.

Good codes should be designed in order to minimize $P(e)$, given the parameters M, n, $\mathcal{P}$, and σ. Shannon showed that the codeword error probability decreases exponentially with n and gave the famous asymptotic result that, for given R, W, and $\mathcal{P}/\sigma^2$, the $P(e)$ can be made arbitrarily small as $n \to \infty$, provided that $R < C$, where

$$C = W\log_2\left(1 + \frac{\mathcal{P}}{\sigma^2}\right) \tag{6}$$

On the contrary, if $R > C$, then $P(e) \to 1$ as $n \to \infty$.

Explicit construction of optimal codes is an open problem and has been analyzed in some very special cases only [1,15].

2.1. Definitions

Let $\{\mu_1, \ldots, \mu_k\}$ be a set of distinct real numbers with $\mu_1 < \mu_2 < \cdots < \mu_k$, and let $\{m_1, \ldots, m_k\}$ be a set of positive integers such that

$$n = \sum_{j=1}^{k} m_j \tag{7}$$

Consider the initial vector with components sorted in ascending order:

$$\mathbf{x}_0 = (\underbrace{\mu_1, \ldots, \mu_1}_{m_1}, \underbrace{\mu_2, \ldots, \mu_2}_{m_2}, \ldots, \underbrace{\mu_k, \ldots, \mu_k}_{m_k}) \tag{8}$$

A variant I permutation code consists of the set of vectors obtained by permuting the components of the initial vector $\mathbf{x}_0$. The total number of codewords in such a code is

$$M_I = \frac{n!}{m_1!m_2!\cdots m_k!} \tag{9}$$

The variant I code with $k = 2$, $m_1 = n - 1$, $m_2 = 1$, and $\mu_1 = 0$ is the well-known PPM or orthogonal modulation.

A variant II permutation code consists of the set of vectors obtained by permuting and applying all possible sign changes to the components of the initial vector $\mathbf{x}_0$. Without loss of generality, we may assume

$$0 \leq \mu_1 < \mu_2 < \cdots < \mu_k \tag{10}$$

The total number of codewords in this code is

$$M_{\mathrm{II}} = \frac{2^h n!}{m_1! m_2! \cdots m_k!} \tag{11}$$

where $h = n - m_1$, if $\mu_1 = 0$ and $h = n$, if $\mu_1 > 0$.

The variant II code with $k = 2$, $m_1 = n - 1$, $m_2 = 1$, and $\mu_1 = 0$ results in the well-known biorthogonal modulation. The variant II code with $k = 1$, $m_1 = n$, and $\mu_1 \neq 0$ yields an n-bit PCM. In this case the points of $\mathcal{S}$ correspond to the 2^n vertices of an n-dimensional hypercube of edge length $2\mu_1$.

It is clear that all codewords of both variant I and II codes lie on a hypersphere of radius $\sqrt{n\mathcal{P}}$ centered at the origin and

$$\mathcal{P} = \frac{1}{n} \sum_{j=1}^{k} m_j \mu_j^2 \tag{12}$$

2.2. Decoding

Let us consider ML decoding of variant I codes. We need to find the minimum of the quantities

$$\|\mathbf{y} - \mathbf{x}_i\|^2 = \|\mathbf{y}\|^2 + \|\mathbf{x}_i\|^2 - 2(\mathbf{y}, \mathbf{x}_i) = \|\mathbf{y}\|^2 + n\mathcal{P} - 2(\mathbf{y}, \mathbf{x}_i), \quad i = 0, \ldots, M-1 \tag{13}$$

where $(\mathbf{y}, \mathbf{x}_i)$ denotes the scalar product of the two vectors. Since $\|\mathbf{y}\|$ is independent of i, the ML decoder may simply maximize the scalar product between the received vector and the codewords:

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathcal{S}} \sum_{k=1}^{n} x_k y_k \tag{14}$$

This maximization problem may be solved as follows. Given the received vector $\mathbf{y}$, replace the smallest m_1 components by the values μ_1, replace the smallest m_2 remaining components with μ_2, and so on until all the components have been replaced.

It is interesting to show the very simple and elegant proof of the optimality of this decoding algorithm given by Slepian. In particular, we want to show that the sum

$$x_{i_1} y_1 + x_{i_2} y_2 + \cdots + x_{i_n} y_n \tag{15}$$

is maximized by the permutation of indices $(i_1, i_2, \ldots, i_n)$, which pairs the largest x to the largest y, the second largest x to the second largest y, and so forth.

For $n = 1$ it is trivially true; then we proceed by induction. For some $n > 1$, let $\overline{x}$ and $\overline{y}$ denote the largest x and the largest y. If $\overline{x}$ is not paired with $\overline{y}$ in (15), then the sum contains the two terms $\overline{x}y' + \overline{y}x'$ for some $x' \leq \overline{x}$ and $y' \leq \overline{y}$. If we swap x' with $\overline{x}$, the sum (15) will decrease; in fact

$$(\overline{x}\overline{y} + x'y') - (\overline{x}y' + \overline{y}x') = (\overline{x} - x')(\overline{y} - y') \geq 0 \tag{16}$$

Hence, pairing $\overline{x}$ with $\overline{y}$ maximizes the sum (15). We now delete $\overline{x}\overline{y}$ from the sum and proceed by induction on the $n - 1$ terms; pairing the second largest x to the second largest y does not reduce the $n - 1$ term sum, and so on.

For variant II codes ML decoding can be performed as follows:

1. Take the absolute value of the components of the received vector $\mathbf{y}$; that is, let
$$\mathbf{y}' = (|y_1|, |y_2|, \ldots, |y_n|)$$
2. Apply the decoder of variant I codes to $\mathbf{y}'$ to make a first decision $\mathbf{x}'$.
3. The final decision is given by
$$\hat{\mathbf{x}} = (\mathrm{sgn}(y_1)x'_1, \mathrm{sgn}(y_2)x'_2, \ldots, \mathrm{sgn}(y_n)x'_n)$$
where $\mathrm{sgn}(x) = +1$, if $x \geq 0$ and $\mathrm{sgn}(x) = -1$, if $x < 0$.

It can be shown that this algorithm is equivalent to solving the maximization problem (14).

The complexity of these decoding algorithms is rather small if compared to the brute-force exhaustive search. In particular, it is enough to perform a sorting algorithm on the n components of the received vector and to keep track of the final index permutation. This permutation uniquely identifies the ML decoded codeword, and the corresponding information bit label may be easily recovered. Sorting can be performed with a complexity of $O(n \log(n))$, whereas exhaustive decoding requires Mn multiplications and $M(n-1)$ additions.

2.3. Decision Regions and Error Probability

Evaluation of the average codeword error probability [4] for permutation codes can be simplified by the following argument. Consider the collection $\mathcal{C}$ of all $n \times n$ permutation matrices, that is, matrices having a single entry equal to one in each row and column and zeros in the remaining positions. When a permutation matrix is applied to the vectors of the codebook of a variant I code, it simply maps the codebook back into itself. In $\mathcal{C}$ we can find a permutation matrix A_{ij} that maps any codeword $\mathbf{x}_i$ into the codeword $\mathbf{x}_j$.

The permutation matrices are also orthogonal matrices so that, when they operate on $\mathcal{S}$, they preserve the distances between the points. Since the decision regions are defined in (5) in terms of distances, the permutation matrix A_{ij} also sends $\mathcal{R}_i$ into $\mathcal{R}_j$. Thus all the decision regions of a variant I code are congruent, and (4) reduces to $P(e) = P(e \mid \mathbf{x}_i)$, which is independent of i.

For variant II codes, a similar argument also enables us to conclude that $P(e) = P(e \mid \mathbf{x}_i)$ is independent of i. In particular, it is enough to replace the collection $\mathcal{C}$ by the collection $\mathcal{O}$ of $2^n n!$, $n \times n$ matrices having a single nonzero entry equal to $+1$ or -1 in each row and column.

We note that this simplification is similar to the one that can be used in evaluating the codeword error probability of linear codes, where it is convenient to consider the case where the all-zero codeword is transmitted. For permutation codes we will focus on $P(e \mid \mathbf{x}_0)$.

Let us now consider in detail the average codeword error probability of variant I codes. First observe that the received vector components have independent Gaussian distributions. The first m_1 components have mean μ_1 and variance σ^2, the next m_2 components have mean μ_2 and variance σ^2, and so on.

Assume that the codeword $\mathbf{x}_0$ was transmitted. To understand when a decoding error appears, let us split the received vector components into k runs of length m_j each:

$$\mathbf{y} = (y_1^{(1)}, y_2^{(1)}, \ldots, y_{m_1}^{(1)}, y_1^{(2)}, y_2^{(2)}, \ldots, y_{m_2}^{(2)}, \ldots, y_1^{(k)}, y_2^{(k)}, \ldots, y_{m_k}^{(k)}) \tag{17}$$

The correct decision will be made if the first m_1 components are smaller than the following m_2 components, the next m_2 components are smaller than the following m_3 components, and so forth. Then we can write

$$P_{\mathrm{I}}(e) = P_{\mathrm{I}}(\mu_1, \mu_2, \ldots, \mu_k) = 1 - P\{\eta_1 \le \xi_2 \le \eta_2 \le \xi_3 \le \cdots \le \eta_{k-1} \le \xi_k\} \tag{18}$$

where, for $j = 1, \ldots, k$

$$\xi_j = \min(y_1^{(j)}, y_2^{(j)}, \ldots, y_{m_j}^{(j)})$$

$$\eta_j = \max(y_1^{(j)}, y_2^{(j)}, \ldots, y_{m_j}^{(j)})$$

and the n independent Gaussian random variables $y_i^{(j)}$ have mean μ_j and variance σ^2.

Note that $P_{\mathrm{I}}(e)$ depends only on the differences of the μ values; thus, for all δ

$$P_{\mathrm{I}}(\mu_1 + \delta, \mu_2 + \delta, \ldots, \mu_k + \delta) = P_{\mathrm{I}}(\mu_1, \mu_2, \ldots, \mu_k) \tag{20}$$

Let $\beta_1 = \mu_1$ and $\beta_i = \mu_i - \mu_{i-1}$, for $i = 2, \ldots, k$. Let

$$\phi(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)} \tag{21}$$

be the probability distribution function of a zero-mean Gaussian random variable with variance σ^2 and

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)\, dz \tag{22}$$

the corresponding cumulative distribution function.

Slepian [12] has shown that for both variants I and II it is possible to bound $P(e)$ as

$$B - \frac{B^2}{2} \le P(e) \le B \tag{23}$$

For variant I codes, we obtain

$$B = B_{\mathrm{I}} = \sum_{i=2}^{k} P_{m_i, m_{i-1}}(\beta_i) \tag{24}$$

and

$$P_{m,n}(\alpha) = P_{n,m}(\alpha) = n \int_{-\infty}^{\infty} \phi(z)[1 - \Phi(z)]^{n-1} \times [1 - \Phi^m(z + \alpha)]\, dz \tag{25}$$

For variant II, when $\mu_1 = 0$, we have

$$B = B_{\mathrm{II}} = \overline{P}_{m_1, m_2}(\beta_i) + \sum_{i=3}^{k} P_{m_i, m_{i-1}}(\beta_i) \tag{26}$$

where

$$\overline{P}_{m,n}(\alpha) = 2m \int_{0}^{\infty} \phi(z)[1 - 2\Phi(-z)]^{m-1} \times \{1 - [1 - \Phi(z - \alpha)]^n\}\, dz \tag{27}$$

and when $\mu_1 \neq 0$

$$B = B_{\mathrm{II}'} = 1 - [1 - \Phi(-\beta_1)]^{m_1} + \sum_{i=2}^{k} P_{m_i, m_{i-1}}(\beta_i) \tag{28}$$

3. EVALUATION

We are now able to consider the problem of selecting good permutation codes for a fixed dimension n. We want to optimize the choice of k, the μ values, and the m values. From (1), (9), and (11) we see that R is independent of the μ values and depends only on k and the m values.

It is natural to fix k and the m values (i.e., fix R) and choose the μ values to minimize $\mathcal{P}$ for some fixed value of $P(e)$.

For variant I codes, the first optimization step to reduce the average signal power is to center the signal constellation $\mathcal{S}$ around its barycenter, by selecting

$$\sum_{j=1}^{k} m_j \mu_j = 0 \tag{29}$$

By imposing this condition to PPM, we obtain the simplex modulation. For variant II codes, $\mathcal{S}$ is already centered around its barycenter.

The optimization problem has been solved numerically [12], and some optimal codes are presented for various n, m values, and k for two values of $P(e) = 10^{-3}$ and $P(e) = 10^{-5}$. A simplified version of the code optimization problem was solved analytically by Biglieri and Elia [2] and independently by Ingemarson [9]. They selected the μ values in order to maximize the minimum Euclidean distance $d_{\min}$, among the points of $\mathcal{S}$, for a fixed average power $\mathcal{P}$. Here, we report the optimal codes for $P(e) = 10^{-5}$ in Table 1.

Figures 1–4 show the codeword error probability of the codes in Table 1 as a function of E_b/N_0, where $E_b = n\mathcal{P}/r$ and $N_0 = 2\sigma^2$. These figures may be interpreted as follows. Given the system constraints T (maximum acceptable

Table 1. Optimized Codes at $P(e) = 10^{-5}$ Found by Slepian [12]

No.	n	m Values	μ Values	$[r]$	γ_{dB}
1	5	I (4,1)	−1.2908, 5.1631	2	1.6
2	5	II (4,1)	0, 6.6558	3	2.2
3	5	II (3,2)	0, 6.7720	5	1.2
4	5	II (2,3)	0, 6.7720	6	0.2
5	5	II (4,1)	4.6540, 11.4152	7	−1.4
6	5	II (2,2,1)	0, 6.7748, 13.3491	7	−2.0
7	5	II (1,3,1)	0, 6.6712, 13.4039	8	−2.3
8	10	I (9,1)	−0.6691, 6.0220	3	2.7
9	10	II (9,1)	0, 6.8907	4	3.3
10	10	II (7,3)	0, 7.1240	9	2.2
11	10	II (4,5,1)	0, 7.1747, 14.1284	16	−0.6
12	10	II (3,5,2)	0, 7.1293, 14.1828	18	−1.6
13	10	II (3,3,3,1)	0, 7.0814, 14.0512, 21.0215	21	−3.7
14	10	II (2,4,2,2)	0, 7.0672, 14.0365, 20.9950	22	−4.4
15	7	II (6,1)	0, 6.7720	3	2.8
16	7	II (5,2)	0, 6.9169	6	2.0
17	25	II (14,11)	0, 7.6409	33	1.8
18	50	II (25,22,3)	0, 8.0046, 15.7001	83	0.6
19	50	II (20,25,5)	0, 7.9966, 15.7951	92	0.0
20	51	II (19,24,6,2)	0, 8.0255, 15.8067, 23.2291	105	−1.5
21	50	II (17,23,8,2)	0, 7.9926, 15.8064, 23.3280	108	−1.7
22	7	II (4,3)	0, 6.9694	8	1.3
23	31	II (23,8)	0, 7.6873	30	2.9

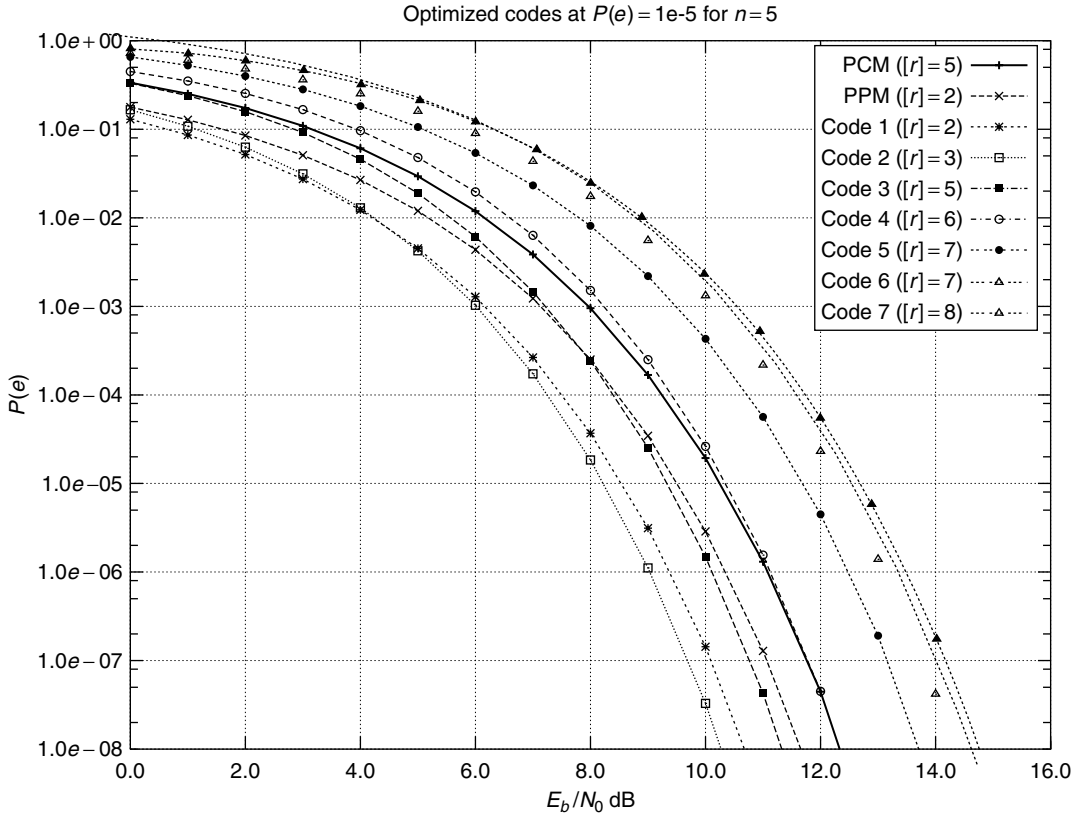


Figure 1. Optimized codes at $P(e) = 10^{-5}$ for $n = 5$.

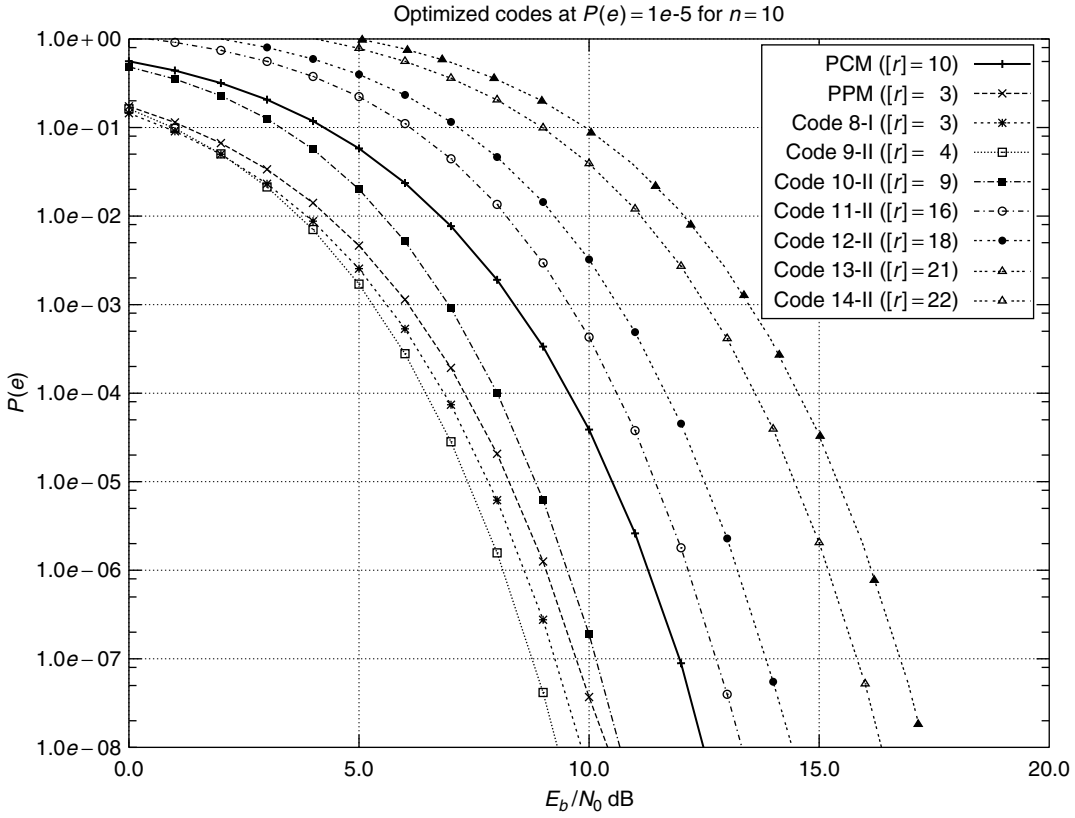


Figure 2. Optimized codes at $P(e) = 10^{-5}$ for $n = 10$.

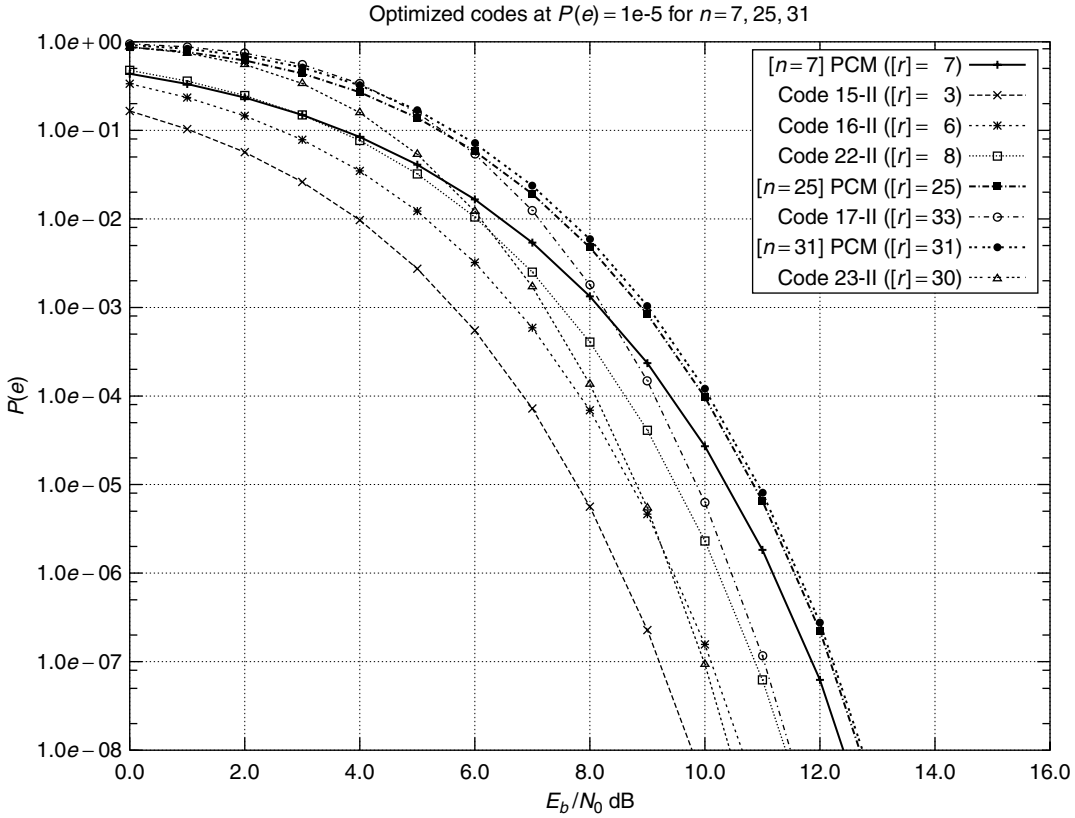


Figure 3. Optimized codes at $P(e) = 10^{-5}$ for $n = 7, 25, 31$.

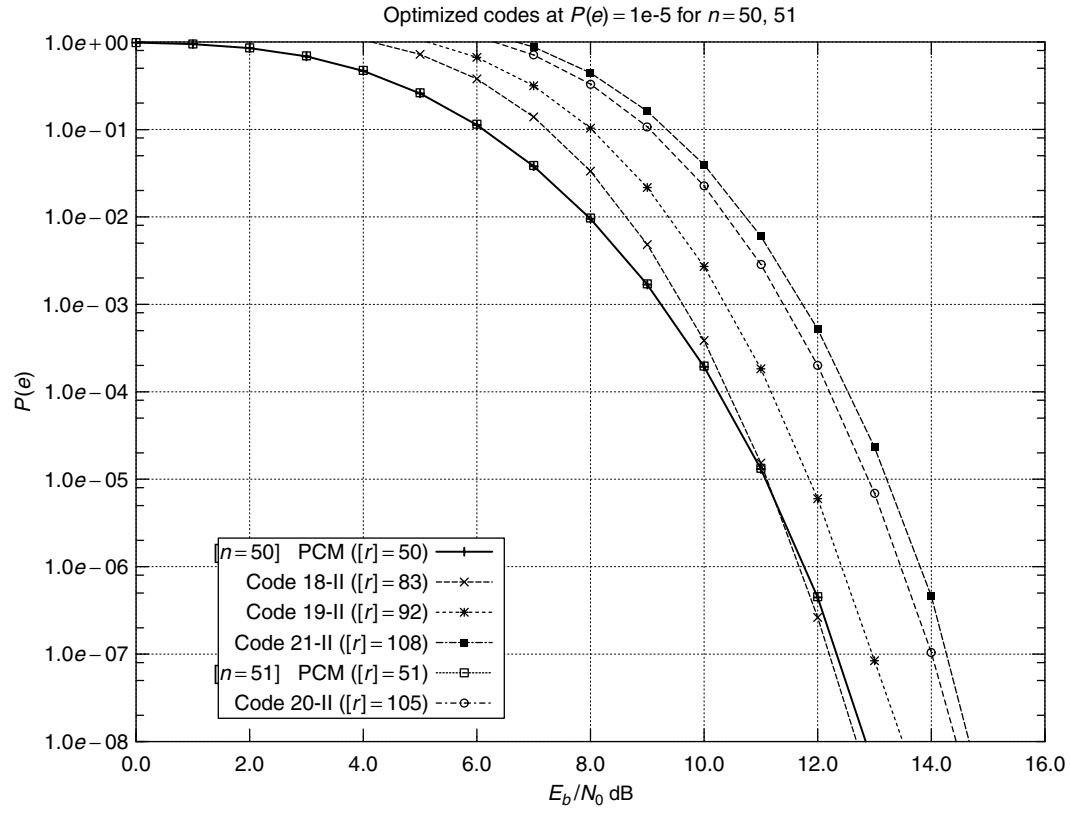


Figure 4. Optimized codes at $P(e) = 10^{-5}$ for $n = 50, 51$.

delay) and W (available bandwidth), we have n; then we can chose among the codes of different rate the one that satisfies our packet error rate $P(e)$ requirement.

For example, code 3 for $n = 5$ enables us to transmit at the same rate of the 5-bit PCM with a lower $P(e)$, corresponding to an asymptotic gain of 1.2 dB (see Fig. 1). Code 10 enables us to transmit at 0.9 the rate of PCM with an asymptotic gain of 2.2 dB (see Fig. 2). Code 23 enables us to transmit at almost at the same rate of PCM with an asymptotic gain of 2.9 dB (see Fig. 3).

Figure 5 shows the performance of the optimal codes given by Slepian [12] at $P(e) = 10^{-3}$. Comparison with Fig. 1 shows that their performance in almost identical in both cases, so we may conclude that the optimization is quite insensitive to the value $P(e)$.

Given an n-dimensional code, we define its asymptotic coding gain with respect to an n bit PCM as

$$\gamma = \frac{d^2_{\min}/E_b}{d^2_{\min,\mathrm{PCM}}/E_{b,\mathrm{PCM}}} \tag{30}$$

We report γ in decibels in the last column of Table 1. These asymptotic values can be approximately verified in all figures at $P(e) = 10^{-8}$, with an accuracy of about 0.5 dB.

We conclude with a few comments on the possible application of permutation codes. In order to implement the transmitter, we need to define an orthonormal basis $\{\psi_j(t)\}_1^n$ of the signal space so that each transmitted signal is given by

$$x(t) = \sum_{j=1}^{n} x_j \psi_j(t) \tag{31}$$

The basis functions must be approximately time- and band-limited. For example, we can select the strictly bandlimited functions

$$\psi_j(t) = \frac{\sin(2\pi Wt - j\pi)}{2\pi Wt - j\pi} \quad j = 1, \ldots, n \tag{32}$$

or the strictly time-limited functions for $0 \le t \le T$

$$\psi_{2j-1}(t) = \sin\left(\frac{2\pi jt}{T}\right)$$
$$\psi_{2j}(t) = \cos\left(\frac{2\pi jt}{T}\right) \quad j = 1, \ldots, \frac{n}{2} \tag{33}$$

Other choices are also possible [e.g., 3–5,11]. The receiver can be implemented using a bank of matched filters, matched to the basis functions in order to obtain the components of the received vector $\mathbf{y}$.

Although permutation codes have a very long history and some very promising coding gains, to the author's knowledge, they have never been used in applications. Nevertheless, we can be expect that they may be exploited in the future for high-speed transmission, due to their very simple decoding algorithm.

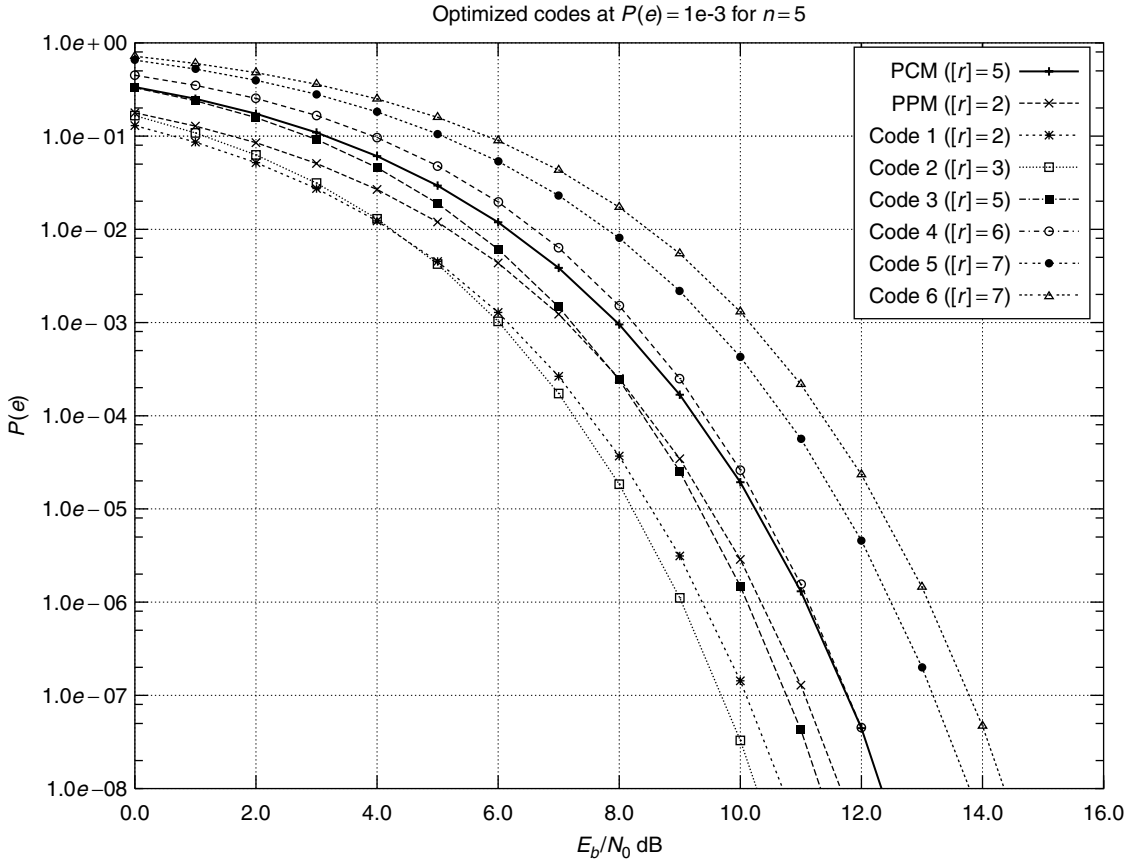


Figure 5. Optimized codes at $P(e) = 10^{-3}$ for $n = 5$.

BIOGRAPHY

Emanuele Viterbo was born in Torino (Turin), Italy, in 1966. He received his baccalaureate degree (Laurea) in Electrical Engineering in 1989 and his Ph.D. in 1995 in Electrical Engineering, both from the Politecnico of Torino, Torino, Italy. From 1990 to 1992 he was with the European Patent Office, The Hague, The Netherlands, as a patent examiner in the field of dynamic recording and in particular in the field of error-control coding. Between 1995 and 1997 he held a postdoctoral position in the Dipartimento di Elettronica of the Politecnico di Torino in Communications Techniques over Fading Channels. Between 1997 and 1998 he was Visiting Researcher in the Information Sciences Research Center of AT&T Research, Florham Park, New Jersey. Since 1998, he has been Assistant Professor at Politecnico di Torino, Dipartimento di Elettronica. Dr. Emanuele Viterbo was awarded a NATO Advanced Fellowship in 1997 from the Italian National Research Council. His current research interests are in lattice codes for the Gaussian and fading channels, algebraic coding theory, digital terrestrial television broadcasting, and digital magnetic recording.

BIBLIOGRAPHY

1. A. V. Balakrishnan, A contribution to the sphere-packing problem of communication theory, *J. Math. Anal. Appl.* **3**: 485–506 (Dec. 1961).
2. E. Biglieri and M. Elia, Optimum permutation modulation codes and their asymptotic performance, *IEEE Trans. Inform. Theory* 751–753 (Nov. 1976).
3. M. Elia, G. Taricco, and E. Viterbo, Optimal energy transfer over bandlimited communication channels, *IEEE Trans. Inform. Theory* **45**(6): 2020–2029 (Sept. 1999).
4. H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty II, *Bell Syst. Tech. J.* **40**: 65–84 (1961).
5. H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty III, *Bell Syst. Tech. J.* **41**: 1295–1336 (1962).
6. J. K. Karlof, Permutation codes for the Gaussian channel, *IEEE Trans. Inform. Theory* **35**(4): 726–732 (July 1989).
7. J. K. Karlof, Decoding spherical codes for the Gaussian channel, *IEEE Trans. Inform. Theory* **39**(1): 60–65 (Jan. 1993).
8. J. K. Karlof and Y. O. Chang, Optimal permutation codes for the Gaussian channel, *IEEE Trans. Inform. Theory* **35**(4): 726–732 (July 1989).
9. I. Ingemarsson, Optimized permutation modulation, *IEEE Trans. Inform. Theory* **36**(5): 1098–1100 (Sept. 1990).
10. D. Slepian, Some further theory of group codes, *Bell Syst. Tech. J.* 1219–1252 (Sept. 1960).
11. D. Slepian and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty I, *Bell Syst. Tech. J.* **40**: 43–64 (1961).
12. D. Slepian, Permutation modulation, *Proc. IEEE* 228–236 (March 1965).
13. D. Slepian, Group codes for the Gaussian channel, *Bell Syst. Tech. J.* **47**(4): 575–602 (April 1968).
14. D. Slepian, On neighbor distances and symmetry in group codes, *IEEE Trans. Inform. Theory* 630–632 (Sept. 1971).
15. M. Steiner, The strong simplex conjecture is false, *IEEE Trans. Inform. Theory* 721–731 (May 1994).

PHOTONIC ANALOG-TO-DIGITAL CONVERTERS

BARRY L. SHOOP
United States Military Academy
West Point, New York

1. INTRODUCTION

Analog-to-digital (A/D) conversion is the process by which an analog signal that is continuous in both time and amplitude is converted to a digital signal that is discrete in time and amplitude. The A/D converter is an important component in any electronic or photonic system that senses the natural environment and processes, stores, displays, or communicates the information using digital techniques. Since the vast majority of signals encountered in nature are analog and the preferred method of processing, storing, and transmitting signals is digital, this interface is generally considered to be the most critical and challenging part of the overall signal acquisition and processing system. Recent advances in both electronics and telecommunication markets as well as continued improvements in sensor resolution has renewed interest in the pursuit of high-speed, high-resolution A/D converters and has focused attention on photonic techniques to provide improvements in this technology area.

In general, the process of A/D conversion can be characterized by the four distinct functional blocks shown in Fig. 1. The analog input signal $x(t)$ is first band-limited to the range $0 \le f_x \le f_B$ (Hz) by a lowpass analog filter with cutoff frequency f_B to protect against aliasing that could occur during the subsequent sampling operation. The sampling operation in a conventional Nyquist rate A/D converter is chosen to satisfy the minimum Nyquist criterion: $f_S = f_N = 2f_B$, where f_S is the sampling frequency, f_N is the Nyquist frequency, and f_B is the constrained signal bandwidth. There are also other alternatives to sampling frequency depending on the specific application. Subsampling below the Nyquist rate is acceptable if the input signal is known to be periodic. Oversampling is another alternative in which $f_S \gg f_N$ and signal processing techniques are subsequently used to achieve improved amplitude resolution through a technique called *spectral noise shaping*. The output from the sampler is $x_n \equiv x(nT_S)$, where T_S is the uniform sampling period $T_S = 1/f_S$. Scalar quantization then maps each continuous-amplitude input x_n to one value in a discrete-amplitude ensemble $q_n \equiv q(nT_S)$. On the basis of the results of this mapping, the digital processor generates a digital codeword that most closely approximates the input analog signal value. The output $y_n \equiv y(nT_S)$ is then the multibit, digital word representing the input analog input value.

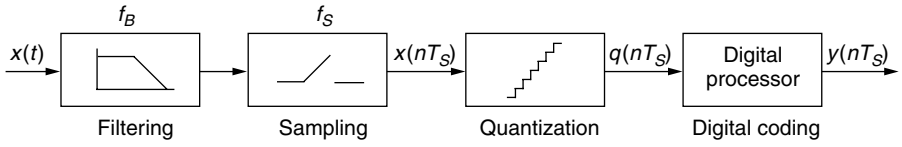


Figure 1. Generic analog-to-digital converter block diagram.

A complete coverage of the subject of photonic A/D conversion can be found in the 2001 book *Photonic Analog-to-Digital Conversion* [1]. For additional background and information on the general subject of A/D conversion, Refs. 2–5 provide an excellent overview.

1.1. Motivation for Photon-Based A/D Conversion

Photonic approaches to A/D conversion provide some distinct performance advantages over their electronic counterparts. Nonlinear optoelectronic devices have demonstrated ultrahigh switching speeds that can be applied to the quantization functionality required in the A/D converter. Mode-locked lasers provide high-speed and accurate optical pulses in the range of 0.5 ps–50 fs that can be used for precision clocks and sampling. These optical clocks provide the capability for ultrafast sampling with extremely low clock jitter. Another advantage specifically associated with optical sampling techniques is the decoupling of the sampled and sampling signals, achieved when the sampling signal is optical and the sampled signal is electronic. Mach–Zehnder interferometers that can be used for optical sampling and modulation have been demonstrated with modulation bandwidths of up to 120 GHz. Photonic techniques also bring the potential for utilization of the full two-dimensional (2D) nature of optics. Many other applications have successfully converted from temporal approaches to spatial approaches and, applying the 2D nature of optical processing, extended the performance of the specific application. Many current approaches to photonic A/D conversion are leveraging this higher dimensionality in an effort to extend converter performance bounds. Many photonic approaches to A/D conversion also produce output digital codes that are Gray codes directly, eliminating the need for additional hardware to produce these coding schemes.

2. HISTORICAL PERSPECTIVE

The application of photonic techniques to the problem of A/D conversion began with optical sampling. In 1970, Steigman and Kuizenga [6] first proposed the use of a mode-locked laser for optical sampling of an electronic signal. Later, in 1975, Taylor [7] applied optical sampling using mode-locked laser pulses to the first optical A/D converter. This optical A/D converter integrated Mach–Zehnder interferometers, avalanche photodetectors, and electronic comparators. In the early 1980s this basic approach was further developed by a group at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory, demonstrating a 4-bit electrooptic A/D converter operating at a sampling frequency of 1 GHz [8]. Jalali and Xie extended this electrooptic A/D converter using a folding architecture [9].

In the early 1990s, Shoop and Goodman proposed the first optical approach to oversampling A/D conversion [10]. This work resulted in a proof-of-concept experimental demonstration of an optoelectronic oversampled A/D converter based on multiple quantum well modulators [11]. Later this work was extended to low-resolution photonic A/D conversion for digital image halftoning based on spatial oversampling and error diffusion using smart pixel technology [12]. A newer approach to photonic A/D has been developed that integrates spatial oversampling, an error diffusion neural algorithm, and spectral noise shaping for high-resolution A/D conversion applications [13,14]. Pace has investigated an alternate approach to photonic oversampled A/D conversion based on a fiber lattice accumulator [15,16].

In the late 1990s, there was renewed interest in photonic A/D conversion, particularly for high-speed A/D applications. The majority of these architectures relied on channelization or interleaving techniques that partition the wide-bandwidth input into N-parallel channels, each operating at $1/N$ of the original sampling rate. This approach allows integration of high-speed optical sampling and lower-speed conventional electronic quantization, taking advantage of each technology's strengths. Twichell and colleagues at MIT Lincoln Laboratory have applied phase-encoded optical sampling and a time-interleaving architecture to wideband photonic A/D conversion, demonstrating a 505-MS/s (megasample/second) system providing 8 bits of resolution [17,18]. Clark and colleagues at the Naval Research Laboratory have investigated time- and wavelength-interleaving for photonic A/D conversion [19,20]. Another interesting approach that falls within the context of a channelized architecture is based on time-stretch preprocessing using an optically dispersive element [21].

Since the mid-1970s, other approaches to photonic A/D converters have also been investigated. Most were novel approaches that were ultimately limited by speed, complexity, or resolution. The interested reader can find details of several of these other approaches to optical A/D conversion in the literature [22–29].

3. REPRESENTATIVE PHOTONIC A/D CONVERTER ARCHITECTURES

There are a variety of different approaches to photonic A/D conversion differing in the fundamental architectures as well as the photonic components used. The following examples of photonic A/D converters are not intended to be inclusive but rather provide an introduction to several representative approaches to the application of photonic architectures and devices to the problem of A/D conversion.

3.1. Electrooptic A/D Conversion

Probably the best known optical A/D conversion technique to date was developed by Taylor in 1975 [7]. He was the

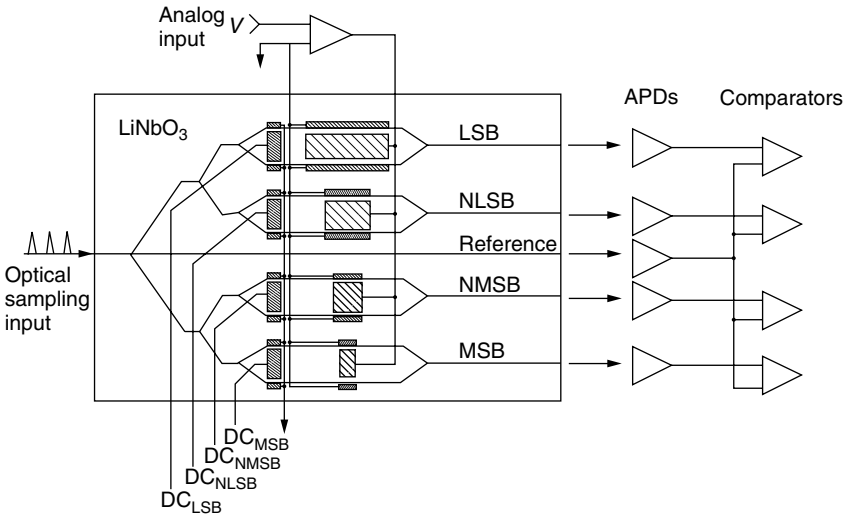


Figure 2. Schematic diagram of a 4-bit electrooptic A/D converter. (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 5.1, p. 124. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

first to recognize the relationship between the periodicity of the output of an interferometric electrooptic modulator with applied voltage and the periodic variation of a binary representation of an analog quantity. A 4-bit implementation of this concept is shown in Fig. 2 [8].

The basic optical element used in this architecture is a planar waveguide version of a Mach–Zehnder interferometric modulator. The interferometer consists of an electrooptic crystal containing a single-mode input optical waveguide that branches at a "Y" to split the optical power into two equal components. The light in the two paths then travels an equal distance before recombining at the second Y and exiting the crystal. The input analog voltage is applied to one arm of the interferometer through coplanar electrodes. In the absence of an applied electric field, the light from the two paths recombines in phase and produces a maximum in the output intensity. With an electric field applied to the electrode, the phase velocity of the light propagating in that arm is changed as a result of the linear electrooptic effect. The output intensity of a single interferometer can be shown to vary as

$$I = I_0 \cos^2\left(\frac{\phi}{2} + \frac{\psi}{2}\right) \tag{1}$$

where ψ is the static phase difference between the two paths and ϕ is the net electrooptic phase difference between the light propagating in the two guides

$$\phi = 2\pi L\left(\frac{\Delta n}{\lambda}\right) = kLV \tag{2}$$

Here, Δn is the refractive index change, V is the applied voltage, L is the modulator length, and k is a constant that depends on the electrooptic parameters of the crystal, the electrode spacing, and optical wavelength. The use of a three-electrode configuration, one between the two waveguides and two outside the waveguides produces an opposite phase shift in the two waveguides and therefore doubles the magnitude of the parameter k. An important parameter of a Mach–Zehnder interferometer is the voltage, which yields a phase shift of $\phi = \pi$

$$V_\pi = \frac{\pi}{kL} \tag{3}$$

In Fig. 2, the analog input signal V is applied in parallel to one arm of each of the four modulators, one for each bit of resolution. The sampling of the analog signal is performed optically, using a series of short optical pulses derived from a pulsed laser source. The optical output from each waveguide modulator is detected by an avalanche photodiode (APD) which converts the optical signal to an electronic signal and also provides amplification. The electronic signal from each modulator is then compared to a reference signal, obtained from the common light source. The output of each comparator is either a binary "1" or "0", depending on whether the modulator output intensity is greater or less than $I_0/2$, respectively. The output of the top modulator represents the least significant bit (LSB) in the digital word, and that of the bottom modulator is the most significant bit (MSB). The output intensity, the threshold, and the corresponding binary representation for each modulator are shown in Fig. 3. The Gray-code representation in Fig. 3 is achieved by controlling the static phase difference in each modulator by applying the appropriate DC biases, labeled as DC_{LSB}, DC_{NLSB}, DC_{NMSB}, and DC_{MSB} in Fig. 2.

In this Gray-code approach, the voltage quantization step size V_Q is equal to one-half the value of V_π for the LSB channel, or

$$V_Q = \frac{\pi}{2kL_N} \tag{4}$$

for an N-bit comparator where L_N is the electrode length for the LSB channel. For each subsequent significant bit, the value of V_π increases by a factor of 2 and therefore the electrode length L_n decreases by a factor of 2. Therefore

$$L_n = 2^{n-N} L_N \tag{5}$$

where $n = 1$ corresponds to the MSB and $n = N$ corresponds to the LSB. The maximum input analog voltage

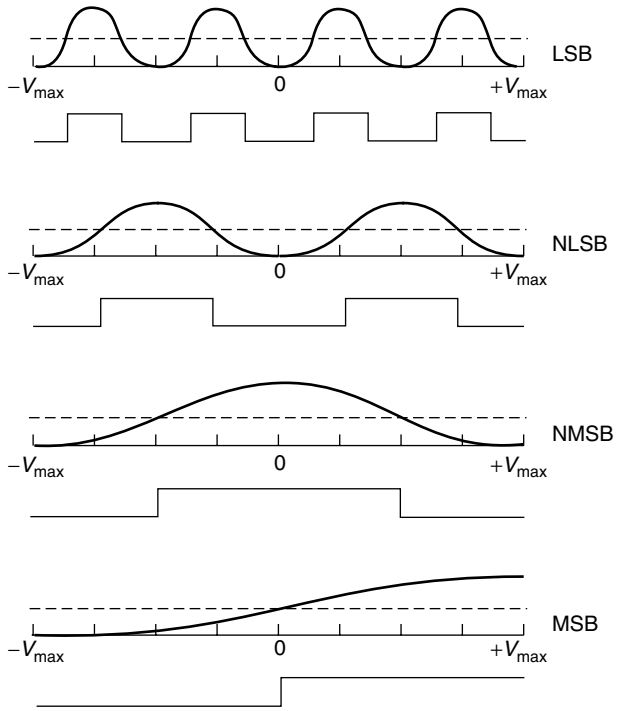


Figure 3. Intensity versus voltage for a 4-bit electro-optic A/D converter with a Gray-code output (the dashed-lines represent the threshold level in each of the four channels). (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 5.2, p. 125. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

that can be converted in this architecture is

$$V_{\max} = 2^{N-1} V_Q \tag{6}$$

This electrooptic A/D converter provides several distinct advantages. This photonic A/D converter is linear in complexity, requiring one additional Mach–Zehnder interferometer for each additional bit of resolution. Another important advantage is the decoupling of the analog sampled signal from the optical sampling signal. This eliminates the distortion effects common to diode bridge sampling circuits, which tend to couple the sampling signal into the converter circuitry. A limitation of this type of converter is that each additional bit of resolution requires a doubling of the electrode length of the LSB modulator. In $LiNbO_3$, this produces a transit-time limitation on performance of approximately 6 bits at 1 GHz. Other electrooptic crystals exist that produce larger refractive index changes and therefore could improve the performance of this transit-time limit. However, most of these crystals also have larger loss mechanisms and therefore would produce loss-limited performance instead.

3.1.1. An Optical Folding-Flash A/D Converter. Jalali and Xie [9] proposed an extension to the electrooptic A/D converter based on a folding architecture. This approach eliminates the geometrical scaling of V_π with amplitude resolution by incorporating an analog encoding technique. Figure 4 is a block diagram of a 4-bit optical folding-flash A/D converter. In contrast to the previous electrooptic A/D converter, the electrode lengths of all the Mach–Zehnder interferometers are identical. An analog encoding scheme is used in this approach to bias all these individual Mach–Zehnder interferometers at different points on the interferometer transfer characteristic.

The optical folding-flash A/D converter provides a solution to the V_π and subsequent electrode length scaling problem of the original electrooptic A/D converter; however, it also introduces some alternate challenges. In this approach, the MSB is susceptible to noise for extreme values of the input. Furthermore, this approach relies on the ability to accurately bias each interferometer in the architecture at several different points on the sinusoidal interferometer transfer characteristic. Hardware complexity is another limitation in this approach. The optical folding-flash A/D converter architecture is exponential in complexity with the number of interferometers

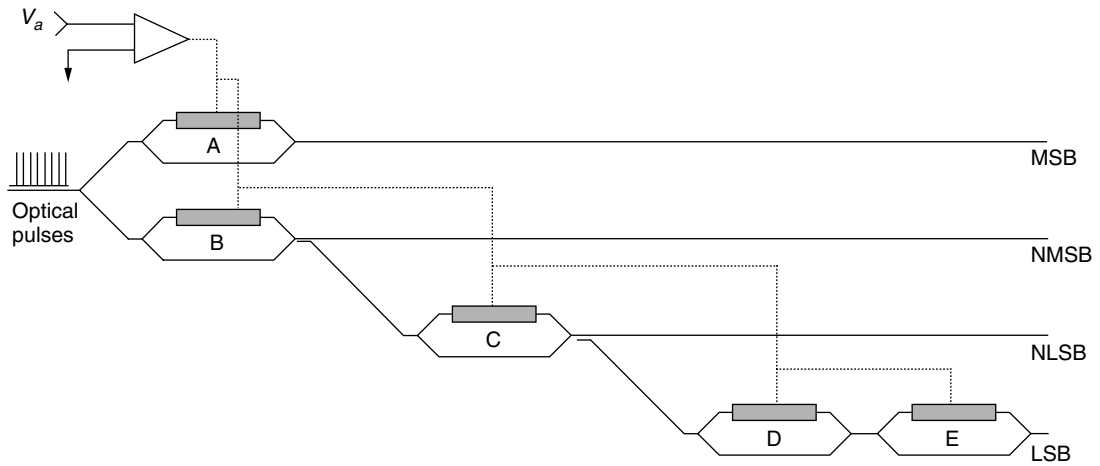


Figure 4. Block diagram of the optical folding-flash A/D converter. (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 5.3, p. 127. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

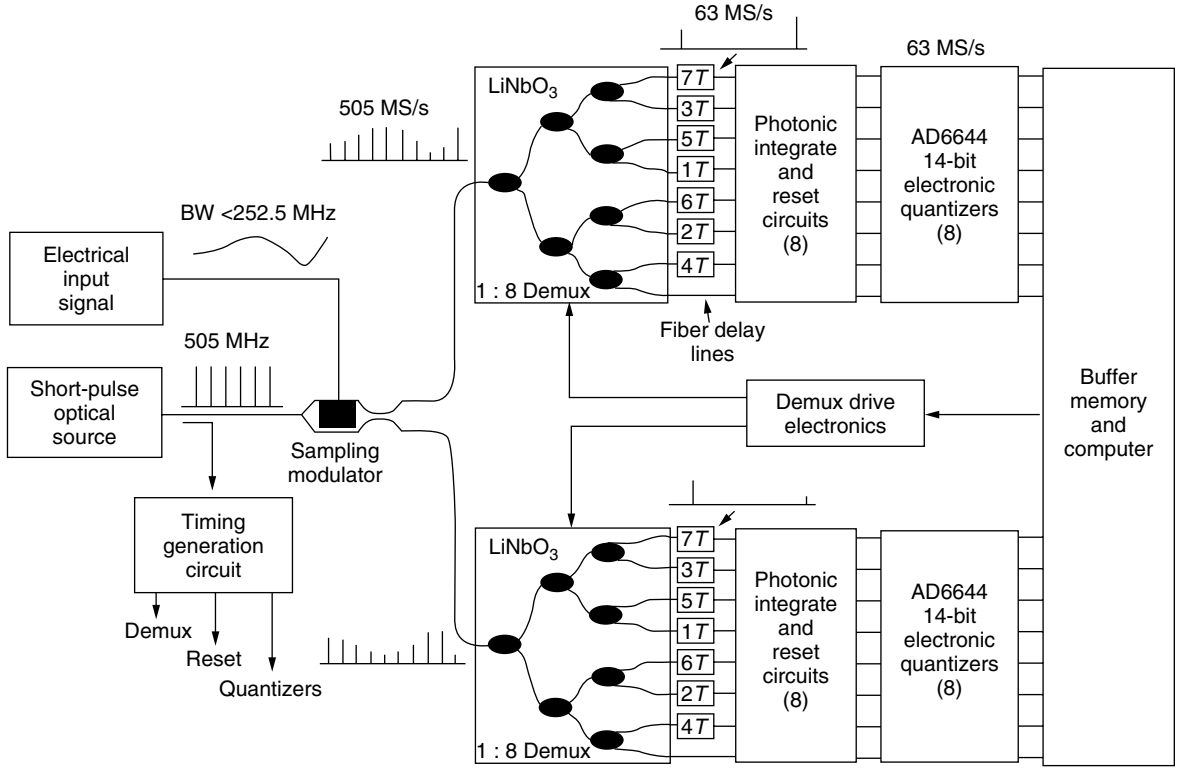


Figure 5. Block diagram of 505-MS/s optically sampled A/D converter.

scaling as $2^{(b-2)} + 1$, where b is the number of bits in the desired resolution.

3.2. Channelized Photonic A/D Conversion

Channelized or interleaved A/D converter architectures partition the input sampled signal into N-parallel channels, each of which operates at $1/N$ of the original sampling rate. In this approach, high-speed optical sampling can be integrated with N-electronic A/D converters in a common architecture that extends the bandwidth performance of the overall photonic A/D converter beyond that of the individual electronic A/D converters.

Mach–Zehnder interferometers can be used for wideband electrooptic modulation and as electrooptic switches for optical demultiplexing. One popular approach to achieving optical time interleaving is to use a tree-structured Mach–Zehnder switching architecture in which an array of Mach–Zehnder interferometric switches is used to provide the demultiplexing functionality. Using 1×2 electrooptic switches, each subsequent stage of the optical time-division demultiplexer is driven by a clock with a switching frequency that is one-half that of the previous stage. Unlike classic Mach–Zehnder interferometers that have a single output, the 1×2 electrooptic switch has a two-channel output. By modulating the refractive index of the interferometer, the output can be switched between the two output channels. Each output channel of the photonic demultiplexer can then be followed by a photodetector stage and an electronic quantizer stage to complete the time-interleaved A/D architecture.

An 8-channel time-interleaved photonic A/D converter operating at 505 MS/s has been demonstrated that employs a wideband electrooptic modulator and 1×2 electrooptic switches to accomplish the 1 : 8 optical demultiplexing [30]. Figure 5 is block diagram of this system [18]. Because the range over which linear modulation occurs in a Mach–Zehnder interferometer is relatively small, this architecture uses a new phase encoding technique [31] to produce linear modulation over a much wider dynamic range. This new approach combines the complimentary outputs from a dual-output electrooptic modulator to improve the linearity of the output.

In the architecture in Fig. 5, the optical source produces 25-ps pulses at the 505-MHz sampling rate that are used to sample the analog electrical input signal. Phase-encoded optical sampling is performed by the dual-output $LiNbO_3$ Mach–Zehnder modulator. The complimentary output of the sampling modulator is then fed to a pair of $LiNbO_3$ 1 : 8 optical time division demultiplexers to distribute the output pulsestreams to an array of photonic integrate and reset circuits and subsequently to electronic quantizers. In this architecture, the 14-bit electronic quantizers operate at 63 MS/s, which is one-eighth of the sampling rate. Fiber delay lines are used to time-align the pulses at the input to the photonic integrate and reset circuits to simplify timing.

The 1 : 8 optical demultiplexers were Ti-indiffused $LiNbO_3$ and employed a high-extinction design to minimize crosstalk between the parallel channels. Each demultiplexing stage consisted of a 1×2 electrooptic switch with an additional extinction modulator at each

output arm. The extinction for a single stage ranged between 36 and 46 dB. The bandwidth of a stage was 600 MHz and the insertion loss for a channel ranged between 6.8 and 8.4 dB. This photonic A/D architecture demonstrated SNR in the Nyquist bandwidth of 51 dB or 8.2 bits of resolution. The spur-free dynamic range (SFDR) was limited to 61 dB by channel-to-channel mismatch and crosstalk errors [32].

Although this approach holds promise for high-speed and high-resolution photonic A/D conversion, a number of technical challenges remain. Incomplete isolation between the individual demultiplexed channels can lead to channel-to-channel crosstalk, which, in turn, will result in undesirable spectral tones in the output frequency spectrum [33]. Consequently, each switch must be accurately designed and fabricated to ensure that optical crosstalk between channels is minimized. Also, the individual clock signals that control the switching of each electrooptic switch must be accurate and synchronous. Inaccuracies in clock timing between channels produce jitter in the demultiplexer which will also contribute to additional mismatch distortion.

3.3. Oversampling Photonic A/D Conversion

Oversampled A/D converters derive their name from the fact that the sampling rate is routinely chosen to be much higher than that required to satisfy the Nyquist criterion. This type of A/D converter trades bandwidth for improved amplitude resolution. In this type of converter, a low-resolution quantizer is embedded in a feedback architecture in an effort to reduce the quantization noise through spectral noise shaping. Here, a large error associated with a single sample is diffused over many subsequent samples and then linear filtering techniques are applied to remove the spectrally shaped noise, thereby improving the overall performance of the converter. Figure 6 shows a generalized block diagram of an oversampled A/D converter.

The analog signal $x(t)$ is again bandlimited to the range $0 \le f_x \le f_B(Hz)$ by an antialiasing filter and is then sampled at a rate $f_S \gg f_N$, where f_S is the sampling frequency, $f_N = 2f_x$ is the Nyquist frequency of the sampled signal, and $f_B \le f_S/2$ is the constrained signal bandwidth. The output of the sampler is then applied to a modulator that provides coarse amplitude quantization and spectral shaping of the quantization noise. The digital postprocessor, which consists of a digital lowpass filter (LPF) and a decimator, removes the quantization noise that was spectrally shaped by the modulator, provides antialiasing protection, and reduces the rate to the original sampled signal's Nyquist rate by trading word rate for word length.

3.3.1. The Modulator. The function of the modulator is to quantize the analog input signal and reduce the quantization noise within the signal baseband. This is accomplished through the use of a low-resolution quantizer, oversampling, negative feedback, and linear filtering.

One common type of modulator used in photonic oversampled A/D architectures is the recursive error diffusion modulator, shown in Fig. 7. Here, $H(z)$ represents the z transform of a causal, unity-gain filter and z^{-1} is a unit sample delay.

For a first-order modulator in which $H(z) = 1$, the modulator output can be shown to be

$$q(u_n) = \underbrace{x_n}_{\text{signal}} + \underbrace{\varepsilon_n - \varepsilon_{n-1}}_{\text{quantization error}} \tag{7}$$

Here, the quantity ε_n is the quantization error that would be seen at the modulator output *if there were no feedback loop*. However, as a result of the negative feedback, the first-order difference or discrete-time derivative of the error, $\varepsilon_n - \varepsilon_{n-1}$, appears at the output instead. By design, this difference signal is concentrated at high frequencies and can be removed by the digital LPF in the postprocessor.

3.3.2. The Digital Postprocessor. The function of the digital postprocessor is to digitally filter and decimate the output of the modulator so that the quantization noise that was spectrally shaped by the modulator is removed through lowpass filtering and the output digital signal is decimated to the Nyquist rate of the original sampled signal. An ideal LPF with cutoff frequency f_B is generally used to characterize upper-bound performance. Since the ideal LPF can be only approximated in practice, other practical filters have been investigated for this application [34]. With more recent advances in microlithography and silicon VLSI (very-large-scale integration technology), the digital postprocessor has

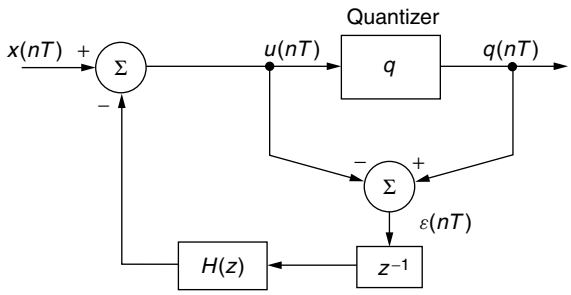


Figure 7. Block diagram of recursive error diffusion modulator.

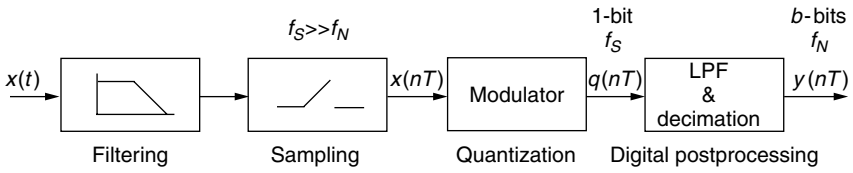


Figure 6. Generalized block diagram of an oversampled A/D converter.

become extremely sophisticated, routinely allowing the realization of filters with orders in excess of 100th-order.

3.3.3. Performance Analysis. Assuming white quantization noise characteristics, an ideal LPF with cutoff frequency f_B, and an Nth-order modulator with a noise shaping characteristic described by

$$H_{\text{ns}}(z) = (1 - z^{-1})^N \tag{8}$$

the spectral distribution of the quantization noise after shaping is the product of the filter shaping function and the spectral density of the quantizer error

$$S_\varepsilon(f) = \underbrace{[|H_{\text{ns}}(z)|^2_{z=e^{j2\pi fT}}]}_{\text{noise shaping}} \cdot \underbrace{\left[\frac{S_q}{f_S}\right]}_{\text{noise power density}} \tag{9}$$

If the quantizer step size is assumed to be Δ and the input signal is sinusoidal with a full-scale input range of $\pm\Delta/2$, the maximum signal-to-quantization noise ratio ($\text{SQNR}_{\max}$) can be computed as

$$\text{SQNR}_{\max}(M, N) = \frac{3}{2} \cdot \left[\frac{2N+1}{\pi^{2N}}\right] \cdot M^{2N+1} \tag{10}$$

where M is the oversampling ratio

$$M = \frac{f_S}{f_N} \tag{11}$$

For comparison, the $\text{SQNR}_{\max}$ of a conventional Nyquist rate uniform quantizer with b-bits resolution can be shown to be

$$\text{SQNR}_{\max}(b) = 3 \cdot 2^{2b-1} \tag{12}$$

Figure 8 shows the theoretical $\text{SQNR}_{\max}(M, N)$ and equivalent resolution for first- through fourth-order oversampled modulators as a function of oversampling ratio. The case of no noise shaping represents the $\text{SQNR}_{\max}$ that can be expected if the same quantizer, embedded in the feedback loop of the oversampled modulator, were simply oversampled and digitally filtered. The slope of this curve is 3 dB per octave and is included only for comparison. The slope of the $N = 1$ curve is 9 dB per octave and that of the $N = 2$ curve is 15 dB per octave, showing the significant advantage achieved by using a noise shaping modulator.

3.3.4. Spectral Characteristics. To better understand the relationship between the spectral shaping of the quantization noise and the input signal, the power spectrum for a fourth-order oversampling modulator output is shown in Fig. 9. This power spectrum assumes a half-scale sinusoidal input and an oversampling ratio of $M = 32$. Here, spectral noise shaping in the low frequencies of the power spectrum improves the A/D converter SNR performance.

3.3.5. The Error Diffusion Neural Network. The concept of spectral noise shaping can be extended to spatial dimensions and can be formulated in the context of a 2D symmetric error diffusion architecture. This extension requires each state variable in Fig. 7 to be represented in matrix–vector notation and can be described in terms of a specialized neural network.

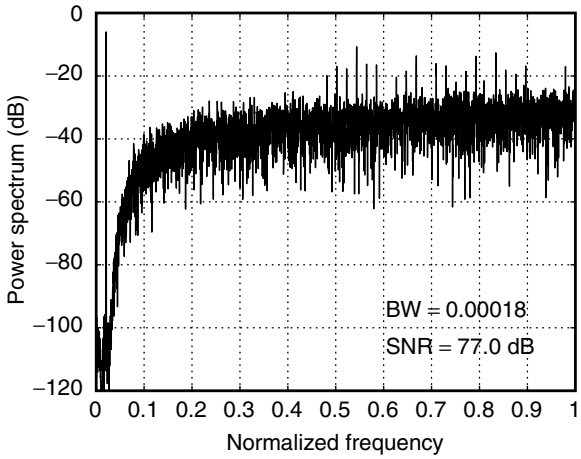


Figure 9. Power spectrum of the output data sequence for a fourth-order oversampling modulator.

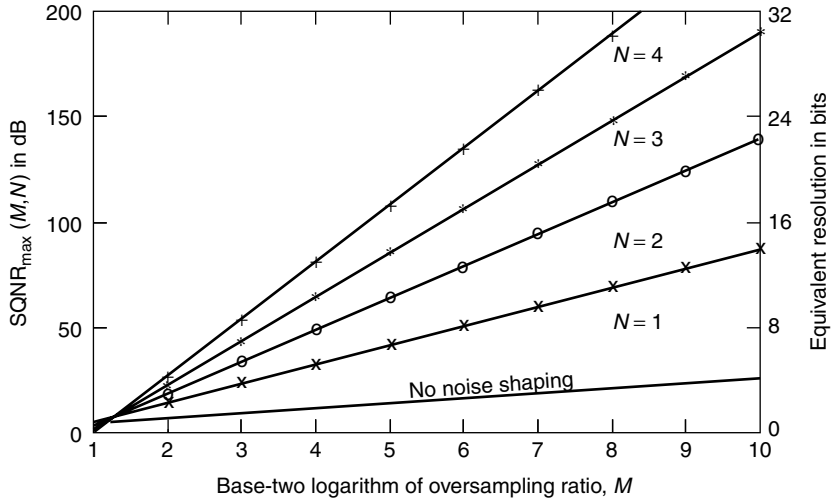


Figure 8. Maximum signal-to-quantization noise ratio of Nth-order modulators.

In equilibrium, the error diffusion neural network satisfies

$$\mathbf{u} = \mathbf{W}(\mathbf{y} - \mathbf{u}) + \mathbf{x} \tag{13}$$

For an $N \times N$ image, $\mathbf{W}$ is an $N^2 \times N^2$ sparse, "circulant without wrapping" matrix derived from the error diffusion filter weights $w_{i,j}$.

An equivalence to the classic Hopfield network can be described by

$$\mathbf{u} = \mathbf{A}(\mathbf{W}\mathbf{y} + \mathbf{x}) \tag{14}$$

where $\mathbf{A} = (\mathbf{I} + \mathbf{W}^{-1})$ and $\mathbf{I}$ is the identity matrix. Effectively, the error diffusion neural network includes a prefiltering of the input image $\mathbf{x}$ by the matrix $\mathbf{A}$ while still filtering the output image $\mathbf{y}$ but now with a new matrix, $\mathbf{AW}$.

The energy function of the error diffusion neural network can be shown to be a quadratic function

$$E(\mathbf{x}, \mathbf{y}) = [\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}]^T[\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}] \tag{15}$$

where $\mathbf{y} \in \{-1, 1\}$ is the output vector of quantized states with one element per pixel and $\mathbf{A} = \mathbf{B}^T\mathbf{B}$. As the error diffusion neural network converges and the energy function is minimized, so, too, is the error between the output and input images.

The error diffusion neural network represents an important class of A/D converter that applies to digital image halftoning. In digital halftoning [35], a continuous-tone input image is converted to a binary output image for purposes of printing, storage, or display.

3.3.6. A Distributed Mesh Feedback Approach to Photonic A/D Conversion. A relatively new approach to photonic A/D conversion uses spatial oversampling techniques, an error diffusion neural network, and a smart pixel [12] hardware implementation. This approach converts a 1D temporal signal to a 2D spatial representation in an effort to leverage the 2D nature of a photonic A/D architecture. In this approach, the input signal is first sampled at a rate higher than that required by the Nyquist criterion and then presented spatially as the input to a 2D error diffusion neural network consisting of $N \times N$ neurons, each representing a pixel in the image space. The neural network processes the input oversampled analog image and produces an $N \times N$ pixel binary or halftoned output image. Decimation and lowpass filtering techniques, common to conventional 1D oversampling A/D converters, digitally sum and average the $N \times N$ pixel output binary image using high-speed digital electronic circuitry. By employing a 2D neural approach to oversampling A/D conversion, each pixel constitutes a simple oversampling modulator, thereby producing a distributed A/D architecture. Spectral noise shaping across the array diffuses the quantization error, thus improving overall SNR performance. The matrix $\mathbf{A}$ in Eq. (14) describes the interconnectivity of the network resulting from local connections through the error diffusion filter. Since $\mathbf{A}$ is full-rank, the network is fully connected. Therefore each quantizer within the $N \times N$ network is embedded in a fully connected, distributed mesh feedback loop that spectrally shapes the overall quantization noise, thereby significantly reducing the effects of component mismatch typically associated with parallel or channelized A/D approaches.

3.3.7. Spectral Characteristics. A representative power spectrum of the output halftoned image of this fully connected distributed mesh feedback architecture is shown in Fig. 10. The input signal used was an $N \times N$ sinusoid with period 2 in both spatial dimension, and the error diffusion filter was a LPF with a 25×25 region of support designed using a Kaiser window with $\alpha = 5$. Approximately 30 dB of low-frequency noise suppression is achieved in this specific approach resulting in approximately a 54-dB SNR. Larger in-band noise suppression has also been achieved with other filter designs.

3.3.8. A Photonic Implementation of the Error Diffusion Neural Network. The functionality necessary to implement the error diffusion neural network consists of a one-bit quantizer, two differencing nodes, and the interconnection and weighting of the error diffusion filter. One photonic implementation of the error diffusion neural algorithm uses smart pixel technology [36]. Smart pixels integrate both electronic processing and individual optical devices on a common chip to take advantage of the complexity of electronic processing circuits and the speed of optical devices. Arrays of these smart pixels bring the advantage of parallelism that optics provides.

The electronic circuitry, used for quantization, weighting, and summation functions, was fabricated in a 0.5 μm complementary metal oxide semiconductor (CMOS) process. In a subsequent processing step, self-electrooptic effect device (SEED) multiple-quantum-well (MQW) modulators are integrated with the CMOS VLSI circuitry using a flip-chip bonding process. The MQW modulators are used in this implementation to provide optical input and output

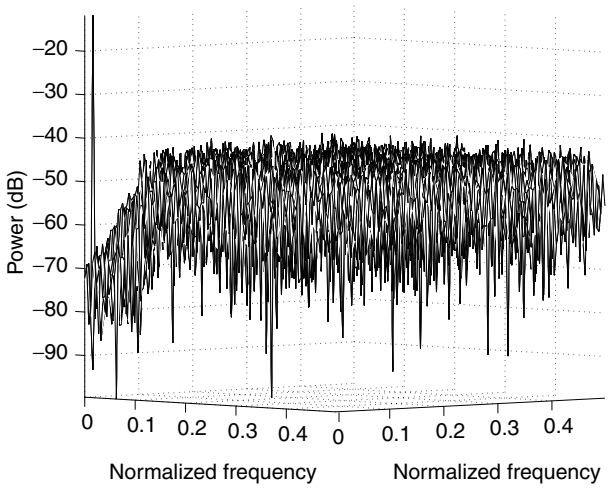


Figure 10. 3D perspective plot of one quadrant of the power spectrum of the fully connected distributed mesh feedback architecture generated with a 2D sinusoidal input of period 2 and a Kaiser error diffusion filter.

functionality. Using this approach, a 5×5 proof-of-concept photonic error diffusion neural network was fabricated.

The electronic circuitry for this CMOS-SEED error diffusion neural network consists of six distinct circuits associated with a single neuron. The standard neuron contains a total of 152 transistors with 23 dedicated to the neural functionality and 3 used for the output SEED driver. The error weighting circuitry for weights 1–5 contain the remaining 126 transistors with weights 1–5 accounting for 22, 20, 20, 28, and 36 transistors, respectively. This photonic implementation of the error diffusion neural network has been experimentally characterized and performed as predicted by both theory and simulation.

Figure 11 shows a photomicrograph of a single neuron of the CMOS-SEED photonic error diffusion neural network. The long rectangular features are the SEED MQW modulators and the background features are the CMOS VLSI transistors and interconnect metallization.

The challenges associated with this distributed approach to photonic A/D conversion include achieving large 2D arrays of smart pixel devices, the speed of convergence of the neural algorithm, and the integration with the electronic digital postprocessing circuitry.

3.4. Time Stretching Using Dispersive Optical Elements

Time stretching utilizes linear group velocity dispersion, most often in optical fibers, to frequency-downshift a signal that has been modulated onto optical pulses. In many of the more recent demonstrations, two long fiberoptic spools of lengths L_1 and L_2 are used with the modulator positioned between the two spools. The stretch factor M is defined as the width of the pulse exiting L_2 compared to that exiting L_1. If the pulsewidth exiting the source is defined as τ_0, and if δ_{τ_1} and δ_{τ_2} are defined as the additional broadening resulting from fiber spools L_1 and L_2, respectively, the stretch factor can be shown to be [37]

$$M = \frac{\tau_0 + \delta_{\tau_1} + \delta_{\tau_2}}{\tau_0 + \delta_{\tau_1}} \approx 1 + \frac{L_2}{L_1} \tag{16}$$

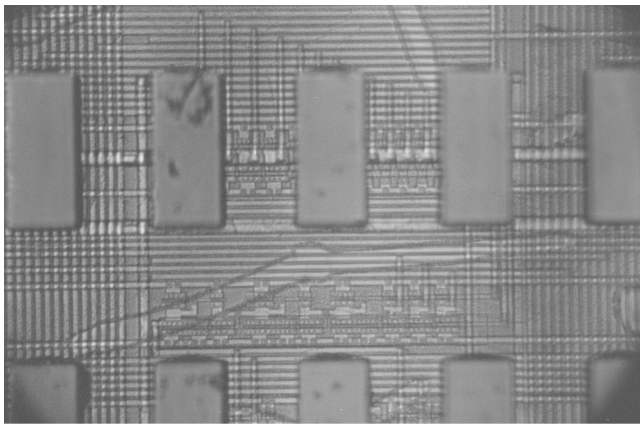


Figure 11. Photomicrograph of a single neuron of the 5×5 error diffusion neural network. (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.4, p. 220. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

if $\tau_0 \ll \delta_{\tau_1}$. In experimental demonstrations to date, L_1 and L_2 have generally been several kilometers in length.

Photonic time-stretch preprocessing techniques have also been proposed as a method to extend the performance of electronic A/D converters [21]. This technique is based on the premise that if an analog signal can be stretched in time, then the effective sampling rate and consequently the input bandwidth of a conventional electronic A/D converter can be increased. This specific technique is best suited to the class of *time-limited signals* such as those used in pulsed radar applications. Figure 12 shows the general concept of time stretching. Here, a modulated optical carrier is introduced to an optically dispersive element that subsequently stretches the modulated envelope of the signal. In Fig. 12, the vertical lines represent individual samples and the sampling interval T that would be required of the original signal is modified by the stretch factor M after the signal passes through the dispersive element. The optically dispersive element can be in a waveguide structure or a fiber. Time stretching using an array waveguide that is a wavelength dispersive element has been applied to photonic A/D conversion [37] and a 30-Gsps, 4-bit time-stretched photonic A/D converter using an 8-Gsps electronic digitizer has been reported [38].

One of the challenges associated with this particular approach to wide-bandwidth A/D conversion is distortion introduced by the nonuniform spectrum of the nonlinear dispersion. Nonuniformities in the spectrum result in temporal modulation of the carrier, which is mixed with the input signal during modulation. This nonuniformity results in a broadband distortion that limits the bandwidth and resolution of this particular approach to A/D conversion.

4. TRENDS IN PHOTON-BASED A/D CONVERSION

In general, trends in photonic A/D conversion can be categorized within the context of individual device and component development and investigation of new and novel architectures.

Individual photonic device development continues to be an active area of research interest. Improving the linearity, dynamic range, insertion loss, and speed of photonic components such as the electrooptic interferometers used for modulation and switching in current photonic A/D converters will directly improve the performance of these converters. Low-noise clock sources are an important part of any photonic A/D converter architecture. In high-speed, high-resolution applications, these sources require low timing and amplitude jitter, high repetition rate, and

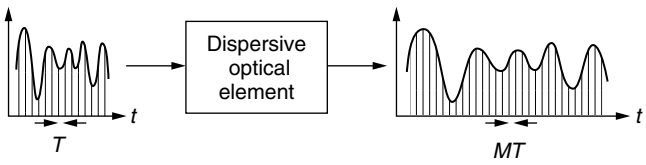


Figure 12. Concept of time stretching using a dispersive optical element.

short pulsewidth. For WDM applications, the clock source also requires large spectral width. For high-speed, high-resolution A/D applications, the clock source specifications are particularly challenging. Because of the importance of low-noise optical clock sources to both photonic A/D converters and other high-speed sampling applications, these sources also continue to be the focus of numerous research efforts.

New approaches and novel architectures continue to be developed to improve both the electronic and photonic contributions to the overall photonic A/D converter performance. Various forms of interleaving architectures will continue to be investigated and improved for application to wideband photonic A/D converters. As optical sampling speeds continue to increase, oversampling architectures will undoubtedly be considered for high-resolution applications. Distributed approaches to photonic A/D conversion can also be expected to be the focus of continued development because of their potential modularity and tolerance to fabrication errors.

A number of other promising approaches are currently under investigation. Details of several of these can be found in Refs. 39–41.

Photonics technology has generally been described as an enabling technology, supporting many of the functions provided by electronics technology. As both electronic and photonic technologies continue to mature, we can expect continued application and improvements to photonic A/D converter technology.

BIOGRAPHY

Barry L. Shoop is an associate professor in the Department of Electrical Engineering and Computer Science at the U.S. Military Academy, West Point, New York. He received his B.S. from the Pennsylvania State University, University Park, Pennsylvania, in 1980, an M.S. from the U.S. Naval Postgraduate School, Monterey, California, in 1986, and a Ph.D. from Stanford University, Stanford, California, in 1992, all in electrical engineering. His research interests are in the area of photonic A/D conversion, optical information processing, image processing, and smart pixel technology. He is a fellow of the OSA, a senior member of the IEEE, and a member of SPIE, Phi Kappa Phi, Eta Kappa Nu, and Sigma Xi.

BIBLIOGRAPHY

1. B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences Vol. 81, Springer-Verlag, Berlin, 2001.
2. D. F. Hoeschele, *Analog-to-Digital and Digital-to-Analog Conversion Techniques*, 2nd ed., Wiley, New York, 1994.
3. B. Razavi, *Principles of Data Conversion System Design*, IEEE Press, Piscataway, NJ, 1995.
4. M. J. Demler, *High-Speed Analog-to-Digital Conversion*, Academic Press, San Diego, CA, 1991.
5. R. van de Plassche, *Integrated Analog-to-Digital and Digital-to-Analog Converters*, Kluwer, Dordrecht, The Netherlands, 1994.
6. A. E. Steigman and D. J. Kuizenga, Proposed method for measuring picosecond pulsewidths and pulse shape in CW more-locked lasers, *IEEE J. Quant. Electron.* **6**: 212–219 (1970).
7. H. F. Taylor, An electrooptic analog-to-digital converter, *Proc. IEEE* **63**(10): 1524–1525 (1975).
8. R. A. Becker, C. E. Woodward, F. J. Leonberger, and R. W. Williamson, Wide-band electrooptic guided-wave analog-to-digital converters, *Proc. IEEE* **72**(7): 802–819 (1984).
9. B. Jalali and Y. M. Xie, Optical folding-flash analog-to-digital converter with analog encoding, *Opt. Lett.* **20**: 1901–1903 (1995).
10. B. L. Shoop and J. W. Goodman, Optical oversampled analog-to-digital conversion, *Appl. Opt.* **31**(26): 5654–5660 (1992).
11. B. L. Shoop and J. W. Goodman, A first-order error diffusion modulator for optical oversampled A/D conversion, *Opt. Commun.* **97**(4): 167–172 (1993).
12. B. L. Shoop, A. H. Sayles, D. A. Hall, and E. K. Ressler, A smart pixel implementation of an error diffusion neural network for digital halftoning, *Int. J. Optoelectron.* **11**: 217–228 (1997).
13. B. L. Shoop, Photonic A/D converters, *Proc. SPIE* **3490**: 252–255 (1998).
14. B. L. Shoop et al., A highly-parallel mismatch tolerant photonic A/D converter, *Proc. Conf. Lasers and Electro-Optics*, OSA Technical Digest, Optical Society of America, Washington, DC, 2001, pp. 64–65.
15. P. E. Pace, S. J. Ying, J. P. Powers, and R. J. Pieper, Integrated optical sigma-delta modulators, *Opt. Eng.* **35**: 1826–1836 (1996).
16. P. E. Pace, S. A. Bewley, and J. P. Powers, Fiber-lattice accumulator design considerations for optical $\sigma\delta$ analog-to-digital converters, *Opt. Eng.* **39**: 1517–1526 (2000).
17. J. C. Twichell and R. J. Helkey, Phase-encoded optical sampling for analog-to-digital converters, *IEEE Photon. Technol. Lett.* **12**: 1237–1239 (2000).
18. P. W. Juodawlkis et al., 505 MS/s photonic analog-to-digital converter, *Proc. Conf. Lasers and Electro-Optics*, OSA Technical Digest, Optical Society of America, Washington, DC, 2001, pp. 63–64.
19. T. R. Clark, J. U. Kang, and R. D. Esman, Performance of a time- and wavelength-interleaving photonic sampler for analog-digital conversion, *IEEE Photon. Technol. Lett.* **11**: 1168–1170 (1999).
20. T. R. Clark and M. L. Dennis, Toward a 100-G sample/s photonic analog-to-digital converter, *IEEE Photon. Technol. Lett.* **13**: 236–238 (2001).
21. A. S. Bhushan, F. Coppinger, and B. Jalali, Time-stretched analog-to-digital conversion, *Electron. Lett.* **34**: 839–840 (1997).
22. Y. Tsunoda and J. W. Goodman, Combined optical A/D conversion and page composition for holographic memory applications, *Appl. Opt.* **16**(10): 2607–2609 (1977).
23. H. K. Liu, Coherent optical analog-to-digital conversion using a single halftone photograph, *Appl. Opt.* **17**(14): 2181–2185 (1978).
24. K. Takizawa and M. Okada, Analog-to-digital converter: A new type using an electrooptic light modulator, *Appl. Opt.* **18**(18): 3148–3151 (1979).

25. N. N. Evtikheiv, D. I. Mirovitskii, N. V. Rostovtseva, and O. B. Serov, Multilayer holographic functional element in an analog-to-digital converter, *Sov. J. Quant. Electron.* **16**(9): 1180–1184 (1986).

26. J. A. Bell et al., Extension of electronic A/D converters to multi-gigahertz sampling rates using optical sampling and demultiplexing techniques, *Proc. 23rd Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1989.

27. A. D. McAulay, Optical analog-to-digital converter using optical logic and table look-up, *Opt. Eng.* **29**(2): 114–120 (1990).

28. Y. Li and Y. Zhang, Optical analog-to-digital conversion using acoustooptic theta modulation and table lookup, *Appl. Opt.* **30**(30): 4368–4371 (1991).

29. J. U. Kang, M. Y. Frankel, and R. D. Esman, Highly parallel pulsed optoelectronic analog-digital converter, *IEEE Photon. Technol. Lett.* **10**: 1626–1628 (1998).

30. P. W. Juodawlkis et al., Time-interleaved optical sampling for analog-to-digital converters (in press), *IEEE J. Lightwave Technol.*

31. J. C. Twichell and R. Helkey, Optical sampling for analog-to-digital converters, in *Lincoln Laboratory, Solid State Research, Quarterly Technical Report*, Lexington, MA, 1996, Vol. ESC-TR-96-096, pp. 28–33.

32. R. C. Williamson et al., Effects of crosstalk in demultiplexers for photonic analog-to-digital converters, *IEEE J. Lightwave Technol.* **19**: 230–236 (2001).

33. R. C. Williamson et al., Effects of crosstalk in demultiplexed photonic analog-to-digital converters, *Proc. Conf. Lasers and Electro-Optics,* OSA Technical Digest, Optical Society of America, Washington, DC, 2000, pp. 625–626.

34. E. B. Hogenauer, An economical class of digital filters for decimation and interpolation, *IEEE Trans. Acoust. Speech Signal Process.* **29**(2): 155–162 (1981).

35. R. A. Ulichney, *Digital Halftoning*, MIT Press, Cambridge, MA, 1987.

36. C. DeCusatis, D. Clement, and R. Lasky, eds., *Handbook of Fiber Optic Data Communication*, Academic Press, San Diego, CA, 1997.

37. F. Coppinger, A. S. Bhushan, and B. Jalali, Optoelectronic time-stretch and its application to analog-to-digital conversion, *IEEE Trans. Microwave Theory Tech.* **47**: 1309–1314 (1999).

38. A. S. Bhushan, P. Kelkar, F. Coppinger, and B. Jalali, 30 G sample/s 4-bit time-stretch analog-to-digital converter, *Proc. Conf. Lasers and Electro-Optics,* OSA Technical Digest, Optical Society of America, Washington, DC, 2000, pp. 623–624.

39. R. Urata et al., High-speed sample and hold using low temperature grown GaAs MSM switches for photonic A/D conversion, *IEEE Photon. Technol. Lett.* **13**: 717–719 (2001).

40. J. Cai and G. W. Taylor, Optoelectronic thyristor-based photonic smart comparator for analog-to-digital conversion, *IEEE Photon. Technol. Lett.* **10**: 1295–1297 (1999).

41. H. Sakata, Photonic analog-to-digital conversion by use of nonlinear Fabry-Perot resonators, *Appl. Opt.* **40**(2): 240–248 (2001).

POLARIZATION MODE DISPERSION MITIGATION

HENNING BÜLOW
Optical Systems
Stuttgart, Germany

With the increase of channel bit rate to 10 and 40 Gbps (gigabits per second) and in conjunction with unrepeated link lengths of hundred of kilometers and beyond, polarization mode dispersion (PMD) might become visible as a degrading property of the transmission link. Since PMD effects are drifting with time and differ for each wavelength channel, various dynamically adapting PMD compensators (PMDC) have been proposed that process either the optical signal field at link output or the electrical signal in the receiver after detection.

1. IMPACT OF PMD ON TRANSMISSION

The PMD of the optical transmission link arises mainly as a result of the residual optical birefringence varying along the fiber length, which is induced during the production of or cabling of the fiber. Nevertheless, other optical components or subsystems within the optical signal path such as optical amplifiers, dispersion-compensating fiber modules, or wavelength multiplexers might also add to the link PMD.

The effect of all these cascaded birefringences can again be regarded as an optical birefringence having a pair of principal states of polarization (PSP) e_- and e_+, which are orthogonally polarized and exhibit different group delays τ_- and τ_+. Unlike the scenario with fixed birefringence, they vary with the optical frequency ω. The difference of the group delay between fast and slow PSP is denoted by the differential group delay (DGD) $\Delta\tau = \tau_+ - \tau_-$ (see Fig. 1). The PMD is described by the dispersion vector $\Omega(\omega) = \Delta\tau e_-$, which has the length $\Delta\tau(\omega)$ and is oriented into the direction of the Stokes vector $e_-(\omega)$, which represents the state of polarization on the Poincaré sphere of the fast PSP at the output of the link [1]. Evaluation of Ω around the signal center frequency ω_0 exhibits that, as long as the signal bandwidth $\Delta\omega_S$ is sufficiently narrow with respect to the link PMD, the transmitter signal spreads among fast and slow PSP and thus suffers from a dual-path propagation due to the DGD (see Fig. 1). This leads to a temporal broadening of the bit beyond a bit period T and thus to intersymbol interference (ISI).

Since the DGD as well as the PSP vary with both wavelength and time (see Fig. 2), PMD is quantified by

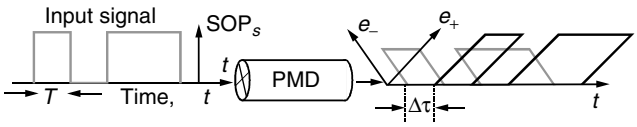


Figure 1. Optical data signal at input and output of a fiber having a first-order PMD with a DGD $\Delta\tau$.

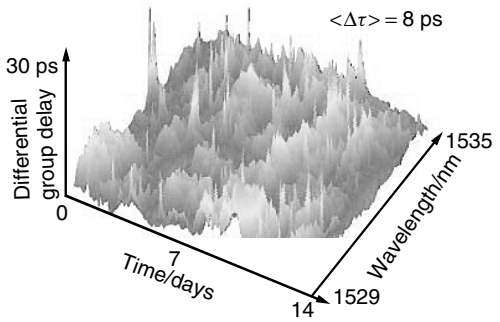


Figure 2. DGD of an installed fiber (246 km length, 8 ps PMD) measured over a timespan of 14 days and a wavelength range of 6 nm.

statistical means: the mean value of the DGD $\langle\Delta\tau\rangle$ and the probability density function of the actual DGD $\Delta\tau$, which is given by the Maxwellian distribution

$$\rho(\Delta\tau) = \sqrt{\frac{6}{\pi}}\frac{3\Delta\tau^2}{\langle\Delta\tau\rangle^3}\exp\left[-\frac{3\Delta\tau^2}{2\langle\Delta\tau\rangle^2}\right]$$

Thus, an actual DGD $\Delta\tau$ of beyond 3.1 $\langle\Delta\tau\rangle$ occurs with a probability of 10^{-5}. As indicated by the autocorrelation function (ACF) of PMD, which decays below 0.5 for $\Delta\omega > 2/\langle\Delta\tau\rangle$, the PMD cannot be considered as constant anymore within increasing signal bandwidth $\Delta\omega_S$ and increasing link PMD $\langle\Delta\tau\rangle$. Then the probability of a signal distortion due to second- and higher-order PMD increases. Second-order PMD denotes a constant variation $d\Omega/d\omega$ of the PMD vector, and higher orders $(n+1)$ denote nonvanishing derivatives $d^n\Omega/d\omega^n$ at the signal center frequency. The impact of second-order PMD on the signal is a depolarization proportional to $de_-/d\omega$ and to $\Delta\tau$ [2], which denotes a cross coupling of signal fields between the orthogonal PSPs, and the PMD-induced chromatic dispersion proportional to $\pm 0.5 d\Delta\tau/d\omega$. The induced dispersion exhibits opposite signs in both PSPs and adds to the chromatic dispersion of the link. A signal distortion induced by PSP rotation $de_-/d\omega$ has the highest likelihood of the two second-order contributions. It generates distortions similar to over- and undershoots in the data signal detected by the receiver photodiode [2].

Variations of $\Delta\tau$ and e_- with time are induced mainly by environmental temperature changes that lead to a variation of the signal distortion and thus to a fluctuating bit error rate (BER). The BER limit that is tolerated, such as 10^{-12} or for forward error correction (FEC) support 10^{-4},

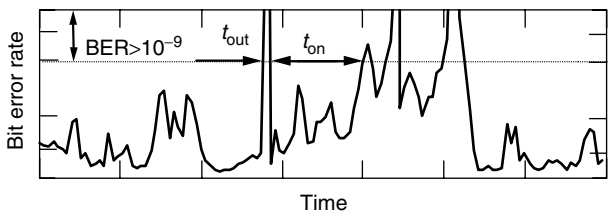


Figure 3. Temporal BER evolution for the transmission over a fiber link with PMD (14-day timespan).

might be exceeded (see Fig. 3). Times of poor BER mean an outage for the transmission system. An outage event has a duration t_{out}. Because of the statistical nature of PMD, the impact on a transmission system is quantified by the cumulative probability (CP) for a PMD-induced outage. CP can be defined by the ratio of accumulated outage time and system operation time for long observation timespans. Besides the outage probability, the mean outage duration $\langle t_{out}\rangle$ and the mean time between outages $\langle t_{on}\rangle$ have been proposed to describe the impact of PMD on a system. That is 56 min for $\langle t_{out}\rangle$ and 2.56 years (outage rate 0.39 yea^{-1}) for $\langle t_{on}\rangle$ have been deduced from PMD measurements of buried fiber [3]. On the other hand, PMD fluctuations, also on the timescale of a few milliseconds, have been observed for aerial cables or fiber exposed to mechanical vibrations. If these fast fluctuations need to be compensated, they determine the upper speed limit of a PMDC.

Alternatively, the robustness of a transmission system to PMD can be quantified by the PMD limit $\langle\Delta\tau\rangle_{lim}$, which is the maximum PMD that leads to a specific outage probability CP given that the system operates with a certain optical signal-to-noise ratio (OSNR) margin. Typical values for CP and margin are 10^{-5} and 2 dB, respectively. $\langle\Delta\tau\rangle_{lim}$ amounts to about ~15% of the bit period T for a non-return-to-zero (NRZ) signal.

In order to assess the relevance of PMD for a transmission system, the value of the length-related PMD, the PMD coefficient τ', needs to be related to $\langle\Delta\tau\rangle_{lim}$ and the link length L. Fiber PMD $\langle\Delta\tau\rangle_{fiber}$ and all component and subsystem PMDs $\langle\Delta\tau\rangle_i$ contribute to the total link PMD$_{tot}$ according to $\langle\Delta\tau\rangle_{tot}^2 = \langle\Delta\tau\rangle_{fiber}^2 + \Sigma\langle\Delta\tau\rangle_i^2$. The maximum link length is $L < (\langle\Delta\tau\rangle_{lim}^2 - \Sigma\langle\Delta\tau\rangle_i^2)/\tau'^2$. With more recently manufactured fiber having a low PMD coefficient of $\leq$0.08 ps/$\sqrt{}$km, and assuming a dispersion compensation fiber PMD coefficient of 0.15 ps/$\sqrt{}$km, an amplifier PMD of 0.5 ps and 80 km amplifier spacing, 40 Gbps transmission over long-haul distances is affected by PMD beyond 1400 km. In fiber production PMD received only minor attention until the mid-1990s. Therefore links incorporating older fibers might exhibit PMD coefficients of $\geq$0.5 ps/$\sqrt{}$km. This means that a few hundreds of kilometers of link might reach the PMD limit even at 10 Gbps.

2. PMD COMPENSATION

A slight improvement in the robustness of PMD can be obtained by keeping the receiver threshold in the middle of the eye diagram degraded by ISI. This automatic threshold adjustment leads to the abovementioned PMD limit of $0.15T$. Moreover, specific modulation formats such as return to zero (RZ) tolerate a slightly higher bit broadening due to PMD and thus increase the limit to approximately $0.18T$. The tolerance to PMD can further be increased if there is room to allocate a power margin (OSNR margin) for PMD of much more than 1 or 2 dB. This margin can be obtained by reducing span loss or by incorporating forward error correction (FEC).

However, in general only a limited margin will be reserved for PMD, and a higher PMD limit must be overcome. Therefore several active compensation

techniques have been proposed to mitigate the degrading effect of PMD. The majority of these approaches can be classified as either optical or electrical compensation techniques that apply optical signal processing within an optical PMD compensator unit, or postdetection electronic signal processing of the photodiode signal by an electrical equalizer within the receiver. Commonly the dynamic adaptation of the compensator unit to the drifting signal distortion is accomplished by three elements arranged in a feedback scheme (see Fig. 4): a signal processing element to reduce the distortion, an element at its output to measure the signal quality and to provide a feedback signal, and an adaptation control algorithm implemented in an electronic processor that tunes the parameters of the signal processor into direction of optimum feedback signal and thus to reduced signal distortion.

3. OPTICAL COMPENSATION

The signal processing element of the simple optical PMD compensator is formed by an electrically tunable polarization controller (PC) based, for example, on the electrooptic effect in lithium niobate, on the elastooptic effect used in fiber squeezer devices, or on liquid crystal technology. A constant optical birefringence [e.g., polarization-maintaining fiber (PMF)] is attached at its output (see Fig. 4). A typical value of the PMF's DGD is $\Delta\tau_C = 0.6T$. The principle of operation of this basic PMDC is to modify the dispersion vector of the total PMD $\Omega_{\text{tot}} = \Omega_L + \Omega_C$ formed by the link PMD Ω_L and the compensator PMD Ω_C (see Fig. 5), by tuning of the PC in such a way that the signal degradation at compensator output is a minimum [4]. The PC modifies the orientation of Ω_C ($|\Omega_C| = \Delta\tau_C$). Factoring in these dispersion vectors and the Stokes vector SOP_S representing the signal input state of polarization on the Poincaré sphere, all

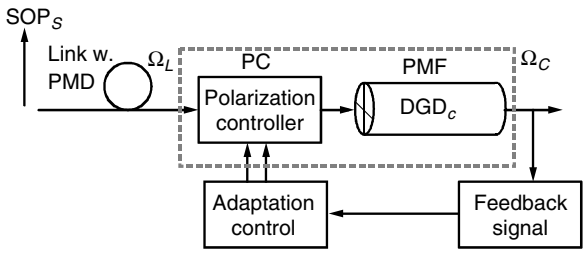


Figure 4. Basic optical PMD compensator.

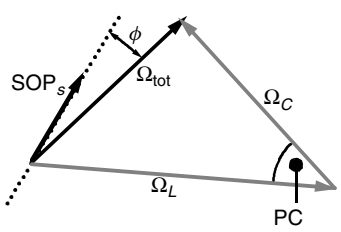


Figure 5. PMD dispersion vectors Ω and the signal state of polarization SOP_S observed at the output of the basic PMD compensator. Ω_L and Ω_C represent the link PMD and compensator PMD, respectively, at signal center frequency.

transformed to PMDC output, the PMD conditions for compensation of first-order PMD at the center frequency can be explained. During operation the compensator will either minimize the total DGD $|\Omega_{\text{tot}}|$ (case A) or the angle ϕ between SOP_S and the total PMD Ω_{tot} (case B). In both cases the degradation effects are minimized: in case A due to minimum delay $|\Omega_{\text{tot}}|$ and in case B due to single-path propagation in one PSP only. A more detailed analysis shows that by detuning of the PC from these ideal first-order conditions — to some extent — higher-order PMD can be mitigated too.

During operation of a PMDC under conditions of drifting link PMD, the adaptation control keeps the signal processor in an extremum of the feedback signal. In this simple PMD compensator structure the decay of a global maximum of the feedback signal to a local maximum limits the efficiency.

Calculations indicate that with the basic PMD compensator, the PMD limit can be extended to about $0.30T$ (2 dB power margin, 10^{-5} outage) for non-return-to-zero (NRZ) signals.

A feedback signal (FS) must satisfy the following the requirements: (1) good correlation with the receiver BER and (2) sufficient sensitivity to detect a detuning of the signal processor from the optimum setting or a variation of the link PMD. As a feedback signal the measurement of signal properties can be used, either in the optical domain such as the degree of polarization (DoP) or, after detection, in the electrical domain, including "spectral line" or eye opening. The DoP is measured at the compensator output by a polarimeter setup. The PMD-induced pulse splitting results in a time-varying state of polarization along the bit pattern, which reduces the measured DoP [5]. The spectral line feedback denotes the analysis of the electrical spectrum at a photodiode illuminated by the compensator output signal. Since PMD generates a notch in the electrical spectrum at the frequency $0.5/\Delta\tau$, the maximization of spectral power measured at, for instance, $0.5/T$ or $0.25/T$ or the weighted sum of both indicate a decreasing residual PMD distortion after compensation and thus an improved receiver BER [6]. A very good correlation between BER and the feedback signal is provided by an electronic performance monitor, also referred to as an "eye monitor" [7]. It extracts a quality measure of the actual eye diagram by bit-synchronous sampling at decision time t_0 and is thus strongly correlated with the BER. The sampling demands a valid clock that might not always be extractable from the signal during times of strong signal distortion as might occur during startup of the compensator when the accommodation to the actual distortion is not completed.

4. HIGH-ORDER PMD COMPENSATION

In basic PMD compensator with a PMD limit of about $0.30T$, the adaptation control has to optimize the 2 degrees of freedom (DoF) of the polarization controller. An improvement of this basic structure can be achieved by replacing the fixed DGD by an continuously tunable DGD that is also tuned by the adaptation control. With this increase of the number of DOFs to 3, the PMD limit

is slightly improved [4]. The main advantage of a variable DGD is that the decay of the global optimum to a less efficient local optimum can be avoided during tracking of the drifting link PMD. Realizations of a variable group delay are based on beam optics with mechanically variable gaps or stacks of birefringent elements with electrically controlled polarization switches in between.

In order to address higher orders of PMD and increase the PMD limit of the compensator, improved optical signal processing can be achieved by a cascade of two or more basic compensators [6]. This leads to structures with ≥ 4 DoF controlled by maximization of one single feedback signal. The increase of the PMD limit beyond $0.40T$ was calculated for a two stage structure [4,8]. However, in general an increasing number of DoFS improves the effectiveness of a PMDC, but it seems that the sensitivity of the feedback signal to the variation of an individual tuning parameter decreases. This slows down the accommodation time for a changing PMD. Moreover, with increasing number of DoFs, the danger of being trapped in suboptimal local maxima increases. Therefore the search for efficient high-order PMDC is continued [8].

5. ANALYSIS OF COMPENSATION

For experimental or numerical assessment of compensator performance, the statistics of the system-relevant quality parameter degradation (BER, power penalty or OSNR penalty) have to be determined. Several hundreds or some thousands of statistically independent samples of actual PMD values of a given PMD $\langle \Delta\tau \rangle$ are applied at a transmission system with PMDC and the BER or power penalties are determined. These PMD samples are generated either in a PMD emulator or in a computer model with all relevant orders. Their occurrence should obey the PMD statistics of a real link (or it should be possible to convert them to the appropriate statistics). The set of BER or penalty values exhibits the same statistics that one would obtain from measurements at a real link after many months or years. For analysis, the relative occurrence that a specific BER or penalty value is exceeded is plotted against this limit (see Fig. 6). Thus, the resulting curves show that with PMDC, specific degradations are exceeded with a lower likelihood than without it and that an improved PMDC (two stages) leads to further reductions of the likelihood. The curves can be extrapolated to more relevant 10^{-5} outage, and a set of curves for different PMDs $\langle \Delta\tau \rangle$ allows us to interpolate the PMD limit $\langle \Delta\tau \rangle_{\text{lim}}$.

6. ELECTRICAL COMPENSATION

Postdetection electronic signal processing is also discussed and applied for PMD mitigation. Adaptive electrical signal processing schemes, well established in lower-rate communications for compensation of channel degradation, are also implemented in an optical receiver. The equalization schemes that are in the scope of application are a feedforward equalizer (FFE) followed by a decision feedback equalizer (DFE). The FFE superimposes differently delayed and weighted replica of the input signal and the

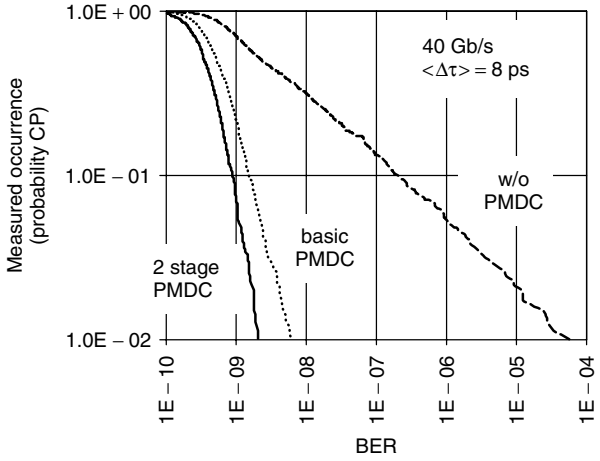


Figure 6. Relative occurrence (y axis; ordinate) of measured system bit error ratios beyond a BER limit (x axis; abscissa) for some hundred transmissions over a statistically changing PMD emulator. The experiments were performed without and with optical PMD compensators.

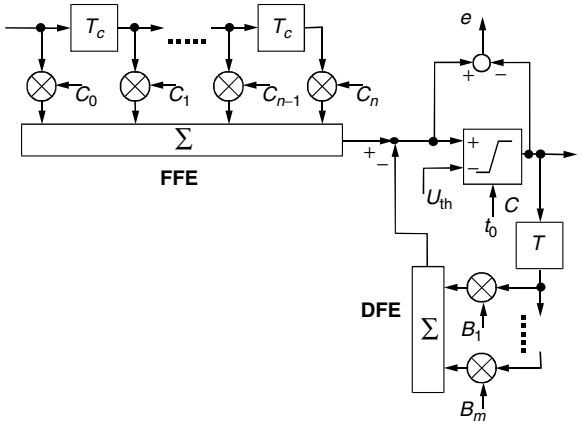


Figure 7. Electronic equalizer for electrical PMD compensation. The electronic signal processing is performed by a feedforward equalizer (FFE) and a decision feedback equalizer (DFE), which also provides an error signal e for the adaptation control.

DFE subtracts the weighted and delayed signal appearing at the decision gate output from the input signal provided by the FFE (see Fig. 7). Tuning to the actual PMD distortion is achieved by adjusting the FFE weights (C_i), also referred to as "tap coefficients," and the DFE feedback weights (B_i). FFE and DFE suitable for 10-Gbps operation have been realized as integrated analog signal processing circuits in high-speed technologies such as SiGe, InP, or GaAs [9]. An alternative realization in a digital signal processing (DSP) scheme based on digitally processing of the signal in an integrated electronic circuit [complementary metal oxide semiconductor integrated circuit (CMOS IC)] after analog-to-digital conversion of bit-synchronous signal samples is also discussed.

Besides these equalization techniques that perform the decision on a bit-by-bit basis, there also exists a concept of optimum detection based on data decisions that are most likely to be correct when utilizing the entire

received signal and the knowledge of the characteristics of the transmission channel. Since in the case of PMD the characteristics are neither known nor stable in time, adaptive mechanisms also have to be applied. The performance limits of this maximum-likelihood sequence estimation (MLSE) under realization constraints for the optical channel is currently under investigation [10]. Constraints are high-speed electronic realization for 10-Gbps signals with reduced complexity, processing of truncated sequences, nonstationary noise, and nonlinear channel due to optical nonlinearity of the fiber and to square-law detection of the photodiode. Preliminary results confirm the potential for an improved performance for PMD compensation compared to the FFE and a DFE equalizer (see Fig. 8). The MLSE detector will be realized as DSP scheme. The processing is based on the Viterbi algorithm, which is an efficient way to organize the computations of the signal samples.

Different adaptation schemes for the FFE, DFE, and clock-phase alignment are under consideration. Similar to the optical compensator, the eye monitor can be used at the FFE output in conjunction with a gradient algorithm that consecutively tunes the equalizer taps. The implementation of the least-mean-square (LMS) algorithm, well established at lower rates, has also been discussed. An error signal e at the decision time t_0 is generated for adaptation purpose by subtracting the signal at the decision gate output, serving as reference, from the signal at the equalizer output (see Fig. 7). The multiplication of this error signal with samples of signals at different positions within the equalizer generates different independent feedback signals for each FFE tap or DFE feedback tap. This allows for simultaneous adaptation of all tap weights. Moreover, the feedback signals are bipolar and thus automatically indicate the direction of tuning to the optimum. Therefore very fast adaptation within some hundreds of bits is possible in principle. An alternative realization of the LMS scheme is possible in conjunction with channel coding used for forward error correction (FEC). By comparing the uncorrected and the corrected data sequences launched into and appearing at the output of the FEC decoder, respectively, the observed errors can be correlated with specific bit constellations. The error count for a specific constellation provides a measure for the error e.

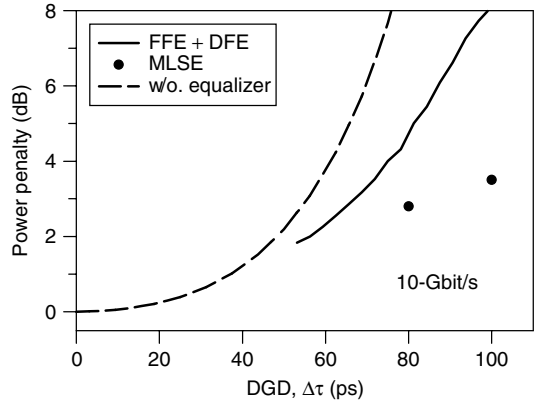


Figure 8. Power penalty (OSNR penalty) of a 10-Gbps optical receiver versus a first-order PMD distortion quantified by the DGD. The penalties are determined numerically for a receiver both without (dashed line) and with (solid line) an equalizer for electrical PMD compensation.

7. OPTICAL VERSUS ELECTRICAL COMPENSATION

Optical PMD compensation by a cascade of two ($0.4T$ PMD limit) or more basic compensators has been demonstrated experimentally. Theoretically, zero penalty can be achieved for an ideal exact compensation of the PMD and operation at bit rates of $\leq$160 Gbps has already been shown in the laboratory. In contrast to these findings, electrical equalization exhibits a residual penalty in the presence of strong PMD distortion and operation has been demonstrated at $\leq$10 Gbps as a result of the speed limits of electronic realization. The remaining penalty is due to the loss of polarization and phase information after square-law detection by the receiver photodiode (see Fig. 8). Mitigation by FFE and DFE is accompanied by a residual penalty of 7 dB for a first-order PMD distortion with a DGD of one bit period and beyond (equal excitation of both PSPs) and in the presence of signal-dependent optical noise (signal independent thermal noise leads to approximately half-residual penalty). This is too high for target applications such as long-distance transmission, which is strongly limited by noise and where only approximately $\leq$2 dB can be allocated for PMD. In this case the FFE and DFE exhibit a reasonable low penalty only if the DGD remains well below a bit period. This value corresponds to a PMD limit in the range of $0.25T$. An improvement is possible by applying MLSE schemes, with the first numerical results indicating a residual penalty in the range of 4 dB [9,10].

Nevertheless, postdetection electronic signal processing improves the receiver performance not only in the presence of PMD but also for ISI stemming from chromatic dispersions or optical nonlinearity such as self-phase modulation (SPM). In addition, it has the potential for effecting a seamless and cost-effective integration in the receiver electronics.

More recent numerical and experimental results have indicated that the efficiency of PMD compensation might be degraded when being used in a WDM environment [11]. As a result of the birefringence induced by other wavelength channels via the Kerr effect of the fiber, a fast polarization modulation can occur at high channel power that is not compensated by current PMDC schemes.

BIOGRAPHY

Henning Bülow received the Dipl.-Ing. degree in electrical engineering in 1985 from the University of Dortmund, Germany, and the Dr. degree in electrical engineering from the University of Berlin in 1988. He joined the Research Center of Alcatel in Stuttgart, Germany, in 1990 as a Research Engineer. At Alcatel he has been studying different aspects of optical communication systems, focusing mainly on erbium-doped fiber amplifiers, optical end electrical time-multiplexed 40-Gbps transmission systems, and assessment of data

transmission in the presence of polarization mode dispersion of the fiber. Since 1998, Dr. Bülow has headed a research team investigating the dynamic mitigation of transmission distortions at 10, 40, and 160 Gbps by adaptive electrical and optical signal processing. He holds over 20 patents in the area of optical communications, has authored or co-authored more than 60 journal and conference publications, and serves on the technical program committee of the Optical Fiber Communication Conference (OFC).

BIBLIOGRAPHY

1. C. D. Poole and J. Nagel, Polarization effects in lightwave systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Communications IIIA*, Academic Press, San Diego, 1997, Chap. 6.
2. C. Francia et al., PMD second-order effects on pulse propagation in single-mode fibers, *IEEE Photon. Technol. Lett.* **10**: 1739–1741 (1998).
3. R. Caponi et al., WDM design issues with highly correlated PMD spectra of buried optical cables, *Technical Digest OFC 2002*, Anaheim, CA (USA), ThI5, 2002.
4. H. Sunnerud et al., A comparison between different PMD-compensation techniques, *IEEE J. Lightwave Technol.* **20**(3): 368–378 (2002).
5. S. Lanne, W. Idler, J.-P. Thiéry, and J.-P. Hamaide, Demonstration of adaptive PMD compensation at 40Gb/s, *Technical Digest OFC 2002*, Anaheim, CA, TuP3, 2001.
6. R. Noé et al., Polarization mode dispersion compensation at 10, 20, and 40 Gb/s with various optical equalizers, *J. Lightwave Technol.* **17**(9): 1602–1615 (1999).
7. F. Buchali et al., A 40 Gb/s eye monitor and its application to adaptive PMD compensation, *Technical Digest OFC 2002*, Anaheim, CA, Proc. WE6, 2002.
8. J. Poirrier, F. Buchali, and H. Bülow, Optical PMD compensation performance: numerical assessment, *Technical Digest OFC 2002*, Anaheim, CA, WI3, 2002.
9. H. Bülow, Electronic equalization of transmission impairments, *Technical Digest OFC 2002*, Anaheim, CA, TuE4, 2002.
10. H. F. Haunstein et al., Design of near optimum electrical equalizer for optical transmission in the presence of PMD, *Technical Digest OFC 2001*, Anaheim, CA, WAA4, 2001.
11. J. H. Lee et al., Impact of nonlinear crosstalk on optical PMD compensation, *Technical Digest OFC 2002*, Anaheim, CA, ThI2, 2002.

POLYPHASE SEQUENCES

So Ryoung Park
Iickho Song
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea

1. INTRODUCTION

Sequences have played an important role in the history of communication systems, especially as the spreading sequences in spread-spectrum (SS) code-division multiple access (CDMA) systems used widely for the most recent personal cellular communications. Among the systems using the SS technique, the direct-sequence (DS) CDMA system expands the bandwidth of a signal by directly multiplying an information symbol by a spreading sequence uniquely assigned to each user, so that a number of users in a cell can share the same frequency band simultaneously. The DSCDMA system is usually preferred over the other SS techniques because it has low implementation cost, can be used with coherent demodulation, and can provide a large capacity in addition to such usual benefits of the SS systems as interference rejection/suppression, multipath mitigation, and security [1–5].

Cellular DSCDMA systems have adopted the multiple or two-layered spreading sequence allocation for flexible system deployment and operation [6]. Orthogonality can be achieved by first multiplying each user's information symbol by a short spreading sequence that is orthogonal to that of any other user in the same cell. This first spreading is followed by a multiplication of a long spreading sequence, which is cell-specific but common to all users in the same cell in the forward link and user-specific in the reverse link. It is possible to provide waveform orthogonality among all users in the same cell while maintaining only mutual randomness between users in different cells. The short spreading sequences are called *channelization sequences* and the long spreading sequences, *scrambling sequences*.

Since different base stations (different mobile users) use different timeshifts of the same sequence in the forward (reverse) link in intercell synchronous operation, the long scrambling sequences are required to have good autocorrelation (AC) properties. The AC property of a sequence is said to be perfect when the AC value is N for the in-phase component and zero for the out-of-phase components, where N is the length of the sequence. This property can support fast symbol synchronization, low multipath interference (MPI), and low intercell interference (ICI). In the intercell asynchronous operation, each cell is assigned to a unique long sequence, which is thus required to have good AC property for fast symbol synchronization and low MPI and to have good crosscorrelation (CC) property for low ICI.

The short spreading sequences in a cell play an important role not only in spreading and despreading but also in the identification of a desired user from interfering users. This is the reason why the short spreading sequence is alternatively called the *signature sequence*. Because the whole frequency band is being used all the time, the bandwidth can be utilized more efficiently (i.e., with narrower equivalent bandwidth per user) in the DSCDMA system than in the conventional systems. The number of users that a system can accommodate is determined by the required signal-to-noise ratio (SNR), which is in turn determined by system design requirements. Note that there is a stringent limitation on the maximum number of users in the time-division multiple access (TDMA) and

frequency-division multiple access (FDMA) systems. On the other hand, the capacity of a CDMA system is softly limited; that is, the maximum number of users is not a clear-cut number. Instead, as more users share the same CDMA channel, the signal quality degrades gradually until it is unacceptable. The acceptable number of users depends on many aspects, including the CC property of the signature sequences in DSCDMA systems. The CC property of signature sequences is linked directly with the multiple access interference (MAI), and consequently with the capacity of the cell. Thus, for low MAI and a large channel capacity, it is desired that the CC value is always (close to) zero.

Traditionally, the cellular DSCDMA systems have utilized the binary maximal length (m) and Gold sequences as the scrambling sequences and Walsh sequences as the channelization sequences. The m sequences have been chosen because they possess the desired randomness and are easily generated via linear feedback shift registers. Gold sequences are an important subclass of the m sequences that can provide good periodic (even) CC. The maximum magnitudes of the periodic AC and CC of Gold sequences are 1 and $1 + 2^{\lfloor (m+2)/2 \rfloor}$, respectively, which is optimum in the sense of the Sidelnikov lower bound for binary sequences. Here, $m \geq 3$ is an integer not equal to a multiple of 4, the length of the sequence is $N = 2^m - 1$, and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x [7]. The Walsh sequences are generated by mapping $\{0, 1\}$ onto $\{-1, 1\}$ for codeword rows of square Hadamard matrices.

In order to obtain sequences having better correlation properties, a number of polyphase sequences have been suggested [8–29]. Using the phase diversity in a chip, polyphase sequences can be so designed as to be more suitable for than binary (two-phase) sequences channelization and scrambling sequences in cellular DSCDMA systems [30,31].

2. SOME PRELIMINARIES

Figure 1 shows the correlation of two sequences in asynchronous DSCDMA systems when the chip synchronization is assumed to be perfect, where k_1 and k_2 denote the user index and $x_j^{(k_i)}$ denotes the jth chip of the k_i(th) sequence. For M-ary phase shift keying (MPSK) systems, the general correlation function [32–34] between two polyphase sequences $\mathbf{x}^{(k_1)}$ and $\mathbf{x}^{(k_2)}$ of length N is

$$\theta_\gamma(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) + \gamma C^*(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l - N) \tag{1}$$

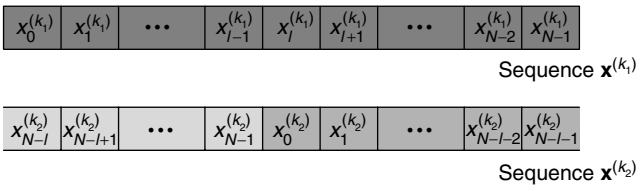


Figure 1. Correlation of two sequences.

where

$$C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = \begin{cases} \sum_{i=0}^{N-1-l} x_i^{(k_1)} \{x_{i+l}^{(k_2)}\}^*, & 0 \leq l \leq N-1 \\ \sum_{i=0}^{N-1+l} x_{i-l}^{(k_1)} \{x_i^{(k_2)}\}^*, & 1-N \leq l < 0 \\ 0, & |l| \geq N \end{cases} \tag{2}$$

is the partial correlation

$$\gamma \in \{W_M^t \mid t = 0, 1, \ldots, M-1\} \tag{3}$$

$$W_A^k = e^{2\sqrt{-1}\pi k/A} \tag{4}$$

and A is a natural number with $*$ denoting the complex conjugate. For binary ($M = 2$) phase shift keying (BPSK) systems, we have two important functions, the *even correlation* (EC)

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) + C^*(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l - N) \tag{5}$$

and the *odd correlation* (OC)

$$\hat{\theta}(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) - C^*(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l - N), \tag{6}$$

from Eq. (1). When $k_1 = k_2$, Eqs. (5) and (6) are called the "even AC (EAC)" and "odd AC (OAC)," respectively, and when $k_1 \neq k_2$, Eqs. (5) and (6) are called the "even CC (ECC)" and "odd CC (OCC)," respectively.

In order to completely analyze the performance of asynchronous DSCDMA systems using a particular class of sequences in an environment of multiple simultaneous users, we should consider not only the EC properties but also the OC properties of sequences. The OC function affects the output of the correlator when the information symbols change over one integration interval, while the EC function affects the output when the information symbols do not change. Thus, when the binary information symbols are equiprobable, both the EC and OC functions are equally important in the system design and performance analysis. Designing sequences with good OC properties is a difficult problem, as observed by other authors [35–37].

3. POLYPHASE SEQUENCES FOR CELLULAR DSCDMA SYSTEMS

3.1. Polyphase Sequences with Perfect EAC

We first consider several polyphase sequences that have perfect EAC properties. Since the good AC properties of such sequences guarantee low interference among the different timeshifts, they are useful as the long scrambling sequences in intercell synchronous cellular DSCDMA systems:

Frank sequence [22] — the sequence of length $N = M^2$ is defined by

$$x_{nM+k} = W_M^{nk}, \quad n, k \in \{0, 1, \ldots, M-1\} \tag{7}$$

Golomb sequence [18,23] — the Golomb sequence of length N is defined by

$$x_n = W_N^{n(n+1)/2}, \quad n = 0, 1, \ldots, N-1 \tag{8}$$

P1 sequence [11] — the P1 sequence of length $N = M^2$ is defined by

$$x_{nM+k} = W_{2M}^{-(M-2n-1)(nM+k)}, \quad n, k \in \{0, 1, \ldots, M-1\} \tag{9}$$

Px sequence [25] — the Px sequence of length $N = M^2$ is defined by

$$x_{nM+k} = \begin{cases} W_{4M}^{(M-2n-1)(M-2k-2)}, & M \text{ even}, \\ W_{4M}^{(M-2n-1)(M-2k-1)}, & M \text{ odd}, \end{cases}$$
$$n, k \in \{0, 1, \ldots, M-1\}. \tag{10}$$

P3 sequence [12] — the P3 sequence of length N is defined by

$$x_n = W_{2N}^{n^2}, \quad n = 0, 1, \ldots, N-1 \tag{11}$$

P4 sequence [13] — the P4 sequence of length N is defined by

$$x_n = W_{2N}^{n(n-N)}, \quad n = 0, 1, \ldots, N-1 \tag{12}$$

In Eqs. (7–12), M and N are natural numbers. The EAC properties of all the sequences listed above are perfect:

$$\theta(\mathbf{x}, \mathbf{x})(l) = \begin{cases} N, & l = 0 \\ 0, & l \neq 0 \end{cases} \tag{13}$$

For example, we obtain Table 1 when $N = 16$. When the length of sequences is 16, the numbers of phases of the Frank, P1, Golomb, Px, P3, and P4 sequences are 4, 8, 16, 16, 32, and 32, respectively. For the sequences shown in Table 1, we have calculated the normalized EAC and OAC functions as shown in Fig. 2; thus we can clearly confirm the perfect EAC property of the sequences. Although each of the six sequences has its own unique defining equation, the EAC properties of the sequences are all the same (at least in the sense of magnitude) and the OAC properties are quite similar to each other. The Frank sequence has the same AC function as the P1 sequence, whose normalized maximum magnitude of the out-of-phase OAC value is 0.1768. The Golomb sequence has the same AC function as do the P3 and P4 sequences, whose normalized maximum magnitude of the out-of-phase OAC value is 0.2310. The maximum magnitude of the out-of-phase OAC value of the Px sequence is the same as that of the Frank sequence, although the OAC functions of these two sequences are different. The peak : second-peak ratio of the Frank, P1, and Px sequences is better than those of the Golomb, P3, and P4 sequences. In addition, the mainlobe : total sidelobe energy ratio of the Px sequence is the lowest among the six sequences.

3.2. Polyphase Sequences with Optimum or Near-Optimum EC

Let $\theta_a = \max_{\mathbf{x}} \max_{l \neq 0} \theta(\mathbf{x}, \mathbf{x})(l)$ and $\theta_c = \max_{\mathbf{x} \neq \mathbf{y}} \max_l \theta(\mathbf{x}, \mathbf{y})(l)$, where $\mathbf{x}$ and $\mathbf{y}$ denote members in a class of polyphase sequences. Then, we have [35]

$$\max\{\theta_a, \theta_c\} \geq \sqrt{N} \tag{14}$$

Thus, if a set of sequences satisfies $\theta_a \leq \sqrt{N}$ and $\theta_c \leq \sqrt{N}$, the sequence is called *optimum* in the sense of the lower bound for polyphase sequences. We now consider some polyphase sequences whose EC functions are (nearly) optimum. These (near) optimum sequences are useful as the long scrambling sequences in intercell asynchronous cellular DSCDMA systems.

3.2.1. Four-Phase Sequence. The generating polynomial of four-phase sequences [16] is a primitive basic irreducible polynomial in $\mathbb{Z}_4[t]$, whose modulo 2 projections are primitive irreducible polynomials in $\mathbb{Z}_2[t]$. If $g(t)$ is a primitive basic irreducible polynomial of degree m, the set of four-phase sequences has period $N = 2^m - 1$ and size $N + 1$. Let the generating polynomial of degree m for the four-phase sequence be

$$g(t) = g_m t^m + g_{m-1} t^{m-1} + \cdots + g_1 t + g_0 \tag{15}$$

where $g_i \in \{0, 1, 2, 3\}$, $g_m \neq 0$, and $g_0 \neq 0$. Then, the recurrence condition of the quaternary sequence $\mathbf{s} = [s_0\, s_1\, \cdots\, s_{N-1}]$ is

$$s_{i+m} = g_{m-1} s_{i+m-1} + g_{m-2} s_{i+m-2} + \cdots + g_1 s_{i+1} + s_i \pmod 4, \text{ for } i \geq 0 \tag{16}$$

and the four-phase sequence $\mathbf{x} = [x_0\, x_1\, \cdots\, x_{N-1}]$ can be obtained from $x_i = W_4^{s_i}$. For example, let $m = 3$ and $g(t) = t^3 + 3t^3 + 2t + 3$. Then, $N = 2^m - 1 = 7$ and the recurrence condition is $t^3 = -3t^2 - 2t - 3 = t^2 + 2t + 1 \pmod 4$. There are $4^m - 1 = 63$ possible nonzero initial

Table 1. Examples of Some Sequences with Perfect EAC When Sequences Length Is 16

Sequence	Value of A in W_A^i (number of phases)	Value of i in W_A^i															
Frank	4	0	0	0	0	0	1	2	3	0	2	0	2	0	3	2	1
P1	8	0	5	2	7	4	3	2	1	0	1	2	3	4	7	2	5
Golomb	16	0	1	3	6	10	15	5	12	4	13	7	2	14	11	9	8
Px	16	9	3	13	7	3	1	15	13	13	15	1	3	7	13	3	9
P3	32	0	1	4	9	16	25	4	17	0	17	4	25	16	9	4	1
P4	32	0	15	28	7	16	23	28	31	0	31	28	23	16	7	28	15

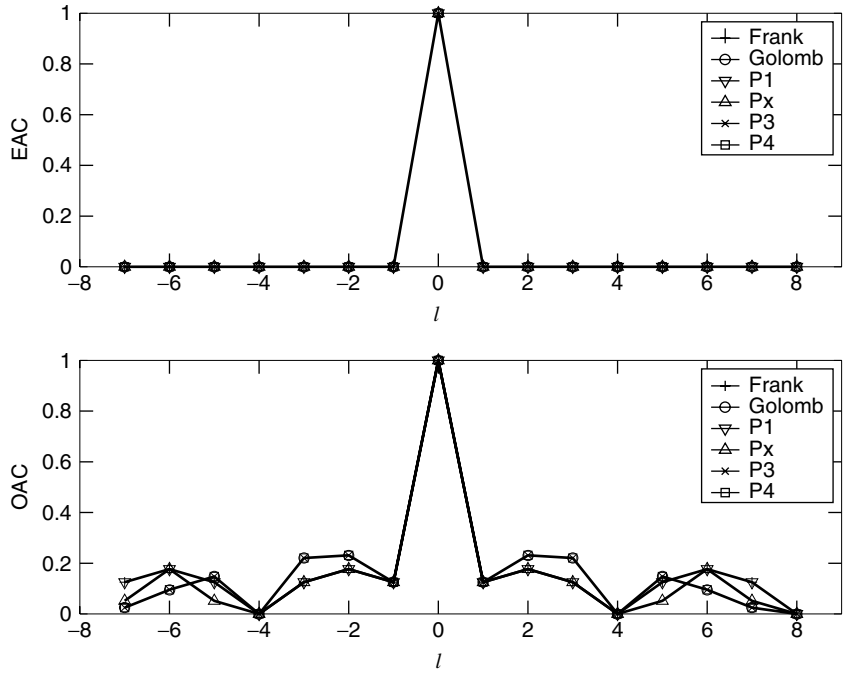


Figure 2. Normalized magnitudes of the AC functions of the sequences in Table 1.

loadings for the mod 4 shift register circuit, but only a collection of $(4^m - 1)/N = N + 2 = 9$ of them will yield cyclically distinct sequences. One of the 9 quaternary sequences is $\mathbf{s} = [1001132]$, and the corresponding four-phase sequence is

$$\mathbf{x} = [W_4^1\ W_4^0\ W_4^0\ W_4^1\ W_4^1\ W_4^3\ W_4^2] \tag{17}$$

The maximum nontrivial correlation magnitude of the four-phase sequence is

$$\max\{\theta_a, \theta_c\} \le 1 + \sqrt{N+1} \tag{18}$$

As in the case of the (binary) Gold sequence, the exact distribution of EC values of the four-phase sequence is known.

3.2.2. Frank–Zadoff–Chu (FZC) Sequence. Let p denote the smallest prime divisor of an odd number N and M_k denote the multiplicative inverse of k mod N, $k = 1, \ldots, p-1$. Then the set $\{\mathbf{x}^{(k)}; k = 1, \ldots, p-1\}$ of $p-1$ FZC sequences [8,10] is defined by

$$\begin{aligned} x_n^{(k)} &= (-1)^{nM_k} W_{2N}^{M_k n^2} \\ &= W_{2N}^{nM_k(n+N)}, \quad n = 0, \ldots, N-1 \end{aligned} \tag{19}$$

When N is prime, there can be as many as $N-1$ FZC sequences. An example of the FZC sequence, when $N = 7$ and $k = 1$, is

$$\mathbf{x}^{(1)} = [W_7^0\ W_7^4\ W_7^2\ W_7^1\ W_7^1\ W_7^2\ W_7^4] \tag{20}$$

The correlation properties of the FZC sequence include

$$\theta(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})(l) = \begin{cases} N, & l = 0 \\ 0, & l \ne 0 \end{cases} \tag{21}$$

for the EAC function, and

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) \le \sqrt{N}, \quad k_1 \ne k_2 \tag{22}$$

for the ECC function.

3.2.3. Generalized Chirplike (GCL) Sequence. Let $\{b_0, b_1, \ldots, b_{m-1}\}$ be a set of m complex numbers all having absolute value 1, and let $\{a_0, a_1, \ldots, a_{N-1}\}$ be a set of $N = sm^2$ numbers defined by

$$a_n = \begin{cases} W_N^{-n^2/2-qn}, & N \text{ even} \\ W_N^{-n(n+1)/2-qn}, & N \text{ odd} \end{cases} \tag{23}$$

where q is an integer and m and s are natural numbers. Then, the GCL sequences [17–21] $\mathbf{x}^{(k)}$, $k = 0, 1, \ldots, N-1$, are defined by

$$x_n^{(k)} = W_N^k a_n b_{n \bmod m}, \quad n = 0, 1, \ldots, N-1. \tag{24}$$

For example, when $s = 2$, $m = 2$ (i.e., $N = 8$), $q = 0$, $k = 0$, and $b_0 = b_1 = 1$, we have

$$\mathbf{x}^{(0)} = [W_{16}^0\ W_{16}^{15}\ W_{16}^{12}\ W_{16}^7\ W_{16}^0\ W_{16}^7\ W_{16}^{12}\ W_{16}^{15}] \tag{25}$$

The EAC function of the GCL sequence is

$$\theta(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})(l) = \begin{cases} N, & l = 0 \\ 0, & l \ne 0 \end{cases} \tag{26}$$

The ECC between two GCL sequences of odd length N, obtained from the two different primitive Nth roots $W_N^{k_1}$ and $W_N^{k_2}$ of unity, is constant if $k_1 - k_2$ is relatively prime to N:

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = \sqrt{N}, \quad N \text{ odd} \tag{27}$$

3.2.4. Lüke Sequence. The construction of the family of Lüke sequences [15] starts with a q-ary m sequence (see Section 3.4) $\mathbf{c} = [c_0\, c_1\, \cdots\, c_{N-1}]$ of length $N = q^r - 1$, where q is prime and $r \geq 2$ is an integer. The Lüke sequence can then be generated by

$$x_n^{(k)} = W_{qN}^{Nc_n + knq}, \quad n, k \in \{0, 1, \ldots, N-1\} \tag{28}$$

For example, when $q = 3$, $r = 2$, $N = 8$, $k = 1$, and $\mathbf{c} = [12022101]$, we have

$$\mathbf{x}^{(1)} = [W_{24}^{8}\ W_{24}^{19}\ W_{24}^{6}\ W_{24}^{1}\ W_{24}^{4}\ W_{24}^{23}\ W_{24}^{18}\ W_{24}^{5}] \tag{29}$$

The EAC function of the Lüke sequence is two-level in magnitude and nearly perfect:

$$\theta(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})(l) = \begin{cases} N, & l = 0 \\ 1, & l \neq 0 \end{cases} \tag{30}$$

Interestingly, the ECC function is also two-level in magnitude:

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = \begin{cases} 0, & l = 0, \\ \sqrt{N+1}, & l \neq 0, \end{cases} \tag{31}$$

for $k_1 \neq k_2$.

3.3. Polyphase Sequences with Perfect ECC

We now consider some polyphase sequences that have perfect ECC properties and are useful as the signature (channelization) sequences in cellular DSCDMA systems.

3.3.1. Park–Park–Song–Suehiro (PS) Sequence. Let $\mathbf{b} = [b_0\, b_1\, \cdots\, b_{N_b-1}]$ be a sequence of length N_b with elements $b_i \in \{W_M^0, \ldots, W_M^{M-1}\}$, $i = 0, 1, \ldots, N_b - 1$, where M is a natural number. Then, the PS sequence [28] $x^{(k)}$ of length $N = KN_b^2$ is defined by

$$x_n^{(k)} = W_N^{nk} \sum_{p=0}^{N_b-1} b_p W_{N_b^2}^{np} \delta(R(n + mp, N_b)) \qquad n = 0, 1, \ldots, N-1, \quad k = 0, 1, \ldots, K-1 \tag{32}$$

where $\delta(\cdot)$ is the Kronecker delta function, K is a natural number, $R(a, b)$ is the remainder when a is divided by b, and m is a natural number less than N_b. If N_b is prime, we have

$$x_n^{(k)} = b_{p_s} W_N^{n(k + Kp_s)} \tag{33}$$

where p_s is the number in $\{0, 1, \ldots, N_b - 1\}$ such that $R(n + mp_s, N_b) = 0$. For example, when $N_b = 2$, $K = 2$, $m = 1$, $k = 1$, and $\mathbf{b} = [W_2^0\ W_2^1]$, we have

$$\mathbf{x}^{(1)} = [W_8^0\ W_8^7\ W_8^2\ W_8^5\ W_8^4\ W_8^3\ W_8^6\ W_8^1] \tag{34}$$

3.3.2. Song–Park (SP) Sequence. The SP sequence [29] of length $N = 2(L + 1)$ is defined by

$$x_n^{(k)} = (-1)^n W_{L+1}^{nk}, \quad n = 0, 1, \ldots, N-1, \tag{35}$$

where the even integer L is the size of the SP sequence. For example, when $L = 2$ and $k = 1$, we have

$$\mathbf{x}^{(1)} = [W_6^0\ W_6^5\ W_6^4\ W_6^3\ W_6^2\ W_6^1] \tag{36}$$

The ECC functions of these two sequences [Eqs. (32) and (35)] are perfect:

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = 0, \quad k_1 \neq k_2 \tag{37}$$

In order to compare the OCC properties of the PS and SP sequences, we show the cumulative frequencies of the normalized maximum magnitude of the OCC (MMO) values in Fig. 3 when the lengths of sequences are almost the same (the lengths of the PS sequence are 36, 126, and 513 and those of the SP sequence are 30, 126, and 510). The distribution curves are obtained by evaluating the MMO for all possible ${}_SC_2$ pairs of sequences, where S is the number of sequences (which depends on N). For example, when $N = 126$, the cumulative frequency for the PS and SP sequences in Fig. 3b is obtained from ${}_{14}C_2$ and ${}_{62}C_2$ values of the MMO, respectively. In this figure, we can clearly see that the MMO of the SP sequence is smaller than that of the PS sequence with high probability. In addition, the cumulative frequency for the SP sequence converges to 1 at a much faster rate than that for the PS sequence.

3.4. Other Polyphase Sequences

Among the other interesting polyphase sequences are the q-phase m sequences, equal OC and EC (EOE) sequences, and generalized Barker sequences.

3.4.1. *q*-Phase *m* Sequences. q-phase m sequences are obtained by expanding the number of phases of binary m sequences. The q-phase m sequence $\mathbf{x}$ of period $N = q^m - 1$ can be defined by $x_i = W_q^{c_i}$, where $\mathbf{c} = [c_0\, c_1\, \cdots\, c_{N-1}]$ is called the "q-ary m sequence." When q is prime, the generating polynomial

$$g(t) = g_m t^m + g_{m-1} t^{m-1} + \cdots + g_1 t + g_0 \tag{38}$$

for q-ary m sequences is a primitive polynomial of degree m. Here, $g_i \in \{0, 1, \ldots, q-1\}$, $g_m \neq 0$, and $g_0 \neq 0$. The recurrence condition of q-ary m sequences $\mathbf{c}$ is

$$c_{i+m} = g_{m-1} c_{i+m-1} + g_{m-2} c_{i+m-2} + \cdots + g_1 c_{i+1} + c_i \pmod q, \text{ for } i \geq 0 \tag{39}$$

For example, let $q = 3$, $m = 2$, and $g(t) = t^2 + 2t + 2$. Then, the recurrence becomes $t^2 = -2t - 2 = t + 1 \pmod 3$. Using the initial loading of 21 in the registers yields the ternary sequence of period $N = 3^2 - 1 = 8$ as $\mathbf{c} = [c_0\, c_1\, \cdots\, c_{N-1}] = [12022101]$, and a three-phase m sequence can be obtained as

$$\mathbf{x} = [W_3^1\ W_3^2\ W_3^0\ W_3^2\ W_3^2\ W_3^1\ W_3^0\ W_3^1] \tag{40}$$

The EAC function of q-phase m sequences is

$$\theta(\mathbf{x}, \mathbf{x})(l) = \begin{cases} N, & l = 0 \\ -1, & l \neq 0 \end{cases} \tag{41}$$

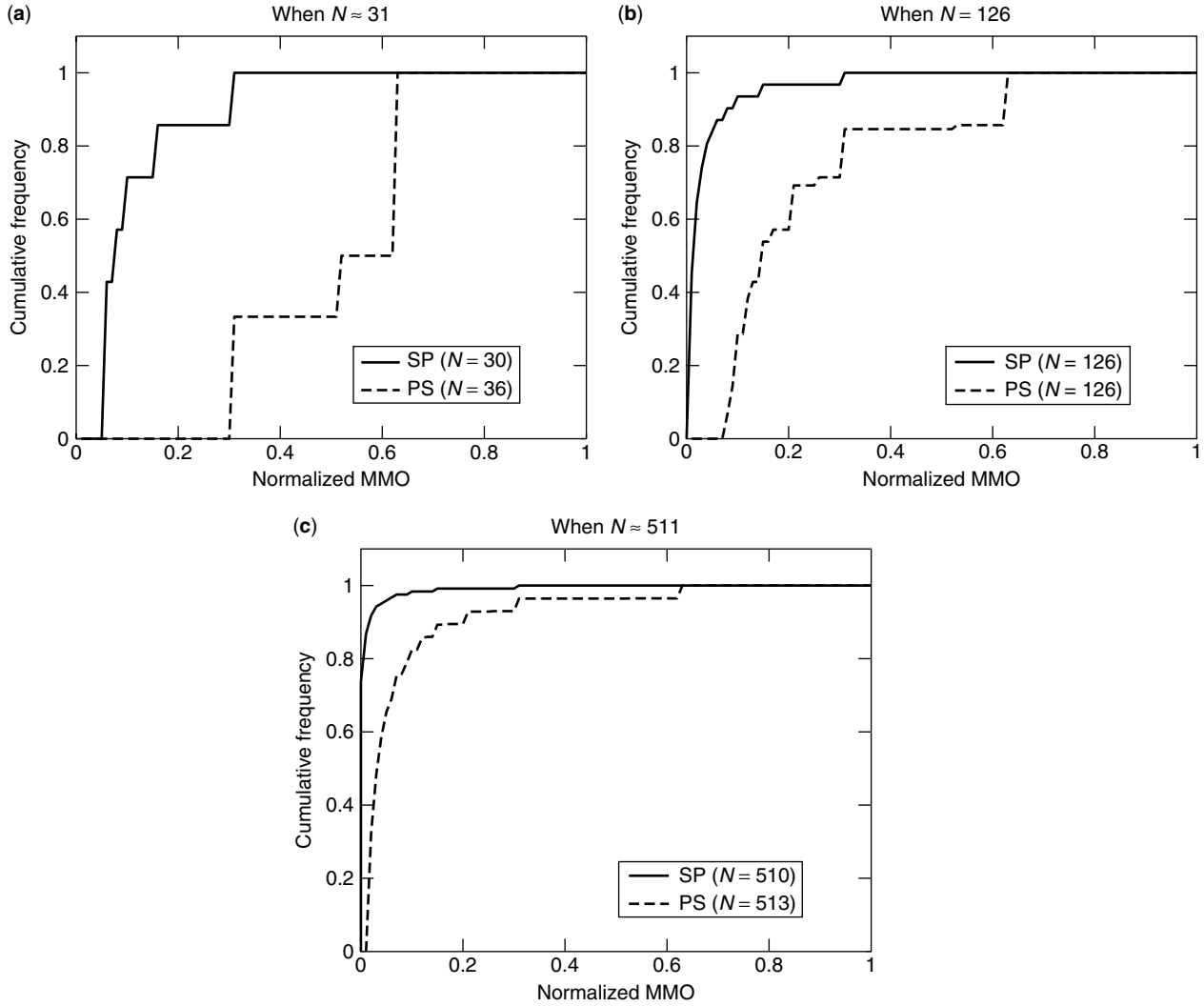


Figure 3. Cumulative frequency of the normalized MMO values of the PS and SP sequences.

3.4.2. EOE Sequences. When $\mathbf{u} = [u_0\, u_1 \cdots u_{N-1}]$ is an arbitrary real-valued sequence of period N,p is an odd integer and β is a real constant satisfying $0 \le \beta < 2\pi$, the complex-valued sequence $\mathbf{x} = [x_0\, x_1 \cdots x_{N-1}]$ defined by

$$x_n = u_n e^{\sqrt{-1}\beta} W_{4N}^{pn}, \quad n = 0, 1, \ldots, N-1 \tag{42}$$

is an EOE sequence [20]. For example, when $N = 7, p = 7$, $\beta = 0$, and $\mathbf{u} = [1\ -1\ -1\quad 1\ -1\ 1\ 1]$ is a Gold sequence, an EOE–Gold sequence with length 7 can be obtained as

$$\begin{aligned}\mathbf{x} &= [W_4^0\ -W_4^1\ -W_4^2\ W_4^3\ -W_4^0\ W_4^1\ W_4^2] \\ &= [W_4^0\ W_4^3\ W_4^0\ W_4^3\ W_4^2\ W_4^1\ W_4^2]\end{aligned} \tag{43}$$

The magnitudes of the OC and EC functions of these sequences are equal:

$$|\theta(\mathbf{x}, \mathbf{y})(l)| = |\hat{\theta}(\mathbf{x}, \mathbf{y})(l)| \tag{44}$$

In addition, we have

$$|\theta(\mathbf{x}, \mathbf{y})(l)| = |\hat{\theta}(\mathbf{x}, \mathbf{y})(l)| \le \max\{\theta(\mathbf{u}, \mathbf{v})(l), \hat{\theta}(\mathbf{u}, \mathbf{v})(l)\} \tag{45}$$

where $\mathbf{y} = [y_0\, y_1 \cdots y_{N-1}]$ denotes an EOE sequence defined with $\mathbf{v} = [v_0\, v_1 \cdots v_{N-1}]$:

$$y_n = v_n e^{\sqrt{-1}\beta} W_{4N}^{pn}, \quad n = 0, 1, \ldots, N-1 \tag{46}$$

3.4.3. Generalized Barker Sequences. A polyphase sequence $\mathbf{x}$ is called a "generalized Barker sequence" [9,24] if the partial AC function satisfies $C(\mathbf{x}, \mathbf{x})(l) \le 1$ for $l \ne 0$. Generalized Barker sequences are widely used as a synchronization sequence because the partial AC property is so good. Let $\mathbf{y} = [y_0\, y_1 \cdots y_{N-1}]$ be a p-phase sequence of length N. Then the q-phase generalized Barker sequence $\mathbf{x}$ of length N is defined by

$$x_n = y_n W_M^n \tag{47}$$

where M is a nonzero integer and q is the least common multiple of p and M. It is easy to see that the partial AC of $\mathbf{x}$ is

$$C(\mathbf{x}, \mathbf{x})(l) = W_M^{-n} C(\mathbf{y}, \mathbf{y})(l), \quad \text{for all } l \tag{48}$$

In particular, we have $|C(\mathbf{x},\mathbf{x})(l)| = |C(\mathbf{y},\mathbf{y})(l)|$ for all l as well as $|x_n| = |y_n|$ since $|W_M^{-n}| = 1$. As a special case, if $|C(\mathbf{y},\mathbf{y})(l)| \le 1$ for all l and $y_n \in \{1,-1\}$, we can obtain a polyphase generalized Barker sequence of length N from a binary Barker sequence of length N. For example, using a binary Barker sequence $\mathbf{y} = [1\ 1\ 1\ -1\ -1\ 1\ -1]$ with length 7, we obtain a four-phase generalized Barker sequence with length 7 as

$$\mathbf{x} = [W_4^0\ W_4^1\ W_4^2\ W_4^1\ W_4^2\ W_4^1\ W_4^0] \tag{49}$$

It has been shown that the binary Barker sequences with odd length $N > 13$ and even length $4 < N < 189884$ do not exist: on the other hand, polyphase sequences of length $N > 13$ satisfying the Barker condition can be found; for example, the 180-phase generalized Barker sequence with length 36 has been reported [24].

4. CONCLUSION

In this article, we have reviewed some polyphase sequences and described their applications to cellular DSCDMA systems based on the correlation properties. The Frank, Golomb, P1, Px, P3, and P4 sequences have a perfect EAC property so that they can be used as scrambling sequences in intercell synchronous cellular DSCDMA systems. The four-phase, FZC, GCL, and Lüke sequences, whose EC functions are (near) optimum in the sense of the lower bound for polyphase sequences, are useful as the scrambling sequences in intercell asynchronous cellular DSCDMA systems. Next, the PS and SP sequences are suitable for the short signature (channelization) sequences in cellular DSCDMA systems because of their perfect ECC properties. Some other polyphase sequences such as the q-phase m, EOE, and generalized Barker sequences are also described briefly.

Acknowledgments
This research was supported by Korea Science and Engineering Foundation (KOSEF) under Grant R01-2000-000-00259-8, for which the authors would like to express their thanks.

BIOGRAPHIES

So Ryoung Park received the B.S. degree in Electronics Engineering from Yonsei University, Seoul, Korea, in 1997, and the M.S.E. and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1999 and 2002, respectively. She was a Research Assistant at the Department of Electrical Engineering, KAIST, from 1997 to 2001 and is now a Research Scientist at the Statistical Signal Processing Laboratory, Department of Electrical Engineering and Computer Science, KAIST. She was the recipient of a Gold Prize at the Samsung Humantech Paper Contest in 2001. Her current research interests are in mobile communications and statistical signal processing with emphasis on spread-spectrum communications.

Iickho Song received the B.S.E. (magna cum laude) and M.S.E. degrees in Electronics Engineering from Seoul National University, Seoul, Korea, in 1982 and 1984, respectively, and the M.S.E. and Ph.D. degrees in Electrical Engineering from University of Pennsylvania, Philadelphia (USA), in 1985 and 1987, respectively. In 1988, he joined KAIST, as an Assistant Professor, where he has been a full Professor since 1998. He is a Fellow and Chartered Engineer of the Institution of Electrical Engineers, and a Senior Member of the Institute of Electrical and Electronics Engineers. He has coauthored a book (Advanced Theory of Signal Detection, Springer, 2002) and more than 150 papers in the area of signal detection and estimation and of code design and acquisition for CDMA systems. He received a number of awards, including the Young Scientists Award presented by the President of the Republic of Korea and four Academic Awards from the Korean Institute of Communication Sciences. He has been listed in many "who's who" books, including *Marquis Who's Who in the World* and *Marquis Who's Who in Science and Engineering*. His research interest include detection and estimation, statistical signal processing, and CDMA communications.

BIBLIOGRAPHY

1. R. A. Scholtz, The origins of spread-spectrum communications, *IEEE Trans. Commun.* **COM-30**: 822–854 (1982).
2. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications: a tutorial, *IEEE Trans. Commun.* **COM-30**: 855–884 (1982).
3. H. Ochsner, Direct-sequence spread-spectrum receiver for communication on frequency-selective fading channels, *IEEE J. Select. Areas Commun.* **JSAC-5**: 188–193 (1987).
4. E. Geraniotis and B. Ghaffari, Performance of binary and quaternary direct-sequence spread-spectrum multiple-access systems with random signature sequences, *IEEE Trans. Commun.* **COM-39**: 713–724 (1991).
5. P. G. Flikkema, Spread-spectrum techniques for wireless communication, *IEEE Signal Process. Mag.* **14**: 26–36 (1997).
6. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* **36**: 48–54 (1998).
7. R. Gold, Maximal recursive sequences with 3-valued recursive crosscorrelation functions, *IEEE Trans. Inform. Theory* **IT-14**: 154–156 (1968).
8. R. L. Frank and S. A. Zadoff, Phase shift pulse codes with good periodic correlation properties, *IRE Trans. Inform. Theory* **IT-8**: 381–382 (1962).
9. S. W. Golomb and R. A. Scholtz, Generalized Barker sequences, *IEEE Trans. Inform. Theory* **IT-11**: 533–537 (1965).
10. D. C. Chu, Polyphase codes with good periodic correlation properties, *IEEE Trans. Inform. Theory* **IT-18**: 531–532 (1972).
11. B. L. Lewis and F. F. Kretschmer, Jr., A new class of polyphase pulse compression codes and techniques, *IEEE Trans. Aerospace. Electron. Syst.* **AES-17**: 364–372 (1981).
12. B. L. Lewis and F. F. Kretschmer, Jr., Linear frequency modulation derived polyphase pulse compression codes, *IEEE Trans. Aerospace. Electron. Syst.* **AES-18**: 637–641 (1982).

13. B. L. Lewis and F. F. Kretschmer, Jr., Doppler properties of polyphase coded pulse compression waveform, *IEEE Trans. Aerospace. Electron. Syst.* **AES-19**: 521–531 (1983).

14. N. Suehiro and M. Hatori, Modulatable orthogonal sequences and their application to SSMA systems, *IEEE Trans. Inform. Theory* **IT-34**: 93–100 (1988).

15. H. D. Lüke, Families of polyphase sequences with near-optimal two-valued auto- and cross-correlation functions, *Electron. Lett.* **28**: 1–2 (1992).

16. S. Boztas, R. Hammons, and P. V. Kumar, 4-phase sequences with near-optimum correlation properties, *IEEE Trans. Inform. Theory* **IT-38**: 1101–1113 (1992).

17. B. M. Popović, Generalized chirp-like polyphase sequences with optimum correlation properties, *IEEE Trans. Inform. Theory* **IT-38**: 1406–1409 (1992).

18. N. Zhang and S. W. Golomb, Polyphase sequence with low autocorrelations, *IEEE Trans. Inform. Theory* **IT-39**: 1085–1089 (1993).

19. B. M. Popović, GCL polyphase sequences with minimum alphabets, *Electron. Lett.* **30**: 106–107 (1994).

20. H. Fukumasa, R. Kohno, and H. Imai, Design of pseudonoise sequences with good odd and even correlation properties for DS/CDMA, *IEEE J. Select. Areas Commun.* **JSAC-12**: 828–836 (1994).

21. B. M. Popović, Efficient matched filter for the generalized chirp-like polyphase sequences, *IEEE Trans. Aerospace. Electron. Syst.* **AES-30**: 769–777 (1994).

22. P. Z. Fan and M. Darnell, Aperiodic autocorrelation of Frank sequences, *IEE Proc. Commun.* **142**: 210–215 (1995).

23. E. M. Gabidulin, P. Z. Fan, and M. Darnell, Autocorrelation of Golomb sequences, *IEE Proc. Commun.* **142**: 281–284 (1995).

24. S. W. Golomb and M. Z. Win, Recent results on polyphase sequences, *IEEE Trans. Inform. Theory* **IT-44**: 817–824 (1998).

25. P. B. Rapajic and R. A. Kennedy, Merit factor based comparison of new polyphase sequences, *IEEE Commun. Lett.* **2**: 269–270 (1998).

26. S. I. Park et al., A noise reduction method for a modulated orthogonal sequence under impulsive noise environments, *IEICE Trans. Fund.* **E82A**: 2259–2265 (1999).

27. S. I. Park, S. R. Park, I. Song, and S. Yoon, On the generation and analysis of a modulated orthogonal sequences, *Signal Process.* **80**: 451–464 (2000).

28. S. I. Park, S. R. Park, I. Song, and N. Suehiro, Multiple access interference reduction for QS-CDMA systems with a novel class of polyphase sequences, *IEEE Trans. Inform. Theory* **IT-46**: 1448–1458 (2000).

29. S. R. Park, I. Song, S. Yoon, and J. Lee, A new polyphase sequence with perfect even and good odd crosscorrelation functions for DS/CDMA systems, *IEEE Trans. Vehic. Technol.* **VT-51** (in press).

30. T. M. Lok and J. S. Lehnert, An asymptotic analysis of DS/SSMA communication systems with random polyphase signature sequences, *IEEE Trans. Inform. Theory* **IT-42**: 129–136 (1996).

31. S. R. Park, I. Song, S. Yoon, and S. Y. Kim, A statistical analysis of random polyphase signature sequences in multipath fading DS-CDMA channels, *Signal Process.* **81**: 2461–2477 (2001).

32. M. B. Pursley, Performance evaluation for phase-coded spread-spectrum multiple-access communication, *IEEE Trans. Commun.* **COM-25**: 795–803 (1977).

33. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudo random and related sequences, *Proc. IEEE* **68**: 593–618 (1980).

34. F. W. Sun and H. Leib, Optimal phases for a family of quadriphase CDMA sequences, *IEEE Trans. Inform. Theory* **IT-43**: 1205–1217 (1997).

35. D. V. Sarwate, Bounds on crosscorrelation of sequences, *IEEE Trans. Inform. Theory* **IT-25**: 720–724 (1979).

36. K. G. Paterson and P. J. G. Lothian, Bounds on partial correlations of sequences, *IEEE Trans. Inform. Theory* **IT-44**: 1164–1175 (1998).

37. S. R. Park, I. Song, and H. Kwon, DS/CDMA signature sequences based on PR-QMF banks, *IEEE Trans. Signal Process.* **SP-50** (in press).

POWER CONTROL IN CDMA CELLULAR COMMUNICATION SYSTEMS

MATTI RINTAMÄKI
Helsinki University of Technology
Helsinki, Finland

1. INTRODUCTION

Transmitter power control (TPC) is vital for capacity and performance in cellular communication systems, where high interference is always present as a result of frequency reuse. The basic intent is to control the transmitter powers in such a way that the interference power from each transmitter to other cochannel users (users that share the same radio resource simultaneously) is minimized while preserving sufficient quality of service (QoS) among all users. Cochannel interference management is important in any system employing frequency reuse. However, in CDMA there are interfering users both inside and outside a cell, which makes CDMA interference limited. Thus efficient TPC is essential in CDMA,[1] especially in the uplink (from mobile to base station communication).

Consider the situation depicted in Fig. 1. Mobile stations MS1 and MS2 share the same frequency band, and their signals are separable at the base station (BS) by their unique spreading codes. The link attenuation of MS2 a particular time instant might be several tens of decibels greater than that from MS1 to BS. If power control is not applied, the signal of MS1 will overpower the signal of MS2 at the base station. This is called the *near–far effect*. To mitigate this effect, power control aims to set the

[1] This applies to direct-sequence CDMA (DS-CDMA). In frequency-hopping CDMA (FHCDMA) the intracell interference can be made very small. In this article we concentrate on DS-CDMA, where transmitter power control is more critical. Thus throughout the rest of the text, DS-CDMA is referred to simply as CDMA.

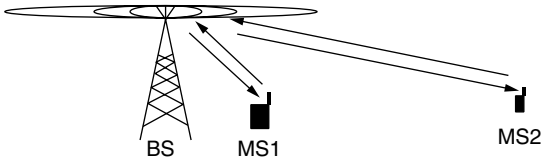


Figure 1. Near–far effect.

transmitter powers of MS1 and MS2 so that both signals are received at the same power level at the base station.

1.1. Radio-Link Attenuation

The attenuation of a radio signal can be modeled as a product of three effects: path loss, shadowing, and multipath fading. *Path loss* is the large-scale distance-dependent attenuation in the average signal power, which can be modeled as r^{-d}, where r is the distance in meters and d is the path loss exponent with typical values ranging from 2 to 5. *Shadowing* is the medium-scale attenuation, which is caused by diffraction and shielding phenomena emanating from terrain variations. This results in relatively slow variations in the mean signal power over a distance of a few tens of wavelengths, caused by reflections, refractions, and diffractions of the signal from buildings, trees, rocks, and other objects. It can be modeled as a lognormally distributed random variable with zero decibels mean and a standard deviation of typically 4–12 dB. *Multipath fading* is the rapid fluctuation in the received signal power that is caused by the constructive and destructive addition of the signals that propagate through different paths with different delays from the transmitter to the receiver. This is usually modeled as a complex Gaussian process, resulting in a Rayleigh-distributed envelope and a classic Doppler spectrum. If a line-of-sight signal is present, the Rice distribution is a better model. Also Nakagami distribution has been widely used to model multipath fading.

With all the three effects included, the attenuation g of the signal power in linear scale can be modeled as

$$g = r^{-d} \times 10^{\zeta/10} \times A_f \tag{1}$$

where ζ is a Gaussian random variable with zero mean and a standard deviation of 4–12 and A_f is a random variable such that $(A_f)^{1/2}$ is either Rayleigh-, Rice- or Nakagami distributed with a classic Doppler spectrum.

As can be understood from above, the signal attenuation is a random variable. Thus, when power control is applied, it must adapt to the changing attenuation of the desired signals, as well as the changing interference conditions, since the attenuations of the cochannel users' signals are also changing, and those signals are power-controlled as well.

1.2. Uplink Versus Downlink Power Control

In CDMA the uplink transmission creates a near–far situation if power control is not used. This occurs because the signals of the different mobile stations propagate through different paths before reaching their serving base station. The task of power control is thus to vary the transmitter powers in order to compensate for the varying channel attenuations, so that the signals from the different mobile stations are received with equal powers at the base station. The requirement of the dynamic range of uplink power control can be of the order of 80 dB.

In downlink there is no near–far situation, since all signals transmitted by a base station propagate through the same path before reaching a mobile station. These signals can be made essentially orthogonal by using proper spreading codes. However, unnecessary high powers can be provided to mobile users near the cell border, thus creating unnecessarily high interference to the neighboring cells. Therefore, downlink power control is used to reduce this interference. The dynamic range of downlink power control is usually much smaller than in uplink, typically on the order of 20–30 dB. This limitation is because a high dynamic range could produce a near–far situation for the mobile stations, since the signals cannot be made perfectly orthogonal.

1.3. Quality Measures for Power Control

A great deal of the work on power control in CDMA cellular systems has focused on how to set the transmitter powers so that all users in the system have acceptable *bit energy-to-interference spectral density ratios* (E_b/I_0). This approach is based on the fairly reasonable assumption that the bit error probability (BEP) at a receiver is a strictly monotonically decreasing function of E_b/I_0. For instance, BEP in an *additive white Gaussian noise* (AWGN) channel decreases very quickly with increasing E_b/I_0. E_b/I_0 is closely related with another measure, namely, the *signal-to-interference ratio* (SIR), such that

$$\frac{E_b}{I_0} = \mathrm{SIR}\frac{W}{r} \tag{2}$$

where W is the transmission bandwidth in hertz and r is the data rate in bits per second (bps). The quantity W/r is called the *processing gain*. When the data rate is fixed, the SIR differs from E_b/I_0 by merely a scaling factor.

In digital communication systems the information to be transmitted is arranged in strings of bits called *frames*, and error correction coding is applied to each frame to further decrease the BEP after decoding. A frame is useless if there are still bit errors in the frame after decoding, and it must be discarded. Hence, depending on the service, a sufficiently low frame erasure rate (FER) must be guaranteed. However, long time delays are needed to obtain reliable estimates of BEP or FER. Since the channel conditions can change very rapidly, these delays might be unacceptably long in practice. Hence most of the attention in the power control field has been on SIR-based algorithms. Also algorithms based on received signal strength have been used, but it has been shown that SIR-based power control offers better performance [1]. The FER information can be used to adjust the target SIR, which a fast TPC algorithm is trying to achieve. This increases system capacity, since a worst-case setting of the SIR target is not required.

In addition to the SIR and FER requirements, the delay or *latency* requirements must be taken into account. For instance, a voice service tolerates a certain amount of data loss but is delay-critical, whereas file downloads do not tolerate bit errors at all (erroneous frames must be retransmitted), but the transmission need not be continuous and must only satisfy some delay limit on the average. The delay tolerances can be taken advantage of in the design of power control algorithms for non-real-time services.

2. THE SIR BALANCING PROBLEM

A widely studied approach to transmitter power control is the SIR balancing problem, namely, how to set the transmitter powers so that all users in the system have equal SIRs. This method is applicable for circuit-switched real-time services such as voice, where the data rate is fixed.

Figure 2 illustrates a simple two-cell CDMA system. We consider only uplink, but the same analysis applies in downlink. The mobile stations MS1 $\cdots$ MS3 share a common frequency band, and their signals are separable at the base stations by their unique spreading codes. The link attenuation from mobile j to base station i is denoted by g_{ij} and is assumed to be fixed in the analysis. In this "snapshot" approach [2,3], it is assumed that the link attenuations change slowly enough compared to the power control dynamics, and can thus be assumed constant. This assumption is reasonable if multipath fading is neglected, but generally leads to optimistic results.

Define a base station assignment function $b(i)$ so that $k = b(i)$ if mobile i is served by base station k. Then the uplink SIR requirement for user i can be expressed as

$$\gamma_i = \frac{g_{b(i)i}p_i}{\sum_{j=1, j\neq i}^{N} g_{b(i)j}p_j + \eta_i} = \frac{p_i}{\sum_{j=1, j\neq i}^{N} \frac{g_{b(i)j}}{g_{b(i)i}}p_j + \frac{\eta_i}{g_{b(i)i}}} \geq \gamma_i^t \qquad (3)$$

where γ_i is the SIR at the receiver of mobile station i, p_i is the transmission power of mobile i, N is the number of mobile stations using the same channel including the intracell and intercell users, η_i is the receiver noise power, and γ_i^t is the uplink SIR requirement for user i. Define the vectors $\mathbf{p} = \{p_i\}$ and $\boldsymbol{\eta} = \{\gamma_i^t \eta_i / g_{b(i)i}\}$ and matrix $\mathbf{H} = \{H_{ij}\}$ with elements $H_{ij} = \gamma_i^t g_{b(i)j}/g_{b(i)i}$ when $i \neq j$ and $H_{ii} = 0$. Now we can put (3) in matrix form:

$$(\mathbf{I} - \mathbf{H})\mathbf{p} \geq \boldsymbol{\eta} \qquad (4)$$

where $\mathbf{I}$ denotes the identity matrix and the inequality holds componentwise. A minimum-power solution corresponds to the case where (4) is satisfied with equality. This is desirable, as it saves energy and thus prolongs the mobile station battery life.

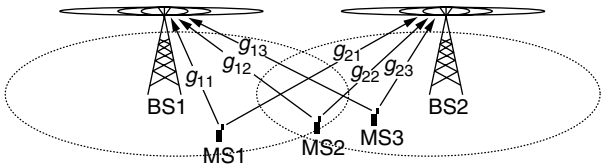


Figure 2. Illustration of a two-cell CDMA system with three mobile stations.

Definition 1. The problem is said to be feasible if there exists a nonnegative power vector $\mathbf{p}$ such that condition in (4) is satisfied.

It has been shown that balancing the SIRs and making the balanced SIR as high as possible maximizes the minimum SIR in all links. Consider the power control problem in (4) with $\gamma_i^t = \gamma^t$. Define matrix $\mathbf{A}$ so that $\mathbf{H} = \gamma^t \mathbf{A}$. If the interference in the receiver is sufficiently high that the receiver noise can be neglected, the problem can be identified as an eigenvalue problem:

$$\mathbf{p} = \mathbf{H}\mathbf{p} \Leftrightarrow \frac{1}{\gamma^t}\mathbf{p} = \mathbf{A}\mathbf{p} \qquad (5)$$

The maximum possible value for γ^t (denoted by γ^*) is equal to the inverse of the maximum (real) eigenvalue of $\mathbf{A}$. The corresponding positive eigenvector $\mathbf{p}^*$ is the power vector achieving this maximum.

If receiver noise cannot be neglected, the optimal power vector is a solution to a set of linear equations:

$$\mathbf{p}^* = \mathbf{H}\mathbf{p}^* + \boldsymbol{\eta} \qquad (6)$$

Proposition 1 [5]**.** The power control problem in (6) is feasible if the largest eigenvalue of the matrix $\mathbf{H}$, denoted by $\rho(\mathbf{H})$, is less than or equal to one.

Note that the case $\rho(\mathbf{H}) = 1$ can only be met if the receiver noise is zero, otherwise infinite transmitter power would be required. Moreover, in practice there always exists an upper limit for the transmitter power.

Using the optimal power vector $\mathbf{p}^*$ results in all users in the system having the same SIR. If the power control problem is not feasible, this results in the disastrous case that none of the users achieve the SIR requirement. To prevent this, a removal strategy must be employed, which removes transmitters from the channel until the power control problem becomes infeasible. The problem is to determine which transmitters to remove. An intuitive approach is to remove those transmitters that produce largest interference, that is, transmitters having the worst link quality. Several removal strategies have been investigated in the literature [4]. Note that a removal of a transmitter from a channel does not necessarily mean that the connection is broken, but it can be handed over to another channel.

3. DISTRIBUTED POWER CONTROL

Solving (6) directly is a centralized method, since it requires the information of all the link attenuations and receiver noise values in the system. This is generally not suitable in real implementations, since it would require extensive signaling overhead. However, it is valuable in determining upper bounds for the performance of distributed algorithms that can be implemented in practice.

A distributed algorithm uses only local measurements to update the transmitter powers. Hence it is more suitable for practical implementation than a centralized algorithm. Since in this case a user does not know the link attenuations, the problem must be iteratively solved. It is thus necessary to find an iteration that depends only on local measurements, and converges to the optimal solution reasonably soon (sooner than the link gains change). Fast convergence can be achieved in two ways: by making the iteration time step smaller, and by designing an iteration with faster convergence property.

A general iterative algorithm to solve the problem in (6) can be found from numerical linear algebra, and is given by [5]

$$\mathbf{p}(k+1) = \mathbf{M}^{-1}\mathbf{N}\mathbf{p}(k) + \mathbf{M}^{-1}\boldsymbol{\eta}, \quad k = 0, 1, \ldots \tag{7}$$

where $\mathbf{M}$ and $\mathbf{N}$ are matrices such that $\mathbf{p}^* = \mathbf{M}^{-1}\mathbf{N}\mathbf{p}^* + \mathbf{M}^{-1}\boldsymbol{\eta}$. By selecting $\mathbf{M}$ and $\mathbf{N}$ properly, the iteration in (7) will converge so that $\lim_{k\to\infty} \mathbf{p}(k) = \mathbf{p}^*$. For example, if we select $\mathbf{M} = \mathbf{I}$ and $\mathbf{N} = \mathbf{H}$, we get

$$\mathbf{p}(k+1) = \mathbf{H}\mathbf{p}(k) + \boldsymbol{\eta}, \quad k = 0, 1, \ldots \tag{8}$$

Looking at this equation componentwise, we have

$$\begin{aligned} p_i(k+1) &= \gamma_i^t \left(\sum_{j=1, j\neq i}^{N} \frac{g_{b(i)j}}{g_{b(i)i}} p_j(k) + \frac{\eta_i}{g_{b(i)i}} \right) \\ &= \frac{\gamma_i^t}{\gamma_i(k)} p_i(k), \quad k = 0, 1, \ldots \end{aligned} \tag{9}$$

where $\gamma_i(k)$ and $p_i(k)$ are the received SIR and transmission power of transmitter i at iteration k, respectively. This is referred to as the *distributed power control* (DPC) algorithm. It was proposed by Foschini and Miljanic [6]. It is totally distributed, since it depends only on local measurements of the SIR.

3.1. Convergence of the Iterative Algorithm

A necessary and sufficient condition for the iteration in (7) to converge is the following [5]. Let $\alpha_1, \alpha_2, \ldots$ be the eigenvalues of the matrix $\mathbf{M}^{-1}\mathbf{N}$. Then the iteration converges if and only if $\max_k |\alpha_k| < 1$. Consider the DPC algorithm in (9). In this case we have $\mathbf{M}^{-1}\mathbf{N} = \mathbf{H}$. Hence the dominant eigenvalue of $\mathbf{H}$, $\rho(\mathbf{H})$, should be less than one. By proposition 1, this ensures that the SIR requirements are satisfied for all users. Hence DPC converges to $\mathbf{p}^*$ whenever the SIR requirements can be satisfied.

The speed of the convergence is very important, since the link attenuations are changing all the time. For the iteration in (7), it can be shown that the smaller the $\rho(\mathbf{M}^{-1}\mathbf{N})$, the faster the convergence. Hence the task is to find $\mathbf{M}$ and $\mathbf{N}$ such that $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$ and as small as possible.

3.2. Convergence Using Standard Interference Functions

A different way of proving convergence of iterative algorithms was developed by Yates [7] and extended by Huang and Yates [8]. There the iteration is formulated by defining an *interference function* $I(\mathbf{p})$ such that

$$\mathbf{p}(k+1) = I(\mathbf{p}(k)) \tag{10}$$

The standard interference function framework gives a *sufficient* but not *necessary* condition for convergence of the iteration in (7). The following definition and proposition summarize this framework.

Definition 2 [7]**.** An interference function $I(\mathbf{p})$ is called "standard" if for all nonnegative power vectors

$$\begin{aligned} &I(\mathbf{p}) > 0 \\ &\mathbf{p} \geq \mathbf{p}' \Rightarrow I(\mathbf{p}) \geq I(\mathbf{p}') \\ &\forall \alpha > 1, \quad \alpha I(\mathbf{p}) > I(\alpha\mathbf{p}) \end{aligned} \tag{11}$$

Proposition 2 [7]**.** If the power control problem is feasible and $I(\mathbf{p})$ is a standard interference function, then for any initial nonnegative power vector $\mathbf{p}$ the iteration in (10) converges to the unique nonnegative fixed point $\mathbf{p}^*$.

3.3. Distributed Constrained Power Control

In all practical systems the transmitter powers are limited so that

$$\mathbf{0} \leq \mathbf{p} \leq \mathbf{p}_{\max} \tag{12}$$

where $\mathbf{0}$ is a vector with all-zero elements and $\mathbf{p}_{\max} = [p_i^{\max}, p_2^{\max}, \ldots, p_N^{\max}]^T$ denotes the maximum transmitter power of each transmitter. To take these limitations into account, the distributed constrained power control (DCPC) algorithm was suggested by Grandhi et al. [9]. It is defined by

$$p_i(k+1) = \min\left(p_i^{\max}, \frac{\gamma_i^t}{\gamma_i(k)} p_i(k) \right), \quad k = 0, 1, \ldots \tag{13}$$

where $p_i^{\max}$ is the maximum allowed transmitter power of transmitter i. With DCPC some transmitters can transmit with the maximum power, thus producing maximum interference to other users, but still not achieving their SIR target. Therefore it might be beneficial to lower the transmission power when link quality is poor. With this in mind, a following more general algorithm has been proposed (see Ref. 5 and the references cited therein) that has DCPC as a special case:

$$p_i(k+1) = \begin{cases} \dfrac{\gamma_i^t}{\gamma_i(k)} p_i(k) & \text{if, } \dfrac{\gamma_i^t}{\gamma_i(k)} p_i(k) \leq p_i^{\max} \\ p_i' & \text{if, } \dfrac{\gamma_i^t}{\gamma_i(k)} p_i(k) > p_i^{\max} \end{cases} \tag{14}$$

where $0 \leq p_i' \leq p_i^{\max}$. It can be shown that this algorithm converges to the optimal power vector $\mathbf{p}^*$ provided that the system in (6) has the optimal solution $\mathbf{p}^*$ in the power range given by (12).

3.4. A Two-User Example

Consider a system with only two mobile stations MS1 and MS2 and two base stations BS1 and BS2. Assume that

MS1 is connected to BS1 and MS2 is connected to BS2. In this case the power control problem is the following:

$$\begin{cases} \gamma_1 = \dfrac{g_{11}p_1}{g_{12}p_2 + \eta_1} \geq \gamma_1^t \\ \gamma_2 = \dfrac{g_{22}p_2}{g_{21}p_1 + \eta_2} \geq \gamma_2^t \end{cases} \Rightarrow \begin{cases} p_1 \geq \gamma_1^t \left(\dfrac{g_{12}}{g_{11}} p_2 + \dfrac{\eta_1}{g_{11}} \right) \\ p_2 \geq \gamma_2^t \left(\dfrac{g_{21}}{g_{22}} p_1 + \dfrac{\eta_2}{g_{22}} \right) \end{cases} \quad (15)$$

This situation is depicted in Fig. 3. The feasible region is shaded in the figure, and the optimal (minimum power) solution $\mathbf{p}^*$ is in the intersection of the two lines. Since $\mathbf{p}^*$ is within the maximum power limits, the problem is feasible. Consider that user 1 raises its target SIR γ_1^t while the link attenuations remain unchanged. It is then necessary for it to also raise its transmitter power as seen from (15). This, in turn, forces user 2 to raise its transmitter power. Thus it can happen that the optimal point $\mathbf{p}^*$ moves outside the maximum power limits, thereby making the problem unfeasible.

4. PRACTICAL ISSUES ON POWER CONTROL

A number of practical issues limit the implementation of the theoretical algorithms:

- *SIR Estimation*. The received SIR is not known exactly at a receiver, but it must be estimated, and thus there will always be some estimation error.
- *TPC Update Rate*. In CDMA one has to deal with the near–far situation, and thus the update rate of the power control algorithm must be sufficiently high that the variations in the link attenuation can be tracked. Typical update rates are from 800 Hz (used in the IS95 system [11]) to 1500 Hz (used in WCDMA [10]).
- *Feedback Information Accuracy*. The information of the SIR at the receiver should somehow be communicated to the transmitter. An accurate representation of the SIR measurement requires several bits, but this requires more signaling overhead. A usual case in practice is that only one bit is used to inform the transmitter to either increase or decrease its transmitter power by a fixed amount, typically by 1 dB (e.g., in IS95).
- *Loop Delay*. The loop delay greatly affects the performance of a power control algorithm. This delay comes from the SIR measurement process, the transmission of the SIR information over the radio channel, the processing of the SIR information to calculate and adjust the transmission power, and the propagation time after which the new transmission power affects the next SIR measurement. Therefore the power update is based on outdated information of the received SIR. If the channel variation is fast in comparison to the loop delay, the TPC algorithm cannot track the variations.
- *Errors in the Transmission of Feedback Information*. To minimize the loop delay, the TPC command bits are sent without error correction coding. Hence the probability of receiving an erroneous command can be relatively high (e.g., 5–10%).

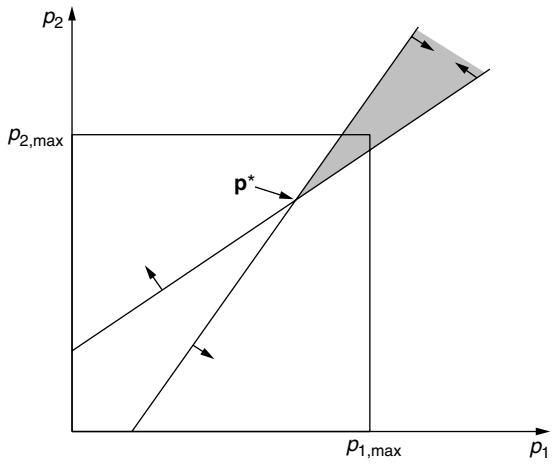


Figure 3. Power control problem for two-user case.

4.1. Open-Loop, Closed-Loop, and Outer-Loop Power Control

Because of the limitations mentioned above, TPC in CDMA is implemented in a slightly different way than the theoretical algorithms.

An intuitive way to compensate for the channel attenuation in the uplink would be to measure the strength of a pilot signal from the downlink, and adjust the transmitter power proportionally to the inverse of this measurement. Since the pilot signal is transmitted at constant power, the variation of its strength gives information of the downlink link attenuation. This is called *open-loop power control*. Unfortunately the center frequencies allocated to up- and down-link transmissions are usually widely separated, and thus the correlation between up- and down-link attenuations is generally weak. Therefore, the transmitter power update of a mobile must be based on feedback information of the received SIR at the base station, forming a closed loop between them. This *closed-loop power control* aims to keep the received uplink signal power level at a specified target. Moreover, the target must also be varied, because the SIR requirement for a given BER is not constant, but depends on the radio propagation conditions. This is the task of the *outer-loop power control*.

Open-loop TPC is generally used for initial power setting, when the two-way communication link is not yet established and closed loop is not possible. *Closed-loop TPC* aims to keep the received SIR at a target value. This is what is happening in the DPC algorithm given in Eq. (9). However, in practice only one bit is used to signal the received SIR information at a fast rate to track the channel variations. The transmitter is commanded to increase its power by a fixed step if the received SIR is below the target, and decrease otherwise. This kind of algorithm is used in IS95. In WCDMA there are some more degrees of freedom, for instance, the possibility to signal a "no change" command when the received SIR is reasonably close to the target, thus reducing the "bang-bang"-effect around the target.

The outer-loop control adjusts the SIR target so that a desired FER is guaranteed. A typical way to do this in practice is to raise the target by a larger step Δ_{up} when a frame is discarded, and to decrease the target with a smaller step Δ_{down} when a frame is correctly received. The relation between the step sizes gives the resulting average FER as $\text{FER} = \Delta_{\text{down}}/(\Delta_{\text{up}} - \Delta_{\text{down}})$ [10].

5. POWER CONTROL IN REAL-TIME VERSUS NON-REAL-TIME AND MULTIRATE SERVICES

The SIR balancing concept has the goal of maximizing the number of users in the system. If the users are real-time users requiring constant bit rates, this is a good strategy for maximizing capacity. However, as the cellular systems evolve to the next generation, the variety of services will be considerably different from those in the previous systems. In addition to the familiar real-time voice, there will be both real-time and non-real-time services with different data rates. Hence, maximizing the data throughput instead of the number of users might be more interesting from the operator's point of view. Of course the delay requirements must be fulfilled, as merely maximizing throughput would be achieved by just letting the user with the best instantaneous link quality to transmit.

Since CDMA is interference-limited, any decrease in the transmission power of one user is directly advantageous for other users because of decreased interference. If the link quality between a transmitter and a receiver is poor, high transmitter power is needed to satisfy the SIR requirements. This produces high interference to other receivers, and their serving transmitters must also increase their powers in order to cope with the increased interference.

In non-real-time services the data rate must only be satisfied on the average sense, and therefore the instantaneous data rate can be considerably varied. This allows the TPC algorithm even to cut off the transmission when the link quality is bad, and to transmit at a high data rate when the link quality is good. Thus, the situation as compared to conventional TPC designed for real-time services is reversed — the transmission power should be small when link quality is low, and vice versa. Since the time dimension can be utilized in the optimization, there is potential for significant capacity gain by minimizing the total transmitted *energy* instead of power. This can be accomplished by scheduling the data transmissions properly [12].

To elaborate, consider a set of users requiring individual data rates. Using (2) and (3), we write the effective data rate of user i as

$$r_i = \frac{W}{(E_b/N_0)_i}\gamma_i(\mathbf{p}) \tag{16}$$

where $\gamma_i(\mathbf{p})$ is the received SIR of user i with the power vector $\mathbf{p}$ and $(E_b/N_0)_i$ is the E_b/N_0 requirement for user i for achieving the data rate r_i. Let the maximum transmission power vector for the users be $\mathbf{p}_{\max} = (p_1^{\max}, p_2^{\max}, \ldots, p_N^{\max})$.

Definition 3 [5,12]**.** A rate vector $\mathbf{r}(\mathbf{p}_{\max}) = (r_1, r_2, \ldots, r_N)$ is instantaneously achievable if there exists a nonnegative power vector $\mathbf{p} \le \mathbf{p}_{\max}$ such that $r_i \le \dfrac{W}{(E_b/N_0)_i}\gamma_i(\mathbf{p})$ for all $1 \le i \le N$.

Definition 4 [5,12]**.** A rate vector $\mathbf{r}^*(\mathbf{p}_{\max}) = (r_1^*, r_2^*, \ldots, r_N^*)$ is achievable in the average sense if it may be expressed as $\mathbf{r}^* = \sum_k \lambda_k r_k$, where $\lambda_k \in [0, 1]$, $\sum_k \lambda_k = 1$, and all the r_k are instantaneously achievable rate vectors.

Thus, different rate vectors can be assigned a fraction of time (or frequency) yielding the required rate vector on the average. Assume that each link i requires a minimum data rate $r_i^{\min}$. Any excess data rate is potentially consumed, and thus paid for, by the user. It is the interest of the operators then to provide as much excess data rate as possible. For non-real-time services, therefore, the following optimization problem is of interest:

$$\begin{aligned} &\max \sum_{i=1}^{N} r_i^*(\mathbf{p}_{\max}) \\ &\text{subject to } r_i^*(\mathbf{p}_{\max}) \ge r_i^{\min}, \forall i \end{aligned} \tag{17}$$

6. POWER CONTROL AND OTHER RADIO RESOURCE MANAGEMENT (RRM)

Optimizing power control alone is not always the best way to enhance capacity. By understanding the relations between power control and other RRM functions, one can design more efficient algorithms by combining them in an ingenious way. For instance, base station assignment is closely related to the power control problem. For the SIR balancing problem, if the mobile stations could always be connected to the base station to which the link quality is optimal, the total transmission power would be minimized. Combined power and rate control is interesting for services with heterogeneous bit rates and quality requirements as discussed in the last section. Other methods that have been considered jointly with power control include smart antennas and beamforming, where there are more degrees of freedom in the optimization of the algorithms. An interested reader is directed to the articles in the Further Reading list for more details.

7. DISCUSSION AND VIEWS INTO THE FUTURE

The ultimate goal of radio resource management is to maximize the network capacity without unduly sacrificing the satisfaction of the users. Efficient transmitter power control is essential in CDMA cellular systems for achieving these goals. The efficiency of TPC depends on its ability to control the interference inherent in wireless multiuser systems. However, there are other methods for combating the multiple access interference in CDMA. One of these methods is *multiuser detection* (MUD). An optimal MUD-based receiver would theoretically eliminate the need for power control completely! In practice,

however, an optimal MUD receiver would be too complex, and suboptimal solutions must be used that are not completely resistant to interference, and TPC can still provide additional gain. Using only TPC provides a much cheaper way of controlling interference. This situation might change in the future, as the microtechnology evolves very rapidly and more efficient chips become available.

BIOGRAPHY

Matti J. Rintamäki received an M.S. degree in electrical engineering from Helsinki University of Technology, Finland, in 2000. Since then he has been a research scientist at Signal Processing Laboratory at HUT, where he has been working on power control algorithms for CDMA systems. His areas of interest are adaptive control and signal processing algorithms, and the design of radio resource management algorithms for wireless communications.

BIBLIOGRAPHY

1. S. Ariyavisitakul, Signal and interference statistics of a CDMA system with feedback power control — part II, *IEEE Trans. Commun.* **42**(2–4): 597–605 (Feb.–April 1994).
2. J. Zander, Performance of optimum transmitter power control in cellular radio systems, *IEEE Trans. Vehic. Technol.* **41**(1): 57–62 (Feb. 1992).
3. S. A. Grandhi, R. Vijayan, D. J. Goodman, and J. Zander, Centralized power control in cellular radio systems, *IEEE Trans. Vehic. Technol.* **42**(4): 466–468 (Nov. 1993).
4. M. Andersin, Z. Rosberg, and J. Zander, Gradual removals in cellular PCS with constrained power control and noise, *ACM/Baltzer Wireless Networks J.* **2**: 27–43 (1996).
5. J. Zander, S.-L. Kim, M. Almgren, and O. Queseth, *Radio Resource Management for Wireless Networks*, Artech House, Norwood, MA, 2001.
6. G. J. Foschini and Z. Miljanic, A simple distributed autonomous power control algorithm and its convergence, *IEEE Trans. Vehic. Technol.* **42**(4): 641–646 (Nov. 1993).
7. R. D. Yates, A framework for uplink power control in cellular radio systems, *IEEE J. Select. Areas Commun.* **13**(7): 1341–1347 (Sept. 1995).
8. C. Y. Huang and R. Yates, Rate of convergence for minimum power assignment in cellular radio systems, *ACM/Baltzer Wireless Networks J.* **1**: 223–231 (1998).
9. S. A. Grandhi, J. Zander, and R. Yates, Constrained power control, *Wireless Pers. Commun.* **1**: 257–270 (1995).
10. H. Holma and A. Toskala, *WCDMA for UMTS, Radio Access for Third Generation Mobile Communications*, Wiley, Chichester, UK, 2000.
11. TIA/EIA Interim Standard-95, *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System*, Telecommunications Industry Assoc., 1993.
12. F. Berggren, S.-L. Kim, R. Jäntti, and J. Zander, Joint power control and intracell scheduling of DS-CDMA nonreal time data, *IEEE J. Select. Areas Commun.* **19**(10): 1860–1870 (Oct. 2001).

FURTHER READING

Bambos N., Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design aspects, *IEEE Pers. Commun.* **5**(3): 50–59 (June 1998) (this article contains a nice review on power control).

Gilhousen K. S. et al., On the capacity of a cellular CDMA system, *IEEE Trans. Vehic. Technol.* **40**(2): 303–312 (May 1991) (an early paper on the capacity of CDMA as a multiple access technology; the need for power control is discussed).

Jäntti R. and S.-L. Kim, Second-order power control with asymptotically fast convergence, *IEEE J. Select. Areas Commun.* **18**(3): 447–457 (March 2000) (a distributed power control algorithm with faster convergence than with the DPC algorithm).

Kim D., On the convergence of fixed-step power control algorithms with binary feedback for mobile communication systems, *IEEE Trans. Commun.* **49**(2): 249–252 (Feb. 2001) (in this paper the author proves the convergence of fixed-step binary-feedback power control algorithms into a specific range of values).

Ulukus S. and R. D. Yates, Stochastic power control for cellular radio systems, *IEEE Trans. Commun.* **46**(6): 784–798 (June 1998) (in this paper the authors take the stochastic nature of the link attenuations and signal measurements into account and develop some power control algorithms whose convergence is proved stochastically).

Viterbi A. J., *Principles of Spread spectrum communication*, Reading, MA: Addison-Wesley, 1995 (a comprehensive book on spread-spectrum technology for commercial wireless applications).

Yener A., R. D. Yates, and S. Ulukus, Interference management for CDMA systems through power control, multiuser detection, and beamforming, *IEEE Trans. Commun.* **49**(7): 1227–1239 (July 2001) (in this paper the authors combine intelligent techniques to achieve higher performance than using the techniques alone).

POWER SPECTRA OF DIGITALLY MODULATED SIGNALS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

In the design of digital communication systems for transmitting digital information through a channel, the spectral characteristics of the information-bearing signals is an important element. The communication system designer must ensure that the signals used for transmitting the information do not violate the bandwidth constraint imposed by the channel or some governmental agency. In this article, the spectral characteristics of several types of digitally modulated signals are described.

In describing the spectral characteristics of digitally modulated signals, it is desirable to classify such signals into two different categories, namely, linearly modulated signals and nonlinearly modulated signals. The class of linearly modulated signals include pulse amplitude modulation or amplitude shift keying (PAM or ASK),

phase shift keying (PSK), and quadrature amplitude modulation (QAM). The class of nonlinearly modulated signals includes continuous-phase modulation (CPM) and the special form of CPM called *continuous-phase frequency shift keying* (CPFSK). CPM and CPFSK are constant-amplitude signals; hence, they are well suited for radiocommunications, where the transmitted signals can be amplified by power amplifiers that can be driven into saturation without introducing nonlinear distortion in the signals.

The spectral characteristics of linearly modulated signals are considered in the next section. Nonlinearly modulated signals are treated in the subsequent section.

2. POWER SPECTRA OF LINEARLY MODULATED SIGNALS

Linear digital modulation methods that include PAM (ASK), PSK, and QAM can be treated in a unified manner. The digitally modulated signals for these methods are usually generated as lowpass signals, which may be expressed mathematically as

$$v(t) = \sum_n I_n g(t - nT) \tag{1}$$

and then translated to bandpass, for transmission over the bandpass channel, by a frequency translation. Hence, the bandpass signal is

$$s(t) = \mathrm{Re}[v(t)e^{j2\pi f_c t}] \tag{2}$$

where f_c is the carrier frequency. In the expression for the lowpass signal given by Eq. (1), the data sequence $\{I_n\}$ has values taken from either a PAM, PSK, or QAM signal-point constellation. In particular, if the modulated signal is M-level PAM, the sequence $\{I_n\}$ takes on values from the set $\{\pm 1, \pm 3, \pm 5, \ldots, \pm(M-1)\}$. When the modulated signal is M-phase PSK, the data sequence $\{I_n\}$ takes values from the set $\{\exp(2\pi m/M), m = 0, 1, \ldots, M-1\}$. A QAM signal is basically a combined amplitude/phase-modulated signal, so the data sequence takes an values from the set $\{A_m e^{j\theta_m}, m = 0, 1, \ldots, M-1\}$. Finally, the signal $g(t)$ in Eq. (1) is a signal pulse that is used to shape the spectrum of the transmitted signal. The rate at which data symbols are transmitted is $1/T$, where T defines the symbol interval.

The digitally modulated signal $v(t)$ is a random process because the data sequence $\{I_n\}$ is random. The power spectrum of $v(t)$ will be determined by first obtaining the autocorrelation function of $v(t)$ and then computing its Fourier transform. The power spectrum of the bandpass signal $s(t)$ is simply obtained from the power spectrum of $v(t)$ by a simple frequency translation. Thus, the autocorrelation function of the bandpass signal $s(t)$, denoted as $\phi_{ss}(\tau)$ is related to the autocorrelation function of the equivalent lowpass signal $v(t)$ through the expression

$$\phi_{ss}(\tau) = \mathrm{Re}[\phi_{vv}(\tau)e^{j2\pi f_c \tau}] \tag{3}$$

where $\phi_{vv}(\tau)$ is the autocorrelation function of the equivalent lowpass signal $v(t)$. The Fourier transform of Eq. (3) yields the desired expression for the power density spectrum $\Phi_{ss}(f)$ in the form

$$\Phi_{ss}(f) = \tfrac{1}{2}[\Phi_{vv}(f - f_c) + \Phi_{vv}(-f - f_c)] \tag{4}$$

where $\Phi_{vv}(f)$ is the power density spectrum of $v(t)$. It suffices to determine the autocorrelation function and the power density spectrum of the equivalent lowpass signal $v(t)$.

The autocorrelation function of $v(t)$ is defined as

$$\begin{aligned}\phi_{vv}(t + \tau; t) &= \frac{1}{2}E[v^*(t)v(t + \tau)] \\ &= \frac{1}{2}\sum_{n=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} E[I_n^* I_m]g^*(t - nT)g(t + \tau - mT)\end{aligned} \tag{5}$$

It is assumed that the sequence of information symbols $\{I_n\}$ is wide-sense stationary with mean μ_i and autocorrelation function

$$\phi_{ii}(m) = \tfrac{1}{2}E[I_n^* I_{n+m}] \tag{6}$$

Hence Eq. (5) can be expressed as

$$\begin{aligned}\phi_{vv}(t + \tau; t) &= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\phi_{ii}(m - n)g^* \\ &\quad \times (t - nT)g(t + \tau - mT) \\ &= \sum_{m=-\infty}^{\infty}\phi_{ii}(m)\sum_{n=-\infty}^{\infty} g^*(t - nT)g \\ &\quad \times (t + \tau - nT - mT)\end{aligned} \tag{7}$$

The second summation in Eq. (7), namely

$$\sum_{n=-\infty}^{\infty} g^*(t - nT)g(t + \tau - nT - mT) \tag{8}$$

is periodic in the t variable with period T. Consequently, $\phi_{vv}(t + \tau; t)$ is also periodic in the t variable with period T:

$$\phi_{vv}(t + T + \tau; t + T) = \phi_{vv}(t + \tau; t) \tag{9}$$

In addition, the mean value of $v(t)$, which is

$$E[v(t)] = \mu_i \sum_{n=-\infty}^{\infty} g(t - nT) \tag{10}$$

is periodic with period T. Therefore, $v(t)$ is a random process having a periodic mean and autocorrelation function. Such a process is called a *cyclostationary process* or a *periodically stationary process in the wide sense*.

In order to compute the power density spectrum of a cyclostationary process, the dependence of $\phi_{vv}(t + \tau; t)$ on the t variable must be eliminated. This can be

accomplished simply by averaging $\phi_{vv}(t+\tau;t)$ over a single period:

$$\begin{aligned}\overline{\phi}_{vv}(\tau) &= \frac{1}{T}\int_{-T/2}^{T/2}\phi_{vv}(1+\tau;t)\,dt \\ &= \sum_{m=-\infty}^{\infty}\phi_{ii}(m)\sum_{n=-\infty}^{\infty}\frac{1}{T}\int_{-T/2}^{T/2}g^*(t-nT)g \\ &\quad\times(t+\tau-nT-mT)\,dt \\ &= \sum_{m=-\infty}^{\infty}\phi_{ii}(m)\sum_{n=-\infty}^{\infty}\frac{1}{T}\int_{-T/2-nT}^{T/2-nT}g^*(t)g(t+\tau-mT)\,dt\end{aligned} \tag{11}$$

We interpret the integral in this equation as the time autocorrelation function of $g(t)$ and define it as

$$\phi_{gg}(t) = \int_{-\infty}^{\infty}g^*(t)g(t+\tau)\,dt \tag{12}$$

Consequently Eq. (11) can be expressed as

$$\overline{\phi}_{vv}(\tau) = \frac{1}{T}\sum_{m=-\infty}^{\infty}\phi_{ii}(m)\phi_{gg}(\tau-mT) \tag{13}$$

The Fourier transform of the relation in Eq. (13) yields the (average) power density spectrum of $v(t)$ in the form

$$\Phi_{vv}(f) = \frac{1}{T}|G(f)|^2\Phi_{ii}(f) \tag{14}$$

where $G(f)$ is the Fourier transform of $g(t)$, and $\Phi_{ii}(f)$ denotes the power density spectrum of the information sequence, defined as

$$\Phi_{ii}(f) = \sum_{m=-\infty}^{\infty}\phi_{ii}(m)e^{-j2\pi fmT} \tag{15}$$

The result in Eq. (14) illustrates the dependence of the power density spectrum of $v(t)$ on the spectral characteristics of the pulse $g(t)$ and the information sequence $\{I_n\}$. Thus, the spectral characteristics of $v(t)$ can be controlled by design of the pulseshape $g(t)$ and by design of the correlation characteristics of the information sequence.

Whereas the dependence of $\Phi_{vv}(f)$ on $G(f)$ is easily understood on observation of Eq. (14), the effect of the correlation properties of the information sequence is more subtle. First, we note that for an arbitrary autocorrelation $\phi_{ii}(m)$ the corresponding power density spectrum $\Phi_{ii}(f)$ is periodic in frequency with period $1/T$. In fact, the equation [Eq. (15)] relating the spectrum $\Phi_{ii}(f)$ to the autocorrelation $\phi_{ii}(m)$ is in the form of an exponential Fourier series with the $\{\phi_{ii}(m)\}$ as the Fourier coefficients. As a consequence, the autocorrelation sequence $\phi_{ii}(m)$ is given by

$$\phi_{ii}(m) = T\int_{-1/2T}^{1/2T}\Phi_{ii}(f)e^{j2\pi fmT}\,df \tag{16}$$

Second, let us consider the case in which the information symbols in the sequence are real and mutually uncorrelated. In this case, the autocorrelation function $\phi_{ii}(m)$ can be expressed as

$$\phi_{ii}(m) = \begin{cases}\sigma_i^2+\mu_i^2 & (m=0)\\ \mu_i^2 & (m+0)\end{cases} \tag{17}$$

where σ_i^2 denotes the variance of an information symbol. When Eq. (17) is used to substitute for $\phi_{ii}(m)$ in Eq. (15), we obtain

$$\Phi_{ii}(f) = \sigma_i^2+\mu_i^2\sum_{m=-\infty}^{\infty}e^{-j2\pi fmT} \tag{18}$$

The summation in Eq. (18) is periodic with period $1/T$. It may be viewed as the exponential Fourier series of a periodic train of impulses with each impulse having an area $1/T$. Therefore Eq. (18) can also be expressed in the form

$$\Phi_{ii}(f) = \sigma_i^2+\frac{\mu_1^2}{T}\sum_{m=-\infty}^{\infty}\delta\left(f-\frac{m}{T}\right) \tag{19}$$

Substitution of Eq. (19) into Eq. (14) yields the desired result for the power density spectrum of $v(t)$ when the sequence of information symbols is uncorrelated:

$$\Phi_{vv}(f) = \frac{\sigma_i^2}{T}|G(f)|^2+\frac{\mu_1^2}{T^2}\sum_{m=-\infty}^{\infty}\left|G\left(\frac{m}{T}\right)\right|^2\delta\left(f-\frac{m}{T}\right) \tag{20}$$

The expression in Eq. (20) for the power density spectrum is purposely separated into two terms to emphasize the two different types of spectral components. The first term is the continuous spectrum, and its shape depends only on the spectral characteristic of the signal pulse $g(t)$. The second term consists of discrete frequency components spaced $1/T$ apart in frequency. Each spectral line has a power that is proportional to $|G(f)|^2$ evaluated at $f=m/T$. Note that the discrete frequency components vanish when the information symbols have zero mean: $\mu_i=0$. This condition is usually desirable for the digital modulation techniques under consideration, and it is satisfied when the information symbols are equally likely and symmetrically positioned in the complex plane. Thus, the system designer can control the spectral characteristics of the digitally modulated signal by proper selection of the characteristics of the information sequence to be transmitted.

As an example that illustrates the spectral shaping from $g(t)$, consider the rectangular pulse shown in Fig. 1. The Fourier transform of $g(t)$ is

$$G(f) = AT\frac{\sin\pi fT}{\pi fT}e^{-j\pi fT}$$

Hence

$$|G(f)|^2 = (AT)^2\left(\frac{\sin\pi fT}{\pi fT}\right)^2 \tag{21}$$

This spectrum is illustrated in Fig. 1. Note that it contains zeros at multiples of $1/T$ in frequency and that it decays inversely as the square of the frequency variable. As a consequence of the spectral zeros in $G(f)$, all except one of

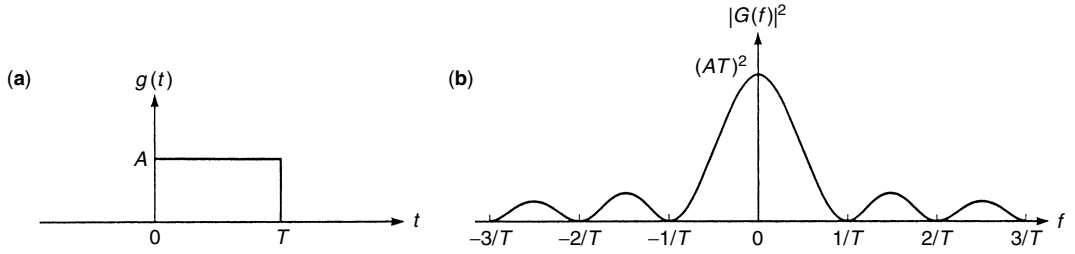


Figure 1. Rectangular pulse and its energy density spectrum.

the discrete spectral components in Eq. (20) vanish. Thus, on substitution for $|G(f)|^2$ from Eq. (21), Eq (20) reduces to

$$\Phi_{vv}(f) = \sigma_i^2 A^2 T \left(\frac{\sin \pi fT}{\pi fT}\right)^2 + A^2 \mu_i^2 \delta(f)$$

To illustrate that spectral shaping can also be accomplished by operations performed on the input information sequence, we consider a binary sequence $\{b_n\}$ from which we form the symbols

$$I_n = b_n + b_{n-1}$$

The $\{b_n\}$ are assumed to be uncorrelated random variables, each having zero mean and unit variance. Then the autocorrelation function of the sequence $\{I_n\}$ is

$$\begin{aligned}\phi_{ii}(m) &= E(I_n I_{n+m}) \\ &= \begin{cases} 2 & (m = 0) \\ 1 & (m = \pm 1) \\ 0 & (\text{otherwise}) \end{cases}\end{aligned}$$

Hence, the power density spectrum of the input sequence is

$$\begin{aligned}\Phi_{ii}(f) &= 2(1 + \cos 2\pi fT) \\ &= 4 \cos^2 \pi fT\end{aligned}$$

and the corresponding power density spectrum for the (lowpass) modulated signal is

$$\Phi_{vv}(f) = \frac{4}{T}|G(fg)|^2 \cos^2 \pi fT$$

Since $\cos^2 \pi fT$ has its first null at $f = 1/2T$, the effect of multiplying $|G(f)|^2$ by $\cos^2 \pi fT$ is to narrow the width of the mainlobe of the signal spectrum.

2. POWER SPECTRA OF CONTINUOUS-PHASE MODULATED SIGNALS

A CPM signal is described mathematically as

$$s(t) = A \cos[2\pi f_c t + \phi(t; \mathbf{I})] \tag{22}$$

where A is the signal amplitude, f_c is the carrier frequency, and $\phi(t; \mathbf{I})$ is the phase of the signal that carries the information. The phase function may be defined as

$$\phi(t; \mathbf{I}) = 2\pi h \sum_{k=-\infty}^{n} I_k q(t - kT), \quad nT \le t \le (n+1)T \tag{23}$$

where $\{I_k\}$ is the data sequence selected from the M-level amplitude alphabet $\{\pm 1, \pm 2, \ldots, \pm(M-1)\}$, h is the modulation index, and $q(t)$ is defined as the integral of a pulse $g(t)$:

$$q(t) = \int_0^t g(\tau)\, d\tau \tag{24}$$

The pulse $g(t) = 0$ for $t < 0$ and $t > LT$, where L is an integer and T is the symbol interval. When $L = 1$, the pulse $g(t)$ is nonzero over a single signal interval, and the CPM signal is called *full-response CPM*. When $L \ge 2$, the pulse $g(t)$ is nonzero over two or more signal intervals and the CPM signal is called partial response. Furthermore $g(t)$ is normalized in area so that

$$q(LT) = \int_0^{LT} g(\tau)\, d\tau = \frac{1}{2}$$

The phase continuous signal may be viewed as having been generated as an FM signal. Suppose that the lowpass data signal $d(t)$ is a PAM signal of the form

$$d(t) = \sum_k I_k g(t - kT) \tag{25}$$

where $\{I_m\}$ is the data sequence of symbols selected from the alphabet $\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$ and $g(t)$ is a pulse with unit area that is nonzero over the interval $0 \le t \le LT$. If $d(t)$ is used to frequency modulate the carrier f_c, then, the phase of the carrier is

$$\begin{aligned}\phi(t; \mathbf{I}) &= 4\pi f_d T \int_{-\infty}^{t} d(\tau)\, d\tau \\ &= 2\pi h \sum_{k=-\infty}^{n} I_k q(t - nT), \quad nT \le t \le (n+1)T\end{aligned} \tag{26}$$

where f_d is the peak frequency deviation and the modulation index is defined as $h = 2f_d T$.

Some examples of the pulseshapes $g(t)$ and $q(t)$ are illustrated in Fig. 2. When $L = 1$ and $g(t)$ is a rectangular pulse, as shown in Fig. 2a, the CPM signal reduces to the special case called *continuous-phase frequency shift keying* (CPFSK). Furthermore, if the modulation index is selected as $h = \frac{1}{2}$, the CPFSK signal is called minimum-shift keying (MSK). The signal pulse $g(t)$ shown in Fig. 2b is called a *raised-cosine pulse*, and the resulting CPM is a full-response CPM signal. On the other hand, the signal pulses $g(t)$ shown in Fig. 2c,d extend over two signal intervals, and hence the CPM signals are partial response.

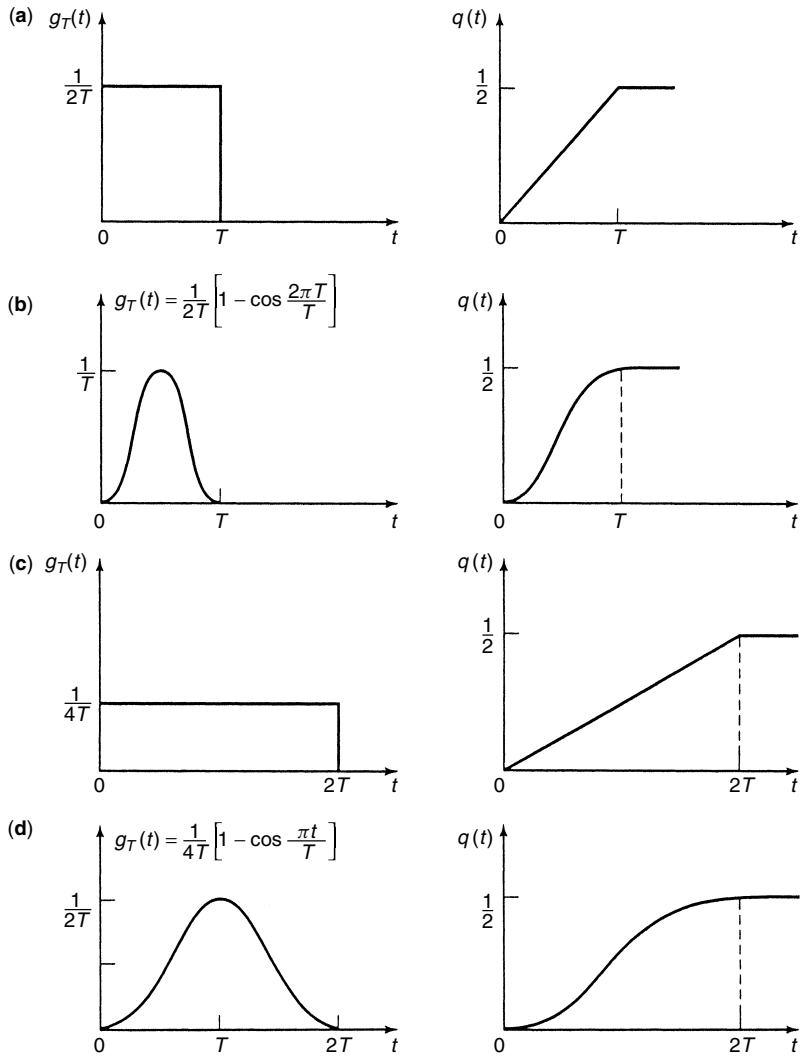


Figure 2. Pulseshapes for full-response (**a**,**b**) and partial-response (**c**,**d**) CPM.

To determine the power spectrum of the CPM signal, the general procedure described in the preceding section may be used. One begins with the equivalent lowpass signal

$$v(t) = e^{j\phi(t;\mathbf{I})} \tag{27}$$

where $\phi(t;\mathbf{I})$ is given by Eq. (26) with $n = \infty$. The autocorrelation function of $v(t)$ is easily shown to be

$$\phi_{vv}(t+t;t) = \frac{1}{2}E\left[\exp\left(j2\pi h \sum_{k=-\infty}^{\infty} I_k[q(t+\tau-kT) - q(t-kT)]\right)\right] \tag{28}$$

The sum of exponents may also be expressed as a product of exponents:

$$\phi_{vv}(t+\tau;t) = \frac{1}{2}E\left(\prod_{k=-\infty}^{\infty} \exp\{j2\pi h I_k[q(t+\tau-kT) - q(t-kT)]\}\right) \tag{29}$$

Although Eq. (29) implies that there are an infinite number of factors in the product, the pulse $g(t) = 0$ for $t < 0$ and $t > LT$, and $q(t) = 0$ for $t < 0$. Consequently only a finite number of terms in the product have nonzero exponents. The next step is to perform the expectation over the data symbols $\{I_k\}$ and then to average the periodic autocorrelation function over the period $(0, T)$:

$$\overline{\phi}_{vv}(\tau) = \frac{1}{T}\int_0^T \phi_{vv}(t+\tau;t)\,dt \tag{30}$$

The final step is to compute the Fourier transform of $\overline{\phi}_{vv}(t)$ to obtain the power spectrum.

Unfortunately, these computations do not yield a closed-form solution except in special cases. When the information symbols are equally likely, the average autocorrelation function is given as

$$\overline{\phi}_{vv}(\tau) = \frac{1}{2T}\int_0^T \prod_{k=1-L}^{[t/T]} \frac{1}{M} \times \frac{\sin 2\pi h M[q(t+\tau-kT) - q(t-kT)]}{\sin 2\pi h[q(t+\tau-kT) - q(t-kT)]}\,dt \tag{31}$$

and the corresponding expression for the power density spectrum is

$$\Phi_{vv}(f) = 2\left[\int_0^{LT} \overline{\phi}_{vv}(\tau)\cos 2p\pi f\tau\, d\tau + \frac{1-\psi(jh)\cos 2\pi fT}{1+\psi^2(jh)-2\psi(jh)\cos 2\pi fT} \times \int_{LT}^{(L+1)T} \overline{\phi}_{vv}(\tau)t\cos 2\pi f\tau\, d\tau - \frac{\psi(jh)\sin 2\pi fT}{1+\psi^2(jh)-2\psi(jh)\cos 2\pi fT} \times \int_{LT}^{(L+1)T} \overline{\phi}_{vv}(\tau)\sin 2\pi f\tau\, d\tau\right] \tag{32}$$

where $\psi(jh)$ is the characteristic function of the information symbols $\{I_k\}$, which is given as

$$\begin{aligned}\psi(jh) &= \sum_{\substack{n=-(M-1)\\ n \text{ odd}}}^{M-1} p_n e^{j\pi hn} \\ &= \frac{1}{M}\sum_{\substack{n=-(M-1)\\ n \text{ odd}}}^{M-1} e^{j\pi hn} \\ &= \frac{1}{M}\frac{\sin M\pi h}{\sin \pi h}\end{aligned} \tag{33}$$

where $p_n = 1/M$ is the probability of each of the M levels. The expression for the power density spectrum given in Eq. 32 must be evaluated numerically.

2.1. Power Density Spectrum of CPFSK

A closed-form expression for the power density spectrum can be obtained from Eq. (32) when the pulseshape $g(t)$ is rectangular and zero outside the interval $[0, T]$. In this case, $q(t)$ is linear for $0 \le 1 \le T$. The resulting power spectrum may be expressed as

$$\Phi_{vv}(f) = T\left[\frac{1}{M}\sum_{n=1}^{M} A_n^2(f) + \frac{2}{M^2}\sum_{n=1}^{M}\sum_{m=1}^{M} B_{nm}(f)A_n(f)A_m(f)\right] \tag{34}$$

where

$$\begin{aligned}A_n(f) &= \frac{\sin \pi[fT - \frac{1}{2}(2n-1-M)h]}{\pi[fT - \frac{1}{2}(2n-1-M)h]} \\ B_{nm}(f) &= \frac{\cos(2\pi fT - \alpha_{nm}) - \psi\cos\alpha_{nm}}{1+\psi^2-2\psi\cos 2\pi fT} \\ \alpha_{nm} &= \pi h(m+n-1-M) \\ \psi \equiv \psi(jh) &= \frac{\sin M\pi h}{M\sin \pi h}\end{aligned} \tag{35}$$

The power density spectrum of CPFSK for $M = 2, 4$ is plotted in Figs. 3 and 4 as a function of the normalized frequency fT, with the modulation index $h = 2f_dT$ as a parameter. Note that only one-half of the bandwidth occupancy is shown in these graphs. The origin corresponds to the carrier f_c. The graphs illustrate that the spectrum of CPFSK is relatively smooth and well confined for $h < 1$. As h approaches unity, the spectra become very peaked and, for $h = 1$ when $|\psi| = 1$, we find that impulses occur at M frequencies. When $h > 1$, the spectrum becomes much broader. In communication systems where CPFSK is used, the modulation index is designed to conserve bandwidth, so that h is selected to be less than unity.

The special case of binary CPFSK with $h = \frac{1}{2}$ (or $f_d = 1/4T$) and $\psi = 0$ corresponds to MSK. In this case, the spectrum of the signal is

$$\Phi_{vv}(f) = \frac{16T}{\pi^2}\left(\frac{\cos 2\pi fT}{1-16f^2T^2}\right)^2 \tag{36}$$

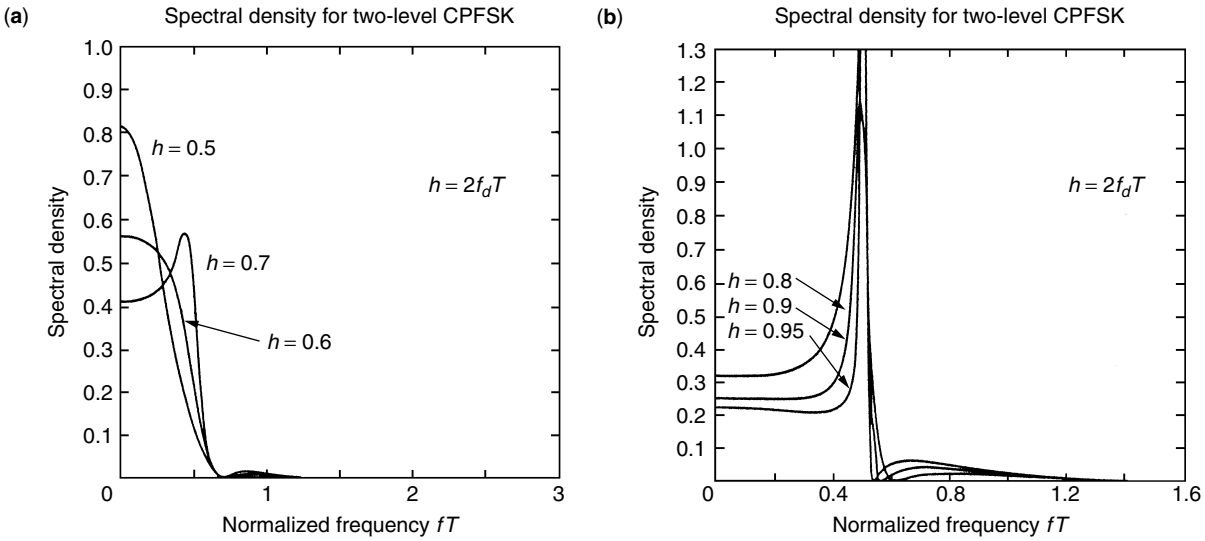


Figure 3. Power spectra for binary CPFSK.

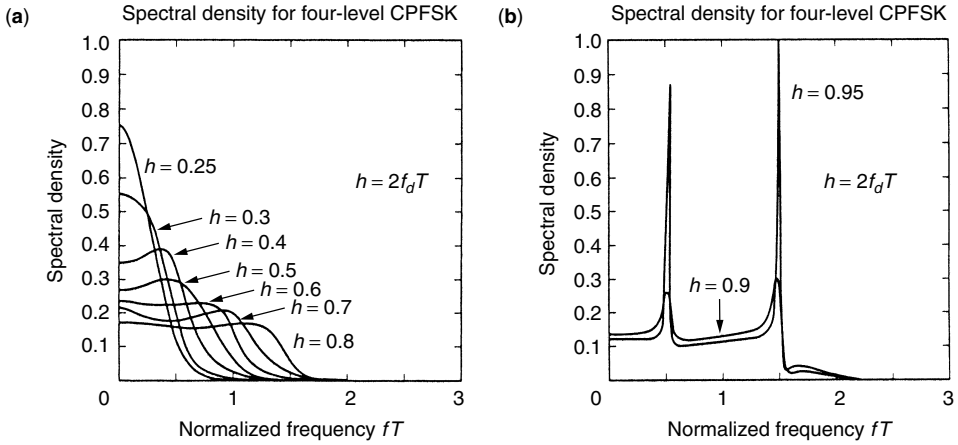


Figure 4. Power spectra for quaternary CPFSK.

2.2. Power Density Spectrum of CPM

The use of smooth pulses such as raised-cosine pulses of the form

$$g(t) = \begin{cases} \dfrac{1}{2LT}\left(1 - \cos\dfrac{2\pi t}{LT}\right) & (0 \le t \le LT) \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

where $L = 1$ for full response and $L > 1$ for partial response, result in smaller bandwidth occupancy and, hence, greater bandwidth efficiency than the use of rectangular pulses. For example, Fig. 5 illustrates the power density spectrum for binary CPM with different partial-response raised-cosine (LRC) pulses when $h = \frac{1}{2}$. For comparison, the spectrum of (MSK) binary CPFSK is also shown. Note that as L increases the pulse, $g(t)$ becomes smoother and the corresponding spectral occupancy of the signal is reduced.

The effect of varying the modulation index in a CPM signal is illustrated in Fig. 6 for the case of $M = 4$ and a raised-cosine pulse of the form given in Eq. (37) with $L = 3$. Note that these spectral characteristics are similar to the ones illustrated previously for CPFSK, except that these spectra are narrower because of the use of a smoother pulseshape.

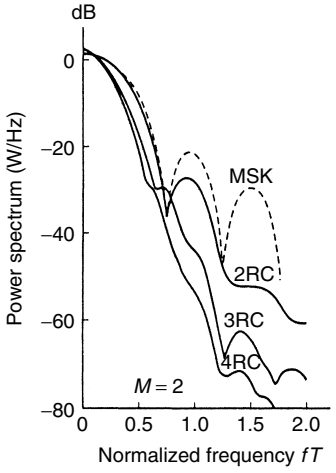


Figure 5. Power density spectrum for binary CPM with $h = \frac{1}{2}$ and different pulseshapes. [From Aulin et al. (1981); © 1981 IEEE.

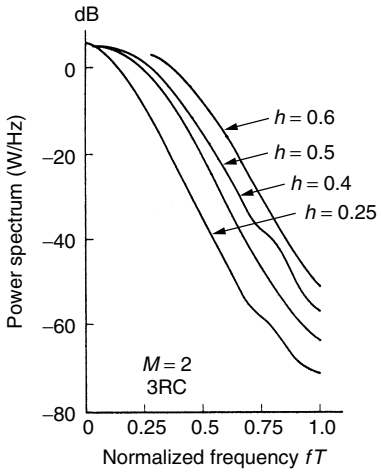


Figure 6. Power density spectrum for $M = 4$ CPM with $3RC$ and different modulation indices. [From Aulin et al.(1981); © 1981 IEEE.]

3. CONCLUDING REMARKS

PAM, PSK, QAM, and CPM signals are described in greater detail in other articles of this encyclopedia as well as in Ref. 1. Additional numerical results on the spectral characteristics of CPM signals can be found in Refs. 2 and 3.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College

of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
2. J. B. Anderson, T. Aulin, and C. W. Sundberg, *Digital Phase Modulation*, Plenum, New York, 1986.
3. R. R. Anderson and J. Salz, Spectra of digital FM, *Bell Syst. Tech. J.* **44**: 1165–1189 (July–Aug. 1965).

POWERLINE COMMUNICATIONS

Halid Hrasnica
Abdelfatteh Haidine
Ralf Lehnert
Dresden University of Technology
Dresden, Germany

1. INTRODUCTION

Powerline communications (PLC) uses the standard electrical power distribution network in parallel with the energy distribution for the transmission of digital data. This avoids the installation of a new telecommunications infrastructure, where a power grid already exists. PLC with low data rates of ~1 kbps (kilobit per second) has been in use since 1990 or 1991. This narrowband technology has been standardized in Europe by CENELEC EN 50065 [1]. It operates in the frequency band from 3 to 148.5 kHz and uses simple digital modulation schemes, such as frequency shift keying (FSK).

With the advent of high-speed digital signal processors, advanced digital modulation techniques can be implemented. As an example, OFDM not only allows a high data rate but is also able to cope with a time-varying transmission channel. This presently allows transmission data rates over the powerline of $\leq$4 Mbps. PLC is therefore able to compete with, For example, digital subscribers lines (DSLS) on the telephone two-wire copper line or with wireless LAN techniques in the home area.

PLC offers an advantage for developing countries, because the power distribution network is already in place and a second network need not be built. In developed countries PLC also opens an opportunity for prospective network operators, who plan to compete with the incumbent operator, the earlier monopolist. Here again, an existing infrastructure can be used, and therefore the often prohibitive costs of new cabling can be avoided.

Within the power grid, PLC technology may be used in the backbone links in the high-voltage area, in the medium-voltage plane in urban areas, in the low-voltage plane for the access to the customer's premises, and, finally, within a household to reach the subscriber's communications terminal. In modern backbone's underground cables, there is usually an integrated fiber offering a nearly unlimited capacity, much higher than current PLC technology. Also the traffic of a large number of subscribers cannot be transported by current PLC technology. The medium voltage feeder network may be a candidate for PLC, but the bandwidth needed for a large number of high-speed customers cannot be supported by state-of-the-art technology. Also many medium- to low-voltage transformers are now reached by a fiber.

PLC has a promising application area and business case in the low-voltage area of the power grid. Here the number of customers per distribution section is limited to approximately 10–50, such that today's gross bandwidth still allows sufficient bandwidth per subscriber. This access network is usually limited in its size, the so-called last mile. In the case of in-home PLC, the size is even further limited (<100 m) and also the number of terminals may correspond to the number of persons in a home.

PLC operates as a shared medium. This means that the total capacity is shared in a statistical manner by all users on the same distribution section. Therefore a MAC protocol is needed to coordinate the sharing of the bandwidth fairly (see Section 5.3).

A major challenge for PLC is electromagnetic compatibility (EMC) (see Section 4.4). This is a bidirectional problem. On one hand, PLC is disturbed by electric noise (spikes, etc.; see Section 4.4). PLC systems also generate

radiation that disturbs other wireless communications. For this reason, countries have issued radiation limits, which constitute the main reason for capacity limits in PLC. In Germany, for example, these limits are defined by the regulatory body RegTP in regulation NB30 [2]. Currently under discussion is the replacement of the actual "chimney" regulation of forbidden frequency bands by a limiting curve within the entire frequency range of 1–30 MHz.

2. APPLICATION OF PLC TECHNOLOGY

2.1. Overview

The application of electrical supply networks in telecommunications has been known since the beginning of the twentieth century. The first *carrier frequency systems* (CFSs) have been operated in high-voltage electrical networks that were able to span distances over 500 km using 10 W signal power [3]. Such systems have been used for internal communication of electrical utilities and realization of remote measurements and control tasks. Also, communication over medium- and low-voltage electrical networks has been implemented by ripple carrier signaling (RCS) systems for realization of load management in electrical supply systems. Internal electrical networks have been used mostly for realization of various automation services within buildings or private houses.

The electrical supply systems consist of three network levels that can be used as a transmission medium for the realization of PLC networks (Fig. 1):

- *High-voltage networks* (110–380 kV) usually connect the power stations with supply regions or very big customers. They span very long distances, allowing the power exchange within a continent. Electrical supply grids of high-voltage networks are realized mostly as overhead lines.
- *Medium-voltage networks* (10–30 kV) supply large regions, cities, and big industrial or commercial complexes. The spanned distances are significantly shorter than those with high-voltage networks. These networks are realized by overhead lines or by underground cables.
- *Low-voltage networks* (230/400 V; in the United States, 110 V) directly supply the end users that are connected either as individual customers or single users belonging to a big customer (e.g., within a business building). Low-voltage networks are usually realized by overhead lines in rural areas and by underground cables within cities. The cable length between the last transformer unit and the customers is varying, but normally not longer than a few hundred meters.

In-home electrical installations belong to the low-voltage network level (Fig. 1). However, internal installations are usually owned by the users. They are connected to the supply network over an electrical power meter unit (M) (see Fig. 2). On the other hand, the rest of the low-voltage network (outdoor) belongs to the electrical supply utilities.

2.2. Narrowband PLC

As a successor to the ANSI X-10 standard, CENELEC has standardized PLC in the frequency range 3–148.5 kHz. This allows mainly private home or business building automation applications with data rates up to 1200 bps. This is the PLC variant of the European installation bus, named Powernet EIB. It became a de facto industry standard with FSK modems and a simple MAC protocol.

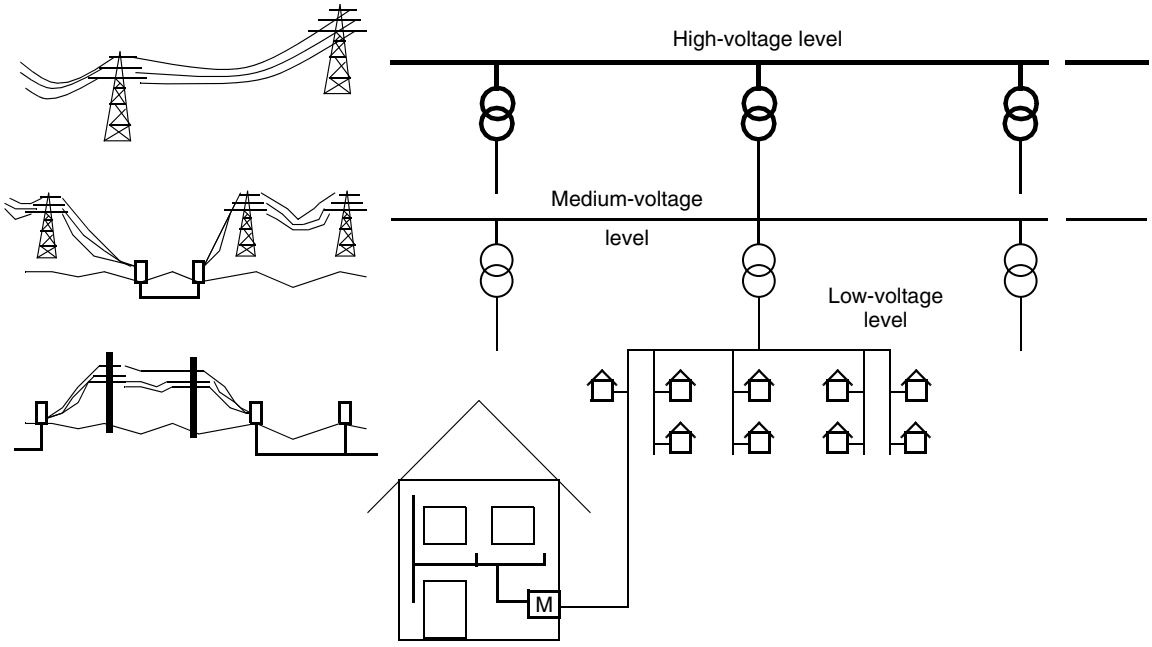


Figure 1. Structure of electrical supply networks.

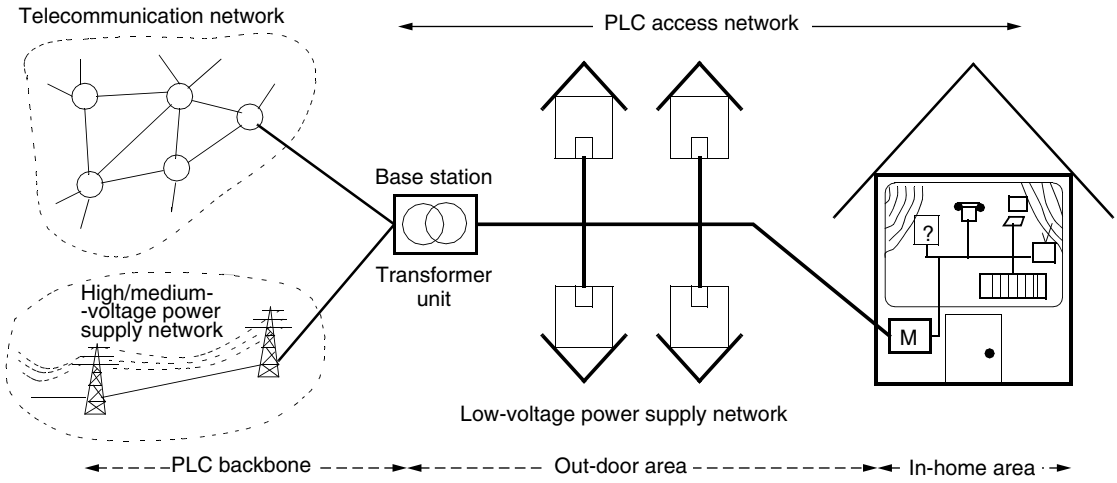


Figure 2. Structure of PLC access networks.

Loss of data may occur especially under high load. This is circumvented by ARQ mechanisms at the higher layers.

2.3. Broadband PLC

Broadband PLC systems provide significantly higher data rates (>2 Mbps) than do narrowband PLC systems. Where the narrowband networks can realize only a small number of voice channels and data transmission with very low bit rates, broadband PLC networks offer realization of more sophisticated telecommunication services: multiple voice connections, high-speed data transmission, transfer of video signals, and narrowband services as well. Therefore, PLC broadband systems are also considered as a capable telecommunication technology.

The realization of broadband communication services over powerline grids offers a great opportunity for cost-effective telecommunication networks without laying of new cables. However, electrical supply networks are not designed for information transfer, and there are some limiting factors in the application of broadband PLC technology. Therefore, the distances that can be covered are limited, as well as the data rates that can be realized by PLC systems. A further very important aspect for application of broadband PLC is its electromagnetic compatibility. For realization of broadband PLC, a significantly wider frequency spectrum is needed (≤30 MHz) than is provided within CENELEC bands. On the other hand, a PLC network acts as an antenna becoming a noise source for other communication systems working in the same frequency range (e.g., various radio services). For this reason, broadband PLC systems have to operate with a limited signal power, which decreases their performance (data rates, distances).

In contrast to narrowband PLC systems, no specified standards apply to broadband PLC networks. The standardization process is currently led by several international bodies. PLCforum [4] is an international association formed to unify and represent the interests of all players engaged in PLC (utility, providers, manufacturers, developers, researchers) and to expedite standardization and regulation on PLC technology worldwide as well as to support its commercialization. HomePlug Powerline Alliance [5] is a similar organization that is oriented to in-home PLC technology, providing both narrowband and broadband applications. Concrete work on technical standardization is done within ETSI (European Telecommunications Standards Institute) and CENELEC.

Current broadband PLC systems provide data rates beyond 2 Mbps in the outdoor arena, which includes middle- and low-voltage supply networks (Fig. 1), and up to 12 Mbps in the in-home area. Some manufacturers have already developed product prototypes providing much higher data rates (~40 Mbps). Middle-voltage PLC technology is usually used for realization of point-to-point connections bridging distances up to several hundred meters. Typical application areas of such systems is connection of LAN networks between buildings or campuses and connection of antennas and base stations of cellular communication systems to their backbone networks. Low-voltage PLC technology is used for realization of the so-called last mile of telecommunication access networks. Because of the importance of telecommunication access, current development of broadband PLC technology is directed mostly to applications in access networks, including the in-home area.

3. PLC ACCESS NETWORKS

3.1. PLC Alternative for Communication Over the "Last Mile"

The access networks are very important for network providers because of their high costs and the possibility for the realization of a direct access to the end users and subscribers. Typically, about 50% of all investments in the telecommunication infrastructure is needed for the realization of the access networks. Following the deregulation of the telecommunication market in a large number of countries, the access networks are still owned by the former monopolistic companies (incumbent providers). New network providers build up their wide-area networks (WANs), but they still have to use the access infrastructure owned by incumbent providers.

Consequently, the new network providers are trying to find a solution to realize their own access network.

Building new access networks (cable or fiberoptic networks, mobile and fixed wireless access systems, satellite networks) is the best way to implement the newest communication technology, which allows realization of attractive services and applications. On the other hand, the realization of new access networks is expensive and, in the case of laying new fiber or cable, takes a long time. The expensive buildup of new communication networks can be avoided by using existing infrastructure. In this case, already existing wireline networks are candidates for connection of subscribers to the telecommunication transport networks. This is possible by using the following infrastructure:

- Telephone networks
- TV cable networks (CATV)
- Power supply networks

Telephone networks usually belong to the former monopolistic companies, and this is a major disadvantage for new network providers to use them to offer services such as ADSL. That is very often the case with CATV networks, too. Additionally, CATV networks have to be made capable for bidirectional transmission, which results in extra costs. Therefore, the usage of power supply networks for communications seems to be reasonable.

3.2. Network Structure

A low-voltage supply network consists of a transformer unit and a number of power supply cables connecting the end users/subscribers (Fig. 2). The transformer unit connects the low-voltage supply network to the medium- and high-voltage levels. A PLC system applied to a low-voltage network uses the power grids as a communication medium and is connected to the backbone communication networks (WAN) via a base station that is usually placed within the transformer unit. The base station may also be located elsewhere on the powerline, such as at a subscriber's premises.

Many utilities supplying electrical power have their own telecommunication networks that can be used as backbone networks for PLC access systems. If this is not the case, a PLC access network can be connected to a conventional telecommunication network. In any case, the transmitted signal from the backbone has to be converted into a form that makes possible its transmission over a low-voltage power supply network. The conversion takes place at the base station of a PLC system.

PLC subscribers are connected to the network via a PLC modem, placed in the electrical power meter unit (Fig. 2, M). The modem converts the signal received from the PLC network in a standard form that can be processed by the conventional communication systems. On the user side, standard communication interfaces (Ethernet, ISDN S_0, etc.) are offered. Within a house, the transmission can be realized via a separate communication network or via an internal electric installation (in-home PLC solution). In this way, a number of communication devices within a house can be directly connected to a PLC access network.

3.3. In-Home PLC Networks

In-home PLC (indoor) systems use the internal electrical infrastructure as a transmission medium. This makes it possible to setup PLC local networks that connect the typical devices used in private homes, including telephones, computers, printers, and video devices. In the same way, small offices can be interconnected by LAN systems realized with PLC technology. Automation services have become more and more popular not only for their application in the industrial and business sector within large buildings but also for private households. The systems providing automation services such as security observation, heating control, and automatic light control, have to connect a big number of end devices, such as sensors, cameras, electromotors, and lights. Therefore, in-home PLC technology seems to be a reasonable solution for the realization of such networks with a large number of end devices, especially within older houses and buildings that do not have an appropriate internal communication infrastructure.

Basically, the structure of in-home PLC networks is not much different from PLC access systems using low-voltage supply networks. There is also a base/main station that controls the in-home PLC network and connects it to the outdoor area. The base station can be placed within the meter unit (Fig. 2), but also in any other suitable place in the in-home PLC network. All devices of an in-home network, are connected via PLC modems like subscribers of a PLC access network.

An in-home PLC network can exist as an independent network covering only a house or a building. However, it excludes usage and control of in-home PLC services from a distance and also access to WAN services from each electrical socket within a house. In-home PLC networks can be connected to a PLC access system and also to an access network realized by any other communication technology.

3.4. Network Elements

PLC systems consists of the following network elements:

- Basic PLC network elements, which exist in every PLC network
  - PLC modem
  - PLC base/master station
- Additional network elements
  - PLC repeater
  - PLC gateway

A PLC modem connects standard communication equipment, used by PLC subscribers, to a powerline transmission medium. The user-side interface can provide various standard interfaces for different communication devices (e.g., Ethernet and USB interfaces for realization of data transmission and S_0 and *a/b* interfaces for telephony). On the other side, the PLC modem is connected to the power grid using a specific coupling method [3] that allows the feeding of communication signals to the powerline medium and its reception. The coupling has to ensure a safe galvanic separation and to act as a highpass filter dividing the

communication signal (>9 kHz) from the electrical power (50 or 60 Hz). To reduce electromagnetic emission from the powerline (Section 4.4), the coupling is realized between two phases in the access area and between a phase and the neutral conductor in the indoor area. The PLC modem implements all functions of the physical layer including modulation and coding. The second communication layer (link layer) is also implemented within the modem, including its MAC (media access control) and LLC (logical link control) sublayers.

A PLC base station (master station) connects a PLC access system to its backbone network (Fig. 2). It realizes the connection between the backbone communication network and the powerline transmission medium. However, the base station does not connect individual subscriber devices, but it may provide multiple network communication interfaces. Usually, the base station controls the operation of a PLC access network. However, the realization of network control or its particular functions can be realized in a distributed manner.

Additional network elements are needed in some PLC systems to provide signal conversion between different network segments. Repeaters (R) make it possible to realize longer network distances, consisting of several network segments (Fig. 3). The segments are separated by using different frequency bands or by different time slots. In the second case, a time slot is used for the transmission within the first network segment, and another slot for the second segment. A repeater does not modify the contents of the transmitted information.

A gateway is used to interconnect a PLC access network and an in-home PLC network. Similar to a repeater, a gateway also converts the signal between the access and in-home frequencies or time slots. It is usually placed near the house meter unit (Fig. 2) providing the division of the access and in-home areas on the logical network level, too. In this case, an in-home network is fully controlled by the gateway and operates independently form the access network. Therefore, the gateway acts as a subscriber of the access network realizing the connection between the in-home and the access areas. Generally, a gateway can be also placed anywhere in a PLC access network to provide both network segmentation (repeater functionality) and network separation on the logical level.

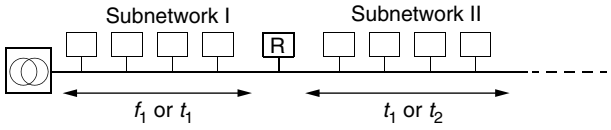


Figure 3. PLC network with repeaters.

4. PLC SYSTEM CHARACTERISTICS

4.1. Network Topology

Low-voltage supply networks are realized by various technologies (different types of cable, transformer units, etc.) according to the existing standards, which differ from country to country. The topologies of low-voltage power supply networks are also different and depend on several factors:

- *Network location* — a PLC network can be installed in a residential, industrial or business area. Furthermore, there is a difference between rural and urban residential areas.
- *Subscriber density* — the number of users/subscribers as well as the user concentration vary from network to network. The subscribers can be placed in family houses (low subscriber density), within houses with several individual customers, in buildings with a larger number of flats or offices, and within apartment or business towers (very high subscriber density).
- *Network span* — the longest distance between the transformer unit and a customer also varies.
- *Network structure* — low-voltage networks usually consist of several network sections/branches, and this number also varies.

Figure 4 shows a possible PLC network structure. There are generally several network sections and branches from the transformer station to the end users. Each branch can have a different topology and connects a variable number of users.

Some characteristic values describing a typical European PLC network, given in Refs. 6 and 7, are listed below — note that the number of users, the number of

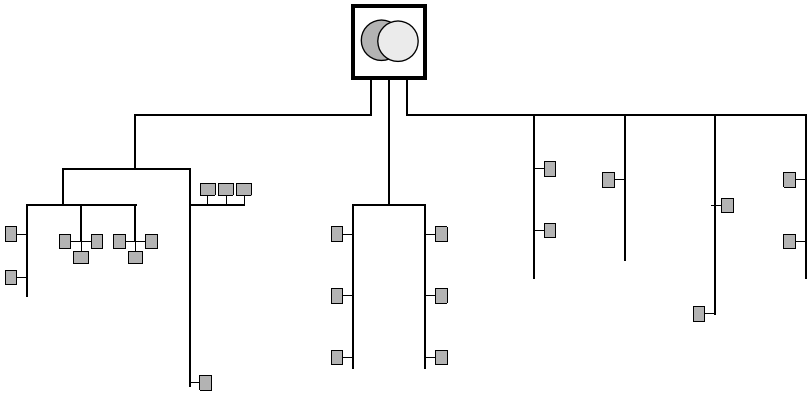


Figure 4. Possible topology of a low-voltage supply network.

potential PLC subscribers, in the United States is significantly lower:

- Number of users in the network: 250–400
- Number of network sections: ~5
- Number of users in a network section: 50–80
- Network span: ~500 m

The PLC base station is usually located within the transformer unit (Fig. 4). However, it could also be placed somewhere else within the network depending on the appropriate interconnection point to the backbone. Each network section/branch may form an individual PLC subnetwork, including a base station and a number of subscribers. Furthermore, a network can be segmented into multiple PLC subnetworks by using repeaters or gateways (Section 3.4). A large PLC network can also cover multiple low-voltage networks. In this case, a number of electrical networks are connected, but only for data transmission (passing only the high-frequency communication signal). Independently of the position of the base station and the kind of network segmentation or concentration, PLC access networks as well as PLC in-home systems physically maintain a tree network topology.

In any network configuration, the communication between PLC subscribers and the WAN is carried out over a base station. This is also the case for internal communication between PLC subscribers. A signal sent by any network station to the base station (uplink transmission direction) reaches all network stations. This is also the case for a signal transmitted in the downlink. Thus, a PLC access network has a star topology with a base station in its center, which leads to a logically bus network structure representing a shared transmission medium.

4.2. Characteristics of the PLC Transmission Channel

The electrical power grid has not been designed to support telecommunication services. To use these lines as an information transmission medium, we need to determine the transfer characteristics, including the attenuation and the impedance. From these parameters, we construct a mathematical channel model that is used in the design of the PLC systems.

The propagation of signals over powerlines introduces an attenuation, which increases with the length of the line and frequency. The attenuation is a function of the powerline characteristic impedance Z_L and the propagation constant γ [8]. These two parameters are defined by the primary resistance R' per unit length, the conductance G' per unit length, the inductance L' per unit length and the capacitance C' per unit length, which are generally frequency dependent, as

$$Z_L(f) = \sqrt{\frac{R'(f) + j2\pi f \cdot L'(f)}{G'(f) + j2\pi f \cdot C'(f)}} \tag{1}$$

and

$$\gamma(f) = \sqrt{(R'(f') + j2\pi f L'(f)) \cdot (G'(f') + j2\pi f C'(f))} \tag{2}$$

$$\gamma(f) = \alpha(f) + j\beta(f) \tag{3}$$

Considering a matched transmission line, which is equivalent to regarding only the wave propagation from source to destination, the transfer function of a line with length l can be expressed by

$$H(f) = e^{-\gamma(f)\cdot l} = e^{-\alpha(f)\cdot l} \cdot e^{-j\beta(f)\cdot l} \tag{4}$$

Investigations and measurements of the fundamental properties of power cables have revealed that $R'(f) \ll 2\pi f L'(f)$ and $G'(f) \ll 2\pi f C'(f)$ is valid in the frequency range from 1 to 30 MHz. Consequently, the dependence of L' and C' on frequency is neglected so that the characteristic impedance Z_L and the propagation constant γ in this frequency range can be expressed as [8]

$$Z_L = \sqrt{\frac{L'}{C'}} \tag{5}$$

and

$$\gamma(f) = \frac{1}{2} \cdot \frac{R'(f)}{Z_L} + \frac{1}{2} \cdot G'(f) \cdot Z_L + j2\pi f \cdot \sqrt{L'C'} \tag{6}$$

$$\gamma(f) = k_1 \cdot \sqrt{f} + k_2 \cdot f + jk_3 \cdot f \tag{7}$$

$$\gamma(f) = \alpha(f) + j\beta(f) \tag{8}$$

where k_1, k_2, and k_3 are constants.

The real part of the propagation constant, describing the cable losses, can be approximated by the equation

$$\alpha(f) = a_0 + a_1 \cdot f^k \tag{9}$$

and with a suitable selection of the attenuation parameters a_0, a_1, and k, the powerline attenuation, representing the amplitude of the channel transfer function, can be defined by the formula [9]

$$A(f, l) = e^{-\alpha(f)\cdot l} = e^{-(a_0 + a_1 \cdot f^k)} \cdot l \tag{10}$$

where l represents the length of the path for the signal wave propagation and k is the exponent of the attenuation factor.

4.3. PLC Channel Model

In addition to the frequency-dependent attenuation that characterizes the powerline channel, deep narrowband notches occur in the transfer function, which may be spread over the whole frequency range. These notches are caused by multiple reflections at impedance discontinuities. The length of the impulse response and the number of the occurred peaks can vary considerably depending on the environment. This behavior can be described by the "echo model" [10]. Transfer characteristics of powerline channels can be regarded as quasistationary, as their changes occur only as a result of changes in the topology and changes in the load situation. Load changes are caused mainly by the connecting or switching of electrical appliances.

Complying with the echo model, each transmitted signal reaches the receiver on N different paths (see Fig. 5). Each path i is defined by a certain delay τ_i and

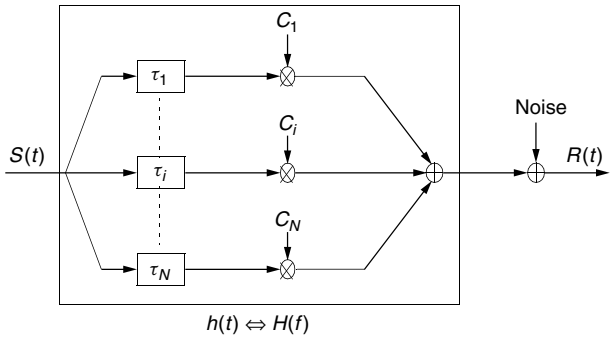


Figure 5. The echo model representing the PLC channel.

a certain attenuation factor C_i. The PLC channel can be described by means of a discrete-time impulse response:

$$h(t) = \sum_{i=1}^{N} C_i \cdot \delta(t - \tau_i) \Leftrightarrow H(f) = \sum_{i=1}^{N} C_i \cdot e^{-j2\pi f \tau_i} \qquad (11)$$

Factoring in the formula of the channel attenuation, the transfer function in the frequency domain can be written

$$H(f) = \sum_{i=1}^{N} g_i \cdot A(f, l_i) \cdot e^{-j2\pi f \tau_i} \qquad (12)$$

where g_i is a weighting factor representing the product of the reflection and transmission factors along the path. The variable τ_i, representing the delay introduced by the path i, is a function of the pathlength l_i.

By replacing the medium attenuation $A(f, l_i)$ by the expression given in Eq. (10), we obtain the final equation defining the PLC channel model, encompassing the parameters of its three main characteristics: attenuation, impedance fluctuations, and multipath effects. This equation is composed of a weighting term, an attenuation term, and a delay term:

$$H(f) = \sum_{i=1}^{N} g_i \cdot e^{-(a_0 + a_1 \cdot f^k) \cdot l_i} \cdot e^{-j2\pi f \tau_i} \qquad (13)$$

4.4. Electromagnetic Compatibility

Broadband PLC systems operate in the high-frequency range from 1.6 to 30 MHz, in order to achieve data rates above 1 Mbps and to avoid the high noise level in the low-frequency range. On the other hand, experiments have shown that PLC systems must overcome an attenuation of about 70 dB, in order to reach from the transformer unit to the subscribers premises [11]. Unlike other communications media, powerlines are electrically asymmetric, as they were not built to transmit information. As a consequence, electromagnetic fields emitted by these lines are high. A solution has to be found to guarantee the coexistence of PLC systems and the radio systems operating in the same spectrum. To solve the problems of electromagnetic compatibility, two solutions are proposed [3]:

1. According to the regulating administration for telecommunications and post (RegTP) in Germany, at present a total spectrum of approximately 7.5 MHz in the frequency range between 0 and 30 MHz may be used principally for PLC. This spectrum is not contiguous, as schematically represented in Fig. 6. It represents some gaps of different width and distributed arbitrarily in the frequency band depicted. This is to secure certain public frequency bands.

At a first glance, it appears feasible to assign the open gaps to PLC services, permitting an increased transmission power spectral density within these chimneys. From such a solution, multicarrier modulation schemes are attractive, particularly those that are able to use narrow gaps with high spectral efficiency, such as orthogonal frequency division multiplex (OFDM). Unfortunately this chimney approach (Fig. 6) encounters some serious problems that render this solution less practical: (a) these gaps are not really free, but they are already dedicated to certain primary users for wireless services, who reserved them for future use; and (b) if the chimneys are very narrow, their use will require the implementation of complex filters.

2. Another proposed solution to accomplish EMC without the chimney approach is a general limitation of radiated fields from powerlines. In March 1999, the RegTP in Germany issued a limitation for the "radiation of telecommunications services in and alongside of cables" (including CATV, xDSL, and PLC). These radiation limitations are part of a plan for the frequency allocation in Germany and are known as NB30.

In comparison with the FCC Part 15, which is the EMC American standard for wire communications, the disadvantages imposed by the German regulations become clearly obvious. At ~2 MHz, for example, the allowed American limits are ~30 dB above the NB30. This means a factor of 1000 in terms of transmitted power or a factor of ~10 in data rate.

4.5. Noise Behavior

Generally, noise in PLC networks can be defined as a superposition of five types, distinguished by their origin, time duration, and spectrum occupancy (see Fig. 7) [12]:

- *Colored background noise* — whose power spectral density (PSD) is relatively low and decreasing with

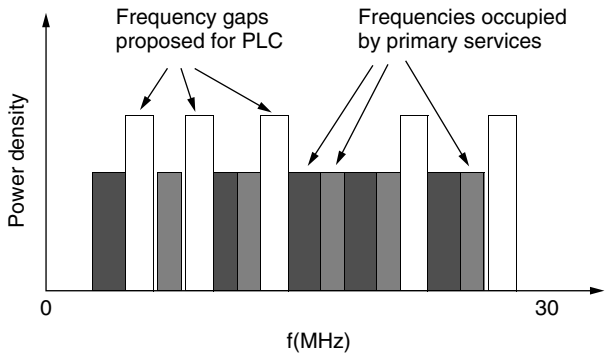


Figure 6. The "chimney approach" for PLC frequency allocation guaranteeing EMC.

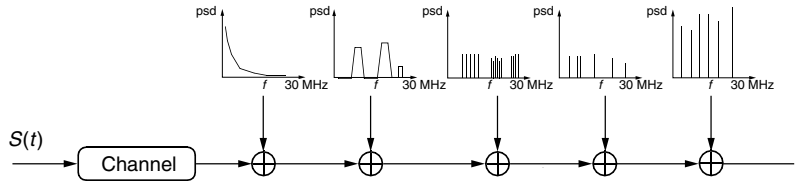


Figure 7. Noise types present in the PLC environment.

frequency. This type of noise is caused mainly by a superposition of numerous noise sources with lower power. Its parameters vary over time in terms of minutes and hours.

- *Narrowband noise* — mostly sinusoidal signals, with modulated amplitudes. This type of noise is caused mainly by ingress of broadcast stations in the medium- and short-wave broadcast bands. Their amplitude generally varies during the daytime.
- *Periodic impulsive noise*, *asynchronous to the main frequency* — impulses that usually have a repetition rate between 50 and 200 kHz, which results in a spectrum with discrete lines with a frequency spacing according to the repetition rate. This type of noise is caused mostly by switching power supplies.
- *Periodic impulsive noise*, *synchronous with the main frequency* — impulses that have a repetition rate of 50 or 100 Hz and are synchronous with the main powerline frequency. They are of short duration (some microseconds) and have a PSD that decreases with frequency. This type of noise is caused by power supplies operating synchronously with the main frequency.
- *Asynchronous impulsive noise* — a type of impulsive noise caused by switching transients in the networks. The impulses have durations of some microseconds up to a few milliseconds with arbitrary arrival times. The PSD of this type of noise can reach values of more than 50 dB above the background noise, which makes it the main cause of occurrence of error in digital transmission.

4.6. PLC Services

PLC access networks together with their backbone provide a bearer service ensuring realization of different teleservices that allow the use of various communication applications [7]. Accordingly, the specification of a PLC transmission system has to include the definition of specific bearer network layers: the physical layer, including modulation and coding methods as well as the data-link layer specifying MAC and LLC sublayers. PLC networks have to offer a large palette of telecommunications services with certain quality requirements to be able to compete with other communications technologies applied to the access area. Therefore, the following four groups of teleservices have to be provided by PLC access networks:

- Connectionless services without QoS guarantees
- Connection-oriented constant-bit-rate (CBR) services, such as telephony
- Connectionless services with QoS guarantees
- Specific PLC services

Present PLC systems provide the connectionless services offering high-speed Internet access to customers. However, PLC networks should also support the classical telephone service, because of its significant penetration in the communications world. Therefore, the manufacturers of PLC equipment have already released systems supporting telephony service, based mostly on VoIP solutions. A special emphasis has to be given to the specific PLC services (home automation, energy management, security, remote functions, etc.), which usually require low data rates and do not require very low transmission delays. So, most of these services can be realized by the connectionless services class without guarantees.

The support of teleservices mentioned above can ensure a competitive position of PLC networks toward other access technologies. However, further development of PLC access networks leads to realization of CBR services with higher data rates and connectionless data services achieving higher QoS requirements.

5. REALIZATION OF PLC NETWORKS

5.1. Specific Performance Problems

The regulatory bodies specify the limits for electromagnetic radiation, which is allowed to be produced by PLC systems operating out of the frequency range defined by the CENELEC standard. In Germany, NB30 directions define very low radiation limits for systems like PLC, which operate in the frequency range up to 30 MHz (see Section 4.4). Accordingly, PLC networks have to operate with a limited signal power to stay within the NB30 limits. Because of the limited signal power, PLC networks are more sensitive to disturbances and are not able to span longer distances.

Well-known error-handling mechanisms can also be applied to PLC systems to reduce the problem of transmission errors caused the disturbances [e.g., forward error correction (FEC) and ARQ]. However, the application of FEC consumes an additional part of the transmission capacity because of the overhead needed for the error detection and correction. ARQ retransmits defective data units, which also consumes a part of the transmission capacity. Additionally, the powerline is a transmission medium that has to be shared by all PLC subscribers.

PLC systems have to compete with other access technologies and to offer a satisfactory QoS and sufficient data rates, but at the same time, to be economically efficient. Therefore, broadband PLC systems have to be provided with the following features:

- Application of efficient modulation schemes ensuring a good utilization of used frequency spectrum and a certain robustness against disturbances

- Realization of suitable coupling methods to reduce electromagnetic radiation
- Implementation of efficient media access control (MAC) protocols to achieve near-maximal utilization of limited PLC network capacity and realization of needed QoS
- Application of optimal error-handling methods to deal with an unfavorable noise scenario consuming a minimum of network resources

5.2. Modulation and Transmission Schemes

Within the PLC systems, the communication is supposed to occur in a channel characterized by frequency-selective phenomena, presence of echoes, and impulsive and colored noise with the superposition of narrowband interferences. This requires that the modulation scheme adopted for PLC effectively face such a hostile environment.

Direct-sequence code-division multiple access (DSCDMA) and orthogonal frequency-division multiplexing (OFDM) are considered as candidates for future broadband PLC [13,14] (see Table 1). As DSCDMA and OFDM permit the separation of the overall transmitted data in many parallel independent substreams, flexible resource management strategies can be implemented. This characteristic is very important in order to cope with channel impairments and to provide fine granularity. This fine granularity is necessary for multimedia services and to achieve a high utilization.

The CDMA technique has the advantages of robustness to narrowband interference and multiple access with low power spectrum density, thus reducing EMC problems. On the other hand, the OFDM technique allows for significant reduction of channel equalizer complexity and increased resistance to narrowband and impulsive noise. Moreover, bit-loading techniques make it possible to achieve a capacity very near the theoretical limit, but at the cost of an increased system complexity, [14].

Overall OFDM seems to be advantageous compared to DSCDMA:

1. The main advantage is obtained by the fact that the channel (the spectrum) is divided into many narrow subchannels. Therefore, equalizing in OFDM is a simple procedure, compared with wideband equalizing.
2. OFDM inherently solves one essential problem associated with high-speed PLC: intersymbol interference (ISI) caused by the multipath delay spread. This is achieved by the introduction of the "guard interval," which is filled by a cyclic prefix.

For channel coding, two variants have been investigated: Reed–Solomon coding [15] and Hamming codes [3]. Several types of interleavers can be implemented with different variants of OFDM, such as OFDM with diversity or adaptive OFDM [15].

5.3. MAC Layer

The MAC layer specifies a multiple-access scheme and a resource-sharing strategy (MAC protocol). Particularly in PLC systems, the MAC layer has to be robust against disturbances and must allow for the use of various telecommunication services [16].

The most widely applied access scheme in PLC networks is TDMA. Because of the disturbances, data packets (e.g., IP packets) are usually segmented into smaller data units whose size is chosen according to the length of a time slot specified by the TDMA scheme. So, if a disturbance occurs, only erroneous data segments are retransmitted. This consumes a smaller network capacity than does retransmission of the entire data packets. The data segmentation ensures a fine network granularity and an easier provision of QoS guarantees.

An effective solution to avoid the influence of narrowband disturbances is to apply FDMA methods, where particular carrier frequencies can be switched off, if they are affected by narrowband disturbances. Therefore, a TDMA/FDMA combination seems to be a reasonable solution for PLC networks (realized by Ascom [17]). PLC networks using OFDM modulation can also provide transmission channels distributed in a frequency spectrum. Unlike FDMA, the transmission channels are realized by a number of subcarriers, which leads to an OFDM access scheme (OFDMA). A further division of the transmission channels in the time domain can also be done according to a slotted nature of OFDM transmission systems combining both multiple-access schemes (OFDMA/TDMA) [18].

MAC protocols for PLC systems have to achieve a maximum utilization of the limited network capacity and to realize time-critical telecommunication services. This can be ensured by reservation of bandwidth that allows particular QoS guarantees needed for various services [16]. In the case of reservation protocols, a part of the network resources is reserved for the reservation procedure: signaling, organized according to a random or dedicated access principle as well as by application of various hybrid solutions [18]. Figure 8 presents the signaling delay caused by the reservation protocols with a signaling procedure organized according to random and dedicated access methods. The protocols based on random access achieve significantly shorter signaling delay, if the transmission requests are relatively infrequent. However, in the case of frequent requests, protocols with dedicated

Table 1. Comparison of PLC Modulation Candidates

	Spectral Efficiency (bps/Hz)	Maximum Data Rate (Mbps)	Robustness against Channel Distortions	Flexibility and Adaptive Features	System Costs (Including Equalizers and Repeaters)	EMC
DSCDMA	<0.1	~ 0.5	Poor	Very poor	Very poor	Excellent
OFDM	$\gg 1$	>10	Excellent	Excellent	Poor	Good

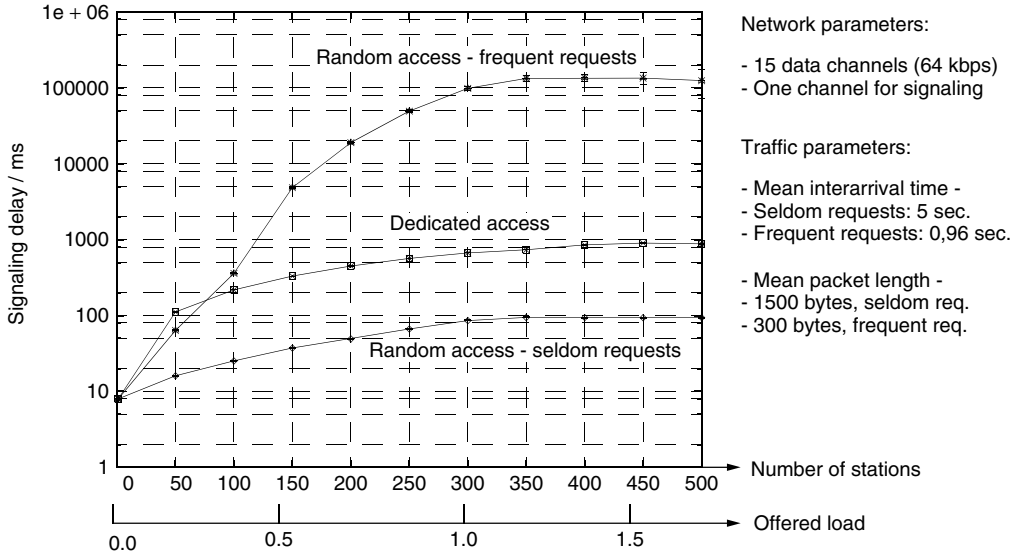


Figure 8. Average signaling delay of different reservation protocols.

access behave much better. So, an optimal solution for the organization for the reservation/signaling procedure can be achieved by a hybrid protocol changing dynamically from random to dedicated access depending on the current situation in the network (network load, number of active stations, etc.). However, the reservation procedure has to be particularly protected against disturbances [19]. Besides reservation MAC protocols, variations of the CSMA/CA protocol ensuring realization of multiple priorities for different services [20] are also widely applied to PLC systems (e.g., Itran technology [21]).

Because of the asymmetric and changing nature of data traffic in the access area, dynamic duplex schemes are used in PLC access networks. This allows the optimal utilization of the network resources, in both downlink and uplink transmission directions according to the current load situation. However, the relatively small PLC network capacity makes it difficult for the simultaneous provision of a required QoS for a high number of subscribers. Therefore, PLC systems have to implement traffic scheduling strategies, including connection admission control (CAC), to limit the number of active subscribers ensuring a satisfactory QoS for currently admitted connections. In the same way, a part of the network resources has to be reserved for capacity reallocation in case of disturbances.

5.4. Error Handling

Because of the special disturbance characteristic, error handling within different network layers warrants careful attention [19]. After a correct dimensioning, it is assumed that the signal-to-noise ratio (SNR) is sufficient to avoid any influence of the background noise. However, the impulsive noise makes it difficult to ensure error-free transmission (Section 4.5). Its influence is reduced by the following methods:

- Setting a sufficient duration of the transmitted symbol within the physical network layer (e.g., duration of an OFDM symbol) eliminates disturbances that are shorter than the symbol duration.
- Channel coding using FEC mechanisms and interleaving allow for the correction of a number of erroneous bits in the case of various kind of disturbances (e.g., single, periodic, and burst errors). However, FEC mechanisms provide overhead, and interleaving increases the transmission delay. Because of the limited network capacity and the demand for low delays, channel coding in PLC networks is organized in a dynamic manner, providing a changing level of protection according to the current noise conditions in the network.
- If the reduction of the BER by channel coding is not sufficient and delay requirements are not too hard (e.g., data traffic), even ARQ mechanisms (retransmission of erroneous data) can be used.

Application of ARQ can improve network performances significantly (Fig. 9). Network utilization in a network without ARQ (only a simple packet retransmission is provided) can achieve a maximum of 50%. Application of go-back-N ARQ retransmitting smaller portions of erroneous packets (disturbed data segments) improves the network utilization by 23%. If the ARQ is additionally adjusted to a per packet reservation method (e.g., ARQ + mechanism [19]), utilization achieves 81%.

In the case of long-term disturbances, a part of the network capacity is unavailable for a longer time. This cannot be improved by FEC and ARQ. Therefore, this part of the transmission capacity (a frequency spectrum) is temporarily not usable (is switched off).

6. SUMMARY

Present powerline communications systems, using electrical power grids as transmission media, provide relatively high data rates (>2 Mbps). PLC can be applied to high-,

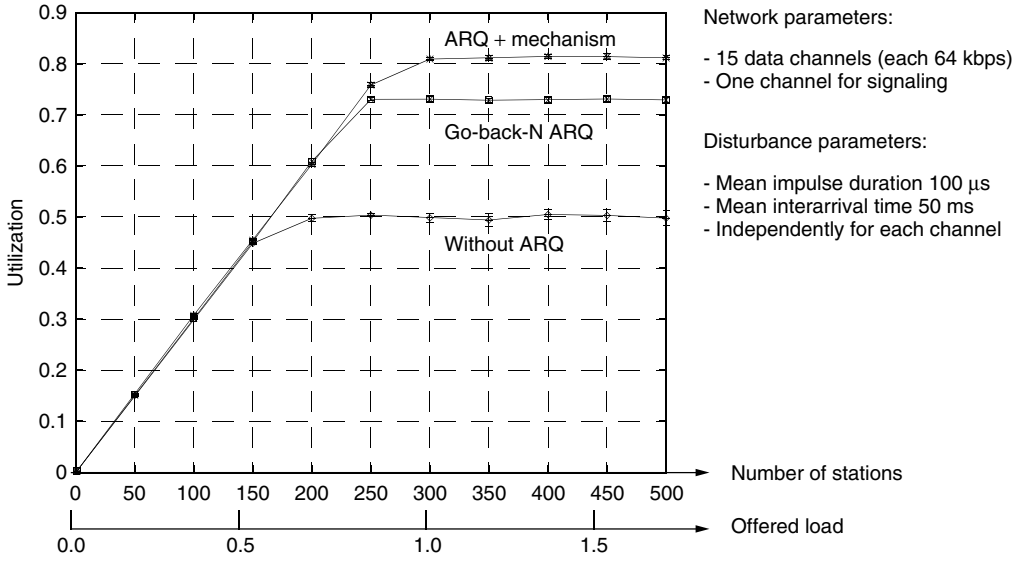


Figure 9. Performances of different ARQ mechanisms — average network utilization.

medium-, and low-voltage supply networks as well as within buildings. PLC technology is now used mainly for access networks and in-home communication networks. This is because of the high cost of the access networks (about 50% of the investments in network infrastructure are needed for the access area) and the liberalization of the telecommunication market in many countries.

Power supply networks are not designed for communications, and they do not present a favorable transmission medium. The PLC transmission channel is characterized by a large, frequency-dependent attenuation, changing impedance, and fading, as well as a strong influence of noise. On the other hand, broadband PLC networks have to operate in a frequency spectrum up to 30 MHz, which is used by various radio services, too. Therefore, the regulatory bodies specify very strong limits regarding the electromagnetic emission from PLC networks to the environment. As a consequence, PLC networks have to operate with a limited signal power, which reduces network distances and data rates, and also increases sensitivity to disturbances.

To reduce the negative impact of powerline transmission medium, PLC systems apply efficient modulation, such as OFDM, which is able to avoid restricted frequency bands. OFDM is also robust against narrowband and impulsive noise and can deal with fading. The problem of longer noise spikes can be solved by well-known error-handling mechanisms (e.g., FEC, ARQ). However, their application consumes a certain portion of the PLC network capacity due to overhead and retransmission. The PLC bandwidth is shared by the subscribers, and therefore any reduction of capacity due to protocol overhead should be minimized. At the same time, PLC systems have to compete with other access technologies and to offer a wide palette of telecommunication services with a satisfactory QoS. Both good network utilization and provision of QoS guarantees can be achieved by an efficient MAC layer. We have shown that reservation-based MAC protocols can achieve these goals efficiently.

ACRONYMS

ADSL	Asymmetric digital subscriber line
ARQ	Automatic repeat request
BER	Bit error rate
CAC	Connection admission control
CATV	Cable TV
CBR	Constant bit rate
CFS	Carrier frequency systems
CSMA/CA	Carrier sensing multiple access with collision avoidance
DS-CDMA	Direct sequence code division multiple access
DSL	Digital subscriber line
EMC	Electromagnetic compatibility
ETSI	European Telecommunications Standards Institute
FDD	Frequency-division duplex
FDMA	Frequency-division multiple access
FEC	Forward error correction
FSK	Frequency shift keying
IP	Internet Protocol
ISDN	Integrated Services Digital Network
LAN	Local-area network
LLC	Logical link control
MAC	Media access control
OFDM	Orthogonal frequency-division multiplexing
OFDMA	OFDM access
PLC	Powerline communications
psd	Power spectral density
QoS	Quality of service
RCS	Ripple carrier signaling
RegTP	Regulating Administration for Telecommunications and Post
TDD	Time-division duplex

TDMA	Time-division multiple access
VBR	Variable bit rate
VoIP	Voice over IP
WAN	Wide-area network

BIOGRAPHIES

Halid Hrasnica, graduated in 1993 at the Faculty for Electrical Engineering — Department for Telecommunications, at the University of Sarajevo — Bosnia and Herzegovina. From 1993 to 1995 he was working in Energoinvest Communications in Sarajevo as developing software engineer for the telephone exchange systems. In 1995 he joined the Chair for Telecommunications at Dresden University of Technology, Germany, as visitor scientist. Since 1996 he is research assistant at Dresden University of Technology and he is currently working toward his Ph.D. He has been involved in several research projects: development of least cost routing strategy, performance analysis and simulation of broadband communications networks. His current research interest is performance analysis of powerline communication networks and investigation on PLC MAC layer and protocols.

Abdelfatteh Haidine received the B.S. in electronics and telecommunications from the University Cadi Ayyad in Marrakech and the MSc in telecommunications from University Chouaib Doukkali El Jadida in 1999, in Morocco. Since 2000, he joined the University of Technology Dresden in Germany as research assistant. He worked in different European and German projects about power line communication networks. His actual research domain is planning and optimization of the access network based on optical fiber and Very high bit Digital Subscriber Line (VDSL) technology.

Ralf Lehnert received both his 1972 diploma degree and the 1979 Ph.D. degree in electrical engineering from Aachen University, Germany. Since 1980 he has been with the basic development department at Philips Communication Systems in Nuremberg, Germany, as the head of a group on applied research in performance evaluation of communication networks, traffic engineering and network planning. In July 1994 he took over the chair for telecommunications, as a full professor in the department of electrical engineering at Dresden University, Germany. His current research interests are in the field of performance evaluation of telecommunication systems, including modeling of B-ISDN networks, network planning and optimization. He has been involved in RACE (Parasol, Exploit, Tribune) and ACTS projects (Expert) all on the subject of ATM.

BIBLIOGRAPHY

1. Cenelec (online) *http://www.cenelec.org* (June 2002).
2. Reg TP — Regulierungsbehörde für Telekommunikation und Post (online), *http://www.regtp.de* (June 2002).
3. K. Dostert, *Powerline Communications*, Prentice-Hall, 2001.
4. PLCforum (online), *http://www.tech.ascom.ch/preview/plc/index.htm* (June 2002).
5. HomePlug Powerline Alliance (online), *http://www.homeplug.org/index_basic.html* (June 2002).
6. O. G. Hooijen, On the channel capacity of the residential power circuit used as a digital communications medium, *IEEE Commun. Lett.* **2**(10): (Oct. 1998).
7. H. Hrasnica and R. Lehnert, *Powerline Communications in Telecommunication Access Area*, VDE World Microtechnologies Congress, MICRO.tec 2000 ETG-Fachtagung und -Forum: Verteilungsnetze im liberalisierten Markt, Sept. 25–27, 2000. Expo 2000, Hannover, Germany.
8. M. Zimmerman and K. Dostert, The low voltage power distribution network as last mile access network — signal propagation and noise scenario in the HF-range, *Int. J. Electron. Commun.* **54**(1): 13–22 (2000).
9. M. Zimmerman, *Energieverteilnetze als Zugangsmedium fuer Telekommunikationsdienste*, Ph.D. dissertation, Karlsruhe Univ. Technology, Shaker Verlag 2000 (in German).
10. H. Philipps, Development of a statistical model for powerline communication channels, *Proc. Int. Symp. Power-Line Communications and Its Applications*, Limerick, Ireland, 2000.
11. H. Dalichau, *EMV-Aspekte von Inhome-PLC-Anlagen, Vergleich des kHz-Bereiches mit dem MHz-Bereich*, EMC Kompendium, 2002 (in German).
12. M. Zimmerman and K. Dostert, An analysis of the broadband noise scenario in powerline networks, *Proc. Int. Symp. Power-Line Communications and Its Applications*, Limerick, Ireland, 2000.
13. S. Tachikawa, M. Nari, and M. Hamamura, Power line data transmission using OFDM and DS/SS systems, *Proc. 6th Int. Symp. Power-Line Communications and Its Applications*, Athens, Greece, 2002.
14. E. Del Re, R. Fantacci, S. Morosi, and R. Seravalle, Comparison of CDMA and OFDM systems for broadband downstream communications on low voltage power grid, *Proc. 5th Int. Symp. Power-Line Communications and Its Applications*, Malmö, Sweden, 2001.
15. T. Waldeck, *Einzel- und Mehrträgerverfahren für die störresistente Kommunikation auf Energieverteilnetz*, Ph.D. dissertation, Karlsruhe Univ. Technology, 1999, Logos Verlag, Berlin, 2000 (in German).
16. H. Hrasnica, A. Haidine, and R. Lehnert, Reservation MAC protocols for powerline communications, *Proc. 5th Int. Symp. Power-Line Communications and Its Applications* (*ISPLC2001*), Malmö, Sweden, April 4–6, 2001.
17. Ascom Powerline Communications, *Powerline System Description*, version 1.2 (March 2002).
18. H. Hrasnica and R. Lehnert, Performance analysis of polling based reservation MAC protocols for broadband PLC access networks, *Proc. 14th Int. Symp. Services and Local Access* (*ISSLS2002*), Seoul, Korea, April 14–18, 2002.
19. H. Hrasnica and R. Lehnert, Performance analysis of error handling methods applied to a broadband PLC access network, *Proc. SPIE Int. Symp. ITCom2002*, Boston, MA, July 29–Aug. 1, 2002.
20. T. Langguth et al., *Performance study of access control in power line communications, Proc. 5th Int. Symp. Power-Line Communications and Its Applications* (*ISPLC2001*), Malmö, Sweden, April 4–6, 2001.
21. Itran Communications Ltd. (online), *http://www.itrancomm.com/index1.html* (June 2002).

PRODUCT CODES

FRANK R. KSCHISCHANG
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

The *product construction*, introduced by Peter Elias in 1954 [1], is one of the simplest ways to combine simple error control codes to get more powerful ones, and any code that results from the product construction is called a *product code*. The codewords of a product code are defined as matrices that are constrained, like crossword puzzles, so that rows and columns form valid codewords in some constituent codes. Decoding typically also proceeds in crossword-puzzle fashion, alternating between the horizontal and vertical constraints in an attempt to find a valid codeword near the received word. Although more complicated codes may offer superior performance, product codes are often attractive for practical implementation — they can be designed to provide excellent performance with reasonably low decoding complexity, and their structure leads to hardware decoder implementations with natural parallelism that is easily exploited.

This article describes product codes, derives some of their basic properties, discusses methods to decode them, and provides a number of examples.

1.1. Linear Codes

We consider only *linear block codes* in this article, so, to establish notation and to be somewhat self-contained, we will start with a brief review of their properties. The basic theory of linear block codes is described in every textbook on coding theory; see, [e.g., 2,3]. While all of the constructions we describe will apply to codes defined over any finite field $\mathbb{F}$, for simplicity we will confine all of our examples to codes defined over the binary field $\mathbb{F}_2 = \{0, 1\}$, with all scalar arithmetic (addition and multiplication) computed modulo 2. For use throughout this article, it is convenient to denote the vector space of $m \times n$ matrices over $\mathbb{F}$ as $\mathbb{F}^{m\times n}$, where m and n are positive integers. The space of $1 \times n$ row vectors is denoted as $\mathbb{F}^n$. If $\mathbf{M}$ is a matrix in $\mathbb{F}^{m\times n}$, then $[\mathbf{M}]_{i,j}$ denotes the component of $\mathbf{M}$ in row i, column j, where $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$.

A block code C of length n over $\mathbb{F}$ is any nonempty subset of $\mathbb{F}^n$. More specifically, C is *linear* if it forms a subspace of $\mathbb{F}^n$, that is, if C satisfies all the axioms that define a vector space, including the requirement that every $\mathbb{F}$-linear combination of codewords is itself a codeword. A code of length n and dimension k is referred to as an $[n, k]$ code.

The simplest example of a binary linear code is the $[k + 1, k]$ single-parity-check (SPC) code, which consists of all binary vectors of length $k + 1$, each having an even number of ones. For example, the $[3, 2]$ SPC code has four codewords: $\{000, 011, 101, 110\}$. It is easy to check that any linear combination of codewords is a codeword; for example, $011 + 101 = 110$.

An *encoder* for a linear code C maps a message of k (input) symbols to a codeword of $n \geq k$ (output) symbols. The ratio of input to output symbols, namely, k/n, is called the *rate* of the code. An encoder is called *systematic* if the message symbols appear directly as the first k symbols of each codeword. A systematic encoder for the binary SPC code would take in k message bits and append a single *check bit*, chosen to make the total number of ones in the codeword an even number.

1.2. Generator and Parity-Check Matrices

An $[n, k]$ code C is a k-dimensional vector space; accordingly, it is always possible to find k linearly independent vectors $g_1, \ldots, g_k$ in C. Any such set is a basis or *generating set* for C, since every codeword v of C may be expressed (uniquely) as an $\mathbb{F}$-linear combination of the generators

$$v = \sum_{i=1}^{k} u_i g_i, \tag{1}$$

where $u_i \in \mathbb{F}$ for all $i \in \{1, \ldots, k\}$. Writing $u = (u_1, \ldots, u_k)$ as a (row) vector with k components, and collecting $g_1, \ldots, g_k$ as the rows of a matrix $\mathbf{G} \in \mathbb{F}^{k\times n}$, we may express (1) in matrix form as $v = u\mathbf{G}$. The matrix $\mathbf{G}$ is referred to as a *generator matrix* for C, and C is equal to the *row space* of $\mathbf{G}$. Since C can have many different bases, it follows that C can have many different generator matrices (all of which, however, must have row space C).

Suppose that C has a generator matrix of the form $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$, where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{P} \in \mathbb{F}^{k\times(n-k)}$ is arbitrary. Then multiplication of an information vector u by $\mathbf{G}$ results in the systematic encoding $u\mathbf{G} = (u, u\mathbf{P})$ of u. In this case, $\mathbf{G}$ is said to be in *systematic form*. Although not every code has a generator matrix in systematic form, the $[3, 2]$ SPC code *does* have a generator matrix in systematic form given by

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

In general, since multiplication by $\mathbf{G}$ effects a one-to-one mapping from $\mathbb{F}^k$ to $\mathbb{F}^n$, it follows that $\mathbf{G}$ always has at least one (in general more than one) right-inverse $\mathbf{G}^{-1} \in \mathbb{F}^{n\times k}$, with the property that $\mathbf{G}\mathbf{G}^{-1} = \mathbf{I}_k$. In other words, from every codeword $c = u\mathbf{G}$, it is always possible to recover a unique message $u = c\mathbf{G}^{-1}$.

An alternative description of a code C arises as follows. Define the scalar ("inner" or "dot") product between two vectors $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$ as $\langle v, w\rangle := \sum_{i=1}^{n} v_i w_i$. For every $[n, k]$ code C, one can always find a set of linearly independent n vectors $h_1, \ldots, h_{n-k}$ such that the scalar product $\langle h_i, v\rangle = 0$ for all $i \in \{1, \ldots, n - k\}$ if and only if v is a codeword of C. Collecting $h_1, \ldots, h_{n-k}$ as the rows of an matrix $\mathbf{H} \in \mathbb{F}^{(n-k)\times n}$, we see that a given vector v is a codeword of C if and only if v satisfies the *parity-check equation* $v\mathbf{H}^T = 0$. The matrix $\mathbf{H}$ is called a *parity-check matrix for* C. As is the case with generator matrices, a code C can have many different parity-check matrices. One possible parity-check matrix

for a code with systematic generator matrix $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$ is $\mathbf{H} = [-\mathbf{P}^T \mid \mathbf{I}_{(n-k)}]$. For example, the $[3, 2]$ SPC code has parity-check matrix $\mathbf{H} = [111]$. In general, it is useful to allow *any* matrix $\mathbf{H}$ with n columns and rank $n - k$ satisfying $v\mathbf{H}^T = 0$ for all $v \in C$ to serve as a parity-check matrix for C, even if $\mathbf{H}$ contains some linearly dependent rows (redundant checks).

1.3. Minimum Hamming Distance

The *Hamming weight* $wt(v)$ of a vector v is defined as the number of positions in which v is nonzero. For example, $wt(110) = 2$. The *Hamming distance* $d(v, w)$ between two vectors v and w of the same length is defined as the Hamming weight of their difference: $d(v, w) := wt(v - w)$. For example, $d(011, 101) = wt(011 - 101) = wt(110) = 2$. It is easily seen that the Hamming distance between two vectors is the number of positions in which the two vectors differ. Thus, since 011 and 101 differ in their first two positions, but not in their last position, we have $d(011, 101) = 2$.

The minimum Hamming distance between pairs of *distinct* codewords in a code C is called the *minimum distance of* C, and is denoted $d_{\min}(C)$. The minimum distance of a code has traditionally been regarded as a parameter of fundamental importance, since this parameter determines the error correction radius $t = \lfloor (d_{\min} - 1)/2 \rfloor$ within which all error patterns are guaranteed correctable by any decoder that maps a received word to the nearest (in the sense of Hamming distance) codeword. In general, it is desirable to make the minimum distance d as large as possible for a given n and k.

For linear codes, since $d(v, w) = wt(v - w)$, the Hamming distance between two codewords v and w is always equal to the weight of some codeword, namely, $v - w$. In other words, every distance (between two codewords) is a weight (of some codeword). On the other hand, since $wt(v) = d(v, 0)$, every weight (of some codeword) is a distance (between two codewords). This equivalence between weight and distance implies that the minimum distance of a linear code C is given by the minimum weight of its nonzero codewords. The $[3, 2]$ SPC code, for example, has minimum distance 2. In general, an $[n, k]$ code with minimum distance d is often referred to as an $[n, k, d]$ code.

For many values of n and k, there are methods known to construct $[n, k]$ codes with large minimum distance, (e.g., see the articles BCH CODES, CYCLIC CODES, FINITE GEOMETRY CODES, HADAMARD CODES, REED–MULLER CODES, and REED–SOLOMON CODES in this encyclopedia).

2. THE PRODUCT CONSTRUCTION

2.1. Definition and Basic Properties

Let A and B be codes over $\mathbb{F}$ of length n_A and n_B, respectively. Whereas the codewords of A and B are vectors, we will now define a code of length $n_A n_B$, the *direct product* of A and B, whose codewords may be viewed as $n_A \times n_B$ matrices over $\mathbb{F}$, namely, as elements of $\mathbb{F}^{n_A \times n_B}$. Every such matrix is readily converted to a vector of length $n_A n_B$ simply by ordering the matrix elements in some way; see Section 2.3.

Definition 1. The direct product $A \otimes B$ is the code consisting of all $n_A \times n_B$ matrices with the property that each matrix column is a codeword of A and each matrix row is a codeword of B.

When A and B are general nonlinear codes, there is no guarantee that $A \otimes B$ is even nonempty. When A and B are linear codes, however, $A \otimes B$ is also linear, with parameters given in the following theorem.

Theorem 1. If A is an $[n_A, k_A, d_A]$ linear code over $\mathbb{F}$ and B is an $[n_B, k_B, d_B]$ linear code over $\mathbb{F}$, then $A \otimes B$ is an $[n_A n_B, k_A k_B, d_A d_B]$ linear code over $\mathbb{F}$.

Before proving this theorem, we give a small example. When A and B both are $[3, 2]$ SPC codes, $A \otimes B$ consists of all the 3×3 matrices in which each row and column contains an even number of ones. The codewords of $A \otimes B$ are listed as follows:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$
$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

There are $2^{2\times 2} = 16$ codewords of length $3 \times 3 = 9$ with minimum nonzero weight $2 \times 2 = 4$, exactly as predicted by Theorem 1. Even though the SPC codes themselves cannot correct errors, their direct product is single-error correcting, since a single error that occurs in row i, column j causes a parity-check violation in that row and column, a situation that is easily detected and corrected by the decoder.

Given any two codewords $a = (a_1, \ldots, a_{n_A}) \in A$ and $b = (b_1, \ldots, b_{n_B}) \in B$, define the $n_A \times n_B$ matrix $a \otimes b$ as the "outer product" $a^T b$ of a and b, specifically, where $[a \otimes b]_{i,j} = a_i b_j$. It is easy to see that each column of $a \otimes b$ is a scalar multiple of a and each row of $a \otimes b$ is a scalar multiple of b. Thus $a \otimes b$ is in fact a codeword of $A \otimes B$. A codeword in $A \otimes B$ of the form $a \otimes b$ is said to be *separable*. In general $A \otimes B$ contains $1 + (|\mathbb{F}|^{k_A} - 1)(|\mathbb{F}|^{k_B} - 1)$ separable codewords, where $|\mathbb{F}|$ denotes the number of elements in $\mathbb{F}$. When k_A and k_B are large, the separable codewords represent a tiny fraction of

all possible codewords, yet, as we will see, the separable codewords generate the entire product code.

Proof of Theorem 1 The code $C = A \otimes B$ contains the zero matrix, and so is clearly a nonempty code of length $n_A n_B$. The linearity of C follows directly from the linearity of codes A and B and the fact that every $\mathbb{F}$-linear combination of codewords gives rise to a corresponding $\mathbb{F}$-linear combination of rows (or columns).

To see that C has minimum distance $d_A d_B$, let v be a nonzero codeword of C. Then v has a nonzero column $a \in A$, whose weight must be at least d_A. Each nonzero component of a also participates in a nonzero row, and each such row is a nonzero codeword of B of weight at least d_B. Thus v has weight at least $d_A d_B$, and so (1) the minimum distance of C is at least $d_A d_B$; on the other hand, if $a \in A$ and $b \in B$ are nonzero codewords of weight d_A and d_B, respectively, then $a \otimes b$ is a codeword of C of weight $d_A d_B$, so (2) the minimum distance of C is at most $d_A d_B$. Statements (1) and (2) together imply that C has minimum distance exactly $d_A d_B$.

The fact that the dimension of $A \otimes B$ is given by $k_A k_B$ follows directly from Theorem 2 (below).

2.2. Encoding of Product Codes

Given generator matrices for A and B, a convenient encoder for $A \otimes B$ is obtained via the following theorem.

Theorem 2. Let code A have generator matrix $\mathbf{G}_A \in \mathbb{F}^{k_A \times n_A}$, let code B have generator matrix $\mathbf{G}_B \in \mathbb{F}^{k_B \times n_B}$, and let $\mathbf{M}$ be an arbitrary matrix in $\mathbb{F}^{n_A \times n_B}$. Then $\mathbf{M} \in A \otimes B$ if and only if $\mathbf{M} = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$, for some $\mathbf{U} \in \mathbb{F}^{k_A \times k_B}$.

Before proving Theorem 2, we remind the reader of the following simple property of matrix multiplication.

Lemma 1. Let $\mathbf{X}$ and $\mathbf{Y}$ be matrices conformable for the product $\mathbf{XY}$. Then each column of $\mathbf{XY}$ is in the column space of $\mathbf{X}$ and each row of $\mathbf{XY}$ is in the row space of $\mathbf{Y}$.

Proof of Theorem 2 Let $\mathbf{U}$ be an arbitrary $k_A \times k_B$ matrix over $\mathbb{F}$, and consider the matrix product $\mathbf{M} = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$. From Lemma 1 (setting $\mathbf{X} = \mathbf{G}_A^T \mathbf{U}$ and $\mathbf{Y} = \mathbf{G}_B$) we see that every row of $\mathbf{M}$ is in the row space of $\mathbf{G}_B$, and hence is a codeword of B. Likewise (setting $\mathbf{X} = \mathbf{G}_A^T$ and $\mathbf{Y} = \mathbf{U}\mathbf{G}_B$) we see from Lemma 1 that every column of $\mathbf{M}$ is in the column space of $\mathbf{G}_A^T$ or equivalently is in the row space of $\mathbf{G}_A$, and hence is a codeword of A. Thus $\mathbf{M}$ is a codeword of $A \otimes B$.

Conversely, let $\mathbf{M}$ be a codeword of $A \otimes B$. Then, since every row of $\mathbf{M}$ is a codeword of B, it follows that $\mathbf{M} = \mathbf{W}\mathbf{G}_B$ for some $n_A \times k_B$ matrix $\mathbf{W}$. Now, since $\mathbf{W} = \mathbf{M}\mathbf{G}_B^{-1}$, where $\mathbf{G}_B^{-1}$ is a right-inverse of $\mathbf{G}_B$, from Lemma 1 it follows that every column of $\mathbf{W}$ is in the column space of $\mathbf{M}$, and hence is a codeword of A. Thus $\mathbf{W} = \mathbf{G}_A^T \mathbf{U}$ for some $\mathbf{U} \in \mathbb{F}^{k_A \times k_B}$, and $\mathbf{M} = \mathbf{W}\mathbf{G}_B = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$.

Observe that the encoder mapping $m: \mathbb{F}^{k_A \times k_B} \to \mathbb{F}^{n_A \times n_B}$ defined by $m(\mathbf{U}) = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$ is linear. The kernel of this mapping [i.e., the set of matrices $\mathbf{U}$ such that $m(\mathbf{U}) = \mathbf{0}$] contains just the zero matrix $\mathbf{U} = \mathbf{0}$. Since the kernel of m is trivial, the mapping m is one-to-one, and C, the image of $\mathbb{F}^{k_A \times k_B}$ under the mapping m, has dimension equal to that of $\mathbb{F}^{k_A \times k_B}$, namely, $k_A k_B$, as claimed in Theorem 1.

When $\mathbf{G}_A = [\mathbf{I}_{k_A} \mid \mathbf{P}_A]$ and $\mathbf{G}_B = [\mathbf{I}_{k_B} \mid \mathbf{P}_B]$, so that A and B have systematic encoders, then the encoder mapping m takes the information matrix $\mathbf{U}$ to the codeword

$$m(\mathbf{U}) = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B = \begin{bmatrix} \mathbf{U} & \mathbf{U}\mathbf{P}_B \\ \mathbf{P}_A^T \mathbf{U} & \mathbf{P}_A^T \mathbf{U} \mathbf{P}_B \end{bmatrix}$$

as illustrated in Fig. 1. The matrix block $\mathbf{U}$ is the information matrix; the block $\mathbf{U}\mathbf{P}_B$ contains "checks on rows," namely, parity-check symbols corresponding to the rows of the message matrix $\mathbf{U}$; the block $\mathbf{P}_A^T \mathbf{U}$ contains "checks on columns," specifically, parity-check symbols corresponding to the columns of the message matrix $\mathbf{U}$; and the block $\mathbf{P}_A^T \mathbf{U} \mathbf{P}_B$ contains "checks on checks," namely, parity-check symbols corresponding to other parity-check symbols. Since $m(\mathbf{U}) = \mathbf{G}_A^T(\mathbf{U}\mathbf{G}_B) = (\mathbf{G}_A^T \mathbf{U})\mathbf{G}_B$, checks on rows and checks on columns may be computed in either order while yielding the identical codeword.

2.3. The Krönecker Product

Let $\mathbf{E}_{i,j}$ be the $k_A \times k_B$ matrix with a one in row i, column j, and zeros in all other positions. Since the set $\{\mathbf{E}_{i,j}: 1 \le i \le k_A, 1 \le j \le k_B\}$ is a basis for $\mathbb{F}^{k_A \times k_B}$, and the encoder mapping m is one-to-one, it follows that $\{m(\mathbf{E}_{i,j}): 1 \le i \le k_A, 1 \le j \le k_B\}$ is a basis for $A \otimes B$. Denote the ith row of $\mathbf{G}_A$ as g_i^A and the jth row of $\mathbf{G}_B$ as g_j^B. Then $m(\mathbf{E}_{i,j}) = \mathbf{G}_A^T \mathbf{E}_{i,j} \mathbf{G}_B = (g_i^A)^T (g_j^B) = g_i^A \otimes g_j^B$. Thus, given a basis $\{g_i^A: 1 \le i \le k_A\}$ for A and a basis $\{g_j^B: 1 \le j \le k_B\}$ for B, the set of outer products $\{g_i^A \otimes g_j^B: 1 \le i \le k_A, 1 \le j \le k_B\}$ is a basis for $A \otimes B$. This proves that separable codewords do indeed generate the entire product code.

It is often useful to "flatten" a product code by converting codewords that are $n_A \times n_B$ matrices into a vectors of length $n_A n_B$. This may be accomplished by establishing a one-to-one correspondence between matrix and vector components. For example, we may define the mapping $s: \mathbb{F}^{n_A \times n_B} \to \mathbb{F}^{n_A n_B}$ that maps a matrix $\mathbf{M}$ to a vector $s(\mathbf{M})$ be assigning $[\mathbf{M}]_{i,j}$ to the $[(i-1)n_B + j]$th component of $s(\mathbf{M})$. In effect, $s(\mathbf{M})$ is obtained by concatenating together the consecutive rows of $\mathbf{M}$; for example

$$s\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}\right) = (110011101)$$

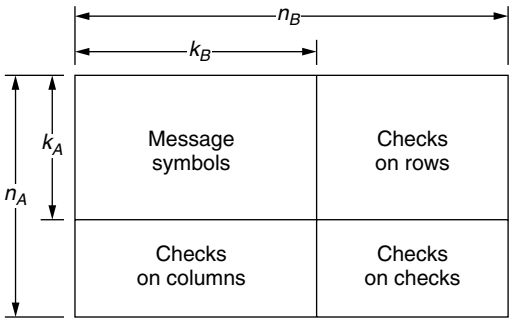


Figure 1. The array formed by the direct product of systematic $[n_A, k_A]$ and $[n_B, k_B]$ codes.

When a and b are vectors in codes A and B, respectively, it is interesting to flatten their outer product. We have

$$\begin{aligned} s(a \otimes b) &= (a_1 b, a_2 b, \ldots, a_{n_A} b) \\ &= a_1 b_1, \ldots, a_1 b_{n_B}, a_2 b_1, \ldots, a_2 b_{n_B}, \\ &\quad \ldots, a_{n_A} b_1, \ldots, a_{n_A} b_{n_B}) \end{aligned}$$

which, in fact, is the *Krönecker product* of the vectors a and b. Recall that the Krönecker product of an $m \times n$ matrix $\mathbf{X}$ with a $p \times q$ matrix $\mathbf{Y}$ is the $mp \times nq$ matrix $\mathbf{X} \otimes \mathbf{Y}$ given in block form as

$$\begin{bmatrix} [\mathbf{X}]_{1,1}\mathbf{Y} & [\mathbf{X}]_{1,2}\mathbf{Y} & \cdots & [\mathbf{X}]_{1,n}\mathbf{Y} \\ [\mathbf{X}]_{2,1}\mathbf{Y} & [\mathbf{X}]_{2,2}\mathbf{Y} & \cdots & [\mathbf{X}]_{2,n}\mathbf{Y} \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{X}]_{m,1}\mathbf{Y} & [\mathbf{X}]_{m,2}\mathbf{Y} & \cdots & [\mathbf{X}]_{m,n}\mathbf{Y} \end{bmatrix}.$$

Note that the symbol $\otimes$ has been burdened to designate the direct product of codes, the outer product of vectors, and now the Krönecker product of matrices. When a and b are vectors, the interpretation of $a \otimes b$ either as an outer product or as a Krönecker product must be made clear.

Observe that the rows of the Krönecker product of $\mathbf{X}$ and $\mathbf{Y}$ are precisely the Krönecker products of all possible pairs of rows, one drawn from $\mathbf{X}$, the other from $\mathbf{Y}$. In light of our observation that the set of all possible outer products $\{g_i^A \otimes g_j^B : 1 \le i \le k_A, 1 \le j \le k_B\}$ of the rows of $\mathbf{G}_A$ and $\mathbf{G}_B$ is a basis for $A \otimes B$, it follows that "flattened code" $s(A \otimes B)$ is generated by the Krönecker product $\mathbf{G}_A \otimes \mathbf{G}_B$. For example, the "flattened" $[9, 4, 4]$ product code from Section 2.1 has generator matrix

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}$$

3. ITERATED PRODUCT CODES AND THE ELIAS CONSTRUCTION

The product construction may be *iterated*, that is, applied to more than two codes. For example, given linear codes A, B, and C of length n_A, n_B, and n_C, respectively, and with generator matrices $\mathbf{G}_A$, $\mathbf{G}_B$, and $\mathbf{G}_C$, respectively, one can define a code $A \otimes B \otimes C$ of length $n_A n_B n_C$. Although one might visualize such codes in terms of structures of higher order than matrices, it is often easier to deal with the "flattened" versions of such codes, by defining $A \otimes B \otimes C$ as the code generated by the Krönecker product $\mathbf{G}_A \otimes \mathbf{G}_B \otimes \mathbf{G}_C$. The Krönecker product is associative:

$$(\mathbf{G}_A \otimes \mathbf{G}_B) \otimes \mathbf{G}_C = \mathbf{G}_A \otimes (\mathbf{G}_B \otimes \mathbf{G}_C)$$

Thus the order in which the Krönecker products are formed does not affect the final outcome. The Krönecker product is *not* commutative, however, so the code generated by $\mathbf{G}_A \otimes \mathbf{G}_B \otimes \mathbf{G}_C$ is not, in general, equal to that generated by $\mathbf{G}_B \otimes \mathbf{G}_A \otimes \mathbf{G}_C$, although they are equivalent.

If $A_1, \ldots, A_L$ is a sequence of codes, where A_i is a linear $[n_i, k_i, d_i]$ code over $\mathbb{F}$, then it follows directly from Theorem 1 that $A_1 \otimes \cdots \otimes A_L$ is a linear $[n_1 n_2 \ldots n_L, k_1\, k_2 \ldots k_L, d_1 d_2 \ldots d_L]$ code.

For example, in his 1954 paper [1], Elias constructed a family of binary codes, each member of which is the direct product of a sequence of extended Hamming codes. Recall that for $m \ge 2$, the extended binary Hamming code H_m of length 2^m is a $[2^m, 2^m - m - 1, 4]$ code. Let $C_{m,L} = H_m \otimes H_{m+1} \otimes \cdots \otimes H_{m+L-1}$. Then $C_{m,L}$ has rate

$$R(m, L) = \prod_{i=m}^{m+L-1} (1 - (i+1)2^{-i})$$

It can be shown that, even when L approaches infinity, the rate $R(m, \infty)$ approaches a nonzero limit. Table 1 tabulates the numerical value of $R(m, \infty)$ for some small values of m. The key point is that the Elias product codes all have finite nonzero rate.

Assuming transmission through a binary symmetric channel with crossover probability p, Elias considered a straightforward strategy for decoding $C_{m,L}$. In the first stage of decoding, decoders for H_m are employed, corresponding to the "rows" of the product code. In the second stage, decoders for H_{m+1} are employed, corresponding to the "columns" of the product code, and so on.

A decoder for the extended Hamming code will successfully correct any single error in a given block. If an even number of errors occur, the Hamming decoder makes no changes to the block. If an odd number of errors greater than 1 occurs, the Hamming decoder will map the received word to a codeword by flipping some bit. This may increase the number of errors (if the flipped bit was correct), or decrease the number of errors (if the flipped bit was in error). Although the exact relationship between the expected number of output errors per block, e_o, and the expected number of input errors per block, e_i, is complicated, it can be shown that if e_i is sufficiently small, then the Hamming decoder for H_m actually *reduces* errors: $e_o < e_i$. In particular, as shown in [2, Sect. 14.83], for large values of m, if $e_i < 2.1779$, then $e_o < e_i$, and if $e_i < 0.6246$, then $e_o < e_i/2$.

Although the ith-stage Hamming decoder introduces statistical dependency among the symbols *within* the decoded word, since each symbol supplied to the $(i+1)$th-stage Hamming decoder is drawn from the output of an independent decoder at stage i, decoders at all stages are supplied with words in which errors in the bits are statistically independent. Provided that at each stage of

Table 1. Limiting Rates for Elias Product Codes

m	2	3	4	5	6	7	8	9
$R(m, \infty)$	0.054	0.215	0.431	0.627	0.772	0.866	0.924	0.958

decoding $e_i < 0.6246$, the expected number of errors per block supplied to the next decoder stage will also satisfy $e_i < 0.6246$, and in fact the number of errors per block will approach zero as the number of stages approaches infinity. Berlekamp [2, Sect. 14.84], provides a table of "threshold values" for the channel crossover probability p that will result in a chain of decreasing error probability at successive decoding stages. When $m = 3$, for example, the threshold is 0.0828, indicating that if $p < 0.0828$, then the family of codes $C_{3,L}$ can be decoded at arbitrarily small error rates when L is large enough. For $m = 4$, the threshold is 0.0402, and for $m = 5$, the threshold is 0.0192.

According to Table 1, $R(3, \infty) = 0.215$. A code operating at the Shannon limit on a binary symmetric channel would be able to tolerate a crossover probability $p = \mathcal{H}^{-1}(1 - 0.215) = 0.234$, while still providing error-free transmission, a value considerably larger than the threshold value 0.0828. [Here $\mathcal{H}$ denotes the binary entropy function $\mathcal{H}(p) := -p \log_2 p - (1-p) \log_2(1-p)$.] Similar conclusions hold for other values of m. Thus, one concludes that the Elias product codes are not capacity-achieving. Nevertheless, they were the first explicitly known family of codes to achieve asymptotically zero error probability at positive code rates. While certain families of irregular low-density parity-check codes (see the article LOW-DENSITY PARITY-CHECK CODES in this encyclopedia) are now known to contain codes with better threshold performance under message-passing decoding than the Elias product codes, it is remarkable indeed that the underlying principles — the use of simple constituent codes with intercommunicating decoders — had already been discovered by Elias as early as 1954.

4. ITERATIVE DECODING OF PRODUCT CODES

The Elias scheme for decoding product codes is strictly sequential. Decoders from one stage pass decoding results on to the next stage, without feedback. In crossword-puzzle terms, this decoding strategy is analogous to looking at the "across" clues, filling in the letters as well as possible, and then looking at the "down" clues, filling in any remaining letters, and then stopping. This leads to a tractable analysis for the decoder; however, as every aficionado of crossword puzzles knows, a far more effective decoding strategy is to alternate or *iterate* between the "across" and "down" clues, since solutions to the "down" clues provide information that are helpful in solving the "across" clues, and vice versa. While the statistical dependence between these interacting decoders makes an exact analysis of the decoder very difficult, simulations of decoder performance show that such iterative decoding can be extremely effective. In fact, it is precisely this sort of iterative decoding that underlies many of the most powerful practical coding schemes known (see, for example, the articles TURBO CODES and LOW-DENSITY PARITY-CHECK CODES in this encyclopedia).

To explain the iterative decoding of product codes, it is helpful to describe these codes in graph-theoretic terms, an approach pioneered by R. M. Tanner [4]. Figure 2 shows a *Tanner graph* for the direct product of codes of length n_A and n_B. A Tanner graph is a bipartite graph containing

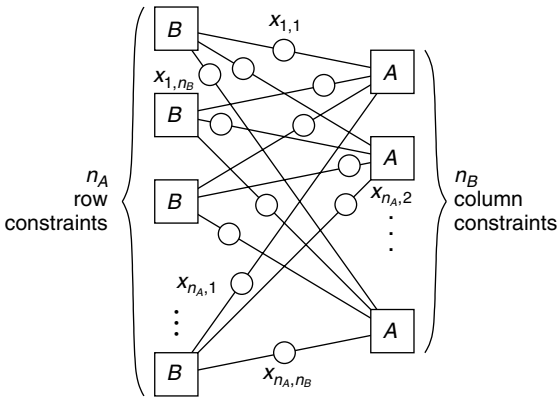


Figure 2. A Tanner graph for the direct product of $[n_A, k_A]$ and $[n_B, k_B]$ codes. The codeword symbol appearing in row i, column j is denoted $x_{i,j}$.

symbol nodes (shown as circles in Fig. 2) and check nodes (shown as squares in Fig. 2). Two nodes are *adjacent* if they are connected by an edge in the graph. In true bipartite fashion, symbol nodes can be adjacent only to check nodes (not to other symbol nodes), and similarly for check nodes.

The symbol nodes represent codeword symbols; in the binary case, each symbol node represents a bit. The purpose of the check nodes is to define the set of *valid configurations* or codewords of the code. They do this locally: each check node imposes a constraint only on the adjacent symbol nodes. While many global configurations may satisfy one or more local configurations, only those global configurations that satisfy *all* the local constraints are deemed to be globally valid configurations.

An example should make this clear. In the product code $A \otimes B$, the symbols in each column of a valid codeword must be a codeword of A and the symbols of each row of a valid codeword must be a codeword of B. Thus there are naturally n_B column constraints (constraining adjacent symbols to codewords of A) and n_A row constraints (constraining adjacent symbols to codewords of B), giving $n_A + n_B$ check nodes in the Tanner graph, as shown in Fig. 2. Each codeword symbol participates in exactly one row and one column, and hence each variable node is adjacent to exactly one column check node and one row check node.

Iterative decoding of this code may be interpreted as a process of *message passing* on the edges of the Tanner graph. We provide here only an intuitive description of the decoding procedure, and refer the reader to, for example, Ref. 5 for a more precise treatment.

The "messages" passed by the algorithm can be interpreted as probabilities or "beliefs" concerning the value of the corresponding symbol node. For example, the beliefs are often expressed as loglikelihood ratios (LLRs), defined as $\lambda(x) = \log(P[x = 0]/P[x = 1])$. When $\lambda(x)$ is large and positive, this indicates a strong belief that x has value 0; when $\lambda(x)$ is large and negative, this indicates a strong belief that x has value 1. A $\lambda(x)$ value that is close to zero indicates that $P[x = 0]$ and $P[x = 1]$ are nearly equal. The initial beliefs are determined by the received word.

Decoding proceeds iteratively. Each symbol node x transmits $\lambda(x)$ to, say, the adjacent-row check nodes.

Each row check node receives n_B incoming messages, and produces n_B outgoing messages that *update* the $\lambda(x)$ values, taking into account the constraints imposed by the structure of B. This updating procedure is sometimes referred to as "soft-in/soft-out" decoding, since the inputs and outputs of the decoder are beliefs, not hard decisions about the values of the bits.

For example, suppose that B is a $[3, 2]$ SPC code, and that a particular row check node receives three messages: two of which represent strong beliefs that the symbols x_1 and x_2 have value 1, and one representing a relatively weaker belief that symbol x_3 is a 1. Since 111 is not a valid local configuration, the most likely explanation is that x_3 is in fact 0. Thus the decoder would update the beliefs, reinforcing the belief that x_1 and x_2 are 1, and suggesting that x_3 is in fact zero.

These updated beliefs are then passed to the column decoders. Each column decoder receives n_A incoming messages, and produces n_A outgoing messages that update the belief values, taking into account the constraints imposed by the structure of A.

These newly updated beliefs could then be passed back to the row decoders, and the whole decoding process would repeat. The process would normally halt when the updated beliefs suggest a valid configuration, or when some predetermined upper limit on the number of iterations is reached.

It is important to note that different row decoders operate independently of one another. In a hardware decoder implementation, it would be possible to operate these decoders fully in parallel. The same is, of course, true for the column decoders. The local soft-in/soft-out decoders are typically implemented using the forward/backward algorithm [5] operating on a trellis representation of the local code. Any codes for which an efficient trellis representation exists would therefore be candidate constituent codes for such a product code. Since the different decoders are interconnected by a simple row/column structure, message passing can be implemented in hardware in straightforward fashion.

The performance of practically decodable product codes can be quite good, with performance on an additive white Gaussian noise channel with BPSK modulation that is within $\sim$2 dB of the Shannon limit at blocklengths on the order of 4000 bits [6]. While Turbo codes and low-density parity-check codes can approach somewhat closer to the Shannon limit, the simplicity of product codes and the prospect of high-speed decoder implementations in hardware make product codes very suitable for a wide variety of practical applications.

BIOGRAPHY

Frank R. Kschischang is a Professor in the Department of Electrical and Computer Engineering at the University of Toronto, where he has been a faculty member since 1991. He received the B.A.Sc. degree with honors from the University of British Columbia in 1985, and the M.A.Sc. and Ph.D. degrees from the University of Toronto in 1988 and 1991, respectively, all in electrical engineering. From October 1997 to October 2000, Dr. Kschischang served as *IEEE Transactions on Information Theory* Associate Editor for Coding Theory. In April 1999, he received the Ontario Premier's Research Excellence Award, and in December 2000, he was named a Canada Research Chair. His research interests lie in the general area of channel coding techniques, particularly in iterative decoding of codes defined on graphs.

BIBLIOGRAPHY

1. P. Elias, Error-free coding, *IRE Trans. Inform. Theory* **PGIT-4**: 29–37 (1954); reprinted in E. R. Berlekamp, ed., *Key Papers in The Development of Coding Theory*, IEEE Press, New York, 1974, pp. 39–47.
2. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
3. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
4. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT-27**: 533–547 (1981).
5. F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. Inform. Theory* **47**: 498–519 (2001).
6. R. M. Pyndiah, Near-optimum decoding of product codes: Block turbo codes, *IEEE Trans. Commun.* **46**: 1003–1010 (Aug. 1998).

PROPAGATION MODELS FOR INDOOR COMMUNICATIONS

F. Landstorfer
G. Woelfle
R. Hoppe
Institut Fuer Hochfrequenztechnik
University of Stuttgart
Stuttgart, Germany

1. INTRODUCTION

The performance of wireless communication systems depends in a fundamental way on the mobile radio channel. In contrast to wired channels that are stationary and easy to design, radio channels show a time-variant behavior which complicates their analysis. Inside buildings the transmission path between transmitter and receiver can vary from simple line-of-sight to one severely obstructed by walls and furniture. As a consequence, predicting the propagation characteristics between two antennas still belongs to the most important tasks for the design and installation of wireless indoor communication systems, ranging from low-bit-rate cordless telephone and cellular systems to high-bit-rate wireless local-area networks (WLANs).

This article reviews and discusses a variety of methods for modeling wave propagation in indoor scenarios. The basic requirements necessary for predicting path loss and other relevant parameters are discussed. Apart from well known and widely used propagation models, new

approaches with minimized computational complexity are also introduced.

1.1. The Mobile Indoor Radio Channel

The mobile radio channel concerning transmission within buildings is characterized by a multipath scenario as shown in Fig. 1. The signal from the transmitting antenna (usually only the downlink is considered as the principle of reciprocity applies) propagates along different paths to the antenna of the (mobile) receiver. In many cases there is no direct line of sight and the only paths connecting transmitter and the receiver penetrate several walls and are reflected, diffracted, and scattered at a number of different obstacles. Since the phases of the waves are randomly distributed, the superposition of these contributions causes constructive and destructive interference (i.e., small-scale fading), which leads to rapidly fluctuating signal levels over very short distances. Figure 2 illustrates this small-scale fading and the slower large-scale signal variation for an indoor radiocommunication system. While the small-scale fading is random, the large-scale variations occur as a result of fundamental changes of the propagation paths (e.g., larger distances, different obstacles). Typically, the local average of the received power is computed by averaging signal measurements over an interval of 10λ to 20λ, which corresponds to movements of the receiver of 1.5–3 m at a frequency of 2 GHz.

1.2. Radiowave Propagation within Buildings

With decreasing wavelength, that is, increasing frequency, wave propagation becomes more and more similar to the propagation of light. A radio ray is assumed to propagate essentially along a straight line and is influenced only by the given obstacles. For a criterion for this type of modeling to be successful, the wavelength should be much smaller than the extensions of the partitions of the building structure. At the frequencies used for wireless indoor communication networks, this criterion is sufficiently fulfilled.

The phenomena that influence radiowave propagation can generally be described by four basic mechanisms: penetration, reflection, diffraction, and scattering. For the practical usage of propagation models in real

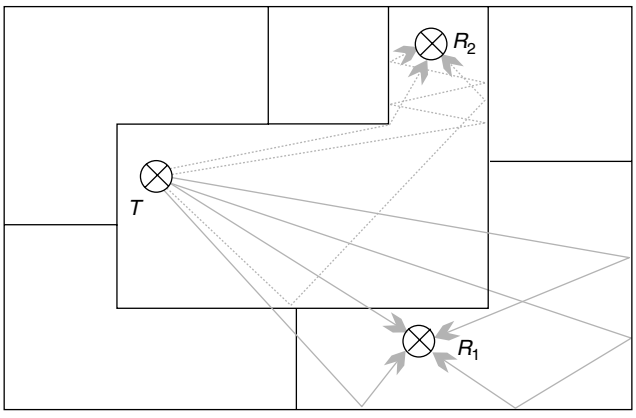


Figure 1. Multipath propagation within buildings.

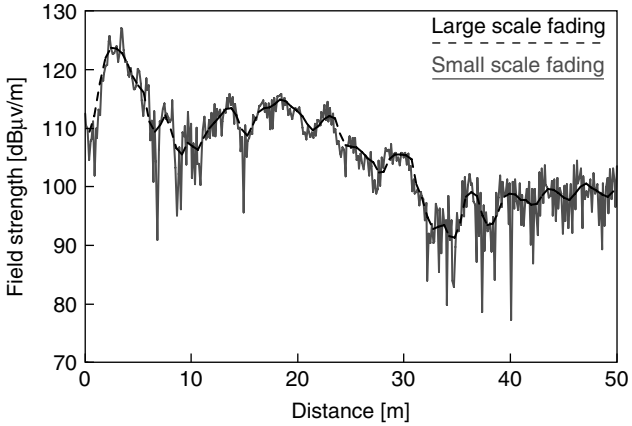


Figure 2. Large-scale and small-scale fading.

scenarios, these mechanisms must be introduced by approximations. This requires a multistage modeling process [1], introduced in the following paragraphs.

In the first step the propagation environment has to be digitized, leading to a database that describes the considered scenario in an adequate way. The second step includes the definition of mathematical approximations for the physical propagation phenomena. On the basis of the solutions for the basic problems, both empirical and deterministic approaches have been developed, forming the third modeling step.

The indoor radio channel differs considerably from that of an outdoor scenario [2,3]. The transmitter–receiver distance is shorter in order to compensate for the high attenuation along the path caused by internal walls and furniture and also because of the lower transmitter power. The short distance implies shorter delay of echoes and consequently a lower delay spread. The temporal variations of the channel are slower than those of mobile antennas moving with an automobile. Nevertheless, the conditions within buildings are time-variant even when transmitter and receiver are fixed; for instance, whether interior doors are open or closed can change the propagation scenario considerably. In general, the propagation within buildings is strongly influenced by the local environment, namely, the layout of the particular building under consideration and the construction materials used for walls, floors, and ceilings. According to the type of building and the corresponding layout, four different categories of indoor environments can be defined as listed in Table 1 [1,4].

As it is the case in outdoor scenarios, there are several important propagation parameters to be predicted. The path loss and the statistical characteristics of the received signal strength are most important for coverage planning applications. The wideband characteristics (delay spread, impulse response) and the time variation are essential for the evaluation of the system performance.

1.3. Properties of Materials

The buildings taken into account within a planning process have a wide variety of walls and obstacles that form the internal and external structure. Hard partitions and soft

Table 1. Different Categories of Indoor Environments [1]

Environment Category	Description	Typical Values for the Delay Spread (ns)
Corridor	Transmitter and receiver along the same corridor (LoS)	≤20
Dense	Small rooms; typically an office where each employee has his own room; mostly non-line-of-sight (NLoS) conditions	10–30
Open	Large rooms; typically an office where one room is shared by several employees; mostly line-of-sight (LoS) or obstructed-line-of-sight conditions (OLoS)	20–50
Large	Environments consisting of very large rooms; typically a factory hall, shopping center, or airport building; mostly LoS or OLoS conditions	50–80

partitions can be distinguished according to the magnitude of the penetration loss for electromagnetic waves [2]. While hard partitions are formed as bearing parts of the building structure (e.g., walls constructed from reinforced concrete or brick, thickness >10 cm), soft partitions may be moved and therefore show lower losses (e.g., plasterboard or plywood, thickness <10 cm). In order to get a more accurate modeling of wave propagation, a detailed description of the electrical properties for all building elements considered is necessary. However, in most cases the partitions are made out of several layers and are not homogenous. Nevertheless, a lot of researchers have collected databases for a great amount of different materials. While the physical behavior is described by the complex permittivity ε, the resulting partition losses as given in Table 2 are of more interest from a practical point of view.

The total loss experienced by electromagnetic waves when penetrating walls can be divided into two independent parts. The first part refers to the penetration of the wall surface and shows no explicit frequency dependence, whereas the second part belongs to the in depth penetration of the wall material. The magnitude of the latter depends on the wavelength and leads to higher losses with increasing frequency (see Table 2).

1.4. Databases for Buildings

The basis for any propagation model is a database that describes the propagation environment. For the purpose of propagation modeling, each building element should be categorized into classes (wall, floor, door, window, etc.) and specified by its coordinates and finally its material properties (thickness and electrical characteristics).

Modern planning tools [5,6] store the building data in a 3D-vector format including all walls, doors, and windows. Usually all elements inside the building are described in terms of plane elements; for example, every wall is represented by a plane and its extent and location is defined by its corners as indicated in Fig. 3. Through the use of

Table 2. Partition Losses of Different Construction Materials

Material Type	Frequency (MHz)	Loss (dB)	Ref.
Reinforced concrete	900	10	1
	1700	16	1
Brick	900	6	1
	1800	10	1
Plywood	900	1.5	1
	1800	2	1
Glass	900	3	1
	1800	4	1
Metal	815	26	2

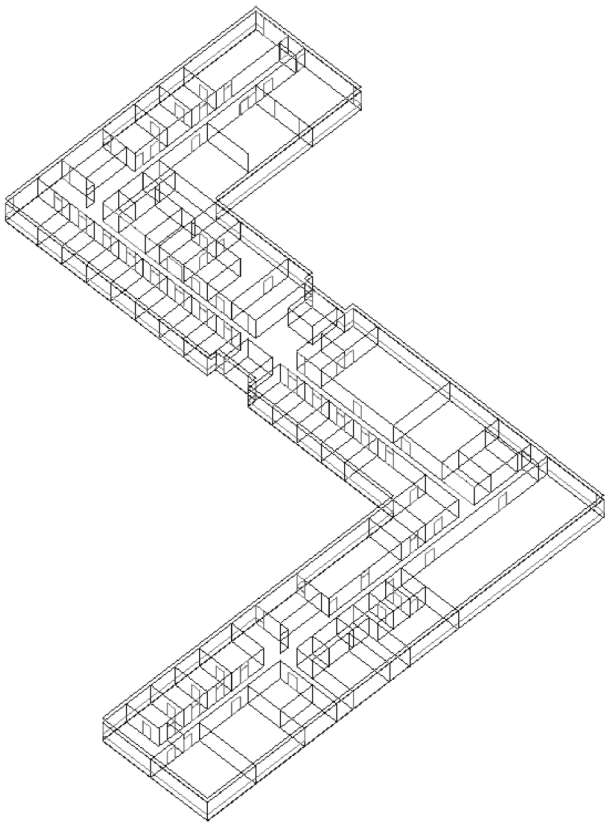


Figure 3. Example of an indoor database.

such building databases, which may be drawn or digitized utilizing standard graphical CAD (computer-aided design) software packages, wireless system designers are able to include accurate representations of building features. With respect to an efficient use it is often possible to import drawing interchange files (DXFs), a very common CAD data format in architecture. Figure 3 shows an example for a three dimensional building data base utilized within a commercial planning tool [6].

2. PROPAGATION MODELS

Propagation models focus on the prediction of the averaged received signal strength, as well as the variability of the signal strength in close vicinity of the particular location. This leads to the distinction between large-scale propagation models that predict the mean signal strength and small-scale propagation models that characterize the rapid fluctuations of the received signal strength over very short distances (a few wavelengths) or short time durations (in the order of seconds) [2].

Similar to this distinction, the propagation models presented here can be divided into three groups: empirical narrowband models, empirical wideband models, and deterministic ones. Empirical narrowband models are expressed in form of simple mathematical equations that give the path loss as an output result. The equations are obtained by fitting the model parameters to measurement results. The empirical wideband models allow the prediction of the wideband characteristics of the channel (e.g., impulse response, delay spread). Deterministic models simulate the propagation of radiowaves in a more physical way. These models predict both narrowband and wideband information of the mobile radio channel inside buildings. Additionally, directional channel properties such as angular spread are readily available, a fact that may be very important for the planning of future systems [7]. Propagation models for the prediction of field strength levels will generally provide only mean or median values, as the small-scale fading within indoor environments is adequately represented by Rayleigh distributions for NLoS (see Table 1) conditions and Rice distributions for clear LoS conditions with K factors up to 15 dB. The long-term fading, which describes the fluctuation of the mean value, can be approximated by a lognormal distribution with standard deviations between 2.7 and 5.3 dB [1].

2.1. Empirical Narrowband Models

Figure 4 shows the two basic approaches to the prediction of the field strength inside buildings. There are empirical models that analyze the direct path between the transmitter and the receiver. A calibration of these empirical models is mandatory, and their computation times are very small. All models of this type are based on the free-space propagation model.

2.1.1. Free-Space Propagation Model. This model is utilized to predict the received power if transmitter and receiver have a clear, unobstructed line-of-sight path. Generally this is not the case within indoor scenarios; nevertheless, this model gives a valuable insight into propagation modeling. The received power of an antenna at a distance d from the radiating transmitter antenna is calculated by the Friis equation:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2} \tag{1}$$

where P_t is the transmitted and P_r is the received power, G_t and G_r are the antenna gains of the transmitter (Tx) and receiver (Rx), respectively, already mention before equation and λ is the free space wavelength. The free space path loss L_{FS}, which represents the signal attenuation as a positive quantity measured in decibels (dB), is defined as difference (in dB) between the transmitted and the received power level:

$$L_{FS} = 10 \log \frac{P_t}{P_r} = 10 \log \left(\frac{4\pi d}{\lambda} \right)^2 - 10 \log G_t - 10 \log G_r \tag{2}$$

Obviously the path loss increases both with distance and with frequency at a rate of 20 dB/decade. With the

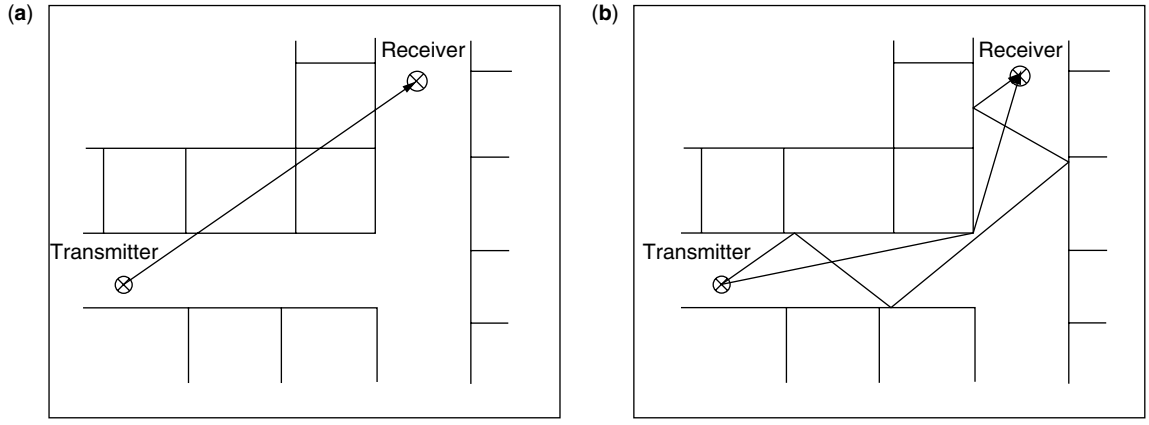


Figure 4. Basic approaches to the modeling of indoor propagation: (**a**) empirical models; (**b**) deterministic models.

introduction of the default loss L_0 (including the antenna gains) at distance d_0, Eq. (2) can be simplified to

$$L_{FS} = L_0 + 2 \cdot 10 \log\left(\frac{d}{d_0}\right) \tag{3}$$

2.1.2. One-Slope Model. The one-slope model analyses the building concerning distances between walls and penetration losses of the walls, but the individual positions of the walls and their material properties are not considered (see Fig. 5). Therefore this model computes the path loss L_{OS} (in dB) similar to the free space loss with adaptable power decay factor n and offset L_0.

$$L_{OS} = L_0 + n \cdot 10 \log\left(\frac{d}{d_0}\right) \tag{4}$$

The walls of the building are not taken into account by the one-slope model; thus no building database is required. With constant values for n and L_0 for every receiver location the prediction leads to path loss values increasing in concentric circles around the transmitter. Consequently, the prediction results are fairly inaccurate and suited only for a rough estimation. For line-of-sight the values of n are in the range between 1.4 and 1.8; in non-line-of-sight scenarios values up to 5.0 are possible [1,2].

2.1.3. Motley–Keenan Model. The model according to Motley and Keenan [8] computes the path loss L_{MK} based on the direct path between transmitter and receiver. In contrast to the one-slope model, in this model the exact locations of the walls, floors, and ceilings are considered. Additional factors for the attenuation of the direct path by partitions account for the shadowing effects.

$$L_{MK} = L_{FS} + k \cdot L_W \tag{5}$$

As shown in Fig. 6, parameter k describes the number of walls intersected by the direct path between transmitter and receiver. A uniform penetration loss L_W (in dB) of all partitions is taken for the computation (see Table 2); thus the individual material properties of the different walls are not considered. This uniform penetration loss as well

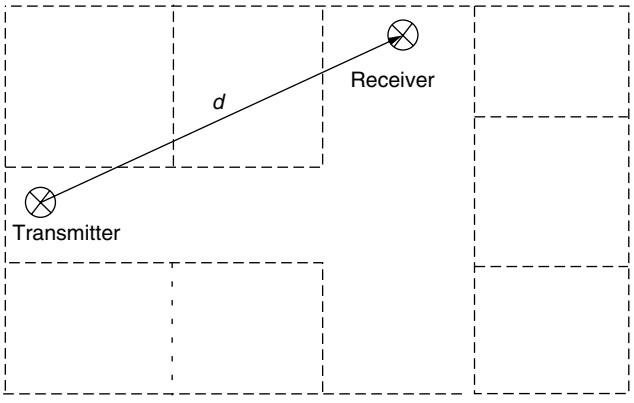


Figure 5. Principle of the one-slope model.

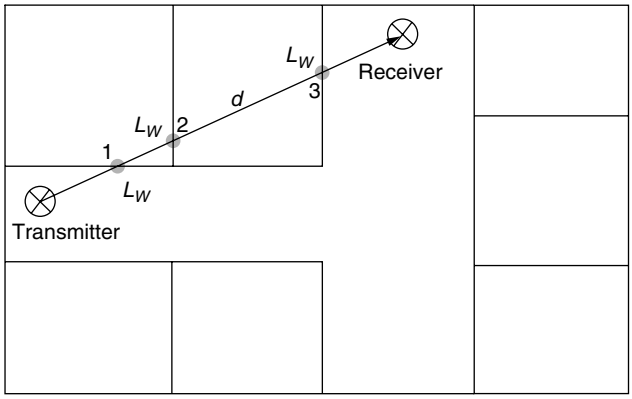


Figure 6. Principle of the Motley–Keenan model.

as the default loss L_0 which is part of the free-space path loss L_{FS} can be calibrated by measurements.

2.1.4. Multiwall Model. The multiwall model [1] gives the path loss L_{MW} as the free-space loss added to losses introduced by the walls and floors penetrated by the direct path between transmitter and receiver (see Fig. 6). In contrast to the Motley–Keenan model, for the multiwall model, individual penetration losses of the walls (depending on their electrical characteristics) are considered for the prediction of the path loss. It has been observed that the total floor loss is a nonlinear function of the number of floors penetrated. This effect is taken into account by introducing an empirical factor b. Hence the multiwall model can be expressed as follows:

$$L_{MW} = L_{FS} + \sum_{i=1}^{N} k_{Wi} \cdot L_{Wi} + k_f^{[(k_f+2)/(k_f+1)-b]} \cdot L_f \tag{6}$$

where k_{Wi} represents the number of penetrated walls of type i and k_f the number of penetrated floors. The parameter L_{Wi} describes the penetration loss of wall type i and L_f the loss between adjacent floors, and N denotes the number of different wall types.

The default loss L_0 as a part of the free-space path loss L_{FS} may be calibrated according to measurement results by evaluating a multiple linear regression technique. It is important to note that the loss factors in the Eq. (6) do not represent physical wall losses but model coefficients that are optimized by the measured path-loss data. Consequently, the loss factors implicitly include the effect of furniture. However, hard partitions such as concrete walls are overemphasized, which leads to pessimistic values of predicted field strength behind these elements as indicated in Fig. 7. Additionally, waveguiding effects are not taken into account within this model; thus its accuracy is only moderate. Nevertheless, this more refined empirical narrowband model has a low dependency on database accuracy and, because of the simple approach, a very short computation time (in the order of seconds).

Within the European cooperation of COST 231, measurements in different environments from several institutions have been evaluated [1] in order to optimize the coefficients of the empirical narrowband models (as given

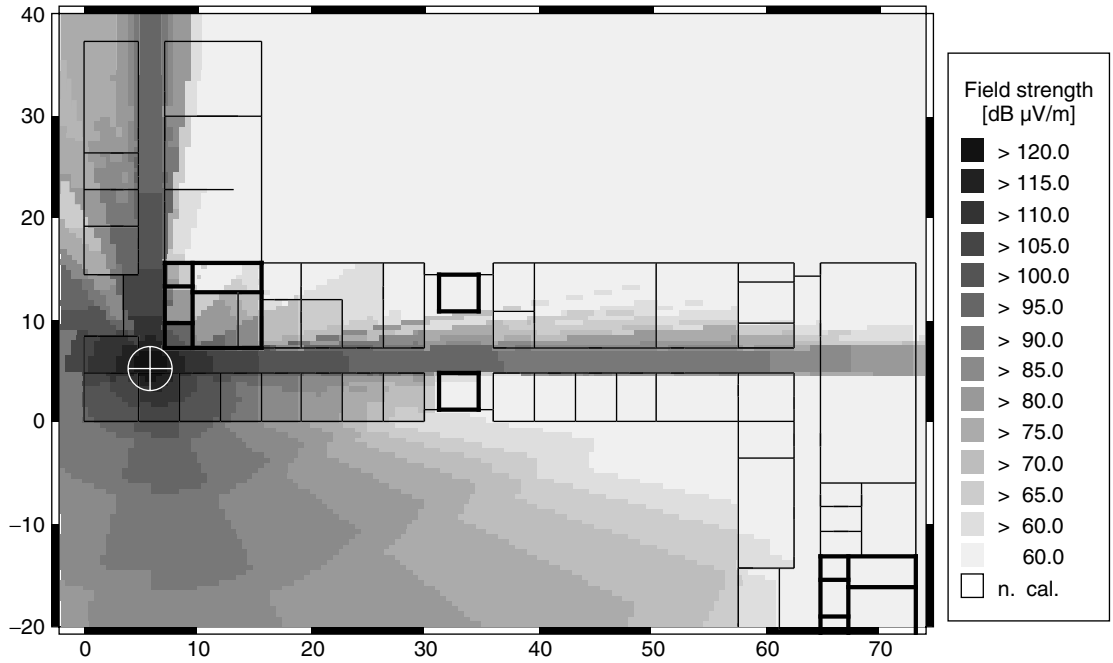


Figure 7. Prediction of the field strength by evaluating the multiwall model at 1800 MHz for a transmitter with 100 mW and omnidirectional Tx-antenna pattern.

Table 3. Investigated Coefficients at 1800 MHz for Empirical Narrowband Models [1]

	One-Slope Model		Multiwall Model			
Environment	L_0 (dB)	n	L_{W1} (dB)	L_{W2} (dB)	L_f (dB)	b
Corridor	39.2	1.4	3.4	6.9	18.3	0.46
Dense						
One floor	33.3	4.0	3.4	6.9	18.3	0.46
Two floors	21.9	5.2				
Multifloor	44.9	5.4				
Open	42.7	1.9	3.4	6.9	18.3	0.46
Large	37.5	2.0	3.4	6.9	18.3	0.46

in Table 3). For the multiwall model two different wall types have been taken into account corresponding to the previous distinction between hard and soft partitions. The coefficients of the multiwall model have been optimized for the category "dense." However, these values can also be utilized within other environments, leading to results close to the free-space model. According to measurements at 900 MHz, the values of the multiwall model should be reduced by 1.5 dB for the soft partition loss and by 3.5 dB for the floor loss. Concerning the one-slope model, the default loss L_0 should be reduced about 10 dB, while the values for the decay index remain constant.

2.2. Empirical Wideband Models

Beyond the planning and deployment of wireless communication networks, propagation models are also utilized for studying and evaluating new radio systems. The propagation models provide a valuable input for so-called channel models. These models reproduce the behavior of the radio channel in order to estimate the performance and the capacity of different system implementations concerning access, modulation, or coding schemes.

Empirical wideband models provide average impulse responses and power delay profiles (PDP) for this purpose. This kind of data can be derived from wideband measurements (e.g., channel sounder or network analyzer) as well as from deterministic propagation models. On the basis of wideband measurements, the delay spread values as given in Table 1 have been determined [1], indicating lowest values in dense environments and increasing values in open and large environments.

Impulse response models are usually defined as tapped delay lines, including a limited number of paths with individual amplitude and delay (both referred to the dominant path). The ITU channel models [9] as given in Table 4 contain six different paths and describe typical impulse responses for large office buildings with an open layout and a Tx-Rx separation less than 100 m. While the tap delay-line models represent typical impulse responses for unlimited bandwidth, the power delay

Table 4. Impulse Response Model for Indoor Office Environments [9]

	Channel A			Channel B		
	Delay Spread = 35 ns			Delay Spread = 100 ns		
Tap #	Delay (ns)	Loss (dB)	Doppler Spectrum	Delay (ns)	Loss (dB)	Doppler Spectrum
1	0	0.0	Flat	0	0.0	Flat
2	50	−3.0	Flat	100	−3.6	Flat
3	110	−10.0	Flat	200	−7.2	Flat
4	170	−18.0	Flat	300	−10.8	Flat
5	290	−26.0	Flat	500	−18.0	Flat
6	310	−32.0	Flat	700	−25.2	Flat

profiles characterize the radio channel with respect to the limited bandwidth of a specific radio system. Typical averaged power delay profiles within buildings show a logarithmic (for LoS and OLoS conditions) to linear (for NLoS conditions) decay on dB scale depending on the specific environment [1].

Because of the slow movements of indoor terminals (in most cases they are more portable than mobile) the Doppler spectrum has maximal values of about 10 Hz for frequencies utilized within personal communication systems (about 2 GHz). For simplicity a flat spectrum is often considered (see Table 4).

2.3. Deterministic Models

Deterministic models utilize physical phenomena in order to describe the propagation of radiowaves. Here, the effect of the actual environment is taken into account more accurately than within empirical models.

2.3.1. Basic Mechanisms. The mobile radio channel in indoor environments is characterized by multipath propagation (see Fig. 1). Dominant propagation phenomena in these scenarios are reflection, transmission, diffraction, and scattering. Because of the multiple reflections, waveguiding in corridors can be observed. Deterministic propagation models are generally based on ray optical techniques. With such an approach it is possible to consider the abovementioned effects as well as the multipath situation within the propagation model (as indicated in Fig. 4). The ray optical models determine all relevant rays between the transmitter and the receiver. Calibration of the deterministic models is not necessary because the predicted values are computed with the Fresnel equations for reflection and transmission and with the universal theory of diffraction (UTD) for diffracted rays [10]. Alternatively, there are empirical equations available for the calculation of the reflection, diffraction, and transmission loss [11].

2.3.1.1. Reflection and Transmission. When an electromagnetic wave propagating in one medium impinges on another medium with different electrical properties, the wave is partially reflected and partially penetrates the medium. Although this phenomenon is valid for the incidence on a perfect dielectric, all the incident energy is reflected back to the first medium if the second medium is a perfect conductor. If there is a smooth boundary between the two materials, the impinging wave is reflected specularly. Concerning the penetration of a dielectric plate, the direction of the penetrated radiowave corresponds to the incident direction if there is the same medium in front of and behind the dielectric plate. The electric field intensities of the reflected and transmitted waves are related to the incident wave through a reflection coefficient. The most common mathematical description of the reflection is the Fresnel reflection coefficient, which is valid for an infinite boundary between two media. The coefficients for reflection and transmission are a function of the material properties, and generally depend on the polarization and the angle of incidence of the propagating wave [10].

2.3.1.2. Diffraction. The diffraction process in ray modeling is the *propagation phenomenon*, which explains the transition from the lit region to the shadowed regions behind obstacles. Although the electrical field strength decreases as a receiver moves deeper into the shadowed region, the diffraction field still exists and often has enough strength to guarantee a sufficient signal. According to the principle of Huygens, the diffraction field is the vector sum of the electric field components of all secondary wavelets in the space around the obstacle. Diffraction by a single wedge can be solved in various ways: empirical formulas, perfectly absorbing wedge, geometric theory of diffraction (GTD) or universal theory of diffraction (UTD) [10]. The advantages and disadvantages of using either of these formulations is difficult to address since it may not be independent of the investigated environments. Indeed, reasonable results are possible with each formulation. The various expressions differ mainly in the approximations being made on the surface boundaries of the wedge considered. However, diffraction around a corner is commonly modeled using the heuristic UTD formulas since they behave well in the lit/shadow transition region and account for the polarization of the incident wave as well as the wedge material. Generally, the wedge diffraction coefficient is inversely proportional to the square root of the frequency, specifically, the coefficient decreases with increasing frequency. Therefore the effect of diffraction can be neglected for millimeter waves ($f > 30$ GHz).

2.3.1.3. Scattering. Rough surfaces and finite surfaces (i.e., surfaces with small extensions in comparison to the wavelength) scatter the incident energy in all directions according to a radiation pattern which depends on the roughness and size of the surface or volume. The dispersion of energy through scattering means a decrease of the energy reflected in specular direction, which can be

taken into account by reducing the reflection coefficient. The consideration of the true dispersion of radio energy in various directions is much more difficult and may be described by radar cross-sectional models. Surface roughness is often tested using the Rayleigh criterion, which defines a critical height of surface protuberances for a given angle of incidence [10].

2.3.2. Path Finding. For the determination of valid rays between transmitter and receiver, which is the most time-consuming part of the ray optical approach, two different principles can be utilized: ray launching and ray tracing.

2.3.2.1. Ray Launching. This approach launches rays in discrete angular increments from the transmitter and determines their path through the database (as indicated in Fig. 8). If there is an intersection between the ray and a wall, the ray is split into a penetrating and a reflecting part and both rays are further launched independently from each other. When the ray impinges on a wedge, new rays are launched on the diffraction cone, which leads to multiple rays. Every time when a ray intersects the prediction plane, the field strength of this ray is added to the already computed field strength at the specific receiver point. The consideration of the path ends if the number of interactions is higher than a given number or if the field strength at the end of the ray is smaller than a given threshold. In order to keep the resolution of the rays independent of the distance to the transmitter, a splitting algorithm may be introduced (ray splitting).

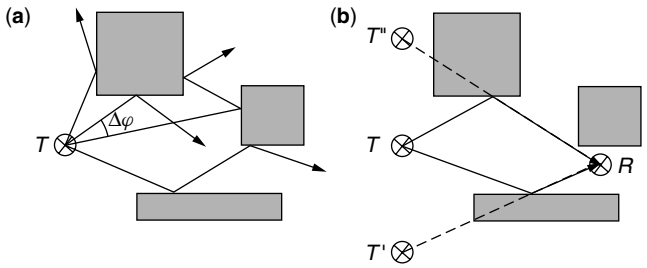


Figure 8. Algorithms for path finding: (**a**) ray launching; (**b**) ray tracing.

2.3.2.2. Ray Tracing (Ray Imaging). The ray tracing algorithm determines valid rays between the transmitter and a given receiver point (as indicated in Fig. 8). For coverage predictions the receiver has to be assigned subsequently to all pixels in the prediction area. Obviously the computation time increases linearly with the number of receiver points and shows an exponential dependence on the number of walls and the number of interactions, respectively. In comparison to the ray launching there is a higher computational effort, but on the other hand, ray tracing obtains a constant resolution and a high degree of accuracy.

Figure 9 shows a prediction with a rigorous 3D ray tracing for the situation already presented in Fig. 7. The differences between empirical and ray optical predictions are obvious especially the waveguiding in corridors due to multiple reflections, the coupling of the waves into the rooms and the diffraction around corners are responsible for the high accuracy of the ray optical models.

2.3.3. Advanced Ray Optical Model Based on Database Preprocessing. The main disadvantage of the deterministic prediction models is their excessive computation time (in the order of hours). Different authors presented ideas to accelerate the path finding, and some of the methods lead to considerable acceleration factors. A further disadvantage of the ray optical propagation models is the abovementioned dependence on the accuracy of the database, in which even small errors in the positions and

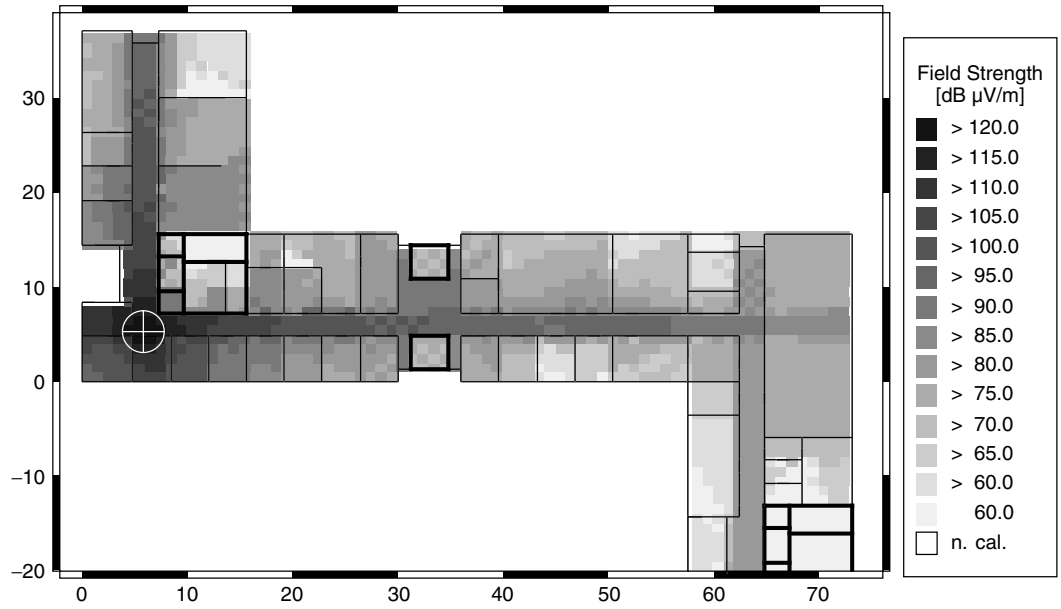


Figure 9. Prediction of the field strength by evaluating the 3D ray tracing at 1800 MHz for a transmitter with 100 mW and omnidirectional Tx-antenna pattern.

materials of the walls or missing parts (e.g., furniture) influence the predicted results. The advanced model [11] presented in this section neglects these disadvantages and is therefore well suited for the planning and deployment of indoor wireless communication networks.

One major application of propagation models is to evaluate the degree of coverage that can be achieved in a radio cell depending on the position of the base station. While the database of the building in question remains the same and only the position of the transmitter changes, the overwhelming part of the different rays remains unchanged; only the rays between the transmitting antenna and primary obstacles or receiving points in line of sight change.

This is the basis for "intelligent database preprocessing." In a first step the walls of the building (or other obstacles) are divided into tiles (reflections and penetrations) and the edges (diffractions) into horizontal and vertical segments. After this, the visibility conditions between these different elements (possible rays) are determined and stored in a file. Figure 10 shows the visibility relation between a central tile and a receiving point. The result of this preprocessing can be represented in the shape of a "visibility tree" as shown in Fig. 11. For a different transmitter location only the uppermost branches in this tree must be computed again. Consequently, all branches below the first interaction layer have to be computed only once, which can be done prior to optimizing the location of the transmitter. The remaining computation time after the preprocessing is many orders of magnitude lower than that needed for the conventional analysis without preprocessing. As a consequence, 3D deterministic models, with their supreme accuracy, can be utilized for all practical applications with computation times in the order of those found with empirical models.

When analyzing rays that contribute to the field strength at the receiving point of a typical indoor situation (as given in Fig. 1), it is obvious that a number of different rays reach the receiver after passing the same sequence of rooms and penetrating the same walls. These rays can be summarized into several dominant paths, each of them characterizing the propagation of a bundle of waves [12]. There is generally more than one dominant path between transmitter and receiver. Figure 12 shows dominant paths for the scenario depicted in Fig. 1. The dominant paths can be deduced using simple algorithms that consider the arrangement of the rooms within the building relative to the transmitter and the receiver [12].

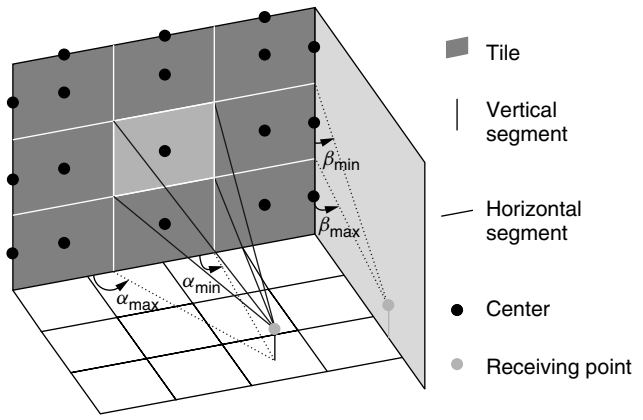


Figure 10. Tiles and segments of a wall with visibility relation indicated.

3. PLANNING TOOLS

With the increasing computational and visualization capabilities of computers, new methods for predicting radio signal coverage have been established, which are based on site specific propagation models and building databases as described in Section 1.4. As planning tools become prevalent, wireless system designers will be able to design and deploy indoor wireless networks for covering the buildings adequately without performing extensive radio measurements. Generally, indoor planning tools support graphical user interfaces in order to generate and visualize building databases, including

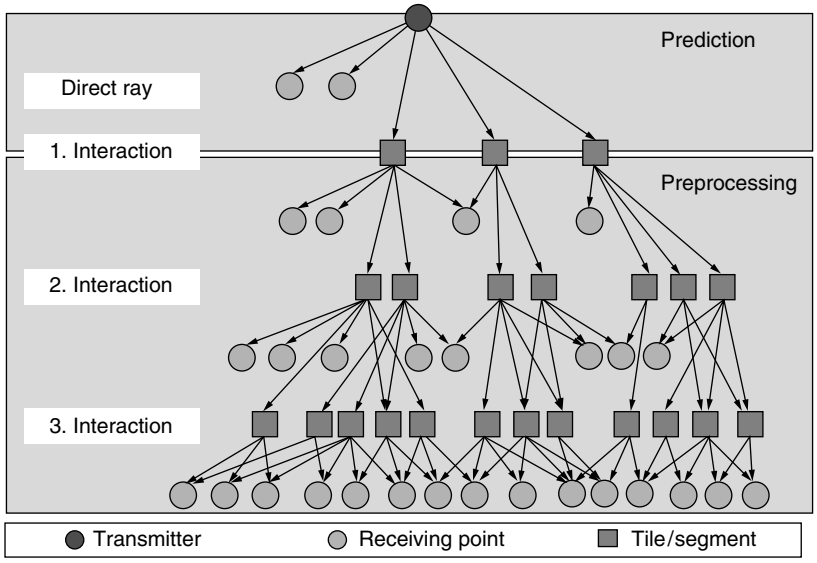


Figure 11. Tree structure of the visibility relations.

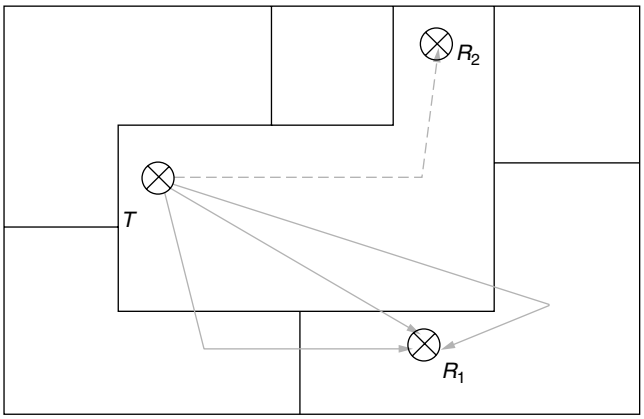


Figure 12. Dominant paths concept.

the material properties as well as to define the characteristics of transmitters and receivers (e.g., antenna patterns). Beyond this, most planning tools provide deterministic and empirical propagation models (as presented in Section 2) for predicting the large-scale path loss in a wide range of building structures and capabilities to visualize and analyze the corresponding results [5,6].

BIOGRAPHIES

Friedrich M. Landstorfer received a Dipl.-Ing. degree in 1964 and a Dr.-Ing. degree in 1967 in electrical engineering from the Technical University of Munich, Germany. In 1971 he became lecturer at the same institution and professor in 1976. In 1986, he moved to Stuttgart to become professor and head of the RF-Institute at the University of Stuttgart, Germany. His research interests include antennas, microwaves, electromagnetic theory, wave propagation in connection with mobile communications, navigation, and electromagnetic compatibility. He received the award of the NTG (Nachrichtentechnische Gesellschaft in VDE, now ITG) in 1977 for the optimization of wire antennas and was chairman of the 21st European Microwave Conference in Stuttgart in 1991. Professor Landstorfer is a member of URSI, was awarded honorary professor of Jiaotong University in Chengdu, China, in 1993, and became fellow of the IEEE in 1995.

Gerd Woelfle received his Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Stuttgart, Germany, in 1994 and 1999, respectively. In his Ph.D. thesis he analyzed the indoor multipath propagation channel and developed adaptive models for indoor propagation. In 1999, he joined AWE Communications as a R&D Engineer, where he has been working on propagation models for wireless communication networks. Dr. Woelfle has received several best paper awards on wireless communication conferences. His areas of interest are empirical and deterministic propagation models, network planning tools, and channel modeling. Since 1999, he has been a lecturer for mobile communications at the University of Stuttgart.

Reiner A. Hoppe received his Dipl.-Ing. degree in electrical engineering in 1997 from the University of Stuttgart, Germany. Since this time, he has been a research scientist at the Institute of Radio Frequency Technology (University of Stuttgart), where he is working towards his Ph.D. degree. His research interests include wave propagation modeling and radio network planning especially within urban and indoor scenarios.

BIBLIOGRAPHY

1. E. Damosso, ed., *Digital Mobile Radio towards Future Generation Systems*, Final Report of the COST Action 231, Bruxelles: European Commission, 1998.
2. T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, Upper Saddle River, NJ, 1996.
3. D. Parsons, *The Mobile Radio Propagation Channel*, Pentech Press, London, 1992.
4. D. Molkdar, Review on radio propagation into and within buildings, *IEE Proc. H* (*Microwaves, Antennas and Propagation*) **138**(1): 61–73 (Feb. 1991).
5. S. J. Fortune, *WiSE—a Wireless System Engineering Tool* (online), *http://www.bell-labs.com/innovate98/wireless/wiseindex.html*, July 4, 2001.
6. AWE Communications, *WinProp—Software for Radio Network Planning Within Terrain, Urban and Indoor Scenarios*, (online) *http://www.awe-communications.com* (July 4, 2001).
7. L. M. Correia, ed., *Wireless Flexible Personalised Communications*, Final Report of the COST Action 259, Wiley, Chichester, UK, 2001.
8. A. J. Motley and J. M. Keenan, Personal communication radio coverage in buildings at 900 MHz and 1700 MHz, *IEE Electron. Lett.* **24**(12): 763–764 (June 1988).
9. ITU-R Recommendation M.1225, *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000*, International Telecommunication Union, 1997, pp. 24–28.
10. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.
11. G. Woelfle, R. Hoppe, and F. M. Landstorfer, A fast and enhanced ray optical propagation model for indoor and urban scenarios, based on an intelligent preprocessing of the database, *Proc. 10th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications* (*PIMRC*), F5-3, Osaka, Japan, Sept. 1999.
12. G. Woelfle and F. M. Landstorfer, Dominant paths for the field strength prediction, *Proc. 48th IEEE Int. Conf. Vehicular Technology* (*VTC*), Ottawa, May 1998, pp. 552–556.

PULSE AMPLITUDE MODULATION

BJØRN A. BJERKE
Qualcomm, Inc.
Concord, Massachusetts

1. INTRODUCTION

Pulse amplitude modulation (PAM) is a modulation method used in digital communication systems to facilitate

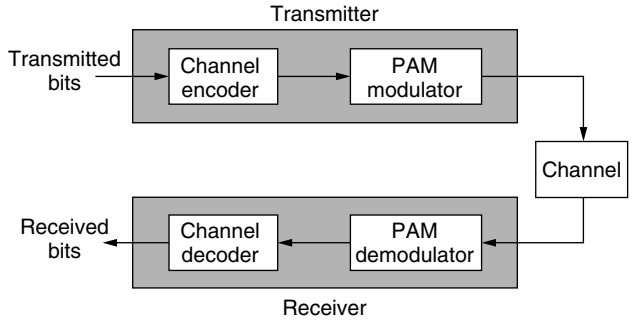


Figure 1. The basic elements of a digital communication system that uses pulse amplitude modulation.

transmission of digital information from a source (the transmitter) to a destination (the receiver). In such systems, the digital information is represented by a sequence of binary digits, or bits. Figure 1 shows the basic elements of a communication system that uses PAM. The transmitter consists of two blocks, namely, a channel encoder and a digital modulator. The channel encoder adds redundancy in a controlled manner to the transmitted bits in order to enable detection and possibly also correction of bit errors. Such errors may occur as a result of disturbances in the transmission system.

The output sequence from the encoder is passed to the PAM modulator, which transforms the discrete-time sequence into a signal that conforms to the limitations of the transmission channel. The channel is the physical medium through which transmissions occur. Most channels are analog in nature, so the modulator outputs analog waveforms that correspond to the particular coded sequence. The physical channel can, for example, be a cable, an optical fiber, or a wireless radio channel. In each of these channels, the signal is affected by disturbances such as thermal noise and interference from various sources. These disturbances corrupt the signal in a number of ways that ultimately may lead to bit errors at the receiver.

The role of the PAM demodulator is to recover the sequence of coded information bits from the corrupted received signal. In so doing, the demodulator often must compensate for the various channel disturbances. Finally, the coded sequence is passed to the channel decoder. Using knowledge of the channel code and the redundancy contained in the received sequence, the decoder attempts to reconstruct the original information bit sequence with as few bit errors as possible.

In this article, we shall concentrate on the modulator/demodulator pair, commonly referred to as the *modem*. The remainder of the article is organized into three sections, each dealing with different aspects of PAM. First, in Section 2, we introduce a convenient mathematical representation of PAM signals. This representation is used in the subsequent sections where we investigate the properties and characteristics of such signals. In Section 3, we discuss the spectral characteristics of PAM signals, and, finally, in Section 4, we discuss the various aspects of signal demodulation and detection, including carrier recovery and symbol synchronization.

2. PAM SIGNAL REPRESENTATION

Many digital communication systems are *bandpass* systems, where the digital information is transmitted over the communication channel by means of carrier modulation. At the transmitter, the information-bearing *lowpass* signal is impressed on the carrier signal by the modulator, thus translating the information signal from *baseband* frequencies to the carrier frequency. At the receiver, the demodulator performs the reverse process of recovering the digital lowpass signal from the modulated carrier, translating the signal from the frequency band of the carrier down to baseband. When representing bandpass signals mathematically, it is often convenient to reduce them to equivalent lowpass signals so that they may be characterized independently of the particular carrier frequency or frequency band in use.

A bandpass signal may be represented as

$$s(t) = \mathrm{Re}\{x(t)e^{j2\pi f_c t}\} \tag{1}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of a complex-valued quantity and $x(t)$ is the equivalent lowpass signal that is impressed on the carrier signal with center frequency f_c. In general, $x(t)$ is a complex-valued signal and it is often referred to as the *complex envelope* of the real signal $s(t)$, but in PAM the lowpass signal is real-valued.

A PAM modulator maps digital information (bits) into analog, finite energy waveforms that differ only in amplitude. The mapping is usually performed by taking groups of k bits at a time and selecting one out of a total of $M = 2^k$ possible such waveforms for transmission over the channel. The signal waveforms may be represented as

$$s_m(t) = \mathrm{Re}\{A_m g(t)e^{j2\pi f_c t}\} = A_m g(t)\cos 2\pi f_c t,$$
$$m = 1, 2, \ldots, M, \quad 0 \le t \le T \tag{2}$$

where $\{A_m, m = 1, 2, \ldots, M\}$ denotes the set of M possible amplitudes, $g(t)$ is a real-valued signal pulse, and T is the duration of a symbol interval. The shape of $g(t)$ may be tailored to achieve a certain spectral shaping of the transmitted signal so that it matches the spectral characteristics of the channel.

The energy of the PAM signal is dependent on the energy in the signal pulse, and is given by

$$E_m = \int_0^T s_m^2(t)\,dt = \int_0^T \left[A_m g(t)\cos 2\pi f_c t\right]^2 dt$$
$$= A_m^2 \int_0^T g^2(t)\cos^2 2\pi f_c t\,dt = \frac{1}{2}A_m^2 E_g \tag{3}$$

where $E_g = \displaystyle\int_0^T g^2(t)\,dt$ denotes the signal pulse energy. When $g(t)$ has a rectangular shape, as shown in Fig. 2, the resulting modulation is also known as *amplitude shift keying* (ASK). In this case, the pulseshape is given as

$$g(t) = \begin{cases} a, & 0 \le t \le T \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

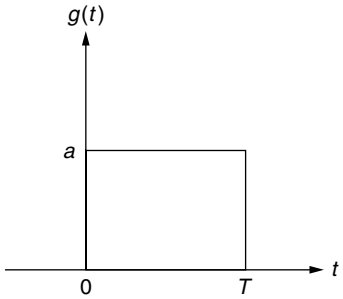


Figure 2. Rectangular signal pulse $g(t)$.

The reciprocal of the symbol interval, $R_s = 1/T$, is known as the symbol rate, namely, the rate at which the modulated carrier changes amplitude. The symbol rate is related to the bit rate $R_b = 1/T_b$ as

$$R_s = \frac{R_b}{k} \tag{5}$$

where k is the number of bits transmitted per symbol and T_b is the duration of a bit interval. The equivalent lowpass representation of the PAM waveform is real-valued and given by

$$x_m(t) = A_m g(t), \quad m = 1, 2, \ldots, M, \quad 0 \le t \le T \tag{6}$$

The signal amplitude can take on the discrete values

$$A_m = (2m - 1 - M)\, d, \quad m = 1, 2, \ldots, M \tag{7}$$

where $2d$ is the distance between adjacent signal amplitudes. Figure 3 illustrates the signal amplitude diagram for binary ($M = 2$) and quaternary ($M = 4$) PAM signals, respectively, as well as examples of possible bit mappings. In the examples shown in the figure, *Gray mapping* is used, where only a single bit differs in adjacent constellation points. Since the signal amplitudes are confined to the real line, we refer to these signals as *one-dimensional*.

Digital PAM may also be used in baseband transmission systems, that is, systems that do not require carrier modulation. In this case, the signals are represented by the equivalent lowpass formulation of Eq. (6). Figure 4 illustrates a four-level baseband PAM signal using rectangular signal pulses with amplitude $a = 1$.

3. SPECTRAL CHARACTERISTICS OF PAM

Now that we have introduced a mathematical representation of pulse-amplitude-modulated (PAM) signals, we are

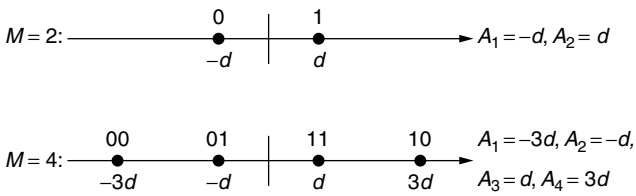


Figure 3. Signal amplitude diagram for binary and quaternary PAM signals.

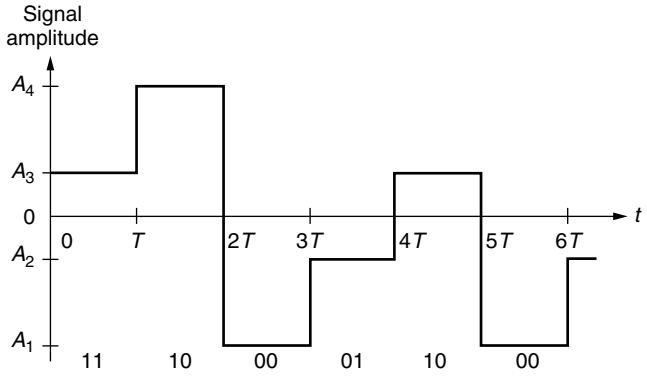


Figure 4. Four-level baseband PAM signal.

ready to discuss their spectral characteristics. Most transmission channels offer a limited bandwidth, due to either physical limitations or regulatory constraints. It is therefore important to determine the spectral content of PAM signals, or any modulated signals, for that matter, to make sure that the bandwidth of the signal does not exceed that of the channel. Also, since bandwidth is such a precious resource, it is important that it is utilized efficiently.

Let us again consider the bandpass signal $s(t)$ introduced in (1). The transmitted information sequence is a random sequence. Consequently, the corresponding PAM signal is a stochastic process whose spectral characteristics are described by its power density spectrum. The power density spectrum, denoted by $\Phi_{ss}(f)$, is obtained by Fourier transforming the autocorrelation function of $s(t)$, which is given by

$$\phi_{ss}(\tau) = \mathrm{Re}\{\phi_{xx}(\tau) e^{j2\pi f_c \tau}\} \tag{8}$$

The Fourier transformation results in

$$\Phi_{ss}(f) = \tfrac{1}{2}[\Phi_{xx}(f - f_c) + \Phi_{xx}(-f - f_c)] \tag{9}$$

where $\Phi_{xx}(f)$ is the power density spectrum of the lowpass equivalent signal $x(t)$. It is evident from (9) that in order to investigate the spectral characteristics of $s(t)$ it is sufficient to consider the power density spectrum of $x(t)$.

The lowpass signal has the general form

$$x(t) = \sum_{n=-\infty}^{\infty} A_n g(t - nT) \tag{10}$$

where the subscript n is a time index. The mean of $x(t)$ and its autocorrelation function are both periodic with period T. Hence, $x(t)$ is a *cyclostationary process*. The autocorrelation function averaged over a single period can be shown to be [1]

$$\overline{\phi}_{xx}(\tau) = \frac{1}{T} \sum_{m=-\infty}^{\infty} \phi_{AA}(m) \phi_{gg}(\tau - mT) \tag{11}$$

where $\phi_{AA}(m)$ is the autocorrelation of the information sequence represented by the amplitudes $\{A_n\}$, and $\phi_{gg}(\tau)$ is the autocorrelation function of the pulse $g(t)$, defined as

$$\phi_{gg}(\tau) = \int_{-\infty}^{\infty} g^*(t) g(t + \tau)\, dt \tag{12}$$

where $*$ denotes the complex conjugate. By taking the Fourier transform of (11), we obtain the power density spectrum

$$\Phi_{xx}(f) = \frac{1}{T}|G(f)|^2 \Phi_{AA}(f) \tag{13}$$

where $G(f)$ is the Fourier transform of $g(t)$ and $\Phi_{AA}(f)$ is the power density spectrum of the information sequence. From (13) we realize that the power density spectrum $\Phi_{xx}(f)$ depends directly on the spectral characteristics of both the information sequence $\{A_n\}$ and the pulse $g(t)$. Consequently, we may shape the spectral characteristics of the PAM signal by manipulating either the transmitter pulseshape or the correlation properties of the transmitted sequence. In the latter case, dependence between signals transmitted in different symbol intervals is introduced in a process known as *modulation coding*, resulting in a signal with memory. However, standard pulse amplitude modulation is a *memoryless* modulation since the mapping of bits into symbol waveforms is performed independently of any previously transmitted waveforms.

We shall assume here that the information symbols in the transmitted sequence are uncorrelated and have zero mean. In this case, the power density spectrum of the transmitted sequence is simply $\Phi_{AA}(f) = \sigma_A^2$. Thus, spectral shaping of the PAM signal is accomplished exclusively by selecting the pulseshape $g(t)$. Let us consider two examples that illustrate the spectral shaping that results from two different pulses. The first is the rectangular pulse introduced earlier in Fig. 2, which is used in amplitude shift keying. The energy density spectrum of the rectangular pulse, given as the square of the magnitude of the Fourier transform $G(f)$, is

$$|G(f)|^2 = (aT)^2 \left(\frac{\sin \pi f T}{\pi f T}\right)^2 \tag{14}$$

The energy spectrum is shown in Fig. 5. We note that the spectrum has a main lobe with a width of $2/T$ and zeros at $f = n/T$, $n = \pm 1, \pm 2, \ldots$. Its tail decays inversely as f^2. The resulting power density spectrum of the lowpass signal is given by

$$\Phi_{xx}(f) = \sigma_A^2 a^2 T \left(\frac{\sin \pi f T}{\pi f T}\right)^2 \tag{15}$$

In the second example, the pulse is a *raised-cosine* pulse as shown in Fig. 6 and mathematically represented by

$$g(t) = \frac{a}{2}\left[1 + \cos \frac{2\pi}{T}\left(t - \frac{T}{2}\right)\right], \quad 0 \le t \le T. \tag{16}$$

The corresponding energy density spectrum is given by

$$|G(f)|^2 = \frac{(aT)^2}{4}\left(\frac{\sin \pi f T}{\pi f T(1 - f^2T^2)}\right)^2 \tag{17}$$

In this case, the main lobe has a width of $4/T$, which is twice the width of the main lobe of the rectangular pulse. Zeros occur at $f = n/T$, $n = \pm 2, \pm 3, \ldots$. However, the tails of the spectrum decay inversely as f^6, which means that the energy outside the main lobe is virtually zero. This is illustrated in Fig. 7, which shows a heavily magnified version of the energy density spectrum.

These two examples serve to illustrate how the spectral shape of the transmitted signal can be tailored to match the spectral characteristics of the channel. The raised-cosine pulse requires a greater bandwidth, but in return the out-of-band energy is much lower than with the rectangular pulse.

4. DEMODULATION AND DETECTION OF PAM SIGNALS

In this section, we describe the various elements of the optimal receiver for PAM signals transmitted over the additive white Gaussian noise (AWGN) channel. Furthermore, we evaluate the performance of this receiver in terms of the symbol error and bit error probabilities. The AWGN channel is the most benign of channels, and the performance that can be achieved on this channel is

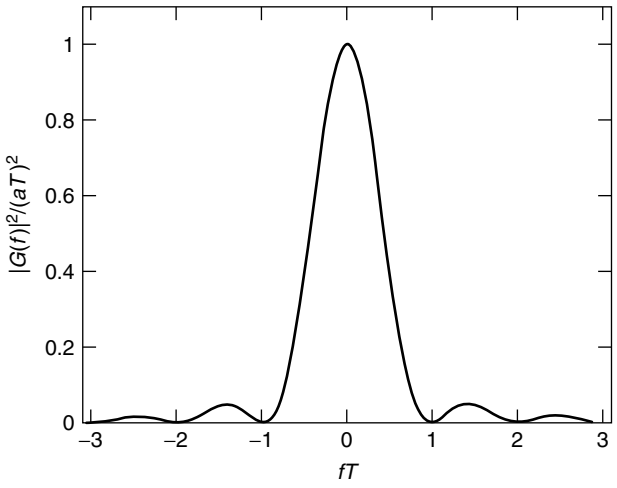


Figure 5. Energy density spectrum $|G(f)|^2$ for a rectangular pulse.

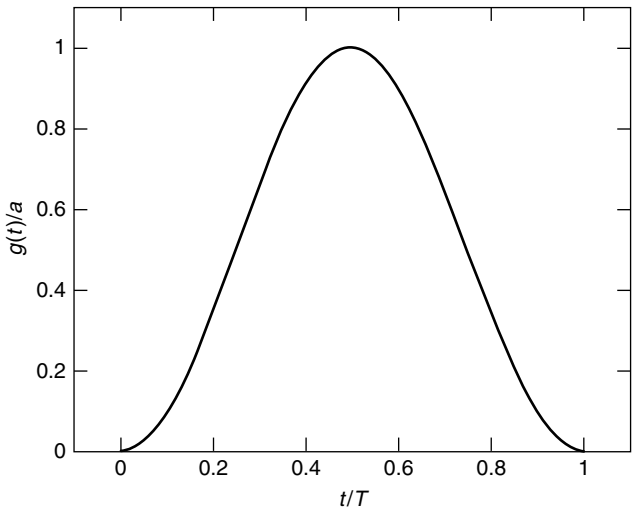


Figure 6. Raised-cosine pulse.

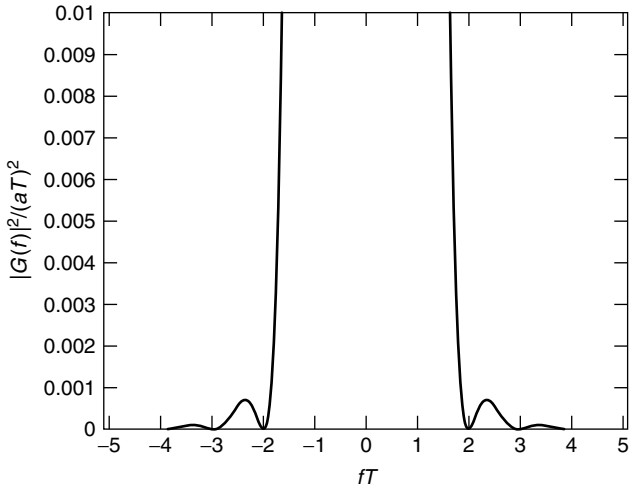


Figure 7. Energy density spectrum $|G(f)|^2$ for a raised-cosine pulse.

often used as a benchmark for evaluating a particular modulation scheme and receiver structure on other types of channels (e.g., fading channels, multipath channels, and channels with various kinds of interference). At the end of this section, we discuss carrier-phase recovery and symbol synchronization for PAM signals.

Let us first consider the input to the receiver, that is, the received signal, which consists of the original transmitted signal as well as white Gaussian noise. As described earlier, in each signaling interval the transmitted signal consists of one of M possible signal waveforms $\{s_m(t), m = 1, 2, \ldots, M\}$. The received signal can therefore be represented as

$$r(t) = s_m(t) + n(t), \quad 0 \le t \le T \tag{18}$$

where $n(t)$ is a sample function of the stochastic AWGN process with power spectral density $N_0/2$. At this point it is useful to introduce the concept of *signal space*. Signal waveforms can be given an equivalent vector representation, where, in general, the M finite energy waveforms are represented by weighted linear combinations of $N \le M$ orthonormal functions $f_n(t)$. Thus, any signal can be represented as a point in the N-dimensional signal space spanned by the basis functions $\{f_n(t), n = 1, 2, \ldots, N\}$. The *Euclidean distance* between these points is a measure of the similarity of the signal waveforms, and consequently dictates how well the receiver will be able to determine which waveform was transmitted. As noted earlier, PAM signals are one-dimensional ($N = 1$) and can therefore be represented simply by the general form

$$s_m(t) = s_m f(t) \tag{19}$$

where the basis function $f(t)$ is defined as the unit energy signal waveform

$$f(t) = \sqrt{\frac{2}{E_g}}\, g(t) \cos 2\pi f_c t \tag{20}$$

and s_m is a point on the real line given as

$$s_m = A_m \sqrt{\frac{E_g}{2}}, \quad m = 1, 2, \ldots, M \tag{21}$$

The Euclidean distance between any pair of signal points is

$$d_{lk} = \sqrt{(s_l - s_k)^2} = d\sqrt{2E_g}\,|l - k|, \quad l, k = 1, 2, \ldots, M, l \ne k \tag{22}$$

where $2d$ is the distance between adjacent signal amplitudes, as defined earlier. The distance between a pair of adjacent signal points is known as the minimum Euclidean distance, and is given by $d_{\min} = d\sqrt{2E_g}$.

A block diagram of an M-ary PAM receiver is shown in Fig. 8. The receiver consists of a demodulator and a detector as well as circuits for carrier recovery and symbol synchronization. The task of the demodulator is to compute the projection r of the received waveform $r(t)$ onto the basis function spanning the one-dimensional signal space. Based on r, the detector then decides which one of the M possible waveforms was transmitted. Since we assume a memoryless modulated signal, the detector can make its decisions separately for each signaling interval. In

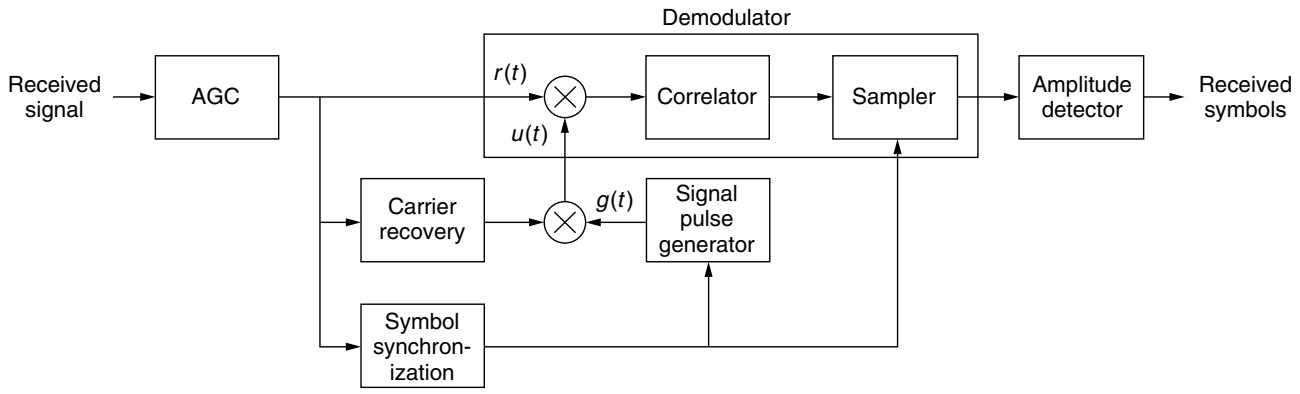


Figure 8. Block diagram of M-ary PAM receiver.

order to perform these tasks, the demodulator output must be sampled periodically, once per symbol interval. Symbol synchronization is therefore required and must be established before demodulation and detection can be performed. Because of the unknown propagation delay from the transmitter to the receiver, the symbol timing must be extracted from the received signal. Carrier-phase offset is another consequence of the unknown propagation delay. In the case of *phase-coherent detection*, this offset must be estimated, and this is the task of the carrier-phase recovery circuit. In the subsequent sections, we examine in more detail the functions of signal demodulation, signal detection, carrier recovery, and symbol synchronization.

4.1. Signal Demodulator

Let us first assume that the carrier phase has been accurately estimated and symbol synchronization has been properly established. The demodulator computes the projection of the received waveform onto the signal space by passing the received signal $r(t)$ through a *correlator*, as shown in Fig. 8. The result of this projection is

$$r = \int_0^T r(t)f(t)\,dt = s_m + n \tag{23}$$

where

$$s_m = \int_0^T s_m(t)f(t)\,dt \tag{24}$$

and n is the projection of the noise onto the signal space, given as the zero-mean random variable

$$n = \int_0^T n(t)f(t)\,dt \tag{25}$$

with variance $\sigma_n^2 = N_0/2$. The correlator output is a Gaussian random variable with mean $\mathrm{E}[r] = s_m$ and variance equal to the noise variance, i.e., $\sigma_r^2 = \sigma_n^2 = N_0/2$. Thus, the probability density function of r, conditional on the mth signal waveform being transmitted, is the Gaussian *likelihood function* given by

$$p(r|s_m) = \frac{1}{\sqrt{\pi N_0}} \exp\left[-\frac{(r-s_m)^2}{N_0}\right] \tag{26}$$

Instead of using a correlator for signal demodulation, a *matched filter* may be used in its place. In general, the matched filter is a bank of N filters whose impulse responses are matched to the N basis functions $\{f_n(t)\}$. In our case, where $N = 1$, the impulse response of the single filter is given by

$$h(t) = f(T-t), \quad 0 \le t \le T \tag{27}$$

The output of the matched filter is

$$\begin{aligned} y(t) &= \int_0^t r(\tau)h(t-\tau)\,d\tau \\ &= \int_0^t r(\tau)f(T-t+\tau)\,d\tau \end{aligned} \tag{28}$$

Sampling this output at $t = T$ yields

$$y(T) = \int_0^T r(\tau)f(\tau)\,d\tau = r \tag{29}$$

which is identical to the output of the correlator. An important property of the matched filter is that it maximizes the output signal-to-noise ratio (SNR) in an AWGN channel. The output SNR is given by

$$\mathrm{SNR}_{\mathrm{out}} = \frac{(\mathrm{E}[r])^2}{\sigma_n^2} \tag{30}$$

4.2. Signal Detector

The correlator (or matched-filter) output contains all the relevant information, namely, the *sufficient statistic*, about the transmitted signal that the detector needs in order to make a decision, assuming that the carrier phase is known at the receiver and symbol synchronization has been established. The detector applies a decision rule that seeks to maximize the probability of a correct decision. More specifically, it decides in favor of the signal which has the maximum a posteriori probability

$$\Pr\{s_m|r\}, \quad m = 1, 2, \ldots, M \tag{31}$$

of the M possible transmitted signals. This decision rule is known as the *maximum a posteriori probability* (MAP) criterion. The a posteriori probabilities can be expressed as

$$\Pr\{s_m|r\} = \frac{p(r|s_m)\Pr\{s_m\}}{p(r)} \tag{32}$$

where $p(r|s_m)$ was given in (26) and $\Pr\{s_m\}$ is the a priori probability that the mth signal will be transmitted. The denominator is independent of which signal is transmitted. When all the M transmitted signals are equally likely, the MAP criterion is reduced to maximizing the likelihood function (26) over the M possible signals, and this is called the *maximum-likelihood* (ML) criterion. For ease of computation, it is common to use the *loglikelihood function* given by

$$\ln p(r|s_m) = -\frac{1}{2}\ln(\pi N_0) - \frac{1}{N_0}(r-s_m)^2 \tag{33}$$

Maximizing (33) is equivalent to finding the signal that minimizes the Euclidean distance metric

$$d(r, s_m) = (r-s_m)^2 \tag{34}$$

In our case, the detector is an amplitude detector, as shown in Fig. 8. An *automatic gain control* (AGC) circuit [2] is added to the front end of the receiver to eliminate short-term channel gain variations that would otherwise affect the amplitude detector. The AGC maintains a fixed average SNR at its output.

4.3. Detector Performance

Armed with the knowledge gained in the previous two sections, we may evaluate the performance of the optimal

receiver for M-ary PAM signals in terms of the symbol error and bit error probabilities. As before, we assume equally probable transmitted signals. First we need to define some quantities that we will use in our calculations. The average energy of the transmitted signals, using the representations given in (3) and (7), is

$$E_{\text{av}} = \frac{1}{M}\sum_{m=1}^{M} E_m = \frac{d^2 E_g}{2M}\sum_{m=1}^{M}(2m-1-M)^2 = \frac{1}{6}(M^2-1)d^2 E_g \tag{35}$$

The average power is given as $P_{\text{av}} = E_{\text{av}}/T$, and the average energy per transmitted bit is $E_b = P_{\text{av}} T_b$.

Now let us consider the Euclidean distance metric given by (34). The detector compares the projection r with all the M signal candidates and decides in favor of the signal with the smallest metric. Equivalently, the detector compares r with a set of $M-1$ thresholds located at the midpoints between adjacent signal amplitude levels. A decision is made in favor of the level that is closer to r. An erroneous decision will occur if the magnitude of the noise is significant enough to make r extend into one of the two regions belonging to the neighboring amplitude levels. This is true for all but the two outermost levels $\pm(M-1)$, where an error can be made in only one direction. Hence, the average probability of a symbol error, P_M, is equal to the probability that the noise component $n = r - s_m$ exceeds one-half of the distance between two thresholds, $d_{\min}/2$:

$$P_M = \frac{M-1}{M}\Pr\left\{|n| > \frac{d_{\min}}{2}\right\} = \frac{2(M-1)}{M}\Pr\left\{n > \frac{d_{\min}}{2}\right\} \tag{36}$$

where

$$\Pr\left\{n > \frac{d_{\min}}{2}\right\} = \frac{1}{\sqrt{\pi N_0}}\int_{d_{\min}/2}^{\infty} e^{-x^2/N_0}\,dx \tag{37}$$

It is straightforward to show that the symbol error probability is equal to

$$P_M = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{d_{\min}^2}{2N_0}}\right) \tag{38}$$

where $Q(\cdot)$ is the standard error function [1]. It is customary to express the error probability as a function of the average SNR per bit, $\gamma_b = E_b/N_0$, and using the definition of $d_{\min}$ and (35), we can easily show that the symbol error probability can be expressed as

$$P_M = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{6k}{M^2-1}\gamma_b}\right) \tag{39}$$

where $k = \log_2 M$. Figure 9 shows the symbol error probability as a function of γ_b for $M = 2, 4, 8, 16$. When Gray coding is used in the mapping of bits into symbols, adjacent symbols differ in only a single bit and in most

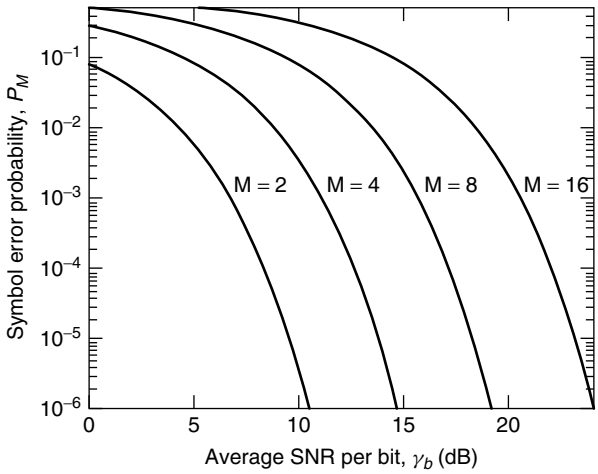


Figure 9. Symbol error probability for M-ary PAM.

cases a symbol error will result in only a single bit error. The bit error probability P_b is therefore approximated as

$$P_b \approx \frac{1}{k} P_M \tag{40}$$

4.4. Carrier-Phase Recovery

If phase-coherent detection is employed, the carrier phase needs to be known or accurately estimated. In most practical systems the phase is estimated directly from the modulated signal. Let us assume that the received signal has the form

$$r(t) = x(t)\cos(2\pi f_c t + \phi) + n(t) \tag{41}$$

where $x(t)$ is the information-bearing lowpass signal and ϕ is the carrier phase. For simplicity, we assume here that the propagation delay is known and we set it equal to zero. As shown in Fig. 8, the carrier recovery circuit computes a phase estimate $\hat{\phi}$, which is then used to generate the reference signal $u(t) = g(t)\sin(2\pi f_c t + \hat{\phi})$ for the correlator.

A *phase-locked loop* (PLL) [2] may be used to provide the maximum-likelihood estimate of the carrier phase. The PLL consists of a multiplier, a loop filter, and a *voltage-controlled oscillator* (VCO), as shown in Fig. 10. Let us for a moment assume that the input to the PLL is an unmodulated carrier signal represented by the sinusoid $\cos(2\pi f_c t + \phi)$. The output of the VCO is another sinusoid

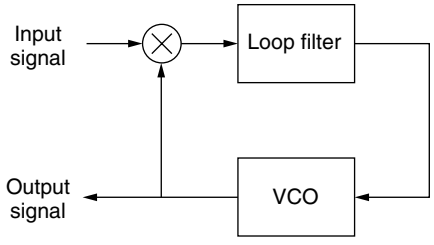


Figure 10. Phase-locked loop.

given by $\sin(2\pi f_c t + \hat{\phi})$. The output of the multiplier is the product of these two signals, given by

$$\cos(2\pi f_c t + \phi)\sin(2\pi f_c t + \hat{\phi}) = \tfrac{1}{2}\sin(\hat{\phi} - \phi) + \tfrac{1}{2}\sin(4\pi f_c t + \phi + \hat{\phi}) \quad (42)$$

The loop filter is a lowpass filter that effectively removes the double-frequency ($2f_c$) component. The output of this filter is a voltage signal that controls the VCO. The resulting PLL tracks the phase of the incoming carrier, seeking to minimize the phase error $\Delta\phi = \hat{\phi} - \phi$. In normal operation, the phase error is small and therefore $\sin(\hat{\phi} - \phi) \approx \hat{\phi} - \phi$.

We now consider two specific methods for estimating the carrier phase of a PAM signal. One widely used method involves employing a square-law device to square the received signal and generate a double-frequency signal that can then be used to drive a PLL tuned to $2f_c$ [1]. Figure 11 illustrates such a *squaring loop*, where, for the sake of clarity, we have ignored the noise term of the received signal. The output of the squarer is

$$r^2(t) = x^2(t)\cos^2(2\pi f_c t + \phi) = \tfrac{1}{2}x^2(t) + \tfrac{1}{2}x^2(t)\cos(4\pi f_c t + 2\phi) \quad (43)$$

This signal is passed through a bandpass filter tuned to $2f_c$, resulting in a periodic signal without the sign information contained in $x(t)$, namely, a phase-coherent signal at twice the carrier frequency. The filtered frequency component at $2f_c$ is used to drive the PLL. Finally, the output signal $u'(t) = \sin(4\pi f_c t + 2\hat{\phi})$ from the VCO is passed through a frequency divider to provide the output signal $u(t) = \sin(2\pi f_c t + \hat{\phi})$. This output has a phase ambiguity of 180° relative to the phase of the received signal, so binary data must be differentially encoded at the transmitter and, consequently, differentially decoded at the receiver. The squaring operation leads to some noise enhancement that results in an increase in the variance of the phase error $\Delta\phi$.

Another well-known carrier-phase estimation method is the *Costas loop* [1], which is illustrated by the block diagram shown in Fig. 12. As before, we ignore the noise term of the received signal for the sake of clarity. In this method, the received signal is multiplied by two versions of the VCO output, one phase-shifted 90° relative to the other. The multiplier outputs are

$$y_{\cos}(t) = r(t)\cos(2\pi f_c t + \hat{\phi}) = \tfrac{1}{2}x(t)\cos(\hat{\phi} - \phi) + \tfrac{1}{2}\cos(4\pi f_c t + \phi + \hat{\phi}) \quad (44)$$

and

$$y_{\sin}(t) = r(t)\sin(2\pi f_c t + \hat{\phi}) = \tfrac{1}{2}x(t)\sin(\hat{\phi} - \phi) + \tfrac{1}{2}\sin(4\pi f_c t + \phi + \hat{\phi}) \quad (45)$$

These signals are passed through lowpass filters that reject the double-frequency terms. An error signal is generated by multiplying the filtered outputs of the multipliers, yielding

$$e(t) = \tfrac{1}{8}x^2(t)\sin(2(\hat{\phi} - \phi)) \quad (46)$$

The error signal is then filtered by the loop filter, and the resulting signal is a voltage signal that controls the VCO.

As in the squaring loop, some noise enhancement occurs that causes the variance of the phase error to increase. Also, the VCO output has a 180° phase ambiguity which necessitates the use of differential encoding and decoding of the binary data.

4.5. Symbol Synchronization

As mentioned at the beginning of Section 4, the output of the correlator or matched filter must be sampled once per symbol interval. Because of the unknown propagation delay τ from the transmitter to the receiver, symbol synchronization must first be established in order to perform the sampling at the right time instants. This is usually accomplished by extracting a clock signal directly from the received signal itself and using it to control the sampling time instants $t_k = kT + \hat{\tau}$, where $\hat{\tau}$ denotes an estimate of the propagation delay.

The maximum-likelihood estimate of the propagation delay can be obtained by maximizing its likelihood

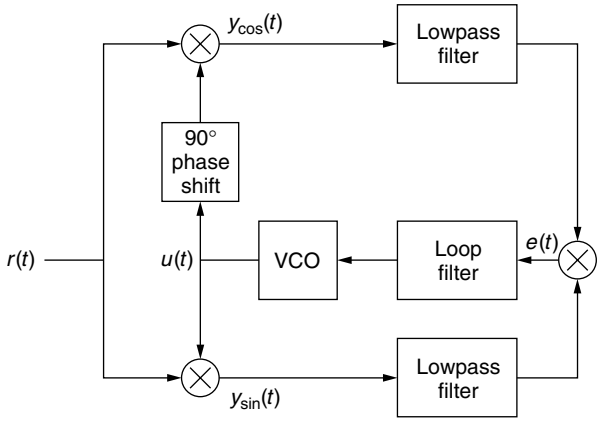


Figure 12. Carrier-phase recovery using a Costas loop.

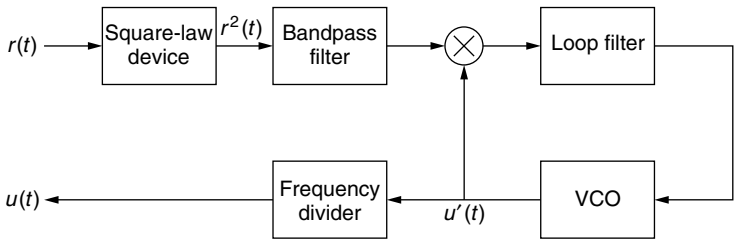


Figure 11. Carrier-phase recovery using a squaring loop.

function. Let us, for simplicity, consider baseband PAM since the procedure is easily extended to carrier-modulated PAM. The received signal is represented by

$$r(t) = s(t;\tau) + n(t) \tag{47}$$

where $s(t;\tau)$ is the (delayed) transmitted signal

$$s(t;\tau) = \sum_k x(t - kT - \tau) = \sum_k A_k g(t - kT - \tau) \tag{48}$$

and $\{A_k\}$ is the transmitted symbol sequence. The likelihood function has the form

$$\begin{aligned}\Lambda(\tau) &= C \int_{T_0} r(t)s(t)\,dt \\ &= C \sum_k A_k y_k(\tau)\end{aligned} \tag{49}$$

where $y_k(\tau)$ is the output of the correlator given by

$$y_k(\tau) = \int_{T_0} r(t)g(t - kT - \tau)\,dt \tag{50}$$

and T_0 is the integration interval. The maximum-likelihood estimate is obtained by averaging $\Lambda(\tau)$ over the probability density function (PDF) of the information symbols, $p(A_k)$, and differentiating the result. In the case of binary PAM, where $A_k = \pm 1$ with equal probability, the PDF is given as

$$p(A_k) = \tfrac{1}{2}\delta(A_k - 1) + \tfrac{1}{2}\delta(A_k + 1) \tag{51}$$

In the case of M-ary PAM, we may approximate the PDF by assuming that the symbols are continuous random variables with a zero-mean, unit-variance Gaussian distribution. In this case, the PDF is given by

$$p(A_k) = \frac{1}{\sqrt{2\pi}} e^{-A_k^2/2} \tag{52}$$

It can be shown that averaging the likelihood function over the Gaussian PDF and taking the natural logarithm yields a loglikelihood function of the form

$$\overline{\Lambda}_L(\tau) \approx \frac{1}{2} C^2 \sum_k y_k^2(\tau) \tag{53}$$

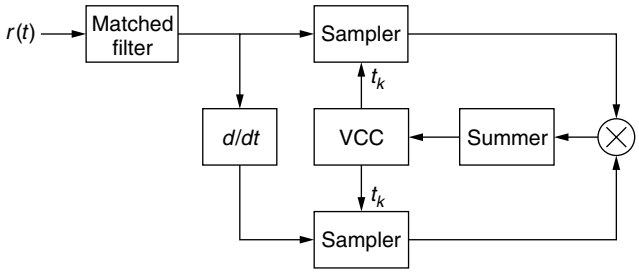


Figure 13. Non-decision-directed symbol synchronization for baseband PAM.

Next, we find the derivative of (53) and set it equal to zero:

$$\frac{d}{dt} \sum_k y_k^2(\tau) = 2 \sum_k y_k(\tau) \frac{dy_k(\tau)}{d\tau} = 0 \tag{54}$$

Figure 13 shows a symbol synchronization circuit that performs timing estimation by using a tracking loop based on Eq. (54). The circuit attempts to move the sampling instant until the derivative is zero, which occurs at the peak of the signal. The summation element serves as the loop filter that drives the voltage-controlled clock (VCC). We note that the structure of the circuit is quite similar to that of the Costas loop discussed earlier in the section on carrier-phase estimation.

A related technique, illustrated in Fig. 14, is known as the *early–late gate* synchronizer. In this technique, the output of the matched filter (or correlator) is sampled 2 times extra per symbol interval, once prior to the proper sampling instant (i.e., $T_- = T - \delta$) and once after the proper sampling instant (i.e., $T_+ = T + \delta$). As an example, let us consider the output of the filter matched to the rectangular symbol waveform introduced in Section 2. The matched-filter output attains its peak at the midpoint $t = T$, as shown in Fig. 15, and this time instant is naturally the optimal sampling time. Since the output is even with respect to the optimal sampling instant, the magnitudes of the samples taken at T_- and T_+ are equal. Hence, the proper sampling instant can be determined by adjusting the early and late sampling instants until the absolute values of the two samples are equal. An error signal is formed by taking the difference between the absolute values of the two samples, and the filtered error signal is used to drive the VCC. If the timing is off, the error signal will be nonzero and the timing signal is either advanced or retarded, depending on the sign of the error.

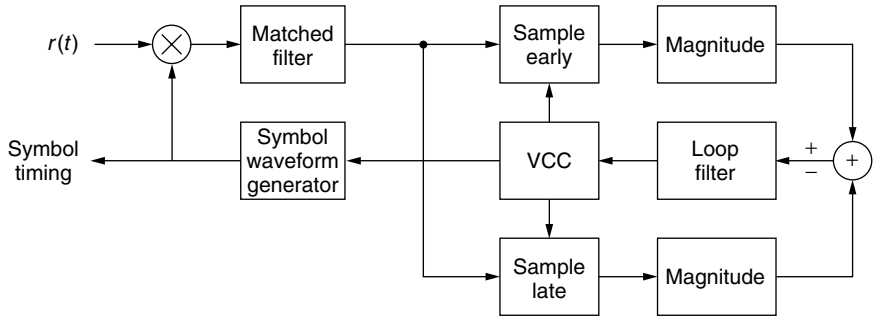


Figure 14. Symbol synchronization using an early–late gate synchronizer.

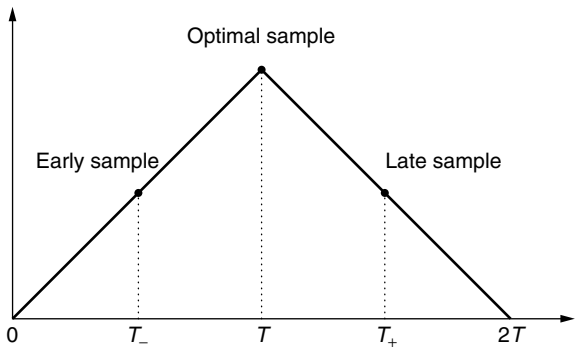


Figure 15. Matched-filter output for a rectangular signal pulse.

5. CONCLUSION

In this article, we have given an overview of the modulation method known as *pulse amplitude modulation*. We started out by introducing a mathematical representation of PAM signals and discussing their spectral characteristics. In particular, we noted that the spectrum of memoryless PAM signals can be manipulated to match the spectral characteristics of the transmission channel by selecting an appropriate signal pulse. Next, we discussed the various elements of PAM receivers for use in AWGN channels, including signal demodulation and detection, carrier-phase recovery, and symbol synchronization. We also discussed the performance of PAM detectors in terms of symbol and bit error probabilities. PAM is a well-established and mature modulation technique and has been treated in numerous articles and textbooks over the years. The interested reader will find comprehensive treatments of the topic in the works listed in the Bibliography.

BIOGRAPHY

Bjørn A. Bjerke received his Siv. Ing. degree in electrical engineering from the Norwegian Institute of Technology (NTH), Trondheim, Norway, in 1995, and his M.S. and Ph.D. degrees from Northeastern University, Boston, Massachusetts, in 1997 and 2001, respectively, both in electrical engineering. He joined Qualcomm, Inc. in 2001 and is currently a senior systems Engineer in their Concord, Massachusetts, R&D unit. His research focuses on physical layer algorithms and architectures for multi-antenna wireless communications, including channel coding/decoding, adaptive modulation, interference cancellation and channel equalization. Dr. Bjerke is a member of Eta Kappa Nu, IEEE, the Norwegian Signal Processing Society and the Norwegian Society of Chartered Engineers.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, 2001.
2. H. Meyr and G. Ascheid, *Synchronization in Digital Communications*, Vols. 1, 2, Wiley, 1990.
3. S. G. Wilson, *Digital Modulation and Coding*, Prentice-Hall, 1996.
4. J. G. Proakis and M. Salehi, *Contemporary Communication Systems Using Matlab*, Brooks/Cole, 2000.
5. B. Sklar, *Digital Communications Fundamentals and Applications*, Prentice-Hall, 1988.
6. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, 1965.
7. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, Prentice-Hall, 1994.
8. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer, 1988.

PULSE POSITION MODULATION

G. CARIOLARO
T. ERSEGHE
Università di Padova
Padova, Italy

1. INTRODUCTION

This article deals with *pulse position modulation* (PPM), which is now introduced in the more general context of *pulse modulations*. In any communication system we deal with some kind of information that we wish to send through a medium (or channel) separating the transmitter from the receiver. The information is rarely in a form that is suitable for direct transmission. In this context we call *modulation* the process of transforming the original signal in such a way that the resulting waveform can be efficiently transferred through the medium and that the original signal can be efficiently reconstructed at the receiver's end.

In *pulse modulation* systems the information is usually carried over a series of regularly recurrent pulses on which we intervene by varying parameters such as amplitude, duration, shape and, of course, temporal position. If the signal that we wish to transmit is a continuous-time function $s(t), t \in \mathbb{R}$, it will be sampled to be conformable to the discrete nature of pulses. In considering the feasibility of pulse modulations, it is important to recognize that the continuous transmission is unnecessary, provided that the continuous function $s(t)$ is band-limited and the pulses occur often enough [1]. That is, pulse modulations are inherently discrete modulations.

The necessary conditions for their feasibility are thus given by the sampling theorem stating, in its basic formulation, that any analog signal $s(t)$ with limited bandwidth range B can be uniquely expressed by samples $s_n = s(nT), n \in \mathbb{Z}$ taken at regular intervals T, where the sampling frequency $F = 1/T$ is at least twice the bandwidth range. The original signal can then be exactly reconstructed from its samples by an interpolating filter. Fortunately, in any physically realizable transmission system this condition is always satisfied and so pulse modulations constitute a feasible way to transfer information.

Nowadays, PPM is mainly used for digital transmission (which does not exclude that the original message may have an analog format converted to digital just on the basis of the sampling theorem), where the discrete sequence s_n takes the values from a finite set of amplitudes (alphabet) and where the size is typically a power of 2. So we have 2-PPM, 4-PPM, 8-PPM, and so on.

The article is organized as follows. In Section 2 we introduce PPM (both analog and digital) in the more general framework of pulse modulations, and in Section 3 we consider the problem of the generation of PPM signals. In these preliminary sections, the signals may be interpreted as deterministic or random functions as well. However, following the lines of modern communication theory, in the subsequent parts a probabilistic methodology becomes mandatory. Thus, in Section 4 we formulate the spectral analysis where the signal is modeled in terms of random processes. In the two final sections, Section 5 and Section 6, we evaluate performances of digital PPM systems in the presence of noise and also their achievable information rates. The article concludes with a brief record of modern applications, mainly in optical communication systems.

We have deliberately omitted the performance evaluation of analog PPM systems. To this regard we suggest to the interest reader the optimal lectures of [2], [3].

2. VARIETIES OF PULSE MODULATIONS

In pulse modulations the unmodulated carrier is usually a periodic repetition of a given pulse $q(t)$

$$v_0(t) = \sum_{n=-\infty}^{+\infty} q(t - nT) \tag{1}$$

and the message is always a discrete-time signal, $s_n = s(nT)$, with sampling spacing T equal to the period of the carrier. The fundamental parameters of the carrier are the *pulse shape* $q(t)$, usually rectangular with duration limited to a fraction of T and the period T that uniquely determines the sampling instants nT, $-\infty < n < +\infty$, which express the repetitiveness of pulses, that is the *synchronism*.

The message s_n can be a sampled version of an analog signal $s(t)$, $t \in \mathbb{R}$, as previously discussed, in which case we will talk of *analog* modulation. Conversely, the message can be digital, in which case the modulation becomes *digital*. The way to reconstruct the original signal is pretty different in these two cases.

Modulation of the carrier is obtained by varying the nth pulse in dependence of the value s_n, that is, the reference pulse $q(t - nT)$ is replaced by a pulse $q(t - nT, s_n)$ dependent on s_n. The modulated signal can thus be written as

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT, s_n) \tag{2}$$

The transmitter has a double role: to generate the modulated pulses $q(t, s_n)$ and to position them at the synchronism instant nT. In the digital case, if $\mathcal{A}$ is the finite set (alphabet) of values that s_n can assume, the set of modulated pulses

$$\{q(t, a) \mid a \in \mathcal{A}\} \tag{3}$$

must be in a one-to-one correspondence with $\mathcal{A}$ to guarantee demodulation. For example, if the alphabet $\mathcal{A}$ has 64 values, the modulator must be capable of generating 64 different pulses. In the analog case, s_n belongs to a continuous set, usually given by a finite interval to assure that pulses do not overlap.

The simplest pulse modulation, and maybe the most common approach, is pulse amplitude modulation (PAM), which varies the pulse amplitude proportionally to the value of s_n, that is, we have $q(t, s_n) = s_n q(t)$ (Fig. 1). Another classical modulation, the one of interest to us, is PPM which modifies the *position* of the pulses, that is, we have

$$q(t, s_n) = q(t - K\, s_n) \tag{4}$$

where K is a suitable constant with the constraint $-\frac{1}{2}T \le K\, s_n < \frac{1}{2}T$ or, alternatively, $0 \le K\, s_n < T$ to guarantee that modulated pulses do not overlap. In PPM the modulated signal has the expression

$$\boxed{v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT - K\, s_n)} \tag{5}$$

A closely related alternative, called pulse duration modulation (PDM) or pulse width modulation (PWM), would instead consider to modify the duration of pulses, that is,

$$q(t, s_n) = q(t/(K\, s_n)) \tag{6}$$

In this case the constraint is in the duration of the modulated pulses, which must be positive and not bigger

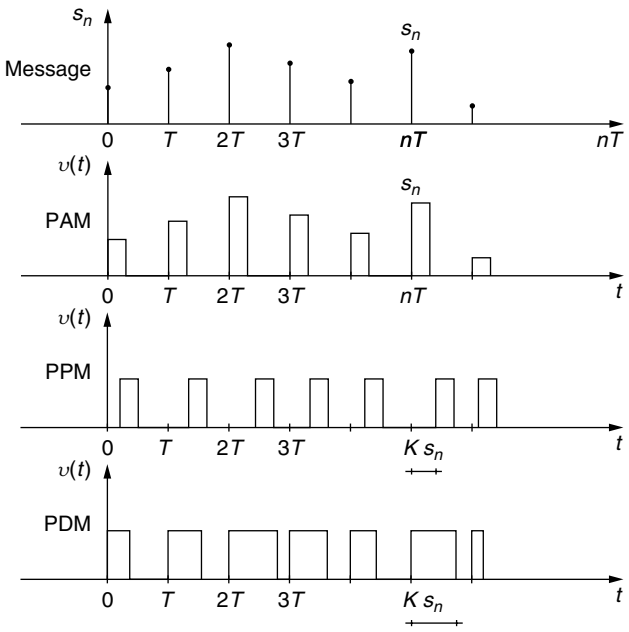


Figure 1. Classical pulse modulations.

than T, so that in (6) we have the further constraint $s_n > 0$; if s_n is not strictly positive, then the expression $K\,s_n$ must be substituted by something of the form of $T_0 + K\,s_n$.

Another worth mentioning modulation format related to PPM is pulse interval modulation (PIM), also known as differential PPM where the information is contained in the relative-distance between successive pulses. The relation to PPM can be seen in Fig. 2, where it is underlined that PIM discards the void portion of the frame after the PPM pulse, thus resulting more efficient in terms of transmission capacity and bandwidth requirements (but not in terms of transmitted-power).

3. GENERATION AND MODELS OF PPM SIGNALS

In a PPM signal the information is confined to the sequence of instants (positions) $nT + K\,s_n$ and the shape of the fundamental pulse $q(t)$ is, in some respects, irrelevant. In the ideal case the pulse may be a delta function, that is

$$v_\delta(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\,s_n) \tag{7}$$

The transition from the ideal PPM signal $v_\delta(t)$ to a PPM with a given pulse shape $q(t)$ is simply obtained by a shaping filter, that is

$$v(t) = v_\delta * q(t) = \sum_{n=-\infty}^{+\infty} q(t - nT - K\,s_n)$$

where $*$ denotes convolution.

2.1. Generation of Analog PPM Signals (general approach)

The traditional generation of analog PPM signals is based on the preliminary generation of a PDM signal. In practice, the principal use of PDM is for the generation and detection of PPM because the latter is superior for message transmission.

A sequence of operations to produce a PPM signal starting from the sampling sequence $s_n = s(nT)$ is depicted in Fig. 3. It is assumed that s_n is unipolar and bounded, that is $0 \le s_n < S_0$. The first operation is a holding of the value s_n for the corresponding time-slot $\mathcal{I}_n = [nT, (n+1)T)$, thus producing a PAM signal $\tilde{s}(t)$ with

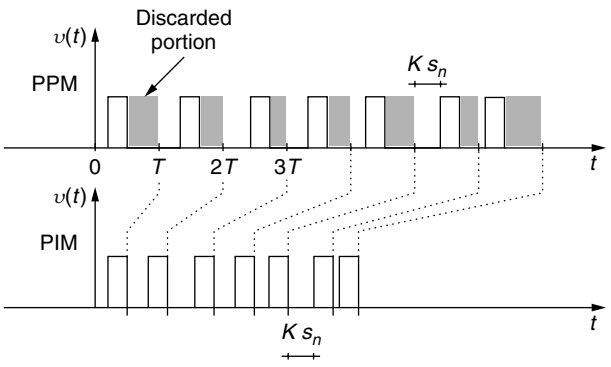


Figure 2. From PPM to PIM.

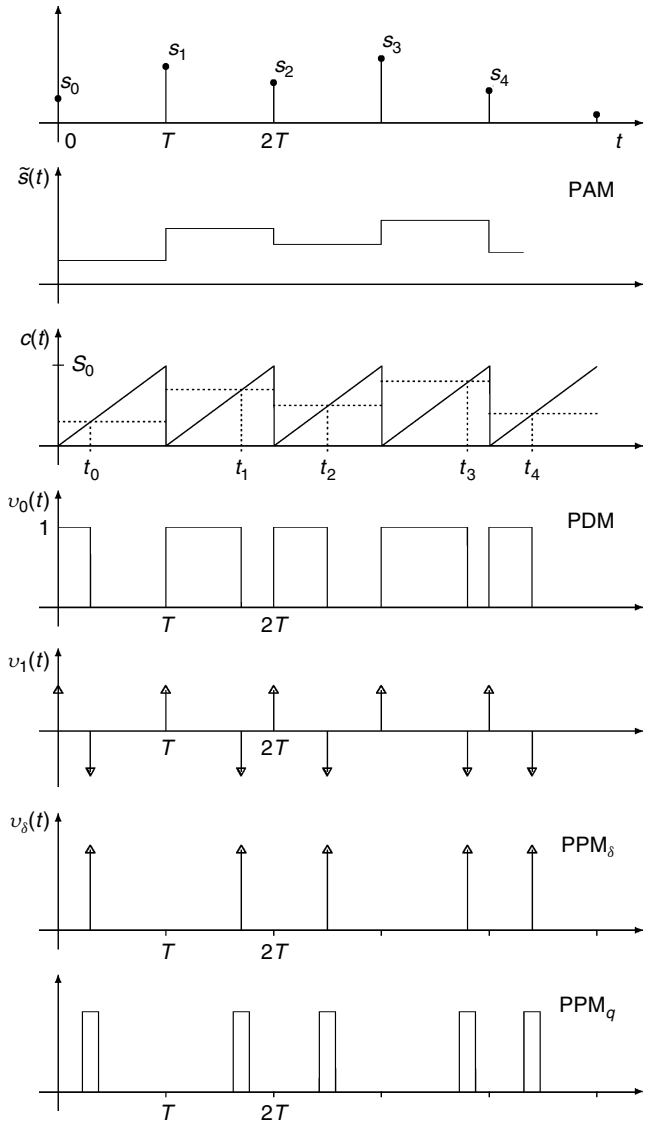


Figure 3. Generation of a PPM signal.

a full duty cycle. In the second step, a triangular carrier $c(t)$, running linearly from $c(nT) = 0$ to $c((n+1)T) = S_0$, is compared with $\tilde{s}(t)$ and a unitary output is produced as long as $\tilde{s}(t)$ is above $c(t)$. In this way, a sequence of rectangular pulses $v_0(t)$ is produced with the nth pulse starting at time nT with a duration proportional to the value of $\tilde{s}(t)$ in $\mathcal{I}_n$, that is $K\,s_n$. Hence, $v_0(t)$ is just a PDM signal. The derivative of $v_0(t)$ produces a sequence of paired delta functions $\delta(t - nT) - \delta(t - nT - K\,s_n)$. With an inverse half-wave rectification which removes the upward delta functions, the PPM signal

$$v_\delta(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\,s_n) \tag{8}$$

is obtained. Finally, a reshaping of these *spikes* gives the standard PPM format.

It is clear that an ideal receiver can recover from the PPM wave the modulating sequence s_n. The natural way

is to first obtain a PDM signal from (8), which can be done in the presence of synchronization which marks the beginning nT of the time slot. Next, the PDM signal is integrated starting from nT and with a finite integration time ($< T$). Hence, a uniform sampling at time just before $(n+1)T$ gives a sequence proportional to s_n. If s_n was obtained by sampling an analog signal $s(t)$, an interpolation finally provides the full recovery of $s(t)$. For digital PPM other more efficient recovery systems are used (see Section 5).

2.2. Analog PPM Signals with Nonuniform Sampling

In the standard analog PPM the modulating sequence s_n is obtained from a continuous-time message $s(t)$ with a sampling at the equally-spaced instants nT (uniform sampling), as remarked by writing $s_n = s(nT)$. There is another format of PPM, which is obtained by a nonuniform sampling. Historically, this form came first for its implementation is easier, at least with analog circuitry.

As a matter of fact, if in the sequence of operations of Fig. 3 we avoid the sampling, that produces the sequence s_n, and the holding, which gives $\tilde{s}(t)$ from s_n, and we feed the comparator directly with the analog message $s(t)$, the subsequent operations work as well to produce a PPM format. But, the times t_n, in correspondence of which the signal values $s_n = s(t_n)$ are taken, are no more equally spaced (Fig. 4). In fact, in the nth time-slot the carrier ramp is given by $c(t) = (t - nT)S_0/T, t \in \mathcal{I}_n$ and it is compared directly with $s(t)$. Hence, the coincidence $s(t) = c(t)$ happens at the instant

$$t_n = nT + K\, s(t_n), \qquad K = S_0/T \tag{9}$$

and the final PPM signal becomes

$$v\delta(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\, s(t_n)) \tag{10}$$

with nonuniform sampling instants determined by (9).

Although more easy to generate (with analog circuitry!), the analysis of this PPM format is very difficult because the instants t_n are determined by an implicit equation, as (9) is. Fortunately, Rowe [2], using some powerful properties of the delta function, was able to obtain a very interesting expression of (10), namely

$$\begin{aligned} v_\delta(t) &= F|1 - K\, s'(t)| \sum_{n=-\infty}^{+\infty} e^{j2\pi nF(t - K\ s(t))} \\ &= F|1 - K\, s'(t)| \left\{ 1 + 2 \sum_{n=1}^{+\infty} \cos 2\pi nF[t - K\, s(t)] \right\} \end{aligned} \tag{11}$$

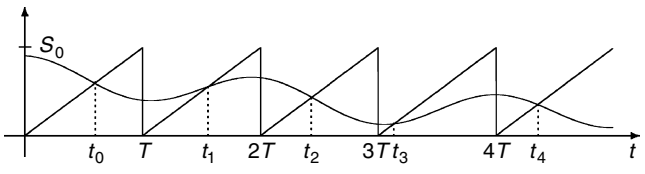


Figure 4. PPM with nonuniform sampling.

where $F = 1/T$. In Appendix A, we give an alternative and easier deduction of this expression. According to (11) the PPM signal $v\delta(t)$ consists of a baseband term $F|1 - K\ s'(t)| = A_0(t)$ and of bandpass terms around the frequencies nF,

$$2\, A_0(t) \cos 2\pi\, [nFt + \varphi_n(t)]$$

which exhibit both an amplitude modulation and a phase modulation with phase deviation $\varphi_n(t) = -nFT_0 s(t)$. Note that with a little signal processing it is possible to obtain a phase or a frequency modulated signal starting from a PPM signal with a nonuniform sampling. As a matter of fact, this possibility was concretely used in the past under the name of *serrasoid* technique.

We remark that (11), although suggestive, should be used with caution, especially in spectral analysis, because in the frequency domain the terms in (11) may have a strong overlapping.

2.3. Generation of Digital PPM Signals

For the generation of digital PPM signals, the general method can be used, but more specific methods, which rely upon the consideration that the permitted positions are finitely many, are possible. Here, we outline a general method, which is valid for all pulse modulations and gives the PPM as a special case.

Let $v(t)$ be the pulse modulated signal

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT, s_n)$$

where the data sequence $s_n = s(nT)$ belongs to an M-ary alphabet $\mathcal{A}_M = \{0, 1, \ldots, M-1\}$. The modulation format is specified by M distinct pulses $q(t, \alpha) = q_\alpha(t)$, which we store in a vector $\mathbf{q}(t) = [q_0(t), q_1(t), \ldots, q_{M-1}(t)]$. In particular, for PPM we have

$$q(t, \alpha) = q_\alpha(t) = q(t - \alpha T_0), \qquad T_0 = T/M$$

so that the vector $\mathbf{q}(t)$ collects M replicas of a same pulse uniformly distributed in the time slot $[0,\ T)$. The key of the method lies on a representation of the M-ary data s_n by a binary word of length M, $\mathbf{b}_n = [b_n(0), b_n(1), \ldots, b_n(M-1)]$, where

$$b_n(\alpha) = \delta_{s_n,\alpha} = \begin{cases} 1 & \text{for } s_n = \alpha \\ 0 & \text{for } s_n \neq \alpha \end{cases} \tag{12}$$

$\delta_{s_n,\alpha}$ being the Kronecker delta function. For instance, for $M = 4$, we find

$$\begin{aligned} \mathbf{b}_n &= [1\ 0\ 0\ 0] \quad \text{when } s_n = 0 \\ &= [0\ 1\ 0\ 0] \quad \text{when } s_n = 1 \\ &= [0\ 0\ 1\ 0] \quad \text{when } s_n = 2 \\ &= [0\ 0\ 0\ 1] \quad \text{when } s_n = 3 \end{aligned}$$

Clearly, the word sequence $\mathbf{b}_n = \mathbf{b}(nT)$ brings the same information of the original data sequence s_n without

ambiguity, and in fact one can uniquely determine s_n from $\mathbf{b}_n$ because

$$s_n = \alpha \quad \Longleftrightarrow \quad b_n(\alpha) = 1$$

that is, the value of s_n at the time nT is given by the position of the 1 in the vector $\mathbf{b}_n$.

Next, consider that the words $\mathbf{b}_n$ feed a bank of interpolating filters with impulse responses $q_\alpha(t)$ (Fig. 5). The input-output relationship of the αth filter is

$$v_\alpha(t) = \sum_{n=-\infty}^{+\infty} b_n(\alpha) q_\alpha(t - nT)$$

and in the nth time slot the output is $q_\alpha(t - nT) = q(t - nT, s_n)$ if $s_n = \alpha$ and 0 otherwise. Hence, the contribution of all filters

$$\sum_{\alpha \in \mathcal{A}_M} v_\alpha(t) = \sum_{n=-\infty}^{+\infty} \sum_{\alpha \in \mathcal{A}_M} b_n(\alpha) q_\alpha(t - nT) = v(t) \tag{13}$$

provides the desired modulated signal $v(t)$. In other words, the word sequence $\mathbf{b}_n$ gives the signaling to the filter-bank to turn on the right filter (one per time) in every time slot.

Note that the scheme of Fig. 5 provides a powerful representation of a pulse modulator because it explicitly separates the nonlinear part of the modulator, given by the word formatting operation, from the linear part, given by the filter bank $\mathbf{q}(t)$. Finally, we remark that (13) can be written in the elegant form

$$v(t) = \sum_{n=-\infty}^{+\infty} \mathbf{b}_n \mathbf{q}(t - nT)$$

provided that $\mathbf{b}_n$ is regarded as a row vector (matrix $1 \times M$) and $\mathbf{q}(t)$ is a column vector (matrix $M \times 1$).

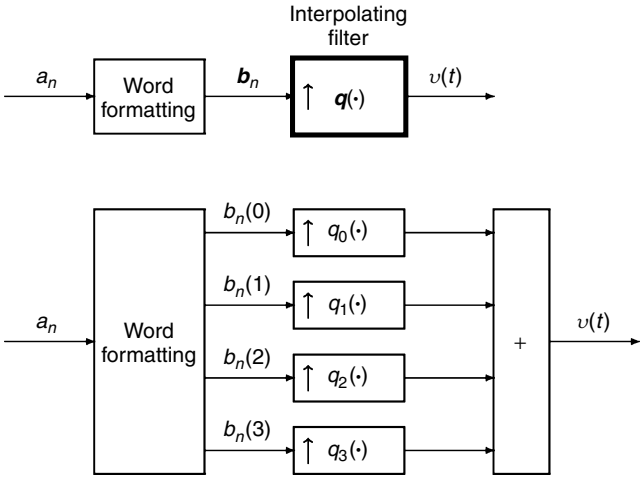


Figure 5. Generation of digital pulse modulated signals.

2.4. Digital PPM Signals Through PAM Signals

In the case of PPM, where the interpolating filters have impulse responses $q_\alpha(t) = q(t - \alpha T_0)$ that are uniformly distributed delayed versions of a reference pulse $q(t)$, the scheme of Fig. 5 can be further simplified to express PPM in terms of PAM modulations.

In particular, we obtain the scheme of Fig. 6, where the word sequence $\mathbf{b}_n$ is mapped by a parallel-to-serial conversion, called de-framing, to the binary sequence c_m (with rate $1/T_0$), where

$$c_{nM+\alpha} = b_n(\alpha)$$

The binary sequence c_m is then interpolated by a filter with impulsive response $q(t)$ to obtain the PPM signal

$$\sum_{m=-\infty}^{+\infty} c_m q(t - mT_0) = \sum_{n=-\infty}^{+\infty} \sum_{\alpha=0}^{M-1} c_{nM+\alpha} q(t - (nM + \alpha)T_0)$$
$$= \sum_{n=-\infty}^{+\infty} \sum_{\alpha=0}^{M-1} b_n(\alpha) q_\alpha(t - nT)$$

which, by use of (13), gives $v(t)$. Again, in Fig. 6 the nonlinear operations are contained in the word formatting, while the final PPM signal is obtained from the binary sequence c_m by a simple PAM modulation.

Simple modifications of the scheme of Fig. 6 let us easily define the discrete version of modulations related to PPM. For example, PDM with $M = 4$ is obtained by redefining the words

$$\begin{aligned} \mathbf{b}_n &= [1\ 0\ 0\ 0] \quad \text{when } s_n = 0 \\ &= [1\ 1\ 0\ 0] \quad \text{when } s_n = 1 \\ &= [1\ 1\ 1\ 0] \quad \text{when } s_n = 2 \\ &= [1\ 1\ 1\ 1] \quad \text{when } s_n = 3 \end{aligned}$$

Similarly we have for PIM, for which the word mapping becomes a variable-length operation,

$$\begin{aligned} \mathbf{b}_n &= [1] \quad \text{when } s_n = 0 \\ &= [0\ 1] \quad \text{when } s_n = 1 \\ &= [0\ 0\ 1] \quad \text{when } s_n = 2 \\ &= [0\ 0\ 0\ 1] \quad \text{when } s_n = 3 \end{aligned}$$

and de-framing must take into account for the variable length of words.

4. SPECTRAL ANALYSIS

As for every other modulation format, spectral analysis is a fundamental tool for understanding the band occupancy of the PPM modulation but also for several other reasons. Surely, PPM produces a *baseband* signal, that is with a

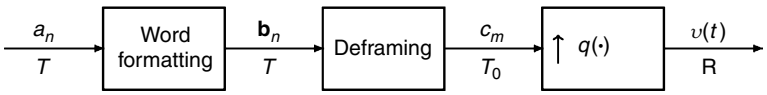


Figure 6. Digital PPM signals generated by a PAM filter.

spectrum displayed around the frequency origin, which is essentially determined by the fundamental pulse $q(t)$, and we also have to expect that the bandwidth is of the order of $1/T_0$, where T_0 is the duration of $q(t)$. However, the exact evaluation of the spectrum is not trivial for two main reasons: 1) PPM is a nonlinear modulation and 2) PPM contains a periodic component, given by the mean value, which determines the presence of spectral lines and that causes a nonstationary behavior.

4.1. Formulation of Spectral Analysis

In spectral analysis signals are modeled as random processes. In particular, we assume that the sample sequence $s_n = s(nT)$ is a discrete-time *stationary* random process. Then, it is easy to show that the PPM signal

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT - s_n) \tag{14}$$

is not stationary, rather it is *cyclostationary*, that is with a statistical description which is periodic with respect to a reference time, being the period given by the sampling period T. We start by showing this for the mean value $m_v(t) = \mathrm{E}[v(t)]$, where $\mathrm{E}[\cdot]$ is the expectation operator. Considering that s_n is stationary, we have

$$m_v(t) = \sum_{n=-\infty}^{+\infty} \mathrm{E}[q(t - nT - s_n)] = \sum_{n=-\infty}^{+\infty} \overline{q}(t - nT)$$

where $\overline{q}(t) = \mathrm{E}[q(t - s_n)]$ is the *average pulse*, which is independent of n because of stationarity. Hence, $m_v(t)$ is the periodic repetition of the average pulse and has period T.

The spectral analysis starts from the *correlation* function of $v(t)$, defined as $\tilde{r}_v(t, \tau) = \mathrm{E}[v(t)v(t + \tau)]$, which is periodic with respect to the reference time t. Then, taking the time average

$$r_v(\tau) = \frac{1}{T} \int_0^T \tilde{r}_v(t, \tau)\, dt \tag{15}$$

the dependence on t is removed. The (average) power spectral density (PSD), that is the quantity of interest in spectral analysis, is then obtained as the Fourier transform of the average correlation (15), namely

$$R_v(f) = \int_{-\infty}^{+\infty} r_v(\tau) e^{-j2\pi f\tau}\, d\tau \tag{16}$$

This is the most common approach to determine the PSD for cyclostationary processes.[1]

In general, the PSD $R_v(f)$ can be decomposed in a *continuous* part $R_v^{(c)}(f)$ and in a *discrete* part $R_v^{(d)}(f)$, which exhibits *spectral lines* (Lebesgue decomposition). Spectral lines are due to the periodic component of the PPM signal, that is by the mean value $m_v(t)$. In fact, it can be shown that $m_v(t)$ has PSD $R_v^{(d)}(f)$ and the deviation $v(t) - m_v(t)$ has PSD $R_v^{(c)}(f)$. For a correct spectral analysis it is important to evaluate the two components separately.

4.2. Evaluation of the PSD

Because PPM is a nonlinear modulation, the evaluation of the PSD requires to know the second-order statistics of the modulating sequence s_n. More specifically, let $\Psi_s(f)$ and $\Psi_s(f_1, f_2; \tau)$ be the characteristic functions of first and second order of s_n, written in terms of "frequency" in place of the usual z variables, namely

$$\begin{aligned} \Psi_s(f) &= \mathrm{E}[e^{j2\pi f s_n}] \\ \Psi_s(f_1, f_2; kT) &= \mathrm{E}[e^{j2\pi(f_1 s_n + f_2 s_{n+k})}] \end{aligned} \tag{17}$$

Let also $\Phi_s(f_1, f_2; f)$ be the (discrete) Fourier transform of $\Psi_s(\cdot; kT)$ with respect to kT, that is

$$\Phi_s(f_1, f_2; f) = \sum_{k=-\infty}^{+\infty} \Psi_s(f_1, f_2; kT) e^{-j2\pi fkT} \tag{18}$$

Then, as shown in Appendix B, the PSD of the PPM signal $v(t)$ is given by

$$\boxed{R_v(f) = F\Phi_s(-f, f; f)|Q(f)|^2} \tag{19}$$

where $F = 1/T$ and $Q(f)$ is the Fourier transform of the fundamental pulse $q(t)$, that is

$$Q(f) = \int_{-\infty}^{+\infty} q(t) e^{-j2\pi ft}\, dt$$

This result holds in general as soon as the modulating sequence s_n is a stationary process. For the spectral separation of the continuous and the discrete parts, further assumptions should be done concerning the behavior of the characteristic function $\Psi(f_1, f_2; kT)$ for $k \to \infty$. Here, we consider the simplest case in which s_n is statistically independent and refer to Ref. [2] for a more general situation.

If s_n and s_{n+k} are independent for $k \neq 0$, then the second part of (17) gives

$$\Psi_s(f_1, f_2; kT) = \begin{cases} \Psi_s(f_1 + f_2) & k = 0 \\ \Psi_s(f_1)\Psi_s(f_2) & k \neq 0 \end{cases} \tag{20}$$

Moreover, at the level of the characteristic function the separation between the discrete and continuous part is such that $\Psi_s(f_1)\Psi_s(f_2)$ determines the discrete part and the difference $\Psi_s(f_1 + f_2) - \Psi_s(f_1)\Psi_s(f_2)$ for $k = 0$ determines the continuous part. So, from (18) we obtain

$$\begin{aligned} \Phi_s^{(c)}(f_1, f_2; f) &= \Psi_s(f_1 + f_2) - \Psi_s(f_1)\Psi_s(f_2) \\ \Phi_s^{(d)}(f_1, f_2; f) &= \Psi_s(f_1)\Psi_s(f_2) \sum_{k=-\infty}^{+\infty} e^{-j2\pi fkT} \\ &= \Psi_s(f_1)\Psi_s(f_2) F \sum_{k=-\infty}^{+\infty} \delta(f - kF) \end{aligned} \tag{21}$$

[1] An equivalent way is the introduction of a *stationary version* of the cyclostationary signal, defined as $v_\vartheta(t) = v(t + \vartheta)$ where ϑ is a random variable uniformly distributed in $[0, T)$ and statistically independent on $v(t)$. It can be shown that $v_\vartheta(t)$ is stationary with correlation given by (15) and PSD given by (16).

where in the last row we used a widely known identity between sequences of exponentials and sequences of delta functions.

Finally, by substituting (21) in (19) we obtain

$$\begin{aligned} R_v^{(c)}(f) &= F|Q(f)|^2(1-|\Psi_s(f)|^2) \\ R_v^{(d)}(f) &= F^2|Q(f)|^2|\Psi_s(f)|^2 \sum_{k=-\infty}^{+\infty} \delta(f-kF) \\ &= F^2 \sum_{k=-\infty}^{+\infty} |Q(kF)|^2|\Psi_s(kF)|^2\delta(f-kF) \end{aligned} \tag{22}$$

where we see the presence of spectral lines at the frequencies that are multiples of the sampling rate $F = 1/T$, at least for those instances where $Q(kF) \neq 0$.

4.3. Spectrum of Digital PPM

In digital PPM the "positions" s_n are the discrete and equally spaced instants $0, T_0, \ldots, (M-1)T_0$ with $T_0 = T/M$ and each position is taken with a given probability $p_i = \mathrm{P}[s_n = iT_0]$. Then, the characteristic function of the first order becomes

$$\Psi_s(f) = \sum_{i=0}^{M-1} p_i e^{j2\pi fiT_0}$$

We now assume equally likely positions, that is $p_i = 1/M$ and a rectangular pulse $q(t) = 1$ for $0 < t < \alpha T_0$, where α is the duty cycle ($0 < \alpha < 1$). Then, the result will be expressed by the very well-known function $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$ and by its (less known) periodic version

$$\mathrm{sinc}_M(x) = \frac{1}{M}\frac{\sin(\pi x)}{\sin(\frac{\pi}{M}x)} \tag{23}$$

which has period M for M odd and period $2M$ for M even. Note that $\mathrm{sinc}^2(k) = 0$ for k an integer that is not a multiple of M and that $\mathrm{sinc}^2(kM) = 1$. The periodic sinc allows us to express the characteristic function in the form

$$\Psi_s(f) = \frac{1}{M}\frac{1-e^{-j2\pi MfT_0}}{1-e^{-j2\pi fT_0}} = e^{-j2\pi(M-1)fT_0}\mathrm{sinc}_M(fMT_0)$$

Conversely, the Fourier transform of the pulse is given by

$$Q(f) = \alpha T_0 e^{-j\pi\alpha T_0 f}\mathrm{sinc}(\alpha T_0 f)$$

Hence, by substitution in (22) we have

$$\begin{aligned} R_v^{(c)}(f) &= F(\alpha T_0)^2\,\mathrm{sinc}^2(\alpha T_0 f)(1-\mathrm{sinc}_M^2(fT)) \\ R_v^{(d)}(f) &= F^2(\alpha T_0)^2 \sum_{k=-\infty}^{+\infty} \mathrm{sinc}^2(k\alpha/M)\,\mathrm{sinc}_M^2(k)\delta(f-kF) \end{aligned}$$

The plot of this PSD is given in Fig. 7 for $M = 8$ and for two values of α. Note that $R_v^{(c)}(f)$ is always zero at $f = 0$ and other zeros fall at frequencies multiple of $F_0 = MF$. An indication of the bandwidth may be given by the side lobe determined by the factor $\mathrm{sinc}^2(f\alpha T_0)$, that is $1/\alpha T_0 = F_0/\alpha$. The spectral lines may be present at frequencies multiples of F_0 whenever $\mathrm{sinc}^2(k\alpha) \neq 0$. Note, in particular, that for $\alpha = 1$ there is only a line at $f = 0$; this is in agreement with the fact that in this case the mean value $m_v(t)$ degenerates to a constant value.

4.4. Spectrum of Analog PPM

When the sampling sequence s_n is continuous the characteristic function is given by

$$\Psi_s(f) = \int_{-\infty}^{+\infty} e^{-j2\pi fa} f_s(a)\,da$$

where $f_s(a)$ is the probability density function of s_n. Assuming a rectangular pulse with duration T_0, we consider two cases: a) s_n is uniform in $[0, T-T_0)$, not extended to $[0, T)$ to avoid collisions, and b) s_n is Gaussian with mean $m_s = (T-T_0)/2$ and variance σ_s^2 with $\sigma_s << T$ so that the probability of collision is negligible.

With s_n uniform the characteristic function is

$$\Psi_s(f) = e^{-j\pi(T-T_0)f}\mathrm{sinc}(f(T-T_0))$$

and hence

$$\begin{aligned} R_v^{(c)}(f) &= FT_0^2\,\mathrm{sinc}^2(fT_0)[1-\mathrm{sinc}^2(f(T-T_0))] \\ R_v^{(d)}(f) &= (FT_0)^2 \sum_{k=-\infty}^{+\infty} \mathrm{sinc}^2(kFT_0) \\ &\quad \times \mathrm{sinc}^2(k(1-FT_0))\delta(f-kF) \end{aligned}$$

which are illustrated at the top of Fig. 8 for $T_0 = \frac{1}{8}T$.

With s_n Gaussian we find that

$$\Psi_s(f) = e^{-j2\pi fm_s}e^{-(2\pi f\sigma_s)^2/2}$$

and hence

$$\begin{aligned} R_v^{(c)}(f) &= FT_0^2\,\mathrm{sinc}^2(fT_0)[1-e^{-(2\pi f\sigma_s)^2}] \\ R_v^{(d)}(f) &= (FT_0)^2 \sum_{k=-\infty}^{+\infty} \mathrm{sinc}^2(kFT_0)e^{-(2\pi kF\sigma_s)^2}\delta(f-kF) \end{aligned}$$

which are illustrated below in Fig. 8 for $T_0 = \sigma_s = \frac{1}{8}T$. We recall that these results hold when the samples s_n are statistically independent.

5. DIGITAL PPM DEMODULATION AND ERROR PROBABILITY

Digital PPM is the common approach in practical applications, mainly because of the digital nature of the information available for transmission. So, in the following we will concentrate on this "digital" format, for which the PPM signal is

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t-nT-s_n)$$

where s_n is an M-ary data sequence with alphabet $\mathcal{A}_M = \{0, T_0, \ldots, (M-1)T_0\}$ with $T_0 = T/M$. The receiver

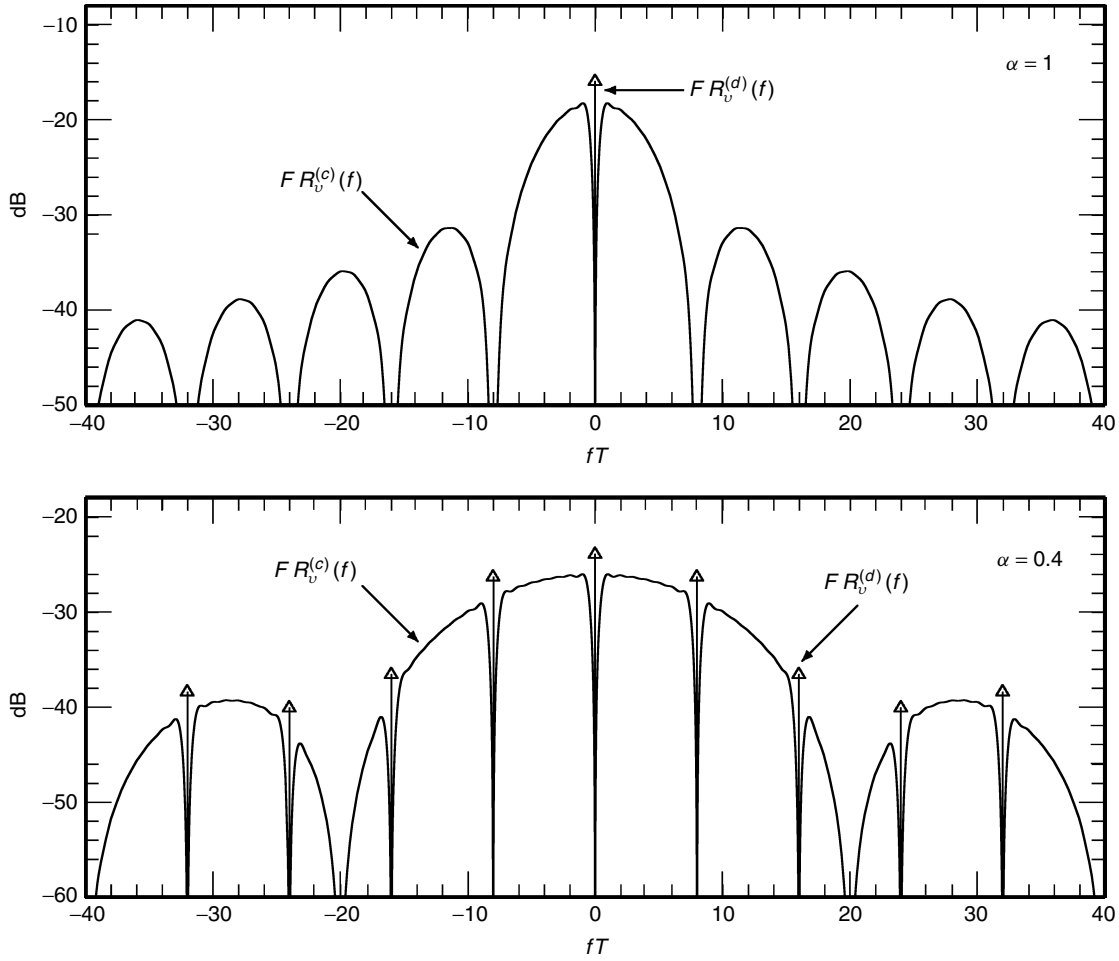


Figure 7. PSD examples for digital PPM and $M = 8$. Spectral lines are highlighted by vertical arrows.

structure is based upon optimal detection strategies and relies on banks of correlators. Fundamental performance measures of the system are given by the error probability evaluation, but also by more sophisticated capacity measures (see Section 6), which set a target for the maximum amount of information that can be sent through the PPM channel with an arbitrarily small probability of error.

5.1. Optimal Detection of Digital PPM Signals

Optimal strategies for detecting digital PPM signals rely on the assumption of ideal propagation and additive white Gaussian noise (AWGN), where the received signal can be modeled as

$$r(t) = Av(t) + \eta(t)$$

with A the attenuation and $\eta(t)$ a stationary Gaussian noise with zero mean and power spectral density $R_\eta(f) = N_0/2$. Optimal detection, that is the one that minimizes the symbol error probability (also known as maximum likelihood), is achieved by use of a matched-filter bank where the demodulator selects the value resulting in the largest cross-correlation between the received signal $r(t)$ and each of the pulses $q_\alpha(t - nT) = q(t - nT - \alpha), \alpha \in \mathcal{A}_M$.

The implementation is illustrated in Fig. 9, where the M-tuple $\varphi_n(\alpha), \alpha \in \mathcal{A}_M$ is generated by a bank of sampling filters that constitute the dual of the interpolating filter bank of Fig. 5. In particular, we have

$$\varphi_n(\alpha) = \int_{-\infty}^{+\infty} r(t) q_\alpha(t - nT)\, dt = \int_{-\infty}^{+\infty} r(t) q(t - nT - \alpha)\, dt$$

where the integration can be limited to a finite region by taking into account that, in practice, $q(t)$ always has limited extension. This extension is usually constrained to $[0, T_0)$ to overcome the superposition of adjacent pulses and, in the following, we consider this assumption (see [4] for more general cases). Finally, the demodulated sequence becomes

$$\hat{s}_n = \arg \max_\alpha \varphi_n(\alpha)$$

The statistical properties of the M-tuple $\varphi_n(\alpha), \alpha \in \mathcal{A}_M$ are usually given under the condition that the transmitted sample s_n is known. By setting $s_n = \beta$, with a little effort one easily derives that

$$\varphi_n(\alpha \mid \beta) = \varphi_n(\alpha \mid s_n = \beta) = \underbrace{AE_q \delta_{\alpha,\beta}}_{\text{expected signal}} + \underbrace{\eta_n(\alpha)}_{\text{Gaussian noise}} \qquad (24)$$

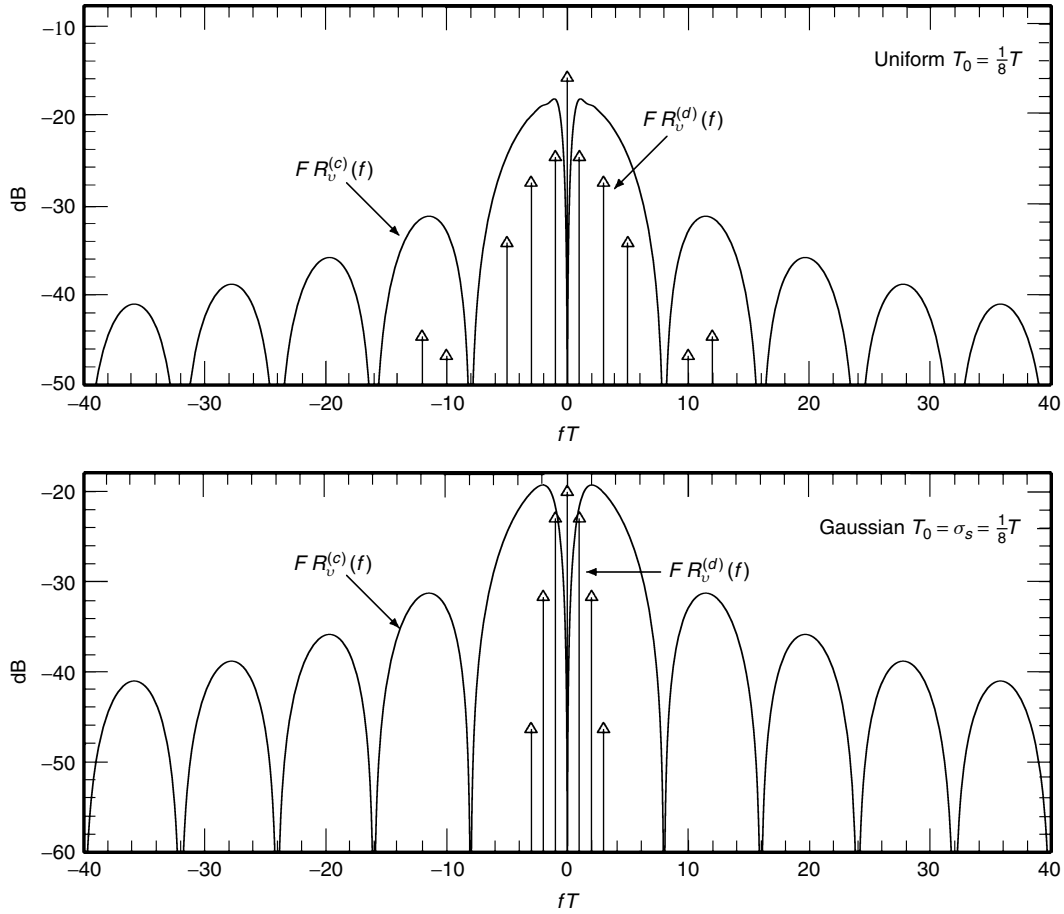


Figure 8. PSD examples for analog PPM: uniform (above) and Gaussian (below). Spectral lines are highlighted by vertical arrows.

where $E_q = \int |q(t)|^2\, dt$ is the energy of the pulse and $\delta_{\alpha,\beta}$ is the Kronecker delta function, giving 1 when $\alpha = \beta$ and 0 otherwise. Note that, the Gaussian noise is the only random term of (24) and, moreover, it is independent on the value of s_n. Because we are dealing with an AWGN channel, for any given instant nT, the random variables $\varphi_n(\alpha \mid \beta)$ are Gaussian and are thus completely specified by their means and covariances. In Appendix C we prove that these Gaussian random variables are independent with mean and variance

$$\begin{aligned} m_{\alpha|\beta} &= \mathrm{E}\,[\varphi_n(\alpha \mid \beta)] = AE_q\delta_{\alpha,\beta} \\ \sigma^2 &= \mathrm{E}\,[(\varphi_n(\alpha \mid \beta) - m_{\alpha|\beta})^2] = \tfrac{1}{2}N_0E_q \end{aligned} \tag{25}$$

where only the mean depends on the transmitted value.

5.2. Symbol Error Probability Evaluation

In the above context, the symbol error probability is given by

$$\begin{aligned} P_e = \mathrm{P}\,[\hat{s}_n \neq s_n] &= \sum_{\alpha\in\mathcal{A}_M} \mathrm{P}\,[\hat{s}_n \neq \alpha \mid s_n = \alpha]\,\mathrm{P}\,[s_n = \alpha] \\ &= \sum_{\alpha\in\mathcal{A}_M} (1 - \mathrm{P}\,[\hat{s}_n = \alpha \mid s_n = \alpha])\,\mathrm{P}\,[s_n = \alpha] \end{aligned}$$

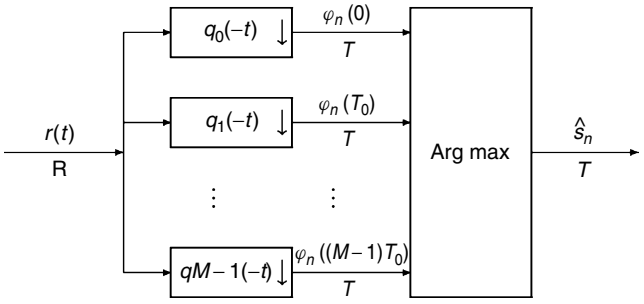


Figure 9. Optimal detection of digital PPM signals.

Because the diagonal transition probabilities

$$\mathrm{P}[\hat{s}_n = \alpha \mid s_n = \alpha] = \mathrm{P}[\alpha = \arg\max_{\gamma} \varphi_n(\gamma \mid \alpha)] = P_c$$

are independent of the value of α, as can be easily derived from the symmetric statistical properties of the M-tuple $\varphi_n(\gamma \mid \alpha)$, $\gamma \in \mathcal{A}_M$, the symbol error probability becomes

$$P_e = 1 - P_c = 1 - \mathrm{P}[\varphi_n(0 \mid 0) = \max_{\gamma} \varphi_n(\gamma \mid 0)] \tag{26}$$

where we deliberately set $\alpha = 0$.

The probability of a correct decision P_c in (26) may be further expressed in compact form as

$$\begin{aligned} P_c &= \int \mathrm{P}[\varphi_n(T_0 \mid 0) < x, \varphi_n(2T_0 \mid 0) < x, \ldots, \varphi_n \\ &\quad \times ((M-1)T_0 \mid 0) < x \mid \varphi_n(0 \mid 0) = x] f_0(x)\, dx \\ &= \int \mathrm{P}[\varphi_n(T_0 \mid 0) < x] \mathrm{P}[\varphi_n(2T_0 \mid 0) < x] \\ &\quad \times \ldots \mathrm{P}[\varphi_n((M-1)T_0 \mid 0) \times < x] f_0(x)\, dx \\ &= \int (\mathrm{P}[\varphi_n(T_0 \mid 0) < x])^{M-1} f_0(x)\, dx \end{aligned}$$

where $f_0(x)$ is the probability density of $\varphi_n(0 \mid 0)$ and where we have taken into account that all $\varphi_n(\alpha \mid 0), \alpha \neq 0$ are statistically independent and identically distributed. Then, by considering the Gaussian distribution of all $\varphi_n(\alpha \mid 0)$, we finally find

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (1 - \Phi^{M-1}(x)) \exp\left[-\frac{1}{2} \left(x - \sqrt{\frac{2\mathcal{E}_s}{N_0}} \right)^2 \right] dx \tag{27}$$

where $\mathcal{E}_s = A^2 E_q$ is the received energy per symbol and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp[-t^2/2]\, \mathrm{d}\, t$ is the normalized cumulative Gaussian distribution function.

Note that, in general, equation (27) must be evaluated numerically, except for the case $M = 2$ for which the exact expression is known to be $P_e = Q(\sqrt{2\mathcal{E}_s/N_0})$ with $Q(x) = 1 - \Phi(x)$ the complementary Gaussian cumulative distribution function.

5.3. Bit Error Probability

Sometimes, it is also desirable to convert the probability of a symbol error into the equivalent probability of a binary digit error. This is a welcome operation if we wish to compare performances of different PPM alphabets and is usually done by considering that M is a power of 2, say $M = 2^k$. So, provided that the M-ary symbol α is mapped into the k-bit word $\mathbf{d}_\alpha = (d_{\alpha,0}, \ldots, d_{\alpha,k-1})$, the bit error probability becomes

$$P_b = \sum_{\alpha \in \mathcal{A}_M} \mathrm{P}[s_n = \alpha] \sum_{\substack{\beta \in \mathcal{A}_M \\ \beta \neq \alpha}} \mathrm{P}[\hat{s}_n = \beta \mid s_n = \alpha] \frac{\mathrm{dist}(\mathbf{d}_\alpha, \mathbf{d}_\beta)}{\log_2 M}$$

where $\mathrm{dist}(\mathbf{d}_\alpha, \mathbf{d}_\beta)$ expresses the Hamming distance. Note that nondiagonal transition probabilities are independent on α and β, that is, $\mathrm{P}[\hat{s}_n = \beta \mid s_n = \alpha] = P_e/(M-1)$ for $\alpha \neq \beta$, because of the symmetric behavior of the channel. So, by further considering equally likely symbols, $\mathrm{P}[s_n = \alpha] = 1/M$, we obtain

$$P_b = \frac{2^{k-1}}{2^k - 1} P_e \tag{28}$$

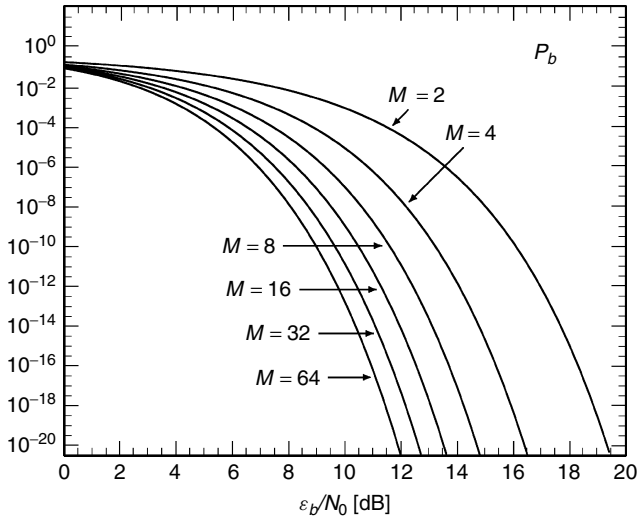


Figure 10. Bit error probability P_b versus SNR per bit $\mathcal{E}_b/N_0$ for different values of M.

that, for large values of M, reaches the limit $P_b \to P_e/2$.

In Fig. 10 we show bit error rate curves for various values of M as a function of the signal-to-noise ratio (SNR) per bit $\mathcal{E}_b/N_0$, where $\mathcal{E}_b = \mathcal{E}_s/\log_2 M$ is the energy-per-bit. Note that, for high values of the SNR, the bit error probability can be drastically reduced by increasing the cardinality M of the PPM alphabet. In particular, it can be demonstrated that as M approaches infinity, the probability of error approaches zero exponentially, provided that $\mathcal{E}_b/N_0 > \ln 2 (= -1.6\ \mathrm{dB})$.

6. INFORMATION RATES

Another fundamental measure in communications is given by the concept of *capacity* (or information rate) which expresses the maximum amount of information that can be sent through a channel with arbitrarily small error probability by use of appropriate coding techniques [5]. One of the main advantages of capacity is that it does not require us to specify the coding technique and represents a target to be (hopefully) met in the construction of error correcting codes. In the following section we investigate the capacity of digital PPM.

6.1. Formulation of Capacity

The basic concept in capacity evaluation is that of *entropy* $H(A)$ of a symbol source A, expressing the amount of information (in bits) carried by each of the symbols of the source. Entropy takes different expressions in dependence on the form of the source output (analog, digital, vectorial).

In digital PPM the source A outputs the digital symbol sequence s_n. We consider s_n to be stationary and with independent and identically distributed symbols with probabilities $p_A(\alpha) = \mathrm{P}[s_n = \alpha]$, in which case the entropy of the PPM source is

$$H(A) = \mathrm{E}[-\log_2 p_A(s_n)] = -\sum_{\alpha \in \mathcal{A}_M} p_A(\alpha) \log_2 p_A(\alpha) \quad [\text{bit/symbol}] \tag{29}$$

with the property $0 \le H(A) \le \log_2 M$, where $H(A) = \log_2 M$ if and only if the symbols are equally likely, that is $p_A(\alpha) = 1/M$.

The symbols s_n are then transmitted through the PPM channel, which includes modulation and demodulation. The output of the PPM channel is thus represented by an output source B that, in the case of optimal detection of each PPM pulse (as in the previous section), produces the symbol sequence $\hat{s}_n$. This approach is called *hard-detection*. Alternatively, one could use the correlation-measures $\varphi_n(\alpha)$ as weights in a Viterbi trellis, in which case we will talk of soft-detection and consider the real-valued vector sequence $\boldsymbol{\varphi}_n = (\varphi_n(0), \varphi_n(1), \ldots, \varphi_n(M-1))$ as the output of the source B. For ease of mathematical treatment, in the following we consider the hard-detection case.

The second quantity of interest is not represented by the entropy of B, rather on the conditional entropy $H(A \mid B)$ expressing the uncertainty on the output value of A once the output value of B is known (e.g., on the value of s_n once $\hat{s}_n$ is known). In hard-detection, the conditional entropy is expressed as

$$H(A \mid B) = \mathrm{E}[-\log_2 p_{A|B}(s_n \mid \hat{s}_n)] = -\sum_{\alpha,\beta \in \mathcal{A}_M} p_{AB}(\alpha, \beta) \log_2 p_{A|B}(\alpha \mid \beta) \qquad (30)$$

where we used the conditional probabilities $p_{A|B}(\alpha \mid \beta) = \mathrm{P}[s_n = \alpha \mid \hat{s}_n = \beta]$ and the joint probabilities $p_{AB}(\alpha, \beta) = \mathrm{P}[s_n = \alpha, \hat{s}_n = \beta]$.

From the source entropy $H(A)$ and the conditional entropy $H(A \mid B)$ it is customary to define the *average information flow* $I(A;B) = H(A) - H(A \mid B)$ expressing the difference between the information $H(A)$ carried by the source and the information $H(A \mid B)$ lost during transmission. In this context, capacity is defined as

$$C = \max_{p_A(\alpha)} I(A;B) = \max_{p_A(\alpha)} H(A) - H(A \mid B) \qquad \text{[bit/symbol]} \qquad (31)$$

that is the maximum value, taken with respect to the source symbol probabilities $p_A(\alpha), \alpha \in \mathcal{A}_M$, of the information flow. Note that, because we always have $H(A \mid B) \le H(A)$, capacity is a positive quantity.

When further taking into account that the symbol rate is $F = 1/T$, we can introduce the related concept of *information rate*

$$R = CF \qquad \text{[bit/s]} \qquad (32)$$

expressing the maximum amount of bit per second that can be sent through the channel with arbitrarily small error probability.

6.2. Information Rates for Digital PPM

In order to derive a compact expression for digital PPM capacity, it is appropriate to use the well-known identity for conditional probabilities $p_{A|B}(\alpha \mid \beta)p_B(\beta) = p_{AB}(\alpha, \beta) = p_{B|A}(\beta \mid \alpha)p_A(\alpha)$, where the meaning of the two new measures $p_{B|A}(\beta \mid \alpha)$ and $p_B(\beta)$ is obvious. With a little effort, this property lets us express the information flow as a function of the transition probabilities $p_{B|A}(\beta \mid \alpha)$ and of the source probabilities $p_A(\alpha)$, and we have

$$I(A;B) = \sum_{\alpha,\beta \in \mathcal{A}_M} p_{B|A}(\beta \mid \alpha)p_A(\alpha) \log_2 \frac{p_{B|A}(\beta \mid \alpha)}{p_B(\beta)} \qquad (33)$$

where

$$p_B(\beta) = \sum_{\gamma \in \mathcal{A}_M} p_{B|A}(\beta \mid \gamma)p_A(\gamma)$$

Moreover, from the results of the previous section we know that the transition probabilities satisfy

$$p_{B|A}(\beta \mid \alpha) = \begin{cases} 1 - P_e & \alpha = \beta \\ P_e/(M-1) & \alpha \ne \beta \end{cases}$$

where P_e is given by (27).

In the present situation, it can be proved that the maximization of (33) occurs when equally likely symbols are used, that is $p_A(\alpha) = 1/M$ in which case we also have $p_B(\beta) = 1/M$, and the capacity thus becomes

$$C = \log_2 M + P_e \log_2 \frac{P_e}{M-1} + (1 - P_e)\log_2(1 - P_e) \qquad (34)$$

while the information rate is expressed by $R = C/T$.

For a fair comparison between modulations employing different alphabet cardinalities M, it is convenient to assume that the source bit rate is fixed to $R_s = \log_2(M)/T$, and in Fig. 11 we show the normalized information rate R/R_s (efficiency) as a function of the bit-to-noise ratio $\mathcal{E}_b/N_0$. We note that for higher values of M the curves display higher efficiency and saturate faster. However, it is perhaps worth recalling that increased values of M correspond to a decrease of the pulse width T_0, and in fact $T_0 = \log_2(M)/(MR_s)$, in which case the available pulse technology could be a limiting factor for the choice of M.

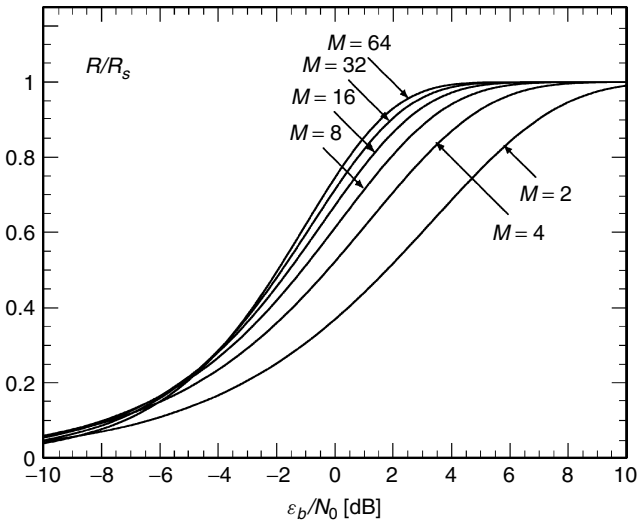


Figure 11. Efficiency R/R_s for hard-detected PPM versus SNR per bit $\mathcal{E}_b/N_0$ for different values of M.

7. APPLICATIONS (MAINLY IN WIRELESS OPTICAL COMMUNICATIONS)

PPM is perhaps the most widely used form of pulse modulation for its efficiency. We have to make a clear distinction between the analog era (the past) and the digital era (the present and the future).

In the past, analog PPM was normally used in the form of nonuniform sampling, also in connection with other modulations, since this approach resulted a simplified circuitry. To this regard the interested reader can refer to [2,3].

More recently, the digital PPM format found remarkable applications mainly in the field of optical communications, where the technology of generating very short pulses perfectly matches with the unipolar nature of PPM. It is worth recalling that PPM found some applications in fiber optics during the transition from copper-cable, where transmission required line encoding strategies with multilevel formats; in place of using multilevel PAM, in fiber optics it was more convenient to use multilevel PPM.

Nowadays, the interest on PPM is mainly restricted to wireless optical communications using infrared links, which require high average-power efficiency to minimize ocular hazards and power consumption. PPM is a technique that achieves very good power efficiency and it is largely used in these applications. For example, the Infrared Data Association (IrDA) has designated 4-PPM as the standard modulation technique for 4-Mb/s optical wireless links working over very short distances (1 m or less) [6]. These are envisioned to connect laptop computers, PDAs, palmtops, printers calculators, and mobile phones. NASA has proposed use of PPM in various free-space applications.

PPM also enters the IEEE 802.11 standard for infrared communications over local area networks (LANs), using 16-PPM schemes for 1 Mb/s links and 4-PPM for 2 Mb/s links. These are diffuse systems for which PPM is the technique that offers the best characteristics for transmission. We may thus emphasize the definitive success of PPM because of its use in many important standards in the digital format.

APPENDIX

Appendix A. Expression of PPM with Nonuniform Sampling

To prove (11) we can follow the sequence of operations of Fig. 3 with $\tilde{s}(t)$ replaced by $s(t)$, as requested in the case of nonuniform sampling.

The PDM signal can be written in the form

$$v_0(t) = u[s(t) - c(t)] \tag{35}$$

where $u(x)$ is the step function: $u(x) = 1$ for $x > 0$ and $u(x) = 0$ for $x < 0$. In fact, (35) gives 1 whenever $s(t) > c(t)$ and 0 otherwise. Considering that the derivative of the step function is the delta function, $du(x)/dx = \delta(x)$, the derivative of $v_0(t)$ is given by

$$v_1(t) = [s'(t) - c'(t)]\delta(s(t) - c(t))$$

Considering that $c(t) = (t - nT)S_0/T, t \in \mathcal{I}_n$ and recalling the property $\delta(\pm Ax) = (1/|A|)\delta(x)$, in the nth time slot we find

$$v_1(t) = [-1 + K\,s'(t)]\delta(t - nT - K\,s(t)), \quad t \in \mathcal{I}_n$$

The inverse half-wave rectification inverts the sign and we find

$$v_\delta(t) = |1 - K\,s'(t)|\delta(t - nT - K\,s(t)), \quad t \in \mathcal{I}_n$$

Hence, the expression of $v_\delta(t)$ for all t is

$$v_\delta(t) = |1 - K\,s'(t)| \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\,s(t))$$

and (11) follows after use of the well-known identity

$$\sum_{n=-\infty}^{+\infty} \delta(t - nT) = F \sum_{n=-\infty}^{+\infty} e^{j2\pi nFt}.$$

Appendix B. Proof of the PSD Expression (19)

We first introduce the PPM expression in Eq. (14) in the correlation definition

$$\begin{aligned}\tilde{r}_v(t,\tau) &= \mathrm{E}[v(t)v^*(t+\tau)] \\ &= \sum_{m,k=-\infty}^{+\infty} \mathrm{E}[q(t - mT - s_m) \\ &\quad \times q^*(t + \tau - (m+k)T - s_{m+k})]\end{aligned}$$

where the conjugate is irrelevant (because the pulse function $q(t)$ is real valued) but useful in the following. By next expressing $q(t)$ as an inverse Fourier transform, $q(t) = \int Q(f)e^{j2\pi ft}df$, and by introducing the characteristic functions (17) and the function defined by (18), we obtain

$$\begin{aligned}\tilde{r}_v(t,\tau) &= \sum_{m,k=-\infty}^{+\infty} \mathrm{E}\left[\int df_1 \int df\; Q(f_1)Q^*(f)\right. \\ &\quad \left. \times e^{j2\pi[f_1(t-mT-s_m)-f(t+\tau-(m+k)T-s_{m+k})]}\right] \\ &= \sum_{m,k=-\infty}^{+\infty} \int df_1 \int df\; Q(f_1)Q^*(f)\Phi_s(-f_1, f; kT) \\ &\quad \times e^{j2\pi[f_1(t-mT)-f(t+\tau-(m+k)T)]} \\ &= \sum_{m=-\infty}^{+\infty} \int df_1 \int df\; Q(f_1)Q^*(f)\Psi_s(-f_1, f; f) \\ &\quad \times e^{j2\pi[f_1(t-mT)-f(t+\tau-mT)]}\end{aligned}$$

which clearly shows that $\tilde{r}_v(t,\tau)$ has period T in t. At this point we evaluate (15)

$$r_v(\tau) = \frac{1}{T}\int df_1 \int df\; Q(f_1)Q^*(f)\Psi_s(-f_1, f; f)A(f_1, f)e^{-j2\pi f\tau}$$

where

$$A(f_1,f)=\sum_{m=-\infty}^{+\infty}\int_0^T e^{j2\pi(f_1-f)(t-mT)}\,dt$$
$$=\int_{-\infty}^{+\infty}e^{j2\pi(f_1-f)u}\,du=\delta(f_1-f)$$

The presence of this delta function allows us to remove the integral with respect to f_1 setting $f_1=f$ elsewhere. Hence,

$$r_v(\tau)=\frac{1}{T}\int df\; Q(f)Q^*(f)\Psi_s(-f,f;f)e^{-j2\pi f\tau}$$

which expresses $r_v(\tau)$ as the inverse Fourier transform of the quantity defined in Eq. (19). Because the inverse Fourier transform is unique, the proof is complete.

Appendix C. Mean and Covariances of $\varphi_a(nT)$

We derive mean and cross-correlation for the Gaussian random variables $\eta_n(\alpha), \alpha\in\mathcal{A}_M$ of (24). The mean value gives

$$\mathrm{E}[\eta_n(\alpha)]=\mathrm{E}\left[\int \eta(t)p(t-nT-\alpha)\,dt\right]$$
$$=\int \mathrm{E}[\eta(t)]p(t-nT-\alpha)\,dt=0$$

because $\eta(t)$ is a zero-mean Gaussian process. This proves the first of Eq. (25) since the expected signal term in Eq. (24) is a deterministic quantity. Covariance of (24) is instead completely determined by the noise terms $\eta_n(\alpha)$ and we have

$$\mathrm{E}[\eta_n(\alpha)\eta_n(\beta)]=\mathrm{E}\left[\int \eta(t)p(t-nT-\alpha)\right.$$
$$\left.\times dt\int \eta(x)p(x-nT-\beta)\,dx\right]$$
$$=\int\int \mathrm{E}[\eta(t)\eta(x)]p(t-nT-\alpha)$$
$$\times p(x-nT-\beta)\,dt\,dx$$
$$=\frac{N_0}{2}\int\int \delta(t-x)p(t-nT-\alpha)$$
$$\times p(x-nT-\beta)\,dt\,dx$$
$$=\frac{N_0}{2}\int p(t-nT-\alpha)p(t-nT-\beta)\,dt \tag{36}$$

where we used the property $\mathrm{E}[\eta(t)\eta(x)]=\frac{1}{2}N_0\delta(t-x)$. By further considering that pulses in the last of (36) are non-colliding for $\alpha\neq\beta$, the Gaussian random variables $\eta_n(\alpha)$ become statistically uncorrelated (hence independent) and the second of (25) follows straightforwardly.

BIOGRAPHIES

Gianfranco Cariolaro (M'66) was born in 1936. He received the degree in Electrical Engineering from the University of Padova, Italy, in 1960, and the Libera Docenza degree in Electrical Communications in 1968 from the same university. He was appointed full professor in 1975 and is currently Professor of Electrical Communications and Signal Theory at the University of Padova. His main research is in the fields of data transmission, images, digital television, multicarrier modulation systems (OFDM), cellular radios, deep space communications, and the fractional Fourier transform. He is author of several books including "Unified Signal Theory" (Torino, Italy: UTET, 2nd Edition, 1996).

Tomaso Erseghe was born in Valdagno, Italy, in 1972. He received the laurea degree in Telecommunication Engineering from the University of Padova, Italy, in 1996, with a thesis on the fractional Fourier transform, and the Ph.D. in Telecommunication Engineering from that same university, in 2002, with a thesis on ultra-wide-band communications. From 1997 to 1999 he worked as an R&D Engineer at Snell & Wilcox a British broadcast equipment manufacturer, in the areas of image restoration and motion compensation. He is now a Post Doc at the University of Padova. His research interests include fractional Fourier transforms, the theory of symmetries with application to the DFT, ultra-wide-band impulse radio, time-hopping constructions. He has also been involved in some research projects sponsored by the European Community.

BIBLIOGRAPHY

1. *Transmission Systems for Communications*, Bell Telephone Laboratories Inc., 4th ed., Feb., 1970.
2. H. E. Rowe, *Signals and Noise in Communication Systems*, D. Van Nostrand Company, Princeton, New Jersey, 1965.
3. H. Schwartz, W. R. Bennet, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill Inc., New York, 1966.
4. J. G. Proakis, *Digital Communications* 3rd ed., McGraw-Hill, New York, 1995.
5. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (July, Oct. 1948).
6. IrDA, Serial Infrared Link Access Protocol (IrLAP)—Version 1.1, 1996.
7. IEEE Std. 802.11, IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, September 1999.

Q

QUADRATURE AMPLITUDE MODULATION

ISRAEL KORN
University of New South Wales
Sydney, Australia

JOHN P. FONSEKA
University of Texas at Dallas
Richardson, Texas

1. INTRODUCTION

Quadrature amplitude modulation (QAM), or quadrature amplitude shift keying (QASK), is a linear method of digital modulation in which M-ary symbols are transmitted by varying the amplitude of two carriers in quadrature. QAM requires coherent detection and linear amplifiers. QAM has an excellent bandwidth efficiency (i.e., the ratio of bit rate to occupied bandwidth is high), a moderate energy efficiency (i.e., moderate values of energy-to-noise ratio per bit for a given bit error probability), and a high degree of complexity at the receiver because there is a need to track the amplitude, phase, and frequency of the carrier as well as the clock of the symbols. QAM is used mainly in modems over telephone lines but its application is growing in satellite communications, coaxial cables and line-of-sight microwave systems. Numerous papers have been published about various aspects of QAM. Reference 1 is a book totally devoted to QAM. Most books (both undergraduate and graduate level) on digital communications contain chapters of various degree of depth on QAM. A partial list of these books is presented in Refs. 2–7. The first paper about QAM was published in September 1960 and at the time of writing this article (January 2002), papers on this topic still appear in the professional literature.

2. QAM SYSTEM

A model of a typical QAM system is shown in Fig. 1.

A sequence of M-ary independent and equiprobable symbols ($M = 2^\mu$, $\mu =$ integer) $\boldsymbol{a} = (a_0 a_1 \ldots)$ is generated by a source at the rate R bauds (symbols/s) so that symbol a_i is produced at time $t = iT$, where $T = 1/R$. Symbol a_i takes values from the set

$$\Omega = \{a(m): m = 1, 2, \ldots, M\} \tag{1}$$

In QAM the symbols are two-dimensional or complex

$$a(m) = a_I(m) + j a_Q(m) \tag{2}$$

where subscripts I and Q denote the in-phase (real) and quadrature-phase (imaginary) components. The symbol set forms a symbol constellation. Several typical but simple symbol constellations are shown in Fig. 2 for $M = 8, 12, 16$. These are examples of (a) square, (b) rectangular, (c) cross, and (d) star constellations.

For rectangular constellations with $M = M_I M_Q$, $a_x(m) = \pm 1, \pm 3, \pm(M_x - 1)$, $x = I, Q$, and for square constellations $M_I = M_Q = \sqrt{M}$. At the destination, we receive the sequence $\hat{\mathbf{a}} = (\hat{a}_0 \hat{a}_1 \ldots)$ where $\hat{a}_i \in \Omega$ and $\hat{a}_i$ may differ from a_i. Hence, the symbol error probability (SEP) is

$$P(e) = P(\hat{a}_i \neq a_i) \tag{3}$$

Each symbol represents μ bits. For example, if $\mu = 4$, the 4 bits represented by a_i are $(b_{\mu i+1}, b_{\mu i+2}, b_{\mu i+3}, b_{\mu i+4})$, $b_j \in \{0, 1\}$. The bit error probability (BEP) is

$$P_b(e) = P(\hat{b}_j \neq b_j) \tag{4}$$

and the bit rate is

$$R_b = \frac{R}{\mu} \tag{5}$$

The BEP and SEP are related by the particular representation or mapping between the symbols and bit sequences. For example, in Gray coding two neighboring symbols in the symbol constellation differ by only 1 bit. In Fig. 3 we illustrate for the case of $M = 4$ two mappings of which (a) is a Gray code and (b) is not. The aim in designing a communication system is to minimize $P(e)$ and $P_b(e)$ when all other factors remain equal.

More complicated and optimal symbol constellation can be found in Refs. 8 and 9. By *optimal* we mean the one producing the minimum SEP or BEP for a given average energy-to-noise ratio. In the process of transmitting the symbols through a physical channel, we impose the sequence of symbols $\boldsymbol{a}$ on a signal that can pass through the channel. For a bandpass channel with transfer function $\tilde{H}_C(f)$ and bandwidth $\tilde{B}_c$ around the frequency f_c (illustrated in Fig. 4 for an ideal channel with rectangular frequency response) the impulse response is

$$\tilde{h}_C(t) = \mathrm{Re}\{2h_C(t)e^{jw_c t}\} = 2h_{CI}(t)\cos(w_c t) - 2h_{CQ}(t)\sin(w_c t) \tag{6}$$

where Re { } denotes the real part of the term in the braces, $\omega_c = 2\pi f_c$ and

$$h_C(t) = h_{CI}(t) + j h_{CQ}(t) \tag{7}$$

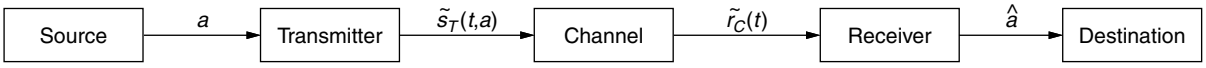


Figure 1. Typical QAM system.

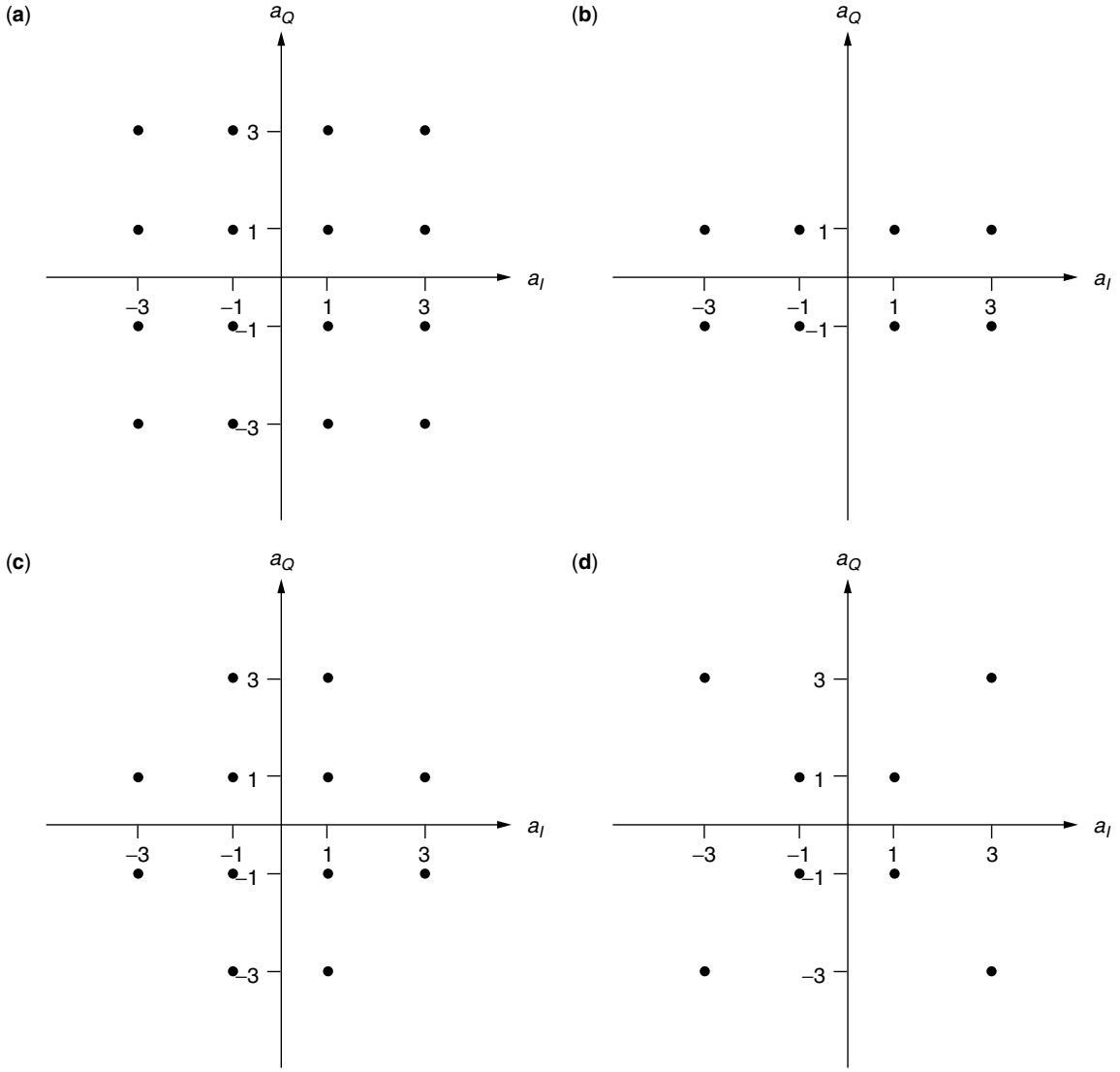


Figure 2. QAM symbol constellations: (**a**) square; (**b**) rectangular; (**c**) cross; (**d**) star.

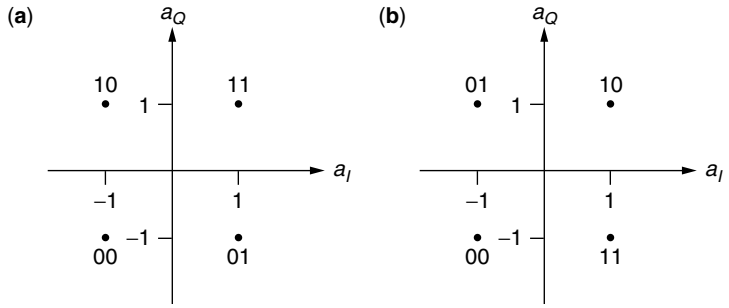


Figure 3. (**a**) Gray code; (**b**) arbitrary code.

is called the *baseband equivalent* or *complex envelope* of $\tilde{h}_C(t)$ with transfer function

$$H_C(f) = \tilde{H}_C(f+f_c)u(f+f_c), \quad u(f) = \begin{cases} 1 & f \geq 0 \\ 0 & f < 0 \end{cases} \tag{8}$$

and bandwidth $B_C = \tilde{B}_C/2$.

The terms $H_c(f)$ and $h_c(t)$ form a Fourier transform pair. Similar notation will be used for other bandpass filters, signals, and noise. Since the channel is bandpass the transmitted signal has to be also bandpass, and is generated in three stages as shown in Fig. 5.

At the first stage, we produce a bandpass signal with a certain desirable baseband shaping pulse, $h_S(t)$ and

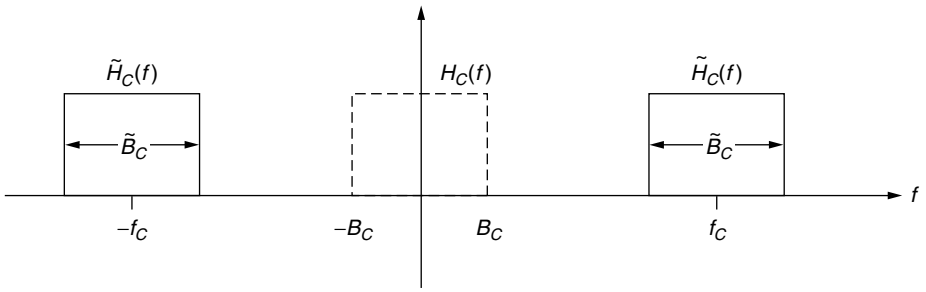


Figure 4. Transfer function of bandpass and baseband equivalent channel.

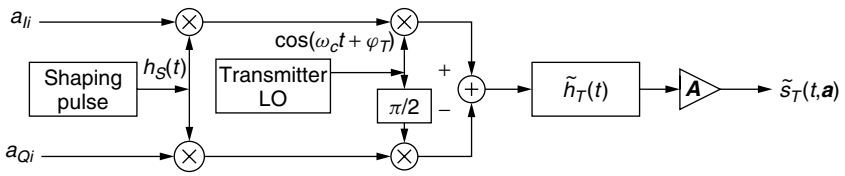


Figure 5. Transmitter of QAM.

two carriers in quadrature $\cos(w_c t + \varphi_T)$, $\sin(w_c t + \varphi_T)$ with carrier frequency f_c and phase φ_T generated by the transmitter local oscillator. This signal is

$$\begin{aligned}\tilde{s}_S(t,\boldsymbol{a}) &= \sum_i [a_{Ii} h_S(t - iT)\cos(w_c t + \varphi_T) \\ &\quad - a_{Qi} h_S(t - iT)\sin(w_c t + \varphi_T)] \\ &= \mathrm{Re}\left\{\sum_i a_i h_S(t - iT) e^{j(w_c t + \varphi_T)}\right\} \end{aligned} \tag{9}$$

with complex envelope

$$s_S(t,\boldsymbol{a}) = \sum_i a_i h_S(t - iT) e^{j\varphi_T} \tag{10}$$

The frequency upconversion from baseband to bandpass is called *modulation*. At the second stage the signal in (9) is filtered by the transmitter bandpass filter with impulse response $\tilde{h}_T(t)$. At the third stage the signal is amplified by a linear amplifier to the required power level. The total amplification of the transmitter is combined into one amplitude A. The transmitted signal is similar to those in (9) and (10) with $H_S(f)$ replaced by the transmitter transfer function $G_T(f) = H_S(f)H_T(f)$ and subscript S replaced by T.

The selection of the pulse shape is of paramount importance in the design of the communication system because it determines the power spectral density (PSD) and hence the bandwidth of transmitted signal. It can be shown [1,2] that the PSD of the signal in (10) is

$$\begin{aligned} S_S(f) &= 0.5A^2\sigma_a^2 |H_S(f)|^2, \\ \sigma_a^2 &= \frac{1}{M}\sum_{m=1}^{M} |a(m)|^2 = \frac{2}{3}(M-1) \end{aligned} \tag{11}$$

where the last equality represents a square constellation. Similarly, the PSD of the signal in (9) is a shifted version of (11) centered at $\pm f_c$.

We shall see later that in order to eliminate intersymbol interference (ISI), which increases the SEP and BEP, only certain pulses (called *Nyquist pulses*) are desirable. Many Nyquist pulses differ in duration and bandwidth. For example a rectangular pulse of duration T

$$h_s(t) = u_T(t) = u(t) - u(t - T) \tag{12}$$

is simple to generate however the resulting bandwidth is very large and more severe filtering is required by $h_T(t)$. In fact the 99% bandwidth, B_{99} of this pulse defined by

$$\int_0^{B_{99}} |H_S(f)|^2\, df = 0.99 \int_0^{\infty} |H_S(f)|^2\, df \tag{13}$$

and that contains 99% of the energy or power of the signal is $B_{99} = 8.65R$. On the other hand the infinite duration pulse $h_S(t)$ (called the raised cosine in frequency) with

$$H_S(f) = \begin{cases} 1 & 0 \le |f| \le (1-\alpha)R/2 \\ \cos\left(\dfrac{\pi}{4\alpha}\left(\left|\dfrac{2f}{R}\right| - 1 + \alpha\right)\right) & (1-\alpha)R/2 \le |f| \\ & \quad \le (1+\alpha)R/2 \\ 0 & (1+\alpha)R/2 \le |f| \le \infty \end{cases} \tag{14}$$

has a 100% bandwidth

$$B = B_{100} = \frac{(1+\alpha)R}{2} \tag{15}$$

where the parameter $0 \le \alpha \le 1$ determines the excess bandwidth beyond the minimum bandwidth of $R/2$. These two pulses are shown in Fig. 6.

For the raised-cosine pulse, no additional filtering is required as long as $B \le B_c$. In designing the system we usually select $B = B_c$ unless guard bands [to reject adjacent-channel interference (ACI) signals with carrier frequencies $f_c \pm \Delta f_c$] are required.

The bandwidth efficiency is defined by the ratio of bit rate and bandpass bandwidth:

$$\eta = \frac{R_b}{\tilde{B}} = \frac{\mu R}{2B} = \frac{\mu}{1+\alpha} \tag{16}$$

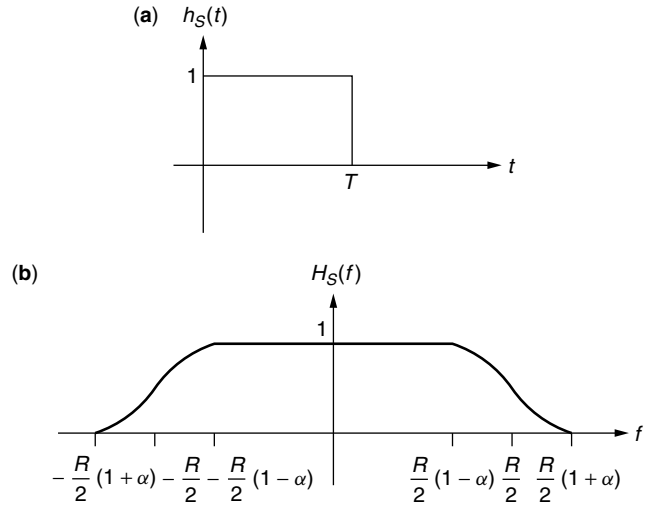


Figure 6. Nyquist pulses: (**a**) rectangular pulse; (**b**) raised cosine pulse in frequency domain with bandwidth $(1+\alpha)R/2$.

Thus we increase the bandwidth efficiency by decreasing α. In practice values of $\alpha \geq 0.15$ are achievable. The bandwidth efficiency as a function of α for $\mu = 2, 4, 6, 8, 10$ (corresponding to $M = 4, 16, 64, 256, 1024$, respectively) is shown in Fig. 7.

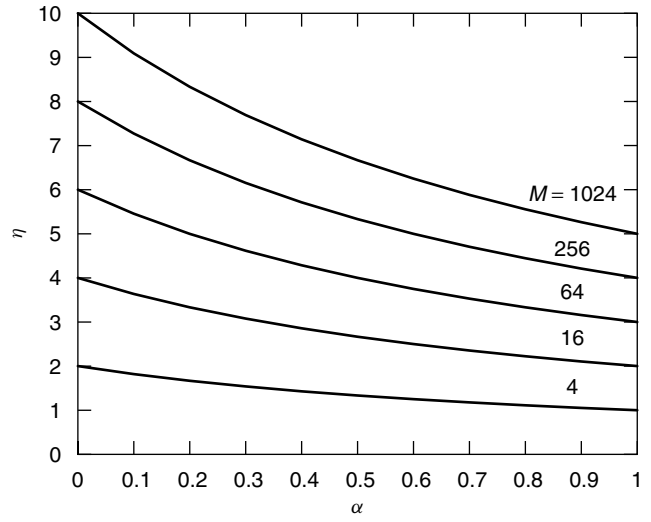


Figure 7. The bandwidth efficiency of QAM as a function of excess bandwidth.

There is a version of QASK called OQASK (offset QASK) or SQASK (staggered QASK) in which the quadrature term in (9) is delayed by $T/2$ [instead of $h_S(t-iT)$, we have $h_S(t-iT-T/2)$]. There is a whole range of shaping pulses, $h_S(t)$ for which we obtain in OQASK with $M=4$ a constant envelope signal for all time: $|\tilde{s}_T(t,\boldsymbol{a})| = c$. This enables the usage of nonlinear amplifiers, which are more efficient than linear amplifiers. In QASK there is only one shaping pulse, namely, a rectangular pulse that gives a constant envelope; however, the resulting bandwidth is large. In OQASK there are many pulses available with a reduced bandwidth. The application of nonlinear amplifiers to QAM with $M > 4$ has been presented in many papers and a comprehensive review can be found in Chap. 7 of Ref. 3.

The channel, a model of which is shown in Fig. 8, is composed of the channel filter, $\tilde{h}_C(t)$ and additive white Gaussian noise (AWGN), $\tilde{n}_C(t)$ with PSD $N_0/2$ for all frequencies. Such a channel is called a AWGN channel.

The channel filter takes into account the channel attenuation and phase shift. The channel output is thus

$$\tilde{r}_C(t) = \tilde{s}_C(t,\boldsymbol{a}) + \tilde{n}_C(t) = \mathrm{Re}\left\{A\sum_i a_i g_C(t-iT)e^{j(w_c t+\varphi_T)}\right\} + \tilde{n}_C(t), \quad G_C(f) = G_T(f)H_C(f) \tag{17}$$

The receiver, a model of which is shown in Fig. 9, is also composed of several stages.

First there is a receiver bandpass filter, $\tilde{h}_R(t)$, whose main task is to eliminate the noise beyond the signal bandwidth and to reduce ACI or spurious interference. The output of this filter is

$$\begin{aligned}\tilde{r}_R(t) &= [\tilde{s}_C(t,\boldsymbol{a}) + \tilde{n}_C(t)] * \tilde{h}_R(t) = \tilde{s}_R(t,\boldsymbol{a}) + \tilde{n}_R(t)\\ &= \mathrm{Re}\left\{\left[A\sum_i a_i g_{CR}(t-iT) + n_R(t)\right]e^{j(w_c t+\varphi_T)}\right\}\\ &= [s_{RI}(t,\boldsymbol{a}) + n_{RI}(t)]\cos(w_c t+\varphi_T)\\ &\quad - [s_{RQ}(t,\boldsymbol{a}) + n_{RQ}(t)]\sin(w_c t+\varphi_T),\\ G_{CR}(f) &= G_C(f)H_R(f)\end{aligned} \tag{18}$$

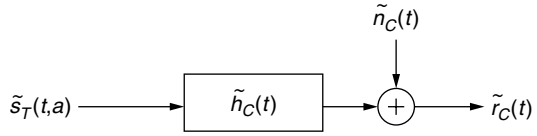


Figure 8. Model of AWGN with filter.

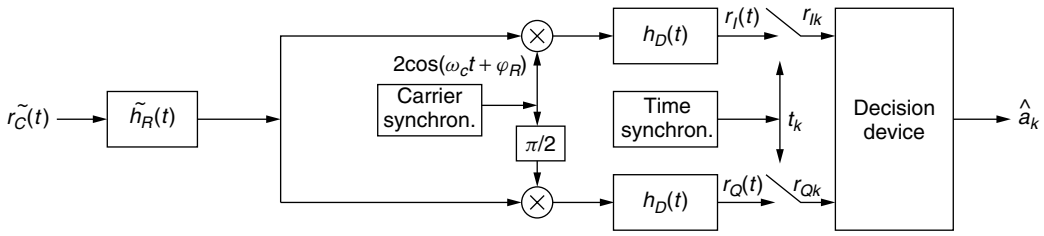


Figure 9. Receiver of QAM.

where $n_R(t)$ is the complex envelope of $\tilde{n}_R(t)$ and is a zero mean complex baseband Gaussian noise process with PSD

$$S_{n_R}(f) = N_0|H_R(f)|^2 \tag{19}$$

In the second stage the output of the receiver filter is multiplied by two carriers in quadrature, $2\cos(w_c t + \varphi_R)$, $2\sin(w_c t + \varphi_R)$ generated by the receiver local oscillator where φ_R tracks φ_T. There is a random phase error

$$\Delta\varphi = \varphi_T - \varphi_R \tag{20}$$

which also takes into account a carrier frequency error. This receiver is called a coherent receiver because it tracks the phase and frequency of the transmitter. Since

$$\begin{aligned} &2\cos(w_c t + \varphi_T)\cos(w_c t + \varphi_R) \\ &\quad = \cos\Delta\varphi + \cos\ (2w_c t + \varphi_T + \varphi_R) \\ &2\sin(w_c t + \varphi_T)\sin(w_c t + \varphi_R) \\ &\quad = \cos\Delta\varphi - \cos\ (2w_c t + \varphi_T + \varphi_R) \\ &2\sin(w_c t + \varphi_T)\cos(w_c t + \varphi_R) \\ &\quad = \sin\Delta\varphi + \sin\ (2w_c t + \varphi_T + \varphi_R) \end{aligned} \tag{21}$$

the multiplier output is the sum of a baseband signal and a bandpass signal with a carrier frequency of $2f_c$, which is easily eliminated by the baseband demodulator filter, $h_D(t)$. The frequency downconversion from bandpass to baseband is called *demodulation* or *detection* and is the inverse of the modulation. The outputs of the detector filters are

$$\begin{aligned} r_I(t) &= s_I(t,\boldsymbol{a}) + n_I(t) = \mathrm{Re}\left\{A\sum_i a_i g(t - iT)e^{j\Delta\varphi}\right\} + n_I(t) \\ r_Q(t) &= s_Q(t,\boldsymbol{a}) + n_Q(t) = \mathrm{Im}\left\{A\sum_i a_i g(t - iT)e^{j\Delta\varphi}\right\} + n_Q(t) \end{aligned} \tag{22}$$

where the transfer function of $g(t)$

$$G(f) = G_R(f)H_D(f) = H_S(f)H_T(t)H_C(f)H_R(f)H_D(f) \tag{23}$$

is the combined effect of all filters in the system on the shaping pulse and $n_I(t)$, $n_Q(t)$ are zero mean, real, baseband, independent, Gaussian noises with identical PSDs

$$S_n = N_0|G_R(f)|^2, \quad G_R(f) = H_R(f)H_D(f) \tag{24}$$

and power or variance

$$P_n = \sigma_n^2 = N_0\int_{-\infty}^{\infty}|G_R(f)|^2\,df \tag{25}$$

In (22) we left the noise unchanged by $\Delta\varphi$ because $n(t)$ and $n(t)e^{j\Delta\varphi}$ are identical, Gaussian processes. The signal in (22)

$$r(t) = r_I(t) + jr_Q(t) = A\sum_i a_i g(t - iT)e^{j\Delta\varphi} + n(t) \tag{26}$$

is sampled at times $t_k = t_0 + kT$. The timing is generated by a time synchronizer from the incoming signal therefore random variations in sampling time are expected. From $r(t_k)$ the decision circuit produces an estimate of symbol a_k, $\hat{a}_k$. We can obtain the result of (26) from the baseband equivalent block diagram shown in Fig. 10, which is composed only of baseband filters and in which the carrier frequency is irrelevant. It can be shown that the receiver presented in Fig. 9 is an optimal receiver provided $g(t)$ is a Nyquist pulse and $g_R(t)$ is a filter matched to $g_C(t)$, namely, $g_R(t) = g_C^*(t_0 - t)$, where x^* denotes the complex conjugate of x.

3. COMPUTATION OF ERROR PROBABILITY

3.1. Computation of SEP

Using the notation

$$r = r(t_k), \quad s = s(t_k,\boldsymbol{a}), \quad n = n(t_k), \quad g_k = g(t_k) \tag{27}$$

we obtain from (26)

$$r = s + n, \quad s = Ag_o e^{j\Delta\varphi}a_k + \mathrm{ISI}, \quad \mathrm{ISI} = A\sum_{i\neq 0} a_{k-i}g_i e^{j\Delta\varphi} \tag{28}$$

Under ideal conditions both $\Delta\varphi = 0$ and $\mathrm{ISI} = 0$. We may also assume that g_0 is real and nonnegative and there is a matched filter; hence

$$r = s + n, \quad s = A_0 a_k, \quad A_0 = Ag_0, \quad s(m) = A_0 a(m) \tag{29}$$

In deciding which symbol is transmitted, we divide the plane of $r = r_I + jr_Q$ into M nonoverlapping regions $\{R_m\}$

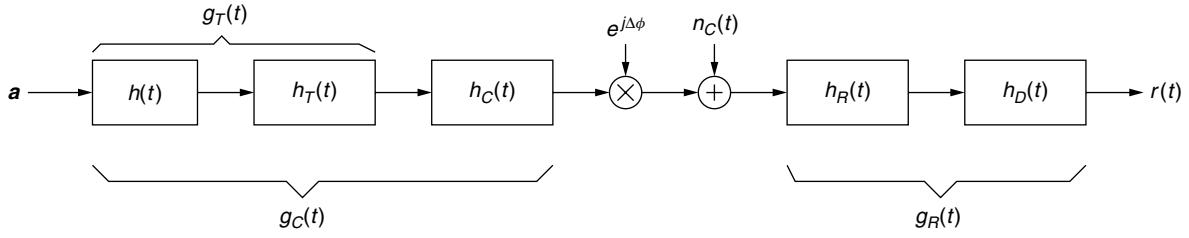


Figure 10. Baseband equivalent of QAM system.

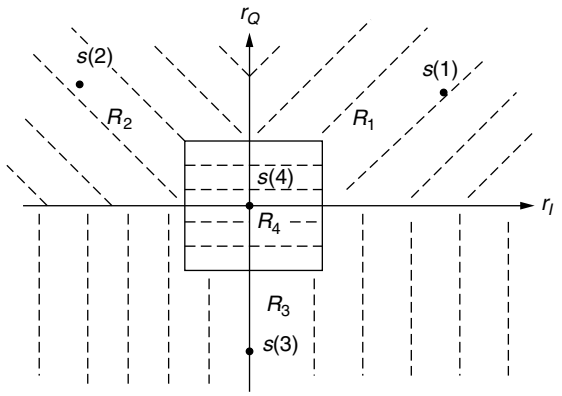


Figure 11. Arbitrary regions of the r plane.

such that if $r \in R_m$ the decision is $\hat{a}_k = a(m)$. This is illustrated in Fig. 11.

The conditional error probability that $\hat{a}_k = a(\hat{m})$ given that $a_k = a(m)$ is computed from

$$P(\hat{a}_k = a(\hat{m}) \mid a_k = a(m)) = P(r \in R_{\hat{m}} \mid a_k = a(m)) = P(\{s(m) + n\} \in R_{\hat{m}}) \tag{30}$$

For equiprobable symbols the SEP is

$$P(e) = P(\hat{a}_k \neq a_k) = \frac{1}{M} \sum_m \sum_{\hat{m} \neq m} P(\hat{a}_k = a(\hat{m}) \mid a_k = a(m)) \tag{31}$$

It can be shown that for optimum decisions we select $r \in R_m$ only if $|r - A_0 a(m)|$ is minimum. These optimum regions are illustrated for two constellations in Fig. 12.

For the square constellations the decision regions are

$$R_m = \{r : s_I(m) - C \leq r_I \leq s_I(m) + C, s_Q(m) - C \leq r_Q \leq s_Q(m) + C\} \tag{32}$$

where C is either A_0 or ∞ depending on whether the regions are finite in both dimensions, finite in one direction or infinite in both dimensions. This is illustrated in Fig. 12a. For square QAM the SEP is [2,5–7]

$$P(e) = 4\left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\sqrt{\frac{3}{M-1}\gamma}\right) - 4\left(1 - \frac{1}{\sqrt{M}}\right)^2 Q^2\left(\sqrt{\frac{3}{M-1}\gamma}\right) \tag{33}$$

where

$$\gamma = \frac{E}{N_0}, \quad E = \frac{A^2 \sigma_a^2 g_0^2}{2} \tag{34}$$

is the energy-to-noise ratio per symbol and

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-0.5y^2}\, dy = \frac{1}{\pi} \int_0^{\pi/2} e^{-0.5x^2/\sin^2\phi}\, d\phi \tag{35}$$

is the standard Q function illustrated in Fig. 13 with lower and upper bounds. $Q^2(x)$ in (33) can be computed from (35) with $\pi/2$ replaced by $\pi/4$ in the limit of the second integral.

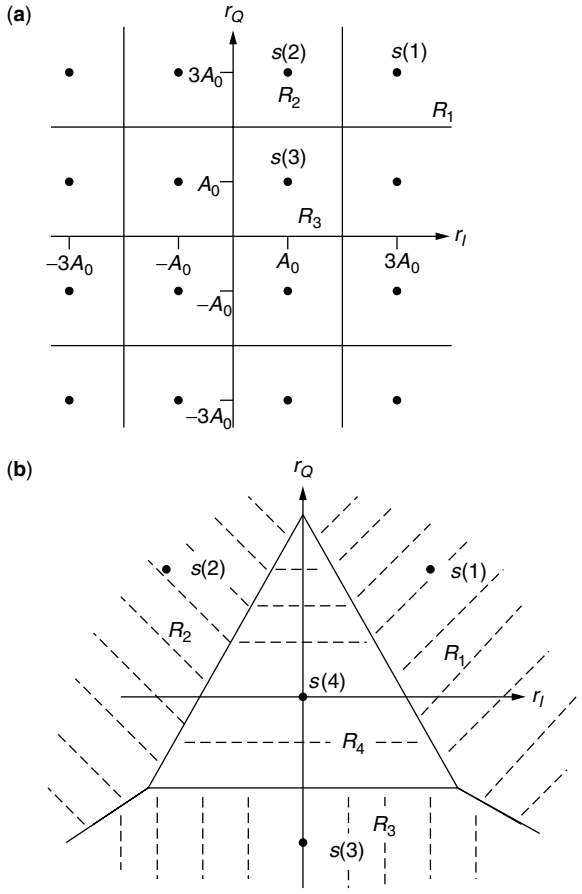


Figure 12. Optimal decision regions: (**a**) square constellation for $M = 16$; (**b**) signal constellation as in Fig. 11.

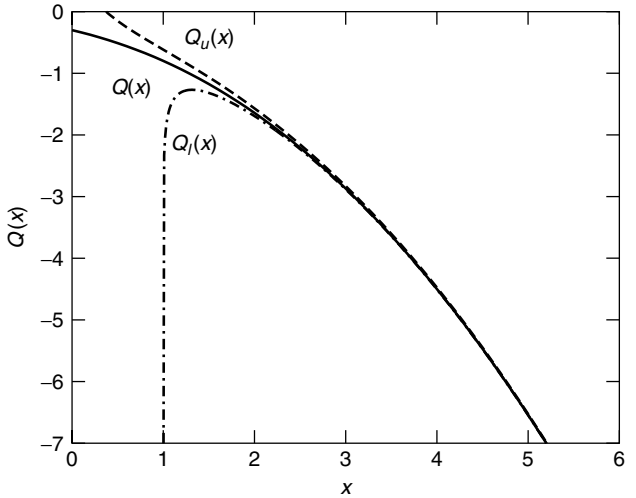


Figure 13. The variations of $Q(x)$, $Q_u(x) = \frac{1}{\sqrt{2\pi x^2}} e^{-(x^2/x)}$ and $Q_1(x) = \frac{1}{\sqrt{2\pi x^2}} \left(1 - \frac{1}{x^2}\right) e^{-(x^2/2)}$.

The SEP as a function of γ is shown in Fig. 14.

The square constellation is optimal for $M = 4$ and very close to optimal [8,9] for other M. For example, to achieve

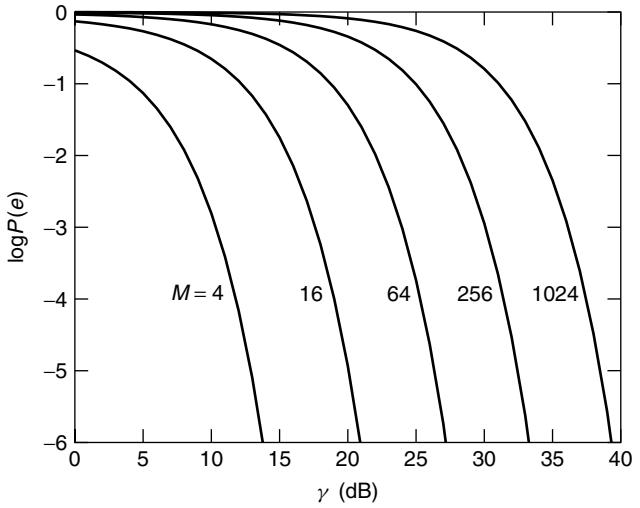


Figure 14. The SEP as a function of energy-to-noise ratio per symbol.

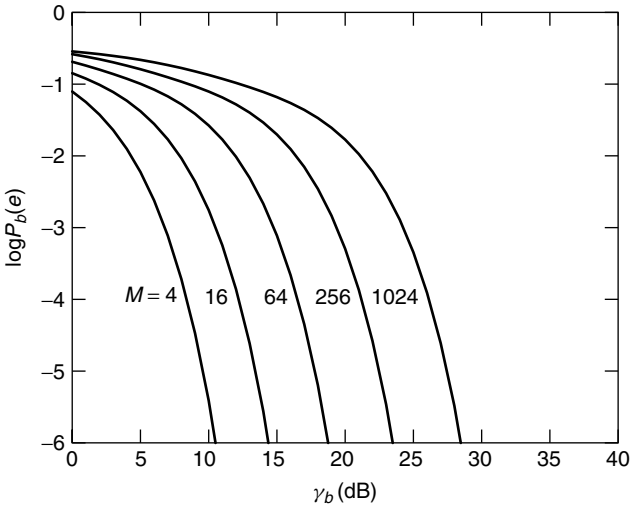


Figure 15. The BEP as a function of energy-to-noise ratio bit.

a SEP of 10^{-5}, γ increases by only 0.1 dB for $M = 16$ and by 0.4 dB for $M = 64$ relative to optimum constellation [2].

3.2. Computation of BEP

When computing the BEP, we have to take into account the mapping between symbols and bits. Let $w_{m\hat{m}}$ be the number of bits by which symbols $a(m)$ and $a(\hat{m})$ differ. Thus, even if the symbols are in error only $w_{m\hat{m}}$ out of the μ bits are in error. The BEP is

$$P_b(e) = \frac{1}{M}\sum_m \sum_{\hat{m}\neq m} \frac{w_{m\hat{m}}}{\mu} P(\hat{a}_k = a(\hat{m}) \mid a_k = a(m)) \geq \frac{P(e)}{\mu} \tag{36}$$

and the lower bound is obtained by substituting $w_{m\hat{m}} = 1$. For square constellations and Gray coding, we can compute $P_b(e)$ precisely:

$$P_b(e) = \sum_{i=1}^{I_M} c_i Q(\sqrt{d_i \gamma_b}), \quad \gamma_b = \frac{E_b}{N_0} \tag{37}$$

For example, for $M = 4$, $I_M = c_1 = 1$, $d_1 = 2$ and for $M = 16$, $I_M = 3$, $c_1 = 0.75$, $c_2 = 0.5$, $c_3 = -0.25$, $d_1 = 0.8$, $d_2 = 7.2$, $d_3 = 20$. An excellent approximation is given by [16]

$$P_b(e) \cong 4\left(1 - \frac{1}{\sqrt{M}}\right) \frac{Q\left(\sqrt{\frac{3\mu\gamma_b}{M-1}}\right) + Q\left(\sqrt{\frac{27\mu\gamma_b}{M-1}}\right)}{\mu} \tag{38}$$

In Fig. 15 we show the BEP as a function of γ_b.

3.3. Optimal Receiver and Transmitter Filters

It follows from (29) that there is no ISI if

$$g_i = \begin{cases} g_0 \neq 0 & i = 0 \\ 0 & i \neq 0 \end{cases} \tag{39}$$

where using the Inverse Fourier transform and assuming $t_0 = 0$ (there is no loss in generality in that because the delay t_0 can be introduced in the final stage so that the matched filter is physically realizable)

$$\begin{aligned} g_i = g(iT) &= \int_{-\infty}^{\infty} G(f) e^{j2\pi fiT}\, df = \sum_{k=-\infty}^{\infty} \int_{-f_N+kR}^{f_N+kR} G(f) e^{j2\pi fiT}\, df \\ &= \sum_{k=-\infty}^{\infty} \int_{-f_N}^{f_N} G(f' + kR) e^{j2\pi(f'+kR)iT}\, df' \\ &= \int_{-f_N}^{f_N} G_{\sum}(f) e^{j2\pi fiT}\, df \end{aligned} \tag{40}$$

where $f_N = R/2$, $R = 1/T$, and

$$G_{\sum}(f) = \sum_k G(f + kR) = T \sum_{i=-\infty}^{\infty} g_i e^{-j2\pi fiT} \tag{41}$$

is a periodic function of f with period R. Thus, if there is no ISI, then

$$G_{\sum}(f) = g_0 T \tag{42}$$

A function $G(f)$ which satisfies this (Nyquist) condition is called a *Nyquist function* $G_N(f)$. No function $G(f)$ with bandwidth less than $R/2$ can satisfy (42). The raised-cosine family with bandwidth $f_N(1+\alpha)$ satisfies this condition [2]. The square root of $G_N(f)$ is denoted by

$$H_N(f) = \sqrt{G_N(f)} \tag{43}$$

The pulse in (14) is $H_N(f)$ for the raised cosine with $g_0 T = 1$.

For a matched filter we can select

$$\begin{aligned} G_C(f) &= H_C(f) G_T(f) = H_N(f), \\ G_R(f) &= H_R(f) H_D(f) = H_N(f) \end{aligned} \tag{44}$$

Knowing one of the filters in the receiver or transmitter, we can find the other filters, which are now the optimal filters. In deriving (44) (although not stated explicitly), we

optimized the filters under the condition of fixed energy, E, at receiver input. If we optimize the filters with the condition of fixed energy at the transmitter the results are different [7] if $H_C(f) \neq 1$ in the range $|f| \leq R_N(1+\alpha)$.

4. QAM IN FADING CHANNELS

There are many fading channels, particularly in mobile communications. The simplest fading channel is a flat and slow fading channel [6]. In such a channel the random channel filter does not distort the signal and the variations in the channel are slow relative to the symbol rate. In such a channel only the amplitude A and phase φ_T are random. Since A is random, the energy-to-noise ratios and their square roots

$$\gamma = \frac{E}{N_0}, \quad \gamma_b = \frac{E_b}{N_0}, \quad \alpha = \sqrt{\gamma}, \quad \alpha_b = \sqrt{\gamma_b} \tag{45}$$

are random variables with probability density function (PDF) $p_x(x)$, where $x = \gamma, \gamma_b, \alpha, \alpha_b$. The SEP and BEP computed in Section 3 are based on the assumption that the receiver tracks the amplitude and phase and knows their values. Without knowledge of A, for example, the decision regions $\{R_m\}$ are useless. Thus the formulas for the SEP and BEP in Section 3 are conditional on the value of γ and γ_b, $P_S(e \mid \gamma)$, and $P_b(e \mid \gamma_b)$. In fading channels these formulas have to be averaged over γ and γ_b.

4.1. The Error Probability without Diversity

The SEP follows from (32) and (33)

$$\begin{aligned} P(e) &= \int_0^\infty P(e \mid \gamma) p_\gamma(\gamma)\, d\gamma = 4C_1 \int_0^\infty Q(\sqrt{C_2\gamma}) p_\gamma(\gamma)\, d\gamma \\ &\quad - 4C_1^2 \int_0^\infty Q^2(\sqrt{C_2\gamma}) p_\gamma(\gamma)\, d\gamma \\ &= \frac{4C_1}{\pi} \int_0^\infty \int_0^{\pi/2} e^{-0.5C_2\gamma/\sin^2\phi} p_\gamma(\gamma)\, d\gamma\, d\phi \\ &\quad - \frac{4C_1^2}{\pi} \int_0^\infty \int_0^{\pi/4} e^{-0.5C_2\gamma/\sin^2\phi} p_\gamma(\gamma)\, d\gamma\, d\phi \end{aligned} \tag{46}$$

where $C_1 = 1 - 1/\sqrt{M}$, $C_2 = 3/(M-1)$.

Applying the Laplace transform (LT)

$$\hat{p}_\gamma(s) = \int_0^\infty e^{-s\gamma} p_\gamma(\gamma)\, d\gamma \tag{47}$$

and combining with (46), we obtain

$$\begin{aligned} P(e) &= \frac{4C_1}{\pi} \int_0^{\pi/2} \hat{p}_\gamma\left(\frac{0.5C_2}{\sin^2\phi}\right) d\phi \\ &\quad - \frac{4C_1^2}{\pi} \int_0^{\pi/4} \hat{p}_\gamma\left(\frac{0.5C_2}{\sin^2\phi}\right) d\phi \end{aligned} \tag{48}$$

Instead of the LT, we can use the moment generating function $M_\gamma(s) = \hat{p}_\gamma(-s)$.

There are many models for the fading channel [6]. A popular and versatile model is the Nakagami $-m$ channel (the m here is a parameter of fading and is unrelated to the m used previously in symbol counting) for which

$$\begin{aligned} p_\gamma(\gamma) &= \frac{m^m \gamma^{m-1}}{(\overline{\gamma})^m \Gamma(m)} e^{-m\gamma/\overline{\gamma}} u(\gamma), \; p_\alpha(\alpha) = \frac{2m^m \alpha^{m-1}}{(\overline{\gamma})^m \Gamma(m)} \\ &\quad \times e^{-m\alpha^2/\overline{\gamma}} u(\alpha), \quad 0.5 \leq m \leq \infty \end{aligned} \tag{49}$$

$$\hat{p}_\gamma(s) = \left(\frac{1+s\overline{\gamma}}{m}\right)^{-m} \tag{50}$$

where

$$\Gamma(m) = \int_0^\infty t^{m-1} e^{-t}\, dt \tag{51}$$

is the gamma function, which for integer m is

$$\Gamma(m) = (m-1)! \tag{52}$$

and $\overline{\gamma}$ is the average value of γ. Note that for $m = 1$, $p_\alpha(\alpha)$ is a Rayleigh PDF, for $m = 0.5$ $p_\alpha(\alpha)$ is a half-Gaussian PDF (because $\Gamma(0.5) = \sqrt{\pi}$), and for $m = \infty$ there is no fading, $p_\alpha(\alpha) = \delta(\alpha - \sqrt{\overline{\gamma}})$ where $\delta(x)$ is impulse function. The fading becomes less severe when m increases. Thus the SEP follows from (48) and (50) as

$$\begin{aligned} P(e) &= \frac{4C_1}{\pi} \int_0^{\pi/2} \left(1 + \frac{0.5C_2\overline{\gamma}}{m \sin^2\phi}\right)^{-m} d\phi \\ &\quad - \frac{4C_1^2}{\pi} \int_0^{\pi/4} \left(1 + \frac{0.5C_2\overline{\gamma}}{m \sin^2\phi}\right)^{-m} d\phi \end{aligned} \tag{53}$$

There is a closed-form formula for integer m that can be found in Eq. (8.108) of Ref. 6. In Fig. 16 we present the SEP as a function of $\overline{\gamma}$ of 16-QAM for several values of m. We can see from this figure that to obtain a SEP of 10^{-2} for a Rayleigh channel, we need an average energy-to-noise ratio per symbol of ~28 dB instead of about ~15 dB with no fading.

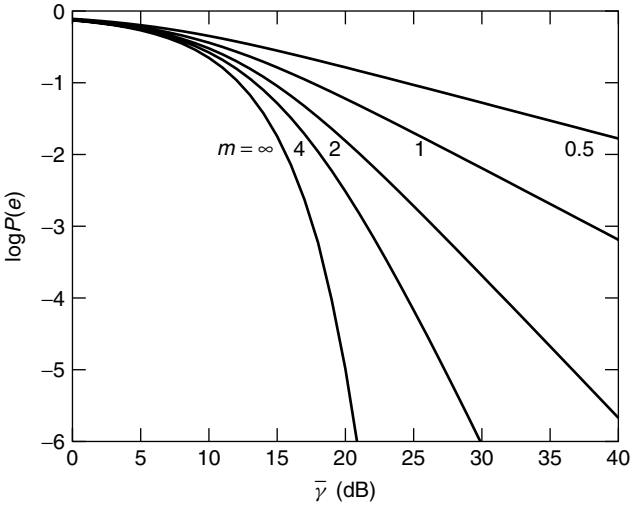


Figure 16. The SEP of 16-QAM as a function of average energy-to-noise ratio per symbol in Nakagami $-m$ fading channel.

The BEP is computed similarly. We cannot use the approximation of (38) unless $\overline{\gamma}_b$ is large. To be precise, we have to now use the conditional BEP (37):

$$P_b(e \mid \gamma_b) = \sum_{i=1}^{I_M} c_i Q(\sqrt{d_i \gamma_b}) \tag{54}$$

After averaging with respect to $p_{\gamma_b}(\gamma_b)$, we obtain

$$P_b(e) = \sum_{i=1}^{I_M} \frac{c_i}{\pi} \int_0^{\pi/2} \hat{p}_{\gamma_b}\left(\frac{0.5 d_i}{\sin^2 \phi}\right) d\phi \tag{55}$$

For $M = 16$ and Nakagami $-m$ fading, the result is

$$\begin{aligned} P_b(e) = {} & \frac{3}{4\pi} \int_0^{\pi/2} \left(1 + \frac{0.4\overline{\gamma}_b}{m \sin^2 \phi}\right)^{-m} d\phi \\ & + \frac{1}{2\pi} \int_0^{\pi/2} \left(1 + \frac{3.6\overline{\gamma}_b}{m \sin^2 \phi}\right)^{-m} d\phi \\ & - \frac{1}{4\pi} \int_0^{\pi/2} \left(1 + \frac{10\overline{\gamma}_b}{m \sin^2 \phi}\right)^{-m} d\phi \end{aligned} \tag{56}$$

There is a closed-form for integer m [see Eq. (5.17a) of Ref. 6]. In Fig. 17 we show the BEP as a function of $\overline{\gamma}_b$ for QAM with $M = 16$ for several values of m.

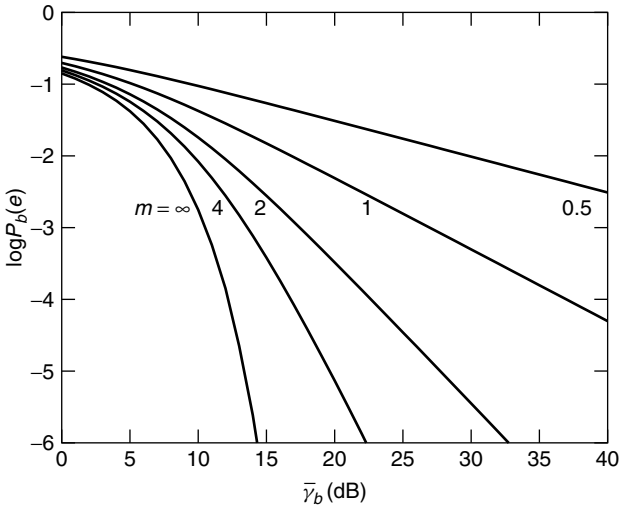


Figure 17. The BEP of 16-QAM as a function of average energy-to-noise ratio per bit for Nakagami $-m$ fading channel.

4.2. The Error Probability with Diversity

In order to reduce the error probability in fading channels, we use *diversity*, where the signal is received simultaneously via K fading channels (by using, e.g., L antenna), which hopefully fade independently so that at least one of the signals has sufficient energy. The received signals are (using complex envelope notation)

$$r_{Cl}(t) = A_l e^{j\varphi_l} \sum_l a_l \cdot g_T(t - \tau_l) + n_{Cl}(t), \quad l = 1, 2, \ldots, L \tag{57}$$

where the amplitudes $\{A_l\}$, phases $\{\varphi_l\}$, and delays $\{\tau_l\}$ are independent random variables that can be tracked by the receiver and $n_{Cl}(t)$ are independent, zero mean, Gaussian, white noises. These signals are combined in an optimal way, called *maximal ratio combining* (MRC) [6] by first aligning the signals in time, using delays $\tau_L - \tau_l$ (τ_L has the maximal delay), then multiplying by $A_l e^{-j\varphi_l}$ and finally processing by $g_R(t)$, which should be a matched filter matched to $g_T(t)$. We have not included the channel filter $h_C(t)$ because we assume flat fading. The block diagram of MRC receiver is shown in Fig. 18.

The decision is based on

$$r(t) = \sum_{l=1}^{L} A_l e^{-j\varphi_l} r_{Ck}(t - \tau_L + \tau_l) \tag{58}$$

For MRC the energy-to-noise ratio per symbol is the sum of the energy-to-noise ratios in each channel [6]

$$\gamma_{\text{MRC}} = \sum_{l=1}^{L} \gamma_l, \quad \gamma_l = \frac{E_l}{N_0} \tag{59}$$

and $\{\gamma_l\}$ are independent random variables. The SEP is

$$\begin{aligned} P(e) = \overline{P(e \mid \gamma_{\text{MRC}})} = {} & \frac{4C_1}{\pi} \int_0^{\pi/2} \overline{e^{-0.5C_2 \gamma_{\text{MRC}}/\sin^2 \phi}}\, d\phi \\ & - \frac{4C_1^2}{\pi} \int_0^{\pi/4} \overline{e^{-0.5C_2 \gamma_{\text{MRC}}/\sin^2 \phi}}\, d\phi \end{aligned} \tag{60}$$

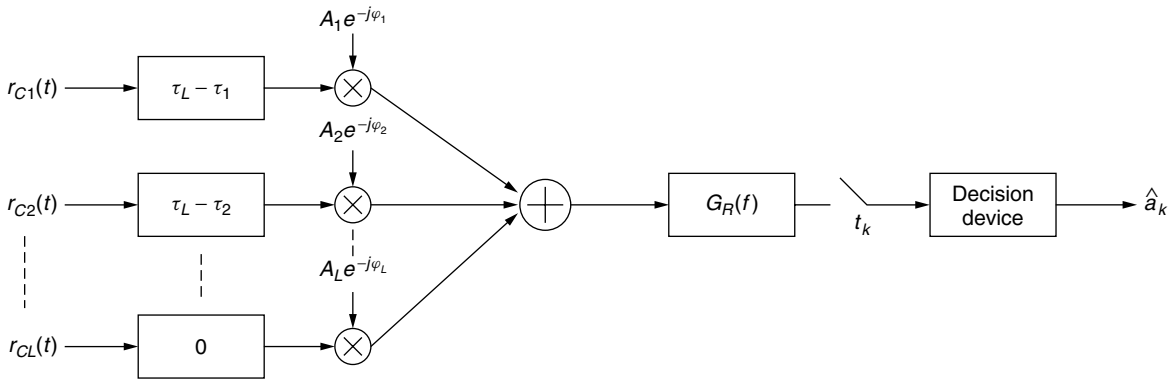


Figure 18. Maximal ratio combining of L fading signals.

where the overbar (vinculum) denotes the average operation. For independent random variables

$$\overline{e^{-C\gamma_{\mathrm{MRC}}}} = \overline{e^{-C\sum_{l=1}^{L}\gamma_l}} = \prod_{l=1}^{L}\overline{e^{-C\gamma_l}} = \prod_{l=1}^{L}\hat{p}_{\gamma_l}(C) = \hat{p}_{\gamma}^{L}(C) \quad (61)$$

where $C = 0.5\,C_2/\sin^2\phi$, $\hat{p}_{\gamma_l}(s)$ is the LT of $p_{\gamma_l}(\gamma_l)$, and the last equality is for identical PDFs for all γ_l,. We thus have

$$P(e) = \frac{4C_1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\hat{p}_{\gamma_l}\left(\frac{0.5C_2}{\sin^2\phi}\right)d\phi - \frac{4C_1^2}{\pi}\int_0^{\pi/4}\prod_{l=1}^{L}\hat{p}_{\gamma_l}\left(\frac{0.5C_2}{\sin^2\phi}\right)d\phi \quad (62)$$

which is a generalization of (48). For Nakagami $-m$ fading with m_l in channel l (62) turns into

$$P(e) = \frac{4C_1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\left(1+\frac{0.5C_2\overline{\gamma}_l}{m_l\sin^2\phi}\right)^{-m_k}d\phi - \frac{4C_1^2}{\pi}\int_0^{\pi/4}\prod_{l=1}^{L}\left(1+\frac{0.5C_2\overline{\gamma}_l}{m_l\sin^2\phi}\right)^{-m_k}d\phi \quad (63)$$

If all $m_l = m$ and $\overline{\gamma}_l = \overline{\gamma}$, we obtain

$$P(e) = \frac{4C_1}{\pi}\int_0^{\pi/2}\left(1+\frac{0.5C_2\overline{\gamma}}{m\sin^2\phi}\right)^{-mL}d\phi - \frac{4C_1^2}{\pi}\int_0^{\pi/4}\left(1+\frac{0.5C_2\overline{\gamma}}{m\sin^2\phi}\right)^{-mL}d\phi \quad (64)$$

which is identical to (53) for $L = 1$ (no diversity). For integer m, the SEP in (64) can be written in a closed form (without integrals). In Fig. 19 we show the SEP of 16-QAM as a function of $\overline{\gamma}$ (energy-to-noise ratio per symbol in one channel) for $m = 1$ (Rayleigh fading) and several values of L. We see from this figure that diversity significantly reduces the SEP.

The corresponding BEP can be calculated in terms of the respective bit quantities

$$\gamma_{b\mathrm{MRC}} = \sum_{l=1}^{L}\gamma_{bl}, \quad \gamma_{bl} = \frac{E_{bl}}{N_0} \quad (65)$$

and replacing (55) by

$$P_b(e) = \sum_{i=1}^{I_M}\frac{c_i}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\hat{p}_{\gamma_{bl}}\left(\frac{0.5d_i}{\sin^2\phi}\right)d\phi = \sum_{i=1}^{I_M}\frac{c_i}{\pi}\int_0^{\pi/2}\hat{p}_{\gamma_b}^{L}\left(\frac{0.5d_i}{\sin^2\phi}\right)d\phi \quad (66)$$

where the last equality is for identical PDFs for all γ_{bl}.

For Nakagami $-m$ fading with identical $m_l = m$ and $\overline{\gamma}_{bl} = \overline{\gamma}_b$, we obtain

$$P_b(e) = \sum_{i=1}^{I_M}\frac{c_i}{\pi}\int_0^{\pi/2}\left(1+\frac{0.5d_i}{m\sin^2\phi}\right)^{-mL}d\phi \quad (67)$$

which can be written in a closed form for integer m.

For 16-QAM, we obtain an equation similar to (56) with the exponential $-m$ replaced by $-mL$. In Fig. 20 we show the BEP of 16-QAM as a function of $\overline{\gamma}_b$ (energy-to-noise ratio per bit in one channel) for $m = 1$ and several values of L.

We conclude that the BEP is also reduced by diversity.

5. CARRIER AND SYMBOL SYNCHRONIZATION

At the receiver of QAM we have to estimate the incoming carrier frequency and phase (this is called *carrier synchronization* or *carrier recovery*) as well as the symbol rate and time delay (this is called *time* or *symbol synchronization* or *clock recovery*). Several books and many papers [11–15] as well as a special (August 1980) issue of

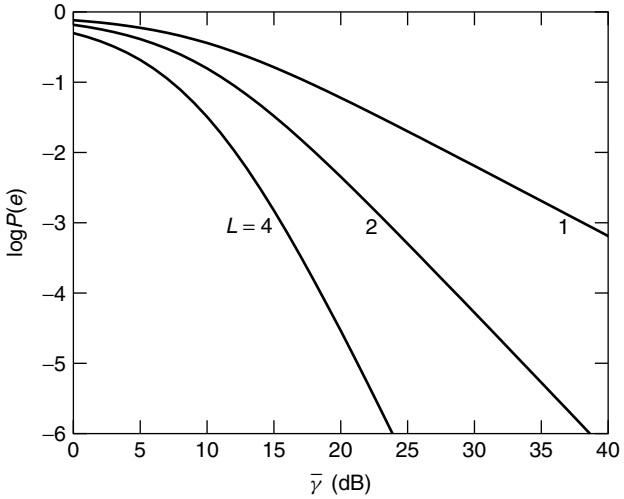


Figure 19. The SEP of 16-QAM with MRC diversity in Rayleigh fading channel.

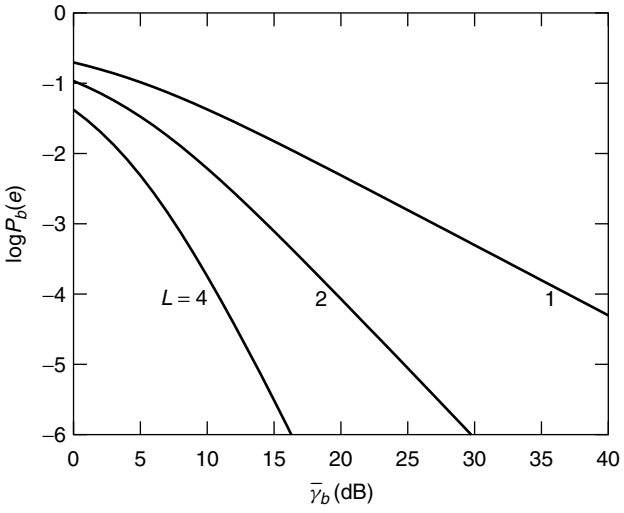


Figure 20. The BEP of 16-QAM with MRC diversity in Rayleigh fading channel.

the *IEEE Transactions on Communications* are dedicated to this problem.

The estimation is performed in two stages: an acquisition stage and a tracking stage. At the acquisition stage we can either use a known sequence of symbols (called *pilot symbols*), which form a preamble to the random data sequence, or we use the unknown data symbols, in which case the acquisition is called blind. At the tracking stage (when the SEP and BEP are already low), we can use the detected symbols $\{\hat{a}_k\}$ instead of the pilot symbols in which case the tracking is called decision directed. The basic device of every synchronizer is a phase-locked loop (PLL) [16].

5.1. Carrier Recovery with Pilot Tone

A block diagram of a PLL is shown in Fig. 21.

We shall assume that the input is a pilot signal corrupted by bandpass noise with PSD $N_0/2$ in the vicinity of the carrier frequency:

$$\tilde{r}_R(t) = A\cos(w_c t + \varphi(t)) + \tilde{n}_R(t) \tag{68}$$

The voltage controlled oscillator (VCO) produces a carrier whose phase is controlled by the input voltage

$$\tilde{s}_v(t) = A_v\cos(w_c t + \hat{\varphi}(t)), \quad \hat{\varphi}(t) = K_v\int_0^t v(t')\,dt' \tag{69}$$

The multiplier output is

$$e(t) = K_m\tilde{s}_v(t)\tilde{r}_R(t) = A_L[\sin(\varphi(t) - \hat{\varphi}(t)) + n(t)] + \text{HFT},$$
$$A_L = \frac{AA_vK_m}{2} \tag{70}$$

where $n(t)$ is zero mean, Gaussian noise with PSD, N_0/A^2, and HFT is a high-frequency term that is eliminated by the loop filter, which is a lowpass filter with transfer function $H_L(f)$. In terms of the phases we obtain the nonlinear circuit in Fig. 22.

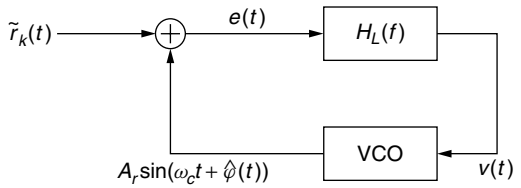


Figure 21. A phase-locked loop.

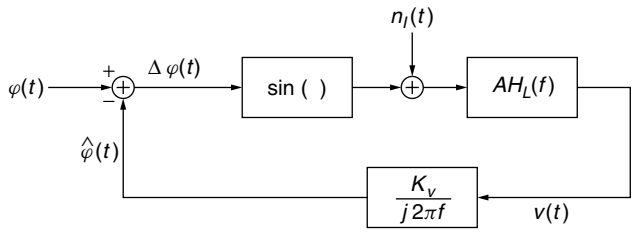


Figure 22. Equivalent circuit of PLL.

The nonlinearity is caused by the term $\sin(\Delta\varphi)$ where

$$\Delta\varphi(t) = \varphi(t) - \hat{\varphi}(t) \tag{71}$$

Note that as long as $|\Delta\varphi(t)| \le \pi/2$, when $\varphi(t)$ increases so does $\Delta\varphi(t)$ and the control signal $v(t)$, hence also $\hat{\varphi}(t)$ which thus tracks $\varphi(t)$. When $\Delta\varphi(t) = 0$, the PLL is locked and $\hat{\varphi}(t) = \varphi(t)$. For small values of $\Delta\varphi(t)$, $\sin(\Delta\varphi) \approx \Delta\varphi$; hence we obtain the linear PLL shown in Fig. 23.

If the input signal has a carrier frequency offset Δf_c, the input phase

$$\varphi(t) = 2\pi f_c t + \varphi(0) \tag{72}$$

may become very large and the PLL may not be able to lock. Therefore during the acquisition stage the frequency of the VCO is swept over a large range until it is locked to $f_c + \Delta f_c$. After locking, the PLL tracks the slow variations in the input phase. There are several methods of sweeping or equivalent operations to obtain frequency locking. These include (1) two switched loop filters, a wideband for acquisition and a narrowband for tracking; (2) nonlinear loop filter with greater sensitivity near lock; and (3) a frequency detector in parallel with a phase detector that is switched off after locking. An example of (3) can be found in Ref. 13.

For the linear PLL in Fig. 24 we can write the equation

$$\hat{\varphi}(t) = [\varphi(t) - \hat{\varphi}(t) + n(t)] * h(t) \tag{73}$$

where

$$H(f) = \frac{KH_L(f)}{j2\pi f} \quad K = A_LK_L \tag{74}$$

is the open-loop transfer function of the PLL. The closed loop transfer function of the PLL is [setting $n(t) = 0$]

$$G_L(f) = \frac{\hat{\Phi}(f)}{\Phi(f)} = \frac{H(f)}{1 + H(f)} = \frac{KH_L(f)}{j2\pi f + KH_L(f)} \tag{75}$$

The noise term that affects $\hat{\varphi}(t)$ has a PSD of $(N_0/A^2)|G_L(f)|^2$ and a variance of

$$\sigma_{\hat{\varphi}}^2 = \frac{2N_0}{A^2}B_L, \quad B_L = \int_0^\infty |G_L(f)|^2\,df \tag{76}$$

where $2B_L$ is the noise bandwidth of the PLL. The signal to noise ratio of the input signal of (68) when taken within the bandwidth $2\tilde{B}_L = 4B_L$ is

$$\text{SNR} = \frac{0.5A^2}{0.5N_0 2\tilde{B}_L} = \frac{0.5A^2}{2N_0B_L} \tag{77}$$

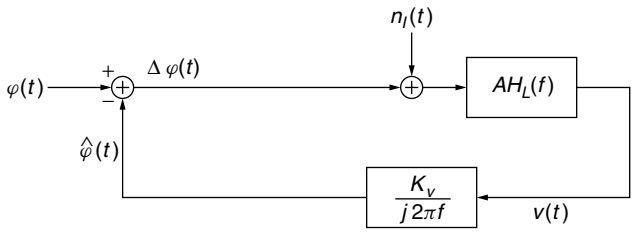


Figure 23. Linear approximation of PLL.

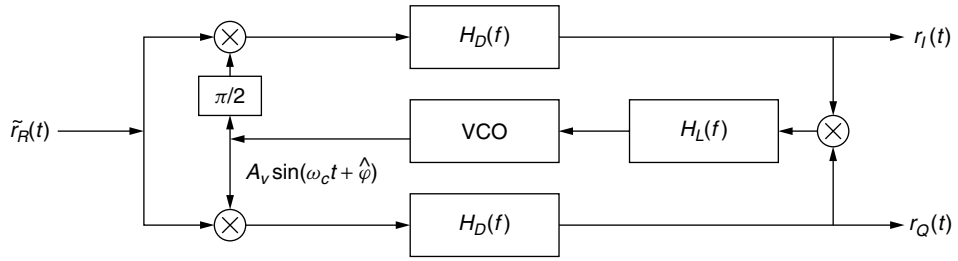


Figure 24. Costas loop.

therefore

$$\sigma_{\hat{\varphi}}^2 = \frac{1}{2\text{SNR}} \tag{78}$$

The variance in (78) is also the variance of the phase $\Delta\varphi$. The nonlinear PLL has been analyzed for the case $H_L(f) = 1$ in Ref. 16, and it has been found that the resulting variance is greater than that of the linear PLL. An alternative to the PLL is the Costas loop, shown in Fig. 24.

For the input in (68), the output of the multipliers are

$$e_I(t) = A_L[\cos(\Delta\varphi) + n_I(t)], \quad e_Q(t) = A_L[\sin(\Delta\varphi) + n_Q(t)] \tag{79}$$

with additional high-frequency terms that are eliminated by the filters. Thus the control voltage (assuming the filters do not change) follows from (79) as

$$v(t) = 0.5K_v A_L^2 \sin\ (2\Delta\varphi) + \text{noise terms} \tag{80}$$

and again $\hat{\varphi}(t)$ will track $\varphi(t)$.

5.2. ML Carrier Recovery with Pilot Symbols or Decision Directed

Here we assume that K symbols $\{a_i\}$ are known. The received signal after the detector filters is

$$r(t) = A\sum_i a_i g(t - iT)e^{j\varphi} + n(t) \tag{81}$$

Taking samples and assuming no ISI, we have

$$r_k = A_0 a_k e^{j\varphi} + n_k, \quad k = 1, 2, \ldots, K, \quad A_0 = Ag_0 \tag{82}$$

The ML estimate of the phase is the minimum of

$$\Lambda_L(\varphi) = \sum_{k=1}^{K} |r_k - A_0 a_k e^{j\varphi}|^2 \tag{83}$$

or equivalently the maximum of

$$\Lambda_L(\varphi) = \text{Re}\left\{\sum_{k=1}^{K} r_k^* a_k e^{j\varphi}\right\}$$

$$= z_I \cos\varphi - z_Q \sin\varphi, \quad z = \sum_{k=1}^{K} r_k^* a_k \tag{84}$$

The maximum is achieved when the derivative is zero, and the solution is

$$\hat{\varphi} = -\tan^{-1}\frac{z_Q}{z_I} \tag{85}$$

It can be shown that $\overline{\hat{\varphi}} = \varphi$; hence the estimate is unbiased. Since z is Gaussian, we can also compute the PDF and variance of $\hat{\varphi}$, which is also $(2KE/N_0)^{-1}$. The block diagram of the QAM system with pilot symbols is shown in Fig. 25, which is very similar to the Costas loop. In the tracking stage we replace the pilot symbol with the estimated symbols $\{\hat{a}_k\}$ which is also shown in the figure.

5.3. Blind Carrier Recovery of QAM Using PLL

We have to create a pilot tone from the QAM signal

$$\tilde{r}_R(t) = A_I(t)\cos(w_c t + \varphi) - A_Q(t)\sin(w_c t + \varphi) + \tilde{n}_R(t) \tag{86}$$

where [assuming for simplicity that $g_{CR}(t)$ is real]

$$A_x(t) = A\sum_i a_{xi} g_{CR}(t - iT), \quad x = I, Q \tag{87}$$

For a symmetric constellation we have

$$\overline{a_{Ii}^p} = \overline{a_{Qi}^p}, \quad \overline{a_{Ii}a_{Qi}} = \overline{a_{Ii}a_{Qi}}, \quad \overline{a_{Ii}} = \overline{a_{Qi}} = 0 \tag{88}$$

therefore

$$\overline{A_I^p(t)} = \overline{A_Q^p(t)}, \quad \overline{A_I(t)A_Q(t)} = \overline{A_I(t)A_Q(t)} = 0 \tag{89}$$

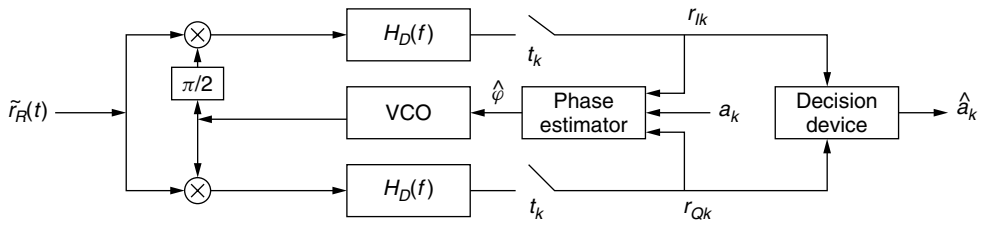


Figure 25. QAM with pilot symbols or decision directed.

The minimum value of p for which we have a pilot tone is 4, namely, $\tilde{s}_R^4(t)$. The pilot tone component is

$$A(t)\cos(4w_c t + 4\varphi(t)), \quad A(t) = 0.25[\overline{A_I^4(t)} - 3(\overline{A_I^2(t)})^2] \tag{90}$$

This component is tracked by the PLL, however to obtain the original carrier we need a frequency divider by 4. The corresponding PLL for QAM is shown in Fig. 26.

Because only a small part of $\tilde{s}_R^4(t)$ contains energy at frequency $4f_c$ and the additional noise terms in the vicinity of this frequency, the variance of the estimated phase is

$$\sigma_{\hat{\varphi}}^2 = \frac{S_L}{2\text{SNR}}, \quad S_L \le 1 + \frac{9}{\text{SNR}} + \frac{6}{\text{SNR}^2} + \frac{1.5}{\text{SNR}^3} \tag{91}$$

where S_L is called a *power loss*. The estimate of the phase can also be computed [14] from discrete samples as in Fig. 25:

$$\hat{\varphi} = 0.25 \arg\left[\overline{a_k^4} \sum_{k=1}^{K} r_k^4\right] \tag{92}$$

with a variance which can be approximated by

$$\sigma_{\hat{\varphi}}^2 = \left(\frac{c_1}{2\gamma} + c_2\right)\frac{1}{K}, \quad \gamma = \frac{E}{N_0} \tag{93}$$

The values of c_1 and c_2 can be computed and are shown in Table 1 for various QAM square and cross-constellations. Note that there is a self-noise represented by c_2 and that both c_1 and c_2 are large for cross-constellations. A reduced constellation power law algorithm [14] in which only r_k with amplitudes that exceeds a certain threshold gives a better estimate.

Note that if $2n\pi$ is added to $4\hat{\varphi}(t)$, the VCO output is unchanged however after the frequency divider $\hat{\varphi}(t)$ has an uncertainty of $n\pi/2$. To resolve this uncertainty, we need again pilot symbols or we have to use differential phase modulation, in which the phase of the symbols in each quarter of the constellation is transmitted as a phase difference instead of an absolute phase. The first 2 bits of each symbol can be represented by this phase.

5.4. Blind ML Carrier Recovery for QAM

The blind ML estimate of the phase maximizes

$$\Lambda_L(\varphi) = \exp -\overline{\frac{1}{2\sigma_n^2}\sum_{k=1}^{K}\left|r_k - A_0 a_k e^{j\varphi}\right|^2}$$

$$= \prod_{k=1}^{K} \exp -\overline{\frac{1}{2\sigma_n^2}\left|r_k - A_0 a_k e^{j\varphi}\right|^2} \tag{94}$$

where the average is over the symbols. For the square constellations, this is equivalent to maximizing

$$\Lambda_L(\varphi) = \sum_{k=1}^{K} \ln\left\{\sum_{i=1}^{\sqrt{M}-1} \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right) \times \cosh\frac{r_{Ik}\cos\varphi(2i-1)A_0}{\sigma_n^2}\right\}$$

$$+ \sum_{k=1}^{K} \ln\left\{\sum_{i=1}^{\sqrt{M}-1} \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right) \times \cosh\frac{r_{Qk}\sin\varphi(2i-1)A_0}{\sigma_n^2}\right\} \tag{95}$$

The Cramer–Rao lower bound (CELB) [15] to any estimate is

$$\sigma_{\hat{\varphi}}^2 \ge \text{CRLB}(\hat{\varphi}) = -\left(\frac{d^2\Lambda_L(\varphi)}{d\varphi^2}\right)^{-1} \tag{96}$$

The CRLB of phase and also of frequency has been derived for QAM [15]. The results are

$$\text{CRLB}(\hat{\varphi}) = \left[2K\gamma F\left(\frac{1}{2\gamma}\right)\right],$$

$$\text{CRLB}(\hat{f}_c) = \left[2K\gamma\frac{K^2-1}{12}F\left(\frac{1}{2\gamma}\right)\right]^{-1} \tag{97}$$

where $F(x)$ is an integral that is evaluated numerically. The values of CRLB $(\hat{\varphi})$ and of $\sigma_{\hat{\varphi}}^2$ for several

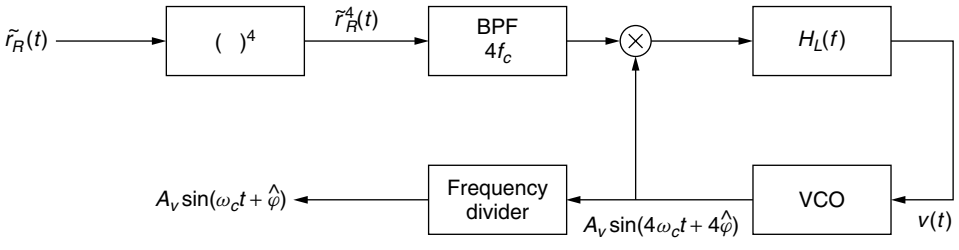


Figure 26. PLL with 4th power signal.

Table 1. Values of Constants Needed in Computation of Phase Error Variance for QAM

M	4	16	32	64	128	256	512	1024	2048	4096
c_1	1	4.24	52.63	5.81	62.07	6.27	64.68	6.39	63.39	6.42
c_2	0	0.06	3.14	0.17	3.79	0.20	3.98	0.21	4.03	0.21

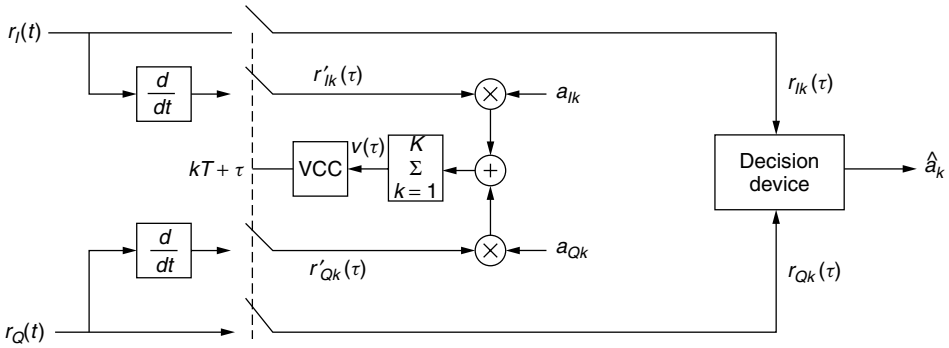


Figure 27. Standard deviation of phase, $\sigma_{\hat{\phi}}$, for 64-QAM and $K = 200$ symbols. (*Source*: F. Riou, B. Cowley, B. Morgan, and M. Rice, Cramer–Rao lower bounds for QAM phase and frequency estimation, *IEEE Trans. Commun.* **COM-49**: 1582–1591, © 2001 IEEE.)

practical systems [histogram algorithm (HA), two-stage conjugate algorithm (2SC), MDE-minimum-distance algorithm (MDA), and power-law estimate (PLE) ($P = 4$)] are shown in Fig. 27 (which is Fig. 8 of Ref. 15) as a function of γ for 64-QAM and $K = 200$. The CWCRLB is $(2K\gamma)^{-1}$. CRLB $(\hat{f}_c)$ is presented in Fig. 28 (which is Fig. 3 of Ref. 15) for several QAM constellations. More figures on this matter can be found in Refs. 14 and 15.

5.5. Time Recovery with Pilot Symbols or Decision Directed

The ML estimator of τ maximizes

$$\Lambda_L(\tau) = \mathrm{Re}\left\{\sum_{k=1}^{K} a_{ki}^{*} r_k(\tau)\right\} = \sum_{k=1}^{K} r_{Ik}(\tau)a_{Ik} + r_{Qk}(\tau)a_{Qk}, r_k(\tau) = r(kT+\tau) \tag{98}$$

Taking the derivative, we have

$$v(\tau) = \frac{d\Lambda_L(\tau)}{d\tau} = \sum_{k=1}^{K} r'_{Ik}(\tau)a_{Ik} + r'_{Qk}(\tau)a_{Qk}, r'_{Xk}(\tau) = \left.\frac{dr(t)}{dt}\right|_{t=kT+\tau} \tag{99}$$

The maximum is achieved when $v(\hat{\tau}) = 0$. The implementation of this equation is shown in Fig. 29. The VCC (voltage controlled clock) is a VCO that produces rectangular pulses instead of a sinusoid. The summer is the digital equivalent of an integrator and forms the loop filter.

In the tracking stage estimated symbols are reliable and can be used instead of the pilot symbols. This is the decision directed mode of the clock recovery.

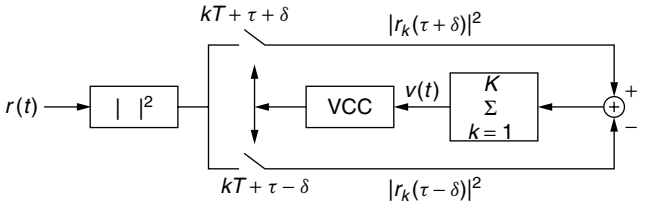


Figure 28. Standard deviation of frequency, $\sigma_{\hat{f}_c}$, for $K = 200$ symbols. (*Source*: F. Riou, B. Cowley, B. Morgan, and M. Rice, Cramer–Rao lower bounds for QAM phase and frequency estimation, *IEEE Trans. Commun.* **COM-49**: 1582–1591, © 2001 IEEE.)

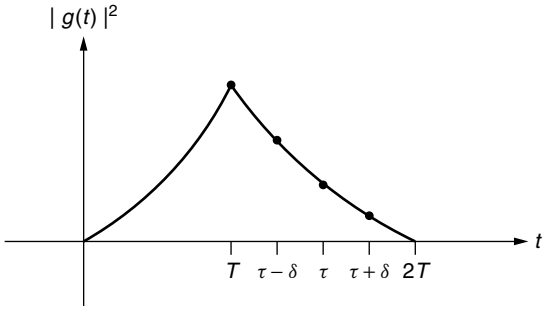


Figure 29. ML clock recovery with pilot symbols or decision directed.

5.6. Blind ML Clock Recovery

Here we maximize an equation similar to (94)

$$\Lambda(\tau) = \exp -\overline{\frac{1}{2\sigma_n^2}\sum_{k=1}^{K}|r_k(\tau) - A_0 a_k|^2} = \prod_{k=1}^{K}\exp -\overline{\frac{1}{2\sigma_n^2}|r_k(\tau) - A_0 a_k|^2} \tag{100}$$

where the average is over the symbols. For QAM with square constellation, this is equivalent to maximizing

$$\Lambda_L(\tau) = \sum_{lk=1}^{K}\ln\left\{\exp -\frac{|r_k(\tau)|^2}{2\sigma_n^2}\sum_{i=1}^{\sqrt{M}-1} \times \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right)\cosh\frac{r_{Ik}(\tau)(2i-1)A_0}{\sigma_n^2}\right\} + \sum_{k=1}^{K}\ln\left\{\exp\left(-\frac{|r_k(\tau)|^2}{2\sigma_n^2}\right)\sum_{i=1}^{\sqrt{M}-1} \times \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right)\cosh\frac{r_{Qk}(\tau)(2i-1)A_0}{\sigma_n^2}\right\} \tag{101}$$

The optimal $\tau$is the solution of

$$\frac{d\Lambda_L(\tau)}{d\tau} = 0\ |_{\tau=\hat{\tau}} \tag{102}$$

An approximation to the ML synchronizer is the E-L (early–late) gate synchronizer, which computes

$$\Delta\Lambda(\tau) = \frac{\Lambda_L(\tau+\delta) - \Lambda_L(\tau-\delta)}{2\delta}, \quad \Lambda_L(\tau) \approx \sum_{k=1}^{K} |r_k(\tau)|^2 \tag{103}$$

A block diagram of this scheme is shown in Fig. 30.

The E-L algorithm is based on the idea that the pulse at the receiver has a peak at the correct sampling time and is symmetric around this point as shown in Fig. 31. Note that the correct sampling time is T and $\Delta\Lambda(T) = 0$ while $\Delta\Lambda(\tau) > 0$ if $\tau < T$ and $\Delta\Lambda(\tau) < 0$ if $\tau > T$. The E-L has to be modified for QAM because certain sequences of symbols will result in false values of $\Delta\Lambda(\tau)$ (see Ref. 1 for elaborations).

A simple blind recovery of the symbol rate is based on the autocorrelation of the signal

$$\begin{aligned} R_s(\tau) &= 0.5\overline{s(t)s^*(t-\tau)} \\ &= 0.5A^2\sigma_a^2 \sum_i g(t-iT)g(t-iT-\tau) \end{aligned} \tag{104}$$

which is a periodic function of t with period T. Therefore a bandpass filter with center frequency $R = 1/T$ or nR can produce a clock frequency directly or after frequency division by n. This circuit is implemented in Fig. 32. The best results are obtained for $\tau = T/2$.

6. SUMMARY

We have described a QAM system and computed the error probability in both Gaussian and fading channels with and without diversity. We presented various methods of carrier and clock recovery. Many topics such as equalization, effect of errors in carrier and clock recovery on the error probability, the error probability of differential phase detection, combined differential phase and differential amplitude for mobile communication, joint phase, frequency and time synchronization, and QAM in spread-spectrum systems have been omitted for lack of space.

BIOGRAPHIES

John P. Fonseka received B.Sc.(Hons.) degree in Electronic and Telecommunication Engineering from University of Moratuwa, Sri Lanka in 1980, M.Eng. in Electrical

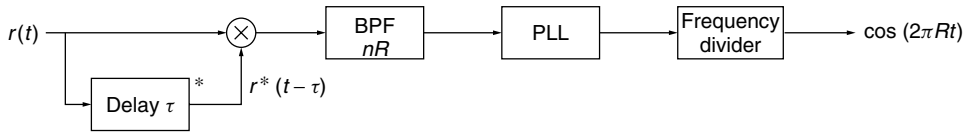


Figure 30. Early–late gate clock recovery.

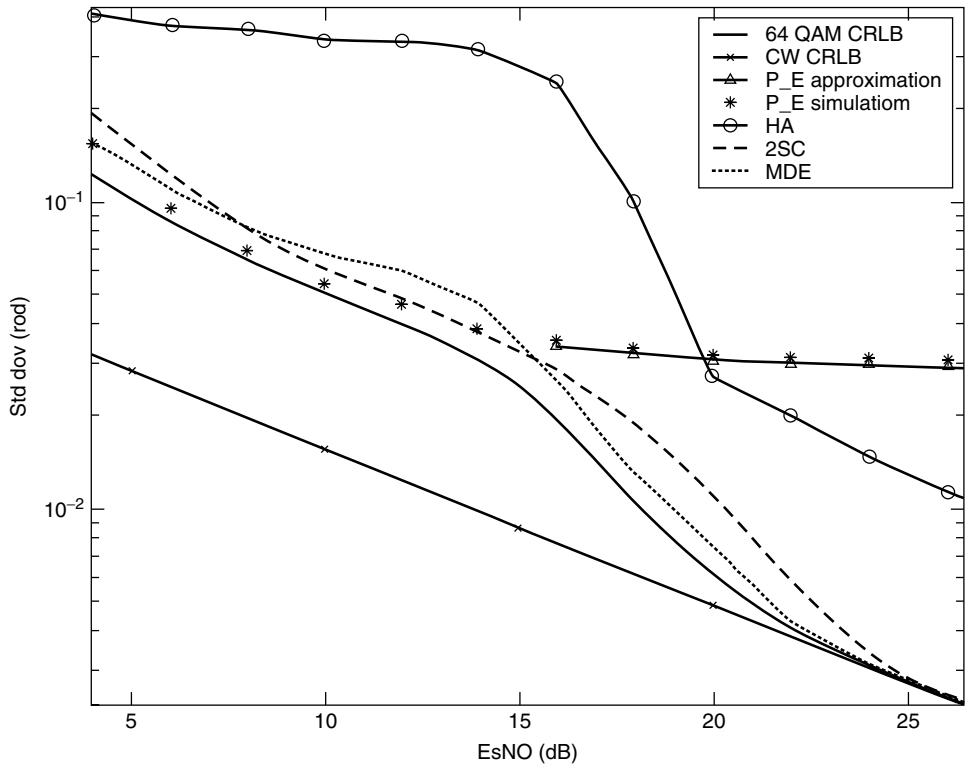


Figure 31. Pulse for early–late clock recovery.

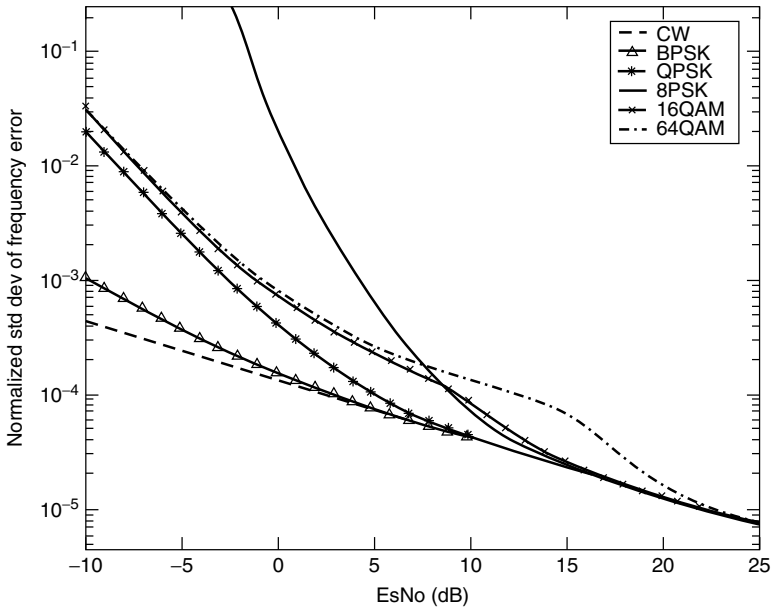


Figure 32. Blind recovery of symbol rate.

Engineering from Memorial University of Newfoundland, Canada in 1985, and Ph.D. in Electrical Engineering from Arizona State University in 1988.

He joined University of Texas at Dallas in August of 1988 as an Assistant Professor, where he is currently serving as a Professor in Electrical Engineering. His research interests include combined coded modulation, signaling through narrowband fading channels, coding theory, telecommunication networks, and optical communications.

Israel Korn received his BSc, MSc, and DSc degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1962, 1964 and 1968 respectively. Since 1978 he has been with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, New South Wales, Australia, where he is now a Visiting Professor or Honorary Visiting Fellow. He was a Visiting Professor at various universities and institutions in the USA, Germany, Denmark Spain and Australia. His research and teaching interests are in the area of Digital Communication with applications to mobile, wireless and personal communications. He has published 87 papers in refereed journals, 45 papers in conference proceedings and two books one of which is, "Digital Communications," Van Nostrand, NY 1985. He has taught various undergraduate and postgraduate course in the general area of Communications at various universities. He was an editor of the IEEE Transactions on Communications (1992–1995) and of Wireless Personal Communications (1992–). Since 1994 he is a Fellow of IEEE.

BIBLIOGRAPHY

1. W. T. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation*, IEEE Press, New York, Pentech Press, London, 1994.
2. I. Korn, *Digital Communication Systems*, Van Nostrand Reinhold, New York, 1985.
3. K. Feher, *Advanced Digital Communication and Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
4. R. E. Ziemer and R. L. Peterson, *Introduction to Digital Communications*, Macmillan, New York, 1992.
5. M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques. Signal. Design and Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
6. M. K. Simon and M. S. Alouini, *Digital Communication over Fading Channels*, Wiley, New York, 2000.
7. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, Boston, 2001.
8. G. J. Foscini, R. D. Gitlin, and S. B. Weinstein, Optimisation of the two-dimensional signal constellations in the presence of Gaussian noise, *IEEE Trans. Commun.* **COM-22**: 28–38 (1974).
9. C. M. Thomas, M. Y. Weidner, and S. H. Durrani, Digital amplitude-phase keying with *M*-ary alphabets, *IEEE Trans. Commun.* **COM-22**: 168–180 (1974).
10. J. Lu, K. B. Letayef, J. C.-I. Cuang, and M. L. Liou, *M*-PSK and *M*-QAM BER computation using signal-space concepts, *IEEE Trans. Commun.* **COM-47**: 181–184 (1999).
11. H. Meyr, M. Moenclaey, and S. A. Fechtel, *Digital Communication Receivers*, Wiley, New York, 1998.
12. U. Mengali and D. D'Andrea, *Synchronization Techniques for Digital Receivers*, Plenum Press, New York, 1997.
13. H. Sari and S. Moridi, New phase and frequency detectors for carrier recovery in PSK and QAM systems, *IEEE Trans. Commun.* **COM-36**: 1035–1043 (1988).
14. C. N. Georghiades, Blind carrier phase acquisition for QAM constellation, *IEEE Trans. Commun.* **COM-45**: 1477–1486 (1997).
15. F. Riou, B. Cowley, B. Moran, and M. Rice, Cramer-Rao lower bounds for QAM phase and frequency estimation, *IEEE Trans. Commun.* **COM-49**: 1582–1591 (2001).
16. A. J. Viterbi, *Principles of Coherent Communications*, McGraw-Hill, New York, 1966.

R

RADIO PROPAGATION AT LF, MF, AND HF

JOHN C. H. WANG
Federal Communications Commission
Washington, District of Columbia

1. INTRODUCTION

Electromagnetic (or radio) waves can be transmitted from a transmitting site to a receiving site by a number of different mechanisms. At the frequencies under study (between about 30 kHz and 30 MHz), the most important ones are ground wave and sky wave. At higher frequencies, the space (or tropospheric) wave becomes more important. The ground (or surface) wave exists when the transmitting and receiving antenna are on or near the surface of the earth. Standard broadcast signals received in daytime are all ground waves.

The sky wave represents energy that travels from the transmitting antenna to the receiving antenna as a result of a "bending" by the earth's upper atmosphere called the *ionosphere*. The ionosphere, which consists of several different layers, begins about 50 km above the earth's surface. Shortwave signals and nighttime medium-wave signals are examples of sky waves. Under certain conditions, the ground-wave and sky-wave components from the same source may be comparable in amplitude but arrive at slightly different times, resulting in interference.

The space (or tropospheric) wave represents energy that travels from the transmitting antenna to the receiving antenna in the earth's troposphere. The troposphere is the lower portion of the earth's atmosphere in which the temperature decreases with increasing altitude. This part of the atmosphere extends to an altitude of about 9 km at the earth's poles and 17 km at the equator. TV and FM signals are examples of space waves. Space wave propagation is beyond the scope of this article.

In the subsequent sections factors affecting propagation will be described. Methods of predicting field strength will also be analyzed and compared. Definitions of the most frequently used terms are given in Section 6.

2. GROUND-WAVE PROPAGATION

2.1. Early Pioneering Studies

At frequencies between about 10 kHz and 30 MHz, ground-wave propagation is possible because the surface of the earth is a conductor. The ground wave is vertically polarized. Any horizontal component of an electric field on the surface of the earth is shorted-circuited by the earth. The earliest work on ground-wave propagation was carried out by Summerfield [1]. His flat-earth theory states that ground-wave field strength, E_g, can be expressed in the form

$$E_g = A\frac{E_0}{d} \tag{1}$$

where E_0 = field strength of wave at the surface of the earth at unit distance from the transmitting antenna, neglecting earth's losses
d = distance to transmitting antenna
A = factor taking into account the ground losses

The field strength E_0 at unit distance in Eq. (1) depends on the power radiated by the transmitting antenna and the directivity of the antenna in the vertical and horizontal planes. If the radiated power is 1 kW and the short vertical antenna is omni-directional in the horizontal plane, then, $E_0 = 300$ mV/m when the distance is 1 km. The reduction factor A is a complicated function of electrical constants of the earth, frequency, and the distance to the transmitter in wavelengths. The reduction is highly frequency-dependent; it increases with increasing frequency. Thus, at LF and MF ground-wave signals can be sufficiently strong for broadcasting service. On the other hand, at HF, ground-wave signals are usually too weak for broadcasting purposes. The Summerfield flat-earth approach, the subsequent Watson transformation [2], and the Bremmer residue series [3] were the important milestones and theoretical advances on which the modern ground-wave theory is still based.

2.2. The Development of Ground-Wave Curves

Intensive efforts to convert the theoretical advances to simple and practical field strength curves took place between 1930 and 1940 [4]. Extensive measurement programs were conducted by many organizations including the Federal Communications Commission. In 1939, the FCC released a complete set of ground-wave curves as an appendix to the Standards for Good Engineering Practice Concerning Standard Broadcasting Stations [5]. These curves and a comprehensive discussion were included in a paper by Norton [6]. Similar but not identical ground-wave curves can also be found in ITU-R Recommendation PN.368-7 [7]. The current FCC curves cover the frequency range of 535–1705 kHz. The ITU-R curves cover a much wider range of frequency, from 10 kHz to 30 MHz. [*Note*: ITU-R, which appears frequently in this article, is the abbreviated name of the *Radiocommunication Study Groups of the International Telecommunication Union*, formerly known as the CCIR (see Section 6). PN denotes propagation in nonionized media. PI, which will appear in the subsequent sections, denotes propagation in ionized media. Numeral 7 after a dash means it is the 7th revised edition. In 2000, ITU-R decided to use the prefix P to replace both PN and PI.]

2.3. Available Software

Currently, there are three computer programs available for calculating ground-wave field strengths. The first

program is called *ITSGW* [8]. The second program, GRWAVE [9], is in ITU-R Recommendation PN.368-7. This program, which takes into account the effects of refraction in an exponential atmosphere, is available from ITU Sales Service, Place des Nations, 1211 Geneva, Switzerland. The third program is called *FCCGW* [10]. FCCGW has been used to generate the metric version of the FCC curves. The FCC curves take into account the effect of refraction by using an effective radius that is $\frac{4}{3}$ times the actual radius of the earth. Refraction is insignificant at distances less than about 100 km. At greater distances, it becomes progressively more significant. It should also be mentioned that GRWAVE is designed for personal computers while FCCGW is designed for mainframe computers. A comparison of these programs has been done [10,11]. It is reported that the three methods give ground-wave field strength predictions sufficiently close in value that they could be considered identical for frequency management purposes.

2.4. Ground-Wave Propagation over Inhomogeneous Terrain

For predicting ground-wave field strengths over paths composed of successive sections of terrain of different conductivities (e.g., land and sea), there are two basic methods available. These are the equivalent-distance (or Kirke) method [12] and the equivalent-field (or Millington) method [13]. The Kirke method has the advantage of simplicity, but in cases where the successive sections show considerable differences in conductivities, it can lead to large error. On the other hand, the Millington method does not suffer from this problem. Furthermore, the Millington method is now no longer as difficult to apply as before, because a simplified graphical solution has been developed by Stokke [14]. The Millington method and the Stokke approximation are presented in ITU-R Recommendation PN.368-7.

2.5. Ground Conductivity

Ground-wave propagation can be considered a reasonably well understood topic. In one area, however, more work is needed. Ground conductivity is a very important factor in calculating ground-wave field strengths. Accurately measured data should be used. Although several maps are available, they present estimates and are not very accurate. A map showing the estimated ground conductivities of the continental United States has been published by the FCC [15]. An atlas of ground conductivities in different parts of the world can be found in ITU-R Recommendation PN.832-2 [16].

Conductivity of seawater is typically 5 siemens per meter (5 S/m) while that of freshwater is about 10 mS/m. Conductivities of rocky land, hills, and mountains vary between 1 and 2 mS/m. Conductivity of rich agricultural land is typically 10 mS/m. Cities and residential areas have a conductivity of about 2 mS/m. In industrial areas, it is even less.

3. THE SKY-WAVE PROPAGATION MECHANISM

3.1. The Solar–Terrestrial System

In 1901, Guglielmo Marconi (1874–1937), a young Italian engineer, succeeded in sending a Morse code message from Cornwall, England, across the Atlantic Ocean to Newfoundland. It is generally believed that the frequency Marconi used was about 1.6 MHz. This revolutionary wireless experiment not only brought him a Nobel Prize later in 1909 but also created a new frontier in the scientific world. Perhaps Oliver Heaviside, an English physicist, gave the earliest satisfactory explanation of his experiment. He theorized that in the earth's upper atmosphere, there is a sufficiently conducting layer [17]. This conducting layer is known as the *ionosphere*, so called because it consists of heavily ionized molecules. To understand sky-wave propagation, it is essential to study the entire solar–terrestrial system, not just the ionosphere alone. In this article, we shall discuss this subject only briefly. For more details, see books by Davies [18] and by Goodman [19]. It should be mentioned that the discussion in this section applies to LF, MF, and HF (low, medium, and high frequency).

3.2. The Ionosphere

The ionosphere consists of three regions (or layers). They are the D, E, and F regions, respectively, in increasing order of altitude.

The D region spans the approximate altitude range of 50–90 km. It exists only in daytime and disappears shortly after sunset. The D region can be considered as an absorber, causing significant signal attenuation. The absorption is frequency-dependent; it decreases with increasing frequency.

The E region spans the approximate altitude range of 90–130 kilometers. This region is important for nighttime low- and medium-frequency propagation at distances greater than about 200 km. The E region exhibits a solar cycle dependence with maximum electron density occurring at solar maximum.

Sporadic E (Es), which has very little relationship with solar radiation, is an anomalous form of ionization embedded within the E region. It can have significant effects on propagation at HF and VHF.

The F region extends upward from about 130 kilometers to about 600 kilometers. The lower and upper portions of the F region display different behaviors at daytime, resulting in a further subdivision into F1 and F2 layers. The F1 layer is the region between 130 and 200 km above the surface of the earth. The F2 layer, which is the highest and the most prominent one, is the principal reflecting region for long-distance high-frequency communications. At night, the F1 layer merges with the F2 layer and the average height of the combined layer (still called the "F2 layer") is about 350 km.

3.3. Solar Activity

The existence of the ionosphere is a direct result of the radiation from the sun, both electromagnetic and corpuscular. The electromagnetic radiation travels toward

the earth at the speed of light, and the entire journey takes about 8.3 min. The ionization process is linked with the intensity of the solar radiation, which in turn varies with factors such as time of day, latitude, and season. Solar activity changes drastically from time to time. Sunspot number is a reasonably good index of the state of solar activity, although several other indices are also available. Sunspots are dark areas on the surface of the sun. Sunspots were, perhaps, first observed by the Chinese in March of 20 A.D. [20] during the Han Dynasty (206 B.C.–220 A.D.). Sunspots appear dark because the temperature is low, only about 3000 K, while the average temperature of the sun is about 6000 K. Sunspots tend to group together and display an 11-year cyclic nature. The astronomic records of the Jin Dynasty (265–418 A.D.) of China [21] indicate that for quite a while in the fourth century, sunspots were observed every 11 years (e.g., 359, 370, 381, and 393 A.D.). Sunspot numbers vary from day to day and year to year. Routine observations have been made since 1749. The cycle beginning in 1755, a year of minimum sunspot number, is considered cycle 1. The ascending portion of a cycle (on average, 4.5 years) is usually much shorter than the descending one (6.5 years). The Zurich (or Wolf) number R is given by

$$R = k(10\,g + s) \tag{2}$$

where g is the number of sunspot groups, s is the number of observed small spots, and k is a correction factor, approximately unity, used to equalize the results from different observations and equipment. The sunspot number is subject to wide variations from month to month and is of little usefulness. Furthermore, it is known that the characteristics of the ionosphere do not follow the short-term variations. In order to achieve a better correlation, some kind of "smoothing" technique is desirable. Consequently, the 12-month smoothed sunspot number (R_{12}) has been adopted and is the most widely used index in ionospheric work today:

$$R_{12} = \frac{1}{12}\left[(0.5)(R_{n+6} + R_{n-6} + \sum_{n-5}^{n+5}(R_n)\right] \tag{3}$$

Thus, by definition, the value of R_{12} is known only 7 months after the recorded observation. R_{12} varies from a minimum of about 10 to a maximum generally of 100–150, although in December 1957 it reached a record high of 239.4.

3.4. Atmospheric Radio Noise

In order to estimate the performance to be expected in a communication system, it is insufficient to consider field strength alone. Equally important is radio noise. There are many different sources of noise: the atmosphere, the receiving system, human activity, the sun, and galaxies. In this article, we emphasize atmospheric noise. For an excellent discussion on all types of noise, see ITU-R Recommendation PI.372-6 [22].

Atmospheric noise is produced mainly by lightning discharges in thunderstorms. Discharges take place between 2 and 4 km above ground. The power released is very great, typically greater than 10 GW [23]. Atmospheric radio noise obeys the same propagation laws as do sky-wave signals. Thus, it travels to distances several thousands of kilometers away. The noise level thus depends on the time of day, season of the year, weather, geographic location, and frequency. In general, atmospheric noise is the highest (1) when the receiver is located near a thunderstorm center, (2) during local summer, (3) during the night, or (4) when the frequency is low. There are three major thunderstorm (hence, noise) centers in the world: the Caribbean, Equatorial Africa, and Southeast Asia. Maps showing the atmospheric noise levels for different parts of the world corresponding to different seasons of the year and different hours of the day have been developed by the CCIR since 1964. The most recent maps can be found in ITU-R Recommendation PI.372-6 [22].

3.5. Magnetic Coordinates

There are several definitions of latitude connected with the geomagnetic field. The centered dipole latitude, or simply the dipole latitude, is an approximation and has been used for ionospheric work for decades. Corrected geomagnetic latitude more accurately represents the real geomagnetic field and should be used when accuracy is desired. Conversion tables from geographical coordinates to corrected latitude are readily available [24].

4. SKY-WAVE PROPAGATION AT LF AND MF

4.1. Results of Six Decades of Worldwide Efforts

This section presents a brief summary of the historical background behind the development of the four currently used LF and MF sky-wave propagation models. For a more detailed presentation, see a paper by Wang [25].

4.1.1. The Cairo Curves. The earliest worldwide efforts to study LF/MF sky-wave propagation began in 1932. At its meeting held in Madrid, the CCIR established a committee, chaired by Balthasar van der Pol of Holland, to study propagation at frequencies between 150 and 2000 kHz. Measurements were made on 23 propagation paths between 1934 and 1937. Consequently, two sky-wave propagation curves were developed. One of the curves is for paths far away from the earth's magnetic poles; this is better known as the *north–south curve*. The other curve is for paths approaching the earth's magnetic poles and is better known as the *east–west curve*. Actually, the former should have been called the *low-latitude curve*; the latter, the *high-latitude curve*. The two curves were formally adopted by the CCIR at the 1938 International Radio Conference, Cairo. Therefore, these curves are known as the *Cairo curves*. In1974 the LF/MF Conference adopted the north–south curve for use in the Asian part of Region 3 [26].

4.1.2. The Region 2 Method (the FCC Clear-Channel Curve). Recognizing the needs for a set of sound engineering standards, the FCC, under the leadership of the late Ken Norton, carried out a sky-wave field

strength measurement program in the spring of 1935, a year of moderate sunspot number. Nighttime signals of all eight clear-channel stations were monitored at 11 widely scattered receiving sites. From these measurements, the FCC clear-channel curve was derived. The 1980 Broadcasting Conference for Region 2 adopted this method for applications in Region 2 [27]. Hence, this method is also known as the Region 2 method. The staff of the FCC has since developed a newer and more accurate method. The newer method, which is being used by the FCC for domestic applications, will be discussed in the subsequent sections.

4.1.3. Comparison of the Two Graphical Methods. Both the Cairo and the FCC curves present field strength as a function of distance only. They do not take into account effects of latitude, sunspot number, and so on. When converted to the same conditions, the two Cairo curves and the FCC clear-channel curve are similar for distances up to about 1400 km. At 3000 km, the north–south curve is about 8 dB greater than the east–west curve; at 5000 km, the difference is about 18 dB. The FCC clear-channel curve falls between the two Cairo curves. The Cairo north–south curve offers good results when applied to low latitudes. When applied to higher latitudes, it tends to overpredict field strength levels [25]. The FCC curve offers reasonable results when applied to temperate latitudes. Neither of these methods is a true worldwide method. The Cairo east–west curve, because it often underestimates field strength levels, has virtually been disregarded.

4.1.4. The Development of Recommendation 435 (the Udaltsov–Shlyuger Method). Recognizing the need for a simple but accurate field strength prediction method for worldwide applications, and in anticipation of a broadcasting conference, the CCIR at its Xth Plenary Assembly (Geneva, 1963) established International Working Party (IWP) 6/4 to undertake such a task. This IWP was first chaired by J. Dixon (Australia), succeeded by G. Millington, P. Knight (UK), and J. Wang (USA). In the late 1960s and early 1970s a number of administrations and scientific organization were very active and collected valuable data. For example, the European Broadcasting Union (EBU), which started its sky-wave studies soon after World War II, reactivated its efforts. Its counterpart in Eastern Europe, the International Organization of Radio and Television (OIRT) was also active. The administration of the former USSR also collected a significant amount of measurements. Results of their analysis together with a new propagation model were published in 1972 [28], although their data have not been made available to the public.

Three international organizations jointly planned and carried out a measurement campaign in Africa between 1963 and 1964. They are the EBU, the OIRT, and the Union of National Radio afd Television Organizations (URTNA). Later, the British Broadcasting Corporation set up seven receiving stations in Africa and signals from two transmitters on the British Ascension Islands were monitored. The BBC project was intended to study polarization coupling loss and sea gain. The Max Planck Institute also conducted measurements at Tsumeb, southwest Africa.

Administrations in ITU Region 3 (parts of Asia, Australia, and New Zealand), in cooperation with the Asian-Pacific Broadcasting Union (ABU), were equally active and productive. Furthermore, the Japanese administration carried out a number of mobile experiments in different parts of the Pacific [29].

While the administrations in the Eastern Hemisphere were busy preparing for the 1974/75 Regional LF/MF Conference for ITU Regions 1 and 3, IWP 6/4 was actively developing a propagation model to be used as part of the technical bases for such a conference. After extensive studies and lengthy deliberations, the IWP was able to agree on the following: the method developed by Udaltsov and Shlyuger [28] was recommended together with the Knight sea gain formula [30] and the Phillips and Knight polarization coupling loss term [31]. Later in 1974, this method was adopted by the CCIR as Recommendation 435 [32]. In the subsequent sections, this method will be called the *Udaltsov–Shlyuger method*. This method, which includes a sound treatment of latitude, appeared to be very promising at that time. When applied to one-hop intra-European paths, good results were obtained [33]. After years of extensive testing against measured data from other parts of the world, however, some major limitations have surfaced. For example, when applied to paths longer than, say, 4000 km, the method has a strong tendency to underestimate field strengths, in many cases by more than 20 dB [25]. Furthermore, Region 2 data do not seem to corroborate the frequency term of this method [34]. Although this method is a great step forward from the two previous methods, it is something short of a true worldwide method.

4.1.5. The Development of Recommendation 1147-1 (the Wang Method). Knowing the clear-channel curve has some limitations and the need for more field-strength data, the FCC initiated a long-term and large-scale measurement program in 1939. Data from more than 40 propagation paths were collected. The measurement lasted for one sunspot cycle. Unfortunately, the midpoint geomagnetic latitude of the majority of the paths range from 45 °N to 56° North, a narrow window of 11°. Recognizing the need for additional data from the low- and the high-latitude areas, the FCC later initiated two separate projects. In 1980, the FCC and the Institute for Telecommunication Sciences of the Department of Commerce jointly began collecting low-latitude data at Kingsville, Texas, and at Cobo Rojo, Puerto Rico. In 1981, the FCC awarded a contract to the Geophysical Institute, University of Alaska. This project called for the acquisition of sky-wave data from the high-latitude areas. The Alaskan project lasted for about 7 years; data representing different levels of solar activity have been successfully collected. On the basis of the enlarged data bank, a new field strength prediction method was developed by Wang [34]. In 1990 the FCC adopted this method for domestic applications. In 1999 the ITU-R adopted this method as Recommendation P.1147-1 [35], replacing the Udaltsov–Shlyuger method. Like the Udaltsov–Shlyuger method, the Wang method

also contains a similar latitude term. This method has essentially linked the Cairo and the FCC clear-channel curves together mathematically. The special case corresponding to geomagnetic latitude of 35° in the Wang method is extremely close to the Cairo curve. The special case corresponding to 45° is very similar to the FCC curve. It works well for long paths and short paths alike.

4.2. Data Bank

The work of IWP 6/4 is now being handled by Working Party 3L (Ionospheric Propagation) of Study Group 3 (Propagation). The most recent version of the data bank consists of field strengths from 417 propagation paths. Great-circle lengths range from 290 to 11,890 km. Signals of the few very short paths have been verified to be sky waves. Frequencies range from 164 to 1610 kHz. Control-point geomagnetic latitudes range from 46.2° south to 63.8° north. This data bank is available to the general public from the ITU Website (*www.itu.int/brsg/sg3/databanks*).

4.3. Characteristics of LF/MF Sky-Wave Propagation

4.3.1. Amplitude Distribution. Nighttime sky-wave field strengths vary greatly from minute to minute and night to night. The within-the-hour short-term variation usually takes the form of Rayleigh distribution. Night-to-night median values of field strengths for a given reference hour are lognormally distributed. For frequency management purposes, the yearly median value of field strength at six hours after sunset is usually used to determine sky-wave (or secondary) service area of a station while the yearly upper-decile value is used to determine interference level. The difference between the annual upper-decile and the median value varies with latitude, from 6 dB in tropical areas [18] to 12 dB or more at high latitudes [36].

4.3.2. Diurnal Variation. At LF, the transition from daytime condition to nighttime condition in winter is very gradual, and field strength does not reach its maximum value until about 2 h before sunrise. The change at sunrise is more rapid. In summer, field strength increases more rapidly at sunset.

At MF, field strength changes very rapidly at sunset as well as sunrise. Field strength reaches its maximum value shortly after midnight or 6 h after sunset. U.S. data suggest that field strength is highly frequency-dependent during transition hours. For example, the signal of a 1530-kHz station is about 15 dB stronger than that of a 700-kHz station at sunrise. At 6 h after sunset, the difference is only about 3 dB in favor of the higher frequency [35].

4.3.3. Seasonal Variation. At LF and in daytime, sky waves propagating in winter are at least 20 dB stronger than in summer. At night, LF signals are strongest in summer and winter and are weakest in spring and autumn. The summer maximum is more pronounced.

At MF and in daytime, sky waves are strongest in winter months. The seasonal variation may exceed 30 dB. At night, MF sky waves propagating at temperate latitudes are strongest in spring and autumn and are weakest in summer and winter. The summer minimum is more pronounced. The overall variation may be as much as 15 dB at the lowest frequencies in the MF band, decreasing to about 3 dB at the upper end of the band. The variation is much smaller in tropical latitudes.

Nighttime high-latitude field strength data collected in Alaska show a pronounced summer minimum and a consistent maximum in April [36]. In a year of minimum sunspot number, the nighttime monthly median field strength of April is typically 10–15 dB greater than the annual median value.

4.3.4. Effect of Sunspots. At LF, the effect of sunspots is virtually nonexistent. At MF, sunspots greatly reduce sky-wave field strength levels. The reduction (L_r) is a function of sunspot number, latitude, distance, and, to a lesser degree, frequency [37].

The effect of sunspots is clearly latitude-dependent. In low-latitude areas (e.g., Central America, Mexico), annual median values of field strengths vary slightly (less than 3 dB) within a sunspot cycle and there is no detectable pattern. A pattern of correlation begins to surface at higher latitudes (e.g., central USA). For example, measured field strengths from a path in the southern parts of the United States (San Antonio, TX to Grand Island, NE; 1200 kHz, 1279 km, 45.1°N) decreased by about 3 dB when the sunspot number reached from minimum to maximum in cycle 18. The correlation becomes more pronounced at still higher latitudes. For example, measured field strengths of a path in the northern United States (Chicago, IL to Portland, OR, 890 kHz, 2891 km, 54°N) decreased by 15 dB in the same cycle. In Alaska, in a year of maximum solar activity, there are virtually no sky waves from northern-tier U.S. stations, although signals can be very strong in a year of low or moderate solar activity [36].

The effect of sunspots has a diurnal variation of its own. In other words, L_r is different at different hours of the night. At 6 h after sunset, L_r is considerably smaller than that at two hours after sunset. For example, consider a path from Cincinnati, OH to Portland, OR (700 kHz, 3192 km, 53.2°N). From 1944 (a year of minimum sunspot number) to 1947 (a year of maximum sunspot number), field strength for the sixth hour after sunset decreased by 7.3 dB; that for the fourth hour after sunset decreased by 13.3 dB, and that for the second hour after sunset decreased by 16.9 dB [37].

4.3.5. Effect of the Magnetic Field. When a radiowave enters the ionosphere in the presence of a magnetic field, it is split into two components: the ordinary and the extraordinary waves. Both are elliptically polarized. The extraordinary wave is then absorbed. Further loss occurs when the wave leaves the ionosphere. Because of the nature of elliptical polarization, only its vertical component normally couples with the receiving antenna. This process is known as polarization coupling loss (L_p). At LF, L_p is negligible. At MF, L_p is negligible in high and temperate latitudes. In tropical areas, however, it can be very large and depends on the direction of propagation relative to that of the earth's magnetic field. In some

extreme cases (e.g., east–west paths in equatorial Africa), polarization coupling losses of more than 20 dB have been observed. This phenomenon is not yet fully understood, and more data are needed. An interim formula, however, has been developed by Phillips and Knight [31].

4.3.6. Influence of Seawater. When at least one terminal of a path is situated near the sea and a significant portion of the path is over seawater, the received signal is significantly stronger than otherwise. This is commonly called sea gain (G_s). Sea gain is a complicated function of several factors, including pathlength (i.e., elevation angle), distance from antenna to the sea, and frequency. Under ideal conditions (elevation angle = 0, antenna is on the coast), sea gain is about 4 dB at LF and about 10 dB at MF. For a more detailed discussion, see a paper by Knight [30].

4.3.7. Propagation at Daytime. Daytime measurements from more than 30 paths are believed to be sky waves and have been studied by Wang [38]. Some trends are briefly stated as follows:

LF Cases. LF sky-wave field strengths at noon can be surprisingly strong, particularly in winter months. Daytime annual median field strength is typically 20 dB lower than its counterpart at night. Daytime upper-decile value is about 13 dB stronger than the median value.

MF Cases. MF sky-wave field strengths at noon display a consistent seasonal variation pattern with maximum occurring in winter months. The average winter-month field strength is about 10 dB stronger than the annual median value. The winter : summer ratio can exceed 30 dB. The annual median value of field strength at noon is about 43 dB lower than its counterpart at 6 h after sunset. Field strength exceeded for 10% of the days of the year is about 13 dB stronger than the median value.

4.4. Comparison of Predicted and Measured Field Strengths

An extensive comparative study using the most recent data bank has been carried out [25]. For each and every propagation path in the data bank, field strengths have been calculated by using the four most popular methods discussed in this article. Polarization coupling loss and sea gain, if applicable, have been included. In a very small number of cases where measurements were taken in a year of maximum sunspot number, the Wang formula [37] for sunspot losses has also been used. Calculated results are compared with measured data, and prediction errors are thus obtained and analyzed. Prediction errors are analyzed from four different viewpoints (path length, latitude, frequency, and geographic regions) and tabulated in detail [25]. In this article, only pathlength and latitude are considered.

4.4.1. Pathlength and Path Accuracy. We arbitrarily define a short path as one whose length is shorter than 2500 km, a medium path between 2500 and 4999 km, and a long path greater than 5000 km. Then there are 267 short paths, 85 medium paths, and 65 long paths. On the RMS (root-mean-square) basis, the prediction errors of the Cairo curve are 6.6 dB for short paths, 7.5 dB for medium paths, and 10.1 dB for long paths. The corresponding errors of the Region 2 method are 6.6, 7.5, and 12.0 dB, respectively. Those of the Udaltsov–Shlyuger method are 5.7, 7.8, and 16.4 dB, respectively. The errors of the Wang method are 5.5, 6.5, and 6.7 dB, respectively. The Wang method is the only method that offers good to excellent results in all distance ranges.

4.4.2. Latitude and Accuracy. In this section we arbitrarily define low latitudes as those between 0° and 35° (geomagnetic), temperate latitudes as those between 35.1° and 50°, and high latitudes as those greater than 50°. Then there are 203 paths in the low-latitude areas, 152 paths in the temperate-latitude areas, and 62 paths in the high latitudes. On the RMS basis, the prediction errors of the Udaltsov–Shlyuger method are 7.9 dB in low-latitude areas, 6.5 dB in temperate-latitude areas, and 14.0 dB in high-latitude areas, respectively. On the other hand, the corresponding errors of the Wang method are 5.7, 5.6, and 4.4 dB, respectively.

In summary, the Wang method is the only method that offers good to excellent results for short and long paths alike, at all frequencies in the LF/MF bands, at all latitudes, and in all regions.

4.5. The Recommended Propagation Model

In this section we recommend and present the Wang method. For a complete step-by-step procedure, see ITU-R Recommendation P.1147-1 [35]. This section is not meant to be self-contained. Only the most important equations are given here.

4.5.1. Annual Median Field Strength. According to the Wang method, the annual median sky-wave field strength at 6 h after sunset, E (in decibels above 1 μV/m), is given by

$$E = P + G + (A - 20\log p) - k\left(\frac{p}{1000}\right)^{0.5} - L_p + G_s - L_r \quad (4)$$

$$k = 2\pi + 4.95\tan^2(\Phi) \quad (5)$$

where P = radiated power in dB above 1 kW
G = transmitting antenna gain (dB)
A = a constant (at LF, $A = 110.2$; at MF, $A = 110$ in Australia and New Zealand; and $A = 107$ in all other places)
p = actual slant distance of the path under study, in kilometers, assuming that average height of E layer is 100 km
Φ = geomagnetic latitude of the midpoint of the path under study in degrees
L_p = polarization coupling loss (dB) [31]
G_s = sea gain (dB) [30]
L_r = loss of field strength due to solar activity (dB) [37]

4.5.2. Upper Decile Field Strength. Field strength exceeded for 10% of the nights of a year $E(10)$, is greater than the annual median value by Δ dB. Then

$$\Delta = 0.2|\Phi| - 2 \tag{6}$$

where Δ is limited between 6.0 and 10 dB.

5. SKY-WAVE PROPAGATION AT HF

5.1. General Description of HF Propagation

HF sky-wave propagation may be represented by rays between the ground and the ionosphere. In the ionosphere, the radiowaves experience dispersion and changes in polarization. The propagation is affected by, among other factors, ionization, operating frequency, ground conductivity and elevation angle. HF waves in the ionosphere undergo continuous refraction (i.e., bending of the ray path). At any given point, refraction is less at lower electron densities, for higher frequencies, and for higher elevation angle. For a given elevation angle, there exists a certain frequency below which the rays will be reflected back to earth. At a higher frequency, the refraction is too low for the rays to be returned to earth. Waves launched vertically may be reflected, if their frequency is below the "critical frequency" (see Section 6).

The apparent height of reflection varies between about 100 and 300 km. Radiowaves that are launched more obliquely travel to greater range. The maximum range attained after one hop arises for rays launched at grazing incidence. For typical E, F1, and F2 layers, it is about, 2000, 3400, and 4000 km, respectively. In HF communication, several propagation paths are often possible between a given transmitter and a given receiver, such as a single reflection from the E region (1E mode), a single reflection from the F region (1F mode), and double reflection from the F region (2F mode). Mode 2F is said to have higher "order" than mode 1F in propagation terms. This feature is known as *multipath*.

At frequencies above the critical frequency, there is an area surrounding the transmitter defined by "skip distance" in which sky wave cannot be received because the elevation angle is too high. The maximum usable frequency (MUF), a very important concept in HF propagation, may be defined as the frequency that makes the distance from the transmitter to a given reception point equal to the skip distance (see also Section 6). The MUF increases with pathlength and decreases with the height of the ionospheric layer. The MUF also undergoes diurnal, seasonal, solar cycle, and geographic variations. The MUF tends to be high during the day and low during the night. Also, the MUF is higher in summer than in winter during the night. Furthermore, the MUF tends to increase with increasing sunspot number. The F2-layer MUF may increase as much as 100% from sunspot minimum and sunspot maximum. The MUF has a very complex geographic variation. The most authoritative presentation of MUF is undoubtedly the CCIR Report 340, *Atlas of Ionospheric Characteristics* [39], which presents world maps of MUF for the F2 layer corresponding to different month of the year, solar activity levels, and distance ranges.

5.2. Fading

5.2.1. Interference Fading. Interference fading results from interference between two or more waves, which travel by different paths to arrive at the receiving point. This type may be caused by interference between multiple reflected sky waves, sky wave, and ground wave. This type of fading may last for a period of a fraction of a second to a few seconds, during which time the resultant field intensity may vary over wide limits.

5.2.2. Polarization Fading. Polarization fading occurs as a result of changes in the direction of polarization of the downcoming wave, relative to the orientation of the receiving antenna, due to random fluctuations in the electron density along the path of propagation. This type of fading also lasts for a fraction of a second to a few seconds.

5.2.3. Absorption Fading. Absorption fading is caused by variation in the absorption due to changes in the densities of ionization, and it may sometimes last longer than one hour.

5.2.4. Skip Fading. Skip fading may be observed at receiving locations near the skip distance at about sunrise and sunset, when the basic MUF for the path may oscillate around the operating frequency. The signal may decrease abruptly when the skip distance increases past the receiving point (or increase with a decrease in the skip distance).

5.3. Regional Anomalies

5.3.1. Tropical Anomalies. In the tropical zone, sky-wave propagation is characterized by the presence of equatorial sporadic E and the spread F. Equatorial sporadic E (Es-q), which appears regularly during daytime in a narrow zone near the magnetic equator, is the principal cause for fading at daytime. In the equatorial zone after local sunset, some irregularities develop in the F-region ionization and are called *spread F*. Under these conditions, the F region increases markedly in height and seems to break up into patchy irregular regions. As a result, a peculiar type of rapid fading, called flutter fading, usually occurs after sunset. Flutter fading is one of the most important factors in the degradation of HF broadcast service in tropical areas. Flutter fading is most pronounced following the equinoxes. Flutter fading correlates negatively with magnetic activity. On magnetically quiet days, it is usually evident, whereas on magnetically disturbed days, it is absent. The fading rate is proportional to the wave frequency and may range between 10 and 300 per minute [19].

5.3.2. High-Latitude Anomalies. At high latitudes, the ionosphere is exposed to the influence of disturbances in interplanetary space and in the magnetosphere, since the magnetic field lines extend far from the earth. Electrically charged particles can move easily along the field lines and perturb the high-latitude ionosphere. The

absorption is inversely proportional to frequency squared. The absorption may be preceded by a sudden ionospheric disturbance (SID) on the sunlit side of the earth, at all latitudes, caused by X rays from solar flares. At HF, absorption can be greater than 100 dB [40,41]. The magnetic storm-related absorption in the sunlit part of the polar cap is much stronger than in the dark side. The average duration of the event is about 2 days, but may be as long as 4 days. It may spread to lower latitudes too.

5.4. Predicting HF Sky-Wave Field Strength

The calculation of HF field strengths is a very complicated process. It requires a computer. In the succeeding section, a survey of existing programs will be presented. In this section, only a brief outline of the calculation procedure is given. The purpose is to illustrate the general procedures and terms involved. For a more detailed presentation, see, for example, ITU-R Recommendation PI.533-4 [42].

The median value of sky-wave field strength for a given mode of propagation, in dB (μV/m), is given by

$$E_{ts} = 136.6 + P_t + G_t + 20\log f - L_{bf} - L_i - L_m - L_g - L_h - 12.2 \quad (7)$$

where P_t = transmitter power in dB relative to 1 kW
G_t = transmitting antenna gain (dB)
f = transmitting frequency (MHz)
L_{bf} = basic free space transmission loss = $32.45 + 20\log f + 20\log p$ (8)
p = slant distance (km)
L_i = absorption loss (dB)
L_m = above MUF loss (dB)
L_g = ground reflection loss (dB)
L_h = auroral and other signal losses

5.5. Performance Prediction Software

A large number of computer programs have been developed for predicting HF circuit performance. The following is a brief list of the programs that are widely used today. For an excellent discussion on this topic, see a paper by Rush [43].

5.5.1. IONCAP. Ionospheric Communications Analysis and Prediction Program (IONCAP) was developed by staff of the Institute for Telecommunication Sciences (ITS) of the National Telecommunications and Information Administration (NTIA), Department of Commerce [44]. The propagation features include refraction bending, scattering on frequencies above the MUF, and sporadic E propagation. The predicted field strength and noise levels can help the designer to determine optimum frequencies, correct antennas, required transmitter powers. IONCAP is available from NTIA/ITS, Department of Commerce, Boulder, CO (USA).

5.5.2. VOACAP. At the request of the Voice of America, IONCAP has been modified and improved [45]. The resultant program, VOACAP, is available from ITS Website *http://elbert.its.bldrdoc.gov/hf.html*.

5.5.3. ITU-R Recommendation 533-4 (REC533). In preparation for the 1984 HF World Administrative Radio Conference (WARC HFBC-84), the CCIR established Interim Working Party 6/12. After extensive deliberations, it adopted the following. For paths shorter than 7000 km, IWP 6/12 adopted a simplified version of the method described in CCIR Report 252-2, similar to IONCAP. For paths longer than 9000 km, the FTZ method [46] was adopted. For in-between paths, a linear interpolation scheme is used. The FTZ method has been known for its simplicity and accuracy when applied to very long paths. Results of the work of IWP 6/12 are documented in Recommendation PI.533-4 [42]. This software is known as REC533, available from the ITU, Geneva, Switzerland; also available from the abovementioned ITS Website.

5.5.4. Input Data and Results of Calculations. In order to use any of the aforementioned programs, the following required input information is usually needed for each given circuit: (1) time of day, month, and year; (2) expected sunspot number; (3) antenna type; (4) geographic locations of the transmitter and receiver; (5) human-made noise level; (6) required reliability; and (7) required signal-to-noise ratio. The results of calculations usually include the following: (1) great-circle and slant distances, (2) angles of departure and arrival, (3) number of hops, (4) time delay of the most reliable propagation mode, (5) the virtual height, (6) MUF and the probability that the frequency exceeds the predicted MUF, (7) median system loss in dB, (8) median field strength in dB above 1 μV/m, (9) median signal power in dBW, (10) median noise power in dBW, (11) median signal/noise ratio in dB, and (12) LUF (the lowest usable frequency).

6. GLOSSARY

CCIR. French acronym for International Radio Consultative Committee (now ITU-R).

Critical frequency fo. The highest frequencies at which a radio wave is reflected by a layer of the ionosphere at vertical incidence. There is usually one such frequency for each ionospheric component (e.g., foE, foF2). The critical frequency is determined by the maximum electron density in that layer. Waves with their frequency below fo will be reflected. As the frequency is increased beyond this, the ray will penetrate the layer.

Fading. The temporary and significant decrease of the magnitude of the electromagnetic field or of the power of the signal due to time variation or the propagation conditions.

Free-space propagation. Propagation of an electromagnetic wave in a homogeneous ideal dielectric medium, which may be considered of infinite extent in all directions.

Frequency band. Continuous set of frequencies in the frequency spectrum lying between two specific limiting frequencies; generally includes many channels.

Low-frequency (LF) band. The part of the spectrum between 30 and 300 kHz. This band is also known as band 5 because the center frequency is 1×10^5 Hz. The corresponding waves are sometimes called the *kilometric* or *long waves*.

Medium-frequency (MF) band. The part of the spectrum between 300 and 3000 kHz. This band is also known as band 6. The corresponding waves are sometimes called the *hectometric* or *medium waves*.

High-frequency (HF) band. The part of the spectrum between 3 and 30 MHz. This band is also known as band 7. The corresponding waves are sometimes called *decametric* or *short waves*.

ITU. International Telecommunication Union.

ITU-R. Radiocommunication Study Groups of the ITU.

ITU Region 1. Africa, Europe, the entire territory of the former USSR, Outer Mongolia, and Turkey.

ITU REGION 2. The Americas and Greenland.

ITU REGION 3. Australia, New Zealand, and all other Asian countries.

Ionosphere. The ionized region of the earth's upper atmosphere.

MUF. Maximum usable frequency.

Basic MUF. The highest frequency by which a radiowave can propagate between given terminals, on a specific occasion, by ionospheric refraction alone. Where the basic MUF is restricted to a particular propagation mode, the values may be quoted together with an indication of that mode (e.g., 2F2 MUF, 1E MUF). Furthermore, it is sometimes useful to quote the ground range for which the basic MUF applies. This is indicated in kilometers following the indication of the mode type [e.g., 1F2(4000) MUF].

Operational MUF (or simply MUF). The highest frequency that would permit acceptable performance of a radio circuit by signal propagation via the ionosphere between giving terminals at a given time under specific working conditions.

Median values of field strengths—yearly median. The median of daily values for the year, usually for a given reference hour.

Multipath propagation. Propagation of the same radio signal between a transmission point and a reception point over a number of separate propagation paths.

Noise

- **Atmospheric noise**. Radio noise produced by natural electric discharges below the ionosphere and reaching the receiving point along the normal propagation paths between the earth and the lower limit of the ionosphere.
- **Human-made noise**. Radio noise having its source in synthetic (human-made) devices.
- **Galactic noise**. Radio noise arising from natural phenomena outside the earth's atmosphere.

Propagation. Energy transfer between two points without displacement of matter.

Reliability. Probability that a specific performance is achieved.

- **Basic reliability**. The reliability of communications in the presence of background noise alone.
- **Overall reliability**. The reliability of communications in the presence of background noise and of known interference.

Skip distance. The minimum distance from the transmitter at which a sky wave of a given frequency will be returned to earth by the ionosphere.

Solar activity. The emission of electromagnetic radiation and particles from the sun, including slowly varying components and transient components caused by phenomena such as solar flares.

Sudden ionospheric disturbance (SID). A sudden marked increase in electron density of the lower ionosphere during daylight hours. This is caused by X-ray emission from the sun.

Transmission loss. The ratio, usually expressed in decibels, for a radio link between the power radiated by the transmitting antenna and the power that would be available at the receiving antenna output.

Basic free-space transmission loss (L_{bf}). The transmission loss that would occur if the antennas were replaced by isotropic antennas located in a perfectly dielectric, homogeneous, isotropic, and unlimited environment [see also Eq. (8)].

Zenith angle. The angle between the sun and the zenith (i.e., directly overhead) at a given geographic location.

BIOGRAPHY

John C. H. Wang received his B.S. degree in electrical engineering in 1959 from the University of Maryland, and his M.S. degree in electrical engineering in 1968 from the University of Pittsburgh, Pennsylvania. He has been with the technical research staff of the Federal Communications Commission in Washington D.C. since 1969. His main area of research is ionospheric propagation and has published more than 30 technical papers. He is also active in the Radiocommunication Study Groups of the Telecommunication Union (ITU-R) and chairs its Working Party on Ionospheric Propagation. His other interests include astronomy and Chinese history. With that unique combination, he has written papers on the sighting of sunspots in ancient China and on the subject of unearthing the star of Bethlehem from Chinese history. He is a fellow of the IEEE.

BIBLIOGRAPHY

1. A. Summerfield, The propagation of waves in wireless telegraphy, *Ann. Physik* **28**: 665–736 (1909).
2. G. N. Watson, The diffraction of radio waves by the earth, *Proc. Roy. Soc. A* **95**: 83–99 (1918).
3. H. Bremmer, *Terrestrial Radio Waves*, Elsevier, Amsterdam, 1949.

4. K. A. Norton, Propagation of radio waves over the surface of the earth and in the upper atmosphere part I, *Proc. IRE* **24**(10): 1367–1387 (1936).
5. Federal Communications Commission, Standards for good engineering practice concerning standard broadcast stations, *Fed. Reg.* (4FR 2862) (July 8, 1939).
6. K. A. Norton, The calculation of ground-wave field intensity over a finitely conducting spherical earth, *Proc. IRE* **29**(12): 623–639 (1941).
7. ITU-R, *Ground-Wave Propagation Curves for Frequencies between 10 kHz and 30 MHz*, Recommendation PN.368-7, Geneva, ITU, 1994.
8. L. A. Berry, *User's Guide to Low-Frequency Radio Coverage Program*, Office of Telecommunications TM 78-247, 1978.
9. S. Rotheram, Ground wave propagation, part 1: theory for short distances; Part 2: Theory for medium and long distances, *Proc. IEEE* **128**(5): 275–295 (1981).
10. R. P. Eckert, *Modern Methods for Calculating Ground-Wave Field Strength over Smooth Spherical Earth*, FCC Report OET R 8601, 1986.
11. E. Haakinson, S. Rothschild, and B. Bedford, *MF Broadcasting System Performance Model*, NTIA Report 88-237, Dept. Commerce, Washington, DC, 1988.
12. H. L. Kirke, Calculation of ground-wave field strength over a composite land and sea path, *Proc. IRE* **37**(5): 489–496 (1949).
13. G. Millington, Ground wave propagation over an inhomogeneous smooth earth, *Proc. IEE* **96**(39) (Pt. III): 53–64 (1949).
14. K. N. Stokke, Some graphical considerations on Millington's method for calculating field strength over inhomogeneous earth, *Telecommun. J.* **42**(Pt. III): 157–163 (1975).
15. H. Fine, An effective ground conductivity map for continental United States, *Proc. IRE* **49**: 1405–1408 (1954).
16. ITU-R, *World Atlas of Ground Conductivities*, Recommendation P.832-2, Geneva, ITU, 1999.
17. O. Heaviside, The theory of electric telegraphy, in *Encyclopedia Britannica*, 10th ed., 1902.
18. K. Davies, *Ionospheric Radio*, Peregrinus, London, 1990.
19. J. Goodman, *HF Communications: Science and Technology*, Van Nostrand, New York, 1992.
20. Ban Ku, *Book of Han*, 99, published for the first time about 92 A.D., reedited and republished under the supervision of Emperor Chien Lung in 1736; available from many publishers, including Yee Wen Press, Taipei.
21. Fang Shyuan Ling, *Book of Jin*, 12, published for the first time about 640 A.D., reedited and republished under the supervision of Emperor Chien Lung in 1736, available from many publishers including Yee Wen Press, Taipei.
22. ITU-R, *Radio Noise*, Recommendation PI.372-6, ITU, Geneva, 1994.
23. F. Horner, Analysis of data from lightning flash counters, *Proc. IEE* **114**: 916–924 (1967).
24. G. Gustafsson, A revised corrected geomagnetic coordinate system, *Arkiv Geofysik* **5**: 595–617 (1970).
25. J. C. H. Wang, An objective evaluation of available LF/MF sky-wave propagation models, *Radio Sci.* **34**(3): 703–713 (1999).
26. International Telecommunication Union, *Final Acts of the Regional Administrative LF/MF Broadcasting Conference (Regions 1 and 3), ITU, Geneva, 1975*, Geneva, 1976.
27. International Telecommunication Union, *Final Acts of the Regional Administrative MF Broadcasting Conference (Region 2) Rio de Janeiro, 1981*, ITU, Geneva, 1982.
28. A. N. Udaltsov and I. S. Shlyuger, Propagation curves of the ionospheric wave at night for the broadcasting range, *Geomag. i Aeronom.* **10**: 894–896 (1972).
29. C. Nemeto, N. Wakai, M. Ose, and S. Fujii, Integrated results of the mobile measurements of MF field strength along the Japan-Antarctica sailing course, *Rev. Radio Res. Lab.* **33**(168): 157–182 (1987).
30. P. Knight, *LF and MF Propagation: An Approximate Formula for Estimating Sea Gain*, BBC Report RD 1975/32, 1975.
31. G. J. Phillips and P. Knight, Effects of polarisation on a medium-frequency sky-wave service, including the case of multihop paths, *Proc. IEE* **112**: 31–39 (1965).
32. CCIR, *Sky-Wave Field Strength Prediction Method for Frequency Range 150 to 1600 kHz*, Recommendation 435, 1974.
33. J. C. H. Wang, P. Knight, and V. K. Lehtoranta, A study of LF/MF skywave data collected in ITU Region 1, in J. M. Goodman, ed., *Proc. 1993 Ionospheric Effects Symp.*, Alexandria, VA, 1993.
34. J. C. H. Wang, Prudent frequency management through accurate prediction of skywave field strengths, *IEEE Trans. Broadcast.* **35**(2): 208–217 (1989).
35. ITU-R, *Prediction of Sky-Wave Field Strength at Frequencies between about 150 and 1700 kHz*, Recommendation P.1147-1, 1999.
36. R. D. Hunsucker, B. S. Delana, and J. C. H. Wang, Medium-frequency skywave propagation at high latitudes: Results of a five-year study, *IEEE Trans. Broadcast.* **35**(2): 218–222 (1989).
37. J. C. H. Wang, Solar activity and MF skywave propagation, *IEEE Trans. Broadcast.* **BC-35**(2): 204–207 (1989).
38. J. C. H. Wang, LF/MF skywave propagation at daytime, *IEEE Trans. Broadcast.* **BC-41**(1): 23–27 (1995).
39. CCIR, *Atlas of Ionospheric Characteristics*, Report 340, ITU, Geneva, 1983.
40. R. D. Hunsucker, B. S. Delana, and J. C. H. Wang, Effects of the 1986 magnetic storm on medium frequency skywave signals recorded at Fairbanks, Alaska, in J. Goodman, ed., *Proc. IES'87*, 1987, pp. 197–204.
41. R. D. Hunsucker, Anomalous propagation behavior of radio signals at high latitudes, in H. Soicher, ed., *AGARD Conf. Proc. P-332*, 1982.
42. ITU-R, *HF Propagation Prediction Method*, Recommendation PI.533-4, 1994.
43. C. M. Rush, Ionospheric radio propagation models and predictions: A mini review, *IEEE Trans.* **AP-34**: 1163–1170 (1986).
44. L. R. Teters, J. L. Lloyd, G. W. Haydon, and D. L. Lucas, *Estimating the Performance of Telecommunication Systems Using the Ionospheric Transmission Channel-Ionospheric Communications Analysis and Predictions Programs User's Manual*, NTIA Report 83-127, NTIS Access No. PB84-111210, 1983.

45. G. Lane, *Signal-to-Noise Prediction Using VOACAP*, Rockwell Collins, Cedar Rapids, IA, 2000.

46. A. Ochs, The forecasting system of the Fernmeldetechischen Zentralamt (FTZ), in V. Agy, ed., *AGARD Conf. Proc. P-49*, 1970.

RATE-DISTORTION THEORY

TOBY BERGER
Cornell University
Ithaca, New York

Coding, in the sense of C. E. Shannon's information theory [1], divides naturally into channel coding and source coding. Source coding exhibits a further dichotomy into lossless source coding and lossy source coding. Lossless source coding is treated in a separate article in this encyclopedia. Here, we discuss the theory of lossy source coding, also widely known as *rate-distortion theory*.[1]

Shannon's principal motivation for including "Section V: The Rate for a Continuous Source" in his celebrated paper was that it provided a means for extending information theory to analog sources. Since all analog sources have infinite entropy, they cannot be preserved losslessly when stored in or transmitted over physical, finite-capacity media.

Shannon's famous formula for the capacity of an ideal band-limited channel with an average input power constraint and an impairment of additive, zero-mean, white Gaussian noise reads

$$C = W \log_2(1 + P/N) \quad \text{bits/s} \tag{1}$$

In this formula — the most widely known and the most widely abused of Shannon's results — P is the prescribed limitation on the average input power, W is the channel bandwidth in positive frequencies measured in Hz, and N is the power of the additive noise. Since the noise is white with one-sided power spectral density N_0 or two-sided power spectral density $N_0/2$, we have $N = N_0 W$. Common abuses consist of applying (1) when

1. The noise is non-Gaussian.
2. The noise is not independent of the signal and/or is not additive.
3. Average power is not the (only) quantity that is constrained at the channel input.
4. The noise is not white across the passband and/or the channel transfer function is not ideally bandlimited.

Abuse (1) is conservative in that it underestimates capacity because Gaussian noise is the hardest additive noise to combat. Abuse (2) may lead to grossly underestimating or grossly overestimating capacity. A common instance of abuse (3) consists of failing to appreciate that it actually may be peak input power, or perhaps both peak and average input power, that are constrained. Abuse (4) leads to an avoidable error in that the so-called "water pouring" result [8], generalizing (1), yields the exact answer when the noise is not white, the channel is not bandlimited, and/or the channel's transfer function is not flat across the band. (See also [3,4].) There is a pervasive analogy between source coding theory and channel coding theory. The source coding result that corresponds to (1) is

$$R = W \log_2(S/N) \quad \text{bits/s} \tag{2}$$

It applies to situations in which the data source of interest is a white Gaussian signal bandlimited to $|f| \leq W$ that has power $S = S_0 W$, where S_0 denotes the signal's one-sided constant power spectral density for frequencies less than W. The symbol, N, although often referred to as a "noise," is actually an estimation error. It represents a specified level of mean squared error (MSE) between the signal $\{X(t)\}$ and an estimate $\{\hat{X}(t)\}$ of the signal constructed on the basis of data about $\{X(t)\}$ provided at a rate of R bits/s. That is,

$$N = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt\, E[\hat{X}(t) - X(t)]^2$$

It was, and remains, popular to express MSE estimation accuracy as a "signal-to-noise ratio," S/N, as Shannon did in Eq. (2). It must be appreciated, however, that $\{\hat{X}(t) - X(t)\}$ is not noise in the sense of being an error process that is independent of $\{X(t)\}$ that nature adds to the signal of interest. Rather, it is a carefully contrived error signal, usually dependent on $\{X(t)\}$, that the information/communication theorist endeavors to create in order to conform to a requirement that no more than R bits/s of information may be supplied about $\{X(t)\}$. In modern treatises on information theory, the symbol "D," a mnemonic for average distortion, usually is used in place of N. This results in an alternative form of (2), namely

$$R(D) = W \log_2(S/D) \quad \text{bits/s} \tag{3}$$

which is referred to as the MSE rate-distortion function of the source.

Equation (3) also suffers widespread abuse which takes the form of applying it to situations in which

1. The signal is non-Gaussian.
2. Distortion does not depend simply on the difference of $\hat{X}(t)$ and $X(t)$
3. Distortion is measured by a function of $\hat{X}(t) - X(t)$ other than its square.
4. The signal's spectral density is not flat across the band.

Again, abuse (1) is conservative in that it results in an overestimate of the minimum rate R needed to achieve a specified MSE estimation accuracy because white Gaussian sources are the most difficult to handle in the sense of bit rate versus MSE. Abuses (2) and (3),

[1] This article is drawn in considerable measure from the first half of a survey paper on lossy source coding by T. Berger and J. Gibson [2] that appeared in the 50th Anniversary Issue of the *IEEE Transactions on Information Theory* in October 1998. There have been some additions, some corrections, some enhancements, and many deletions.

which often stem in practice from lack of knowledge of a perceptually appropriate distortion measure, can result in gross underestimates or overestimates of R. Abuse (4) can be avoided by using a water-pouring generalization of (3) which we shall soon discuss. Toward that end we recast (3) in the form

$$R(D) = \max[0, W \log(S_0 W/D)] \tag{4}$$

This explicitly reflects the facts that (2) the signal spectrum has been assumed to be constant at level S_0 across the band of width W in which it in nonzero, and (2) $R(D) = 0$ for $D \geq S_0 W$, because one can achieve a MSE of $S = S_0 W$ without sending any information simply by guessing that $X(t) = m(t)$, the signal's possibly time-varying deterministic mean value function. The base of the logarithm in Eq. (4) determines the information unit — bits for $\log_2$ and nats for $\log_e$. When we deal with continuously distributed quantities, it is more "natural" to employ natural logs. When no log base appears, assume henceforth that base e is intended.

A basic inequality of information theory is

$$D \geq R^{-1}(C) \tag{5}$$

which we shall refer to as the *information transmission inequality*. It says that if you are trying to transmit data from a source with rate-distortion function $R(D)$ over a channel of capacity C, you can achieve only those average distortions that exceed the distortion-rate function evaluated at C. (The distortion-rate function is the inverse of the rate-distortion function; denoted $D(R)$, it always exists because $R(D)$ always is convex.)

Suppose, for example, that we wish to send data about the aforementioned band-limited white-Gaussian process $\{X(t)\}$ over an average-input-power-limited, ideally band-limited AWGN channel and then construct on the basis of the channel output an approximation $\{\hat{X}(t)\}$ that has the least possible MSE. The source and the channel have the same frequency band $|f| \leq W$. Since $R(D) = W \log_2(S/D)$, the distortion-rate function is

$$D(R) = S 2^{-R/W}$$

so Eqs. (1) and (4) together tell us that

$$D \geq D(C) = S \exp\left[-\frac{W \log(1 + P/N)}{W}\right]$$

or

$$D/S \geq (1 + P/N)^{-1} \tag{6}$$

This tells us that the achievable error power per unit of source power (i.e., the achievable normalized MSE) is bounded from below by the reciprocal of one plus the channel SNR. There happens to be a trivial scheme for achieving equality in Eq. (5) when faced with the communication task in question. It consists of the following steps:

Step 1. Transmit $X(t)$ scaled to have average power P; that is, put $\sqrt{P/S}X(t)$ into the channel.

Step 2. Set $\hat{X}(t)$ equal to the MMSE estimate of $X(t)$ based solely on the instantaneous channel output $\sqrt{P/S}X(t) + N(t)$ at time t.

Since the signal and the channel noise are jointly Gaussian and zero mean, the optimum estimate in Step 2 is simply a linear scaling of the received signal, namely

$$\hat{X}(t) = \alpha[\sqrt{P/S}X(t) + N(t)]$$

The optimum α, found from the requirement that the error of the optimum estimator must be orthogonal to the data, is $\alpha = \sqrt{PS}/(P + N)$. The resulting minimized normalized MSE is easily computed to be

$$D/S = (1 + P/N)^{-1} \tag{7}$$

which means we have achieved equality in Eq. (5).

Thus, the simple two-step scheme of instantaneously scaling appropriately at the channel input and output results in an end-to-end communication system that is optimum. No amount of source and/or channel coding could improve upon this in the MSE sense for the problem at hand. This fortuitous circumstance is attributable to a double coincidence. The first coincidence is that the source happens to be the random process that drives the channel at capacity. That is, the given source, scaled by $\sqrt{P/S}$, is that process of average power not exceeding P which maximizes the mutual information between the input and output of the channel. The second coincidence is that the channel just happens to provide precisely the transition probabilities that solve the MSE rate-distortion problem for the given source. That is, when the channel is driven by the scaled source, its output minimizes mutual information rate with the source over all processes from which one can calculate an approximation to the source that achieves a normalized MSE not in excess of $(1 + P/N)^{-1}$.

We are operating at a saddle point at which the *mutual information* rate is simultaneously maximized subject to the average power constraint and minimized subject to the average distortion constraint. The slightest perturbation in any aspect of the problem throws us out this saddle — unequal source and channel bandwidths, non-Gaussianness of the source or channel, an error criterion other than MSE, and so on. The result of any such perturbation is that, in order to recover optimality, it is in general necessary to code both for the source and for the channel. From 1949 to 1958 no research was reported on rate-distortion theory in the United States or Europe. However, there was a stream of activity during the 1950s at Moscow University by members of Academician *A. N. Kolmogorov's* probability seminar. Kolmogorov saw in Shannon's entropy rate an ideal candidate for the long-sought "invariant" in the challenging isomorphism problem of ergodic theory. Kolmogorov and Sinai [5,7] succeeded in showing that equal entropy rates were a necessary condition for isomorphism of ergodic flows. Years later, Ornstein [9] proved sufficiency within an appropriately defined broad class of random stationary processes comprising all finite-order Markov sources and

their closure in a certain metric space that will not concern us here. With the Moscow probability seminar's attention thus turned to information theory, it is not surprising that some of its members also studied Shannon's Section V, The Rate for a Continuous Source. Pinsker, Dobrushin, Iaglom, Tikhomirov, Oseeyevich, Erokhin, and others made contributions to a subject that has come to be called ε-entropy, a branch of mathematics that is intimately intertwined with what information theorists today call rate-distortion theory. Thus, when invited to address an early information theory symposium, Kolmogorov [8] reported without attribution the exact answer for the ε-entropy of a stationary Gaussian process with respect to the squared L_2-norm. That result, and its counterpart for the capacity of a power-constrained channel with additive colored Gaussian noise, have come to be known as the "water pouring" formulas of information theory. In this generality the channel formula is attributable to [12] and the source formula to Pinsker [13,14].

Accordingly, we shall call them the Shannon–Pinsker water pouring formulas. They generalize the formulas given by Shannon in 1948 for the important case in which the spectrum of the source or of the channel noise is flat across a band and zero elsewhere.

The Shannon–Pinsker water pouring formula for the MSE information rate of a Gaussian source can be derived by using the fact that processes formed by bandlimiting a second-order stationary random processes to nonoverlapping frequency bands are uncorrelated with one another. In the case of a Gaussian process, this uncorrelatedness implies independence. Thus, we can decompose a Gaussian process $\{X(t)\}$ with one-sided spectral density $S(f)$ into independent Gaussian processes $\{X_i(t)\}, i = 0, 1, \ldots$ with respective spectral densities $S_i(f)$ given by

$$S_i(f) = \begin{cases} S(f), & \text{if } i\Delta \le f < (i+1)\Delta; \\ 0, & \text{otherwise} \end{cases}$$

Let us now make Δ sufficiently small that $S_i(f)$ becomes effectively constant over the frequency interval in which it is nonzero, $S_i(f) \approx S_i, i\Delta \le f < (i+1)\Delta$. It is easy to see that the best rate-distortion tradeoff we can achieve for subprocess $\{X_i(t)\}$ is

$$R_i(D_i) = \max[0, \Delta \log(S_i\Delta/D_i)]$$

By additively combining said approximations over all the subprocesses, we get an approximation to $\{X(t)\}$ that achieves an average distortion of

$$D = \sum_i D_i$$

and requires a total coding rate of

$$R = \sum_i R_i(D_i) = \sum_i \max[0,\ \Delta \log(S_i\Delta/D_i)]$$

In order to determine the MSE rate-distortion function of $\{X(t)\}$, it remains only to select those D_i's summing to D which minimize this R. Toward that end we set

$$d(R + \lambda D)/dD_i = 0, i = 0, 1, 2, \ldots,$$

where λ is a Lagrange multiplier subsequently selected to achieve a desired value of D or of R. It follows that the D and R values associated with parameter value λ are

$$\begin{aligned} D_\lambda &= \sum_{\{i:S_i\Delta > (\lambda\Delta)^{-1}\}} (\lambda\Delta)^{-1} + \sum_{\{i:S_i\Delta \le (\lambda\Delta)^{-1}\}} S_i\Delta \\ &= \sum_i \min[(\lambda\Delta)^{-1}, S_i\Delta] \end{aligned}$$

and

$$R_\lambda = \sum_i \max[0, \Delta \log(S_i\Delta/(\lambda\Delta)^{-1})]$$

If we use

$$\gamma = (\lambda\Delta^2)^{-1}$$

as our parameter instead of λ and then let $\Delta \to 0$, a parametric expression for the rate-distortion function emerges. Casting the result in terms of the two-sided spectral density $\Phi(f)$, an even function of frequency satisfying $\Phi(f) = S(f)/2, f \ge 0$ and replacing the parameter γ by $\theta = \gamma/2$, we obtain

$$D_\theta = \int_{-\infty}^{\infty} \min[\theta, \Phi(f)]df \tag{8}$$

$$R_\theta = \int_{-\infty}^{\infty} \max\left[0, \frac{1}{2}\log(\Phi(f)/\theta)\right]df \tag{9}$$

[Some practitioners prefer to use angular frequency $\omega = 2\pi f$ as the argument of $\Phi(\cdot)$; of course, df then gets replaced in Eqs. (8) and (9) by $d\omega/(2\pi)$.]

The parametric representation (8) of the MSE rate-distortion function of a stationary Gaussian source is the source coding analog of the Shannon-Pinsker "water pouring" result for the capacity of an input-power-limited channel with additive stationary Gaussian noise. In a rate-distortion function of a time-discrete Gaussian sequence provided we limit the range of integration to $|f| \le 1/2$ or to $|\omega| \le \pi$. In such cases $\Phi(\omega)$ is the discrete-time power spectral density, a periodic function defined by

$$\Phi(\omega) = \sum_{k=-\infty}^{\infty} \phi(k)\exp(j\omega k)$$

where $\phi(k) = EX_jX_{j\pm k}$ is the correlation function of the source data. Note that when the parameter θ assumes a value less than the minimum[2] of $\Phi(\cdot)$, which minimum we shall denote by D^*, (8a) reduces to $D_\theta = \theta$, which eliminates the parameter and yields the explicit expression

$$R(D) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log[\Phi(\omega)/D]d\omega,\ D \le D^*$$

This may be recast in the form

$$R(D) = \tfrac{1}{2}\log(Q_0/D),\ D \le D^*$$

[2] More precisely, less than the essential infimum.

where

$$Q_0 = \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \Phi(\omega) d\omega\right]$$

is known both as the entropy rate power of $\{X_k\}$ and as its optimum one-step prediction error. In 1959 Shannon delivered a paper at the IRE Convention in New York City entitled "Coding Theorems for a Discrete Source with a Fidelity Criterion" [15]. This paper not only introduced the term "rate-distortion function" but also put lossy source coding on a firmer mathematical footing. Major contributions of the paper are:

- Definition and properties of the rate-distortion function.
- Calculating and bounding of R(D).
- Coding theorems.
- Insights into source-channel duality.

A discrete information source is a random sequence $\{X_k\}$. Each X_k assumes values in a discrete set $\mathcal{A}$ called the source alphabet. The elements of $\mathcal{A}$ are called the letters of the alphabet. We shall assume until further notice that there are finitely many distinct letters, say M of them, and shall write $\mathcal{A} = \{a(0), a(1), \ldots, a(M-1)\}$. Often we let $a(j) = j$ and hence $\mathcal{A} = \{0, 1, \ldots, M-1\}$; the binary case $\mathcal{A} = \{0, 1\}$ is particularly important.

The simplest case, to which we shall restrict attention for now, is that in which:

1. The X_k are independent and identically distributed (i.i.d.) with distribution $\{p(a), a \in \mathcal{A}\}$.
2. The distortion that results when the source produces the n-vector of letters $a = (a_1, \ldots, a_n) \in \mathcal{A}^n$ and the communication system delivers the n-vector of letters $b = (b_1, \ldots, b_n) \in \mathcal{B}^n$ to the destination as its representation of a is

$$d_n(a, b) = n^{-1} \sum_{k=1}^{n} d(a_k, b_k) \tag{10}$$

Here, $d(\cdot, \cdot) : \mathcal{A} \times \mathcal{B} \to [0, \infty)$ is called a single-letter distortion measure. The alphabet $\mathcal{B}$—variously called the reproduction alphabet, the user alphabet, and the destination alphabet—may be but need not be the same as $\mathcal{A}$. We shall write $\mathcal{B} = \{b(0), b(1), \ldots, b(N-1)\}$, where $N < M, N = M$ and $N > M$ all are cases of interest. When Eq. (9) applies, we say we have a *single-letter fidelity criterion* derived from $d(\cdot, \cdot)$.

Shannon defined the rate-distortion function $R(\cdot)$ as follows. First, let $Q = \{Q(b \mid a), a \in \mathcal{A}, b \in \mathcal{B}\}$ be a conditional probability distribution over the letters of the reproduction alphabet given a letter in the source alphabet.[3] Given a source distribution $\{p(j)\}$, we associate with any such Q two nonnegative quantities $d(Q)$ and $I(Q)$ defined by

$$d(Q) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) Q(b|a) d(a, b)$$

and

$$I(Q) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) Q(b|a) \log\left(\frac{Q(b|a)}{q(b)}\right)$$

where

$$q(b) = \sum_{a \in \mathcal{A}} p(a) Q(b|a)$$

The quantities $d(Q)$ and $I(Q)$ are, respectively, the average distortion and the average Shannon mutual information associated with Q.

The rate-distortion function of the i.i.d. source $\{X_k\}$ with letter distribution $\{p(a) = P[X_k = a]\}$ with respect to the single-letter fidelity criterion generated by $d(\cdot, \cdot)$ is defined by the following minimization problem:

$$R(D) = \min_{Q: d(Q) \le D} I(Q) \tag{11}$$

Since the generally accepted object of communication is to maximize mutual information, not to minimize it, many people find the definition of the rate-distortion function counterintuitive. In this regard it often helps to interchange the independent and dependent variables, thus ending up with a distortion-rate function defined by

$$D(R) = \min_{Q: I(Q) \le R} d(Q) \tag{12}$$

Everyone considers that minimizing average distortion is desirable, so no one objects to this definition. Precisely the same curve results in the (D, R)-plane, except that now R is the independent variable instead of D. Distortion-rate functions are more convenient for certain purposes, and rate-distortion functions are more convenient for others. One should become comfortable with both.

Properties of the rate-distortion function include:

(a) $R(D)$ is well defined for all $D \ge D_{\min}$, where

$$D_{\min} = \sum_{a \in \mathcal{A}} p(a) \min_{b \in \mathcal{B}} d(a, b)$$

The distortion measure can be modified to assure that $D_{\min} = 0$. This is done via the replacement $d(a, b) \leftarrow d(a, b) - \min_b d(a, b)$, whereupon the whole rate-distortion curve simply translates leftward on the D-axis by $D_{\min}$.)

(b) $R(D) = 0$ for $D \ge D_{\max}$, where

$$D_{\max} = \min_b \sum_a p(a) d(a, b)$$

$D_{\max}$ is the maximum value of D that is of interest, since $R(D) = 0$ for all larger D. It is the value of D associated with the best guess at $\{X_k\}$ in the absence of any information about it other

[3] Such a Q often is referred to as a test channel. However, it is preferable to call it a test system because it functions to describe a probabilistic transformation from the source all the way to the user—not just across the channel. Indeed, the rate-distortion function has nothing to do with any channel per se. It is a descriptor of the combination of an information source and a user's way of measuring the distortion of approximations to that source.

than *a priori* statistical knowledge. For example, $D_{\max} = 1 - \max_a p(a)$ when $\mathcal{A} = \mathcal{B}$ and $d(a, b) = 1$ if $b \neq a$ and 0 if $b = a$.

(c) $R(D)$ is nonincreasing in D and is strictly decreasing at every $D \in (D_{\min}, D_{\max})$.

(d) $R(D)$ is convex downward. It is strictly convex in the range $(D_{\min}, D_{\max})$ provided $N \leq M$, where $N = |\mathcal{B}|$ and $M = |\mathcal{A}|$. In addition to the ever-present straight-line segment $R(D) = 0, D \geq D_{\max}$, if $N > M$ then $R(D)$ can possess one or more straight line segments in the range $D_{\min} < D < D_{\max}$.

(e) The slope of $R(D)$ is continuous in $(D_{\min}, D_{\max})$ and tends to $-\infty$ as $D \downarrow D_{\min}$. If there are straight-line segments in $(D_{\min}, D_{\max})$ (see (d) above), no two of them share a common endpoint.

(f) $R(D_{\min}) \leq H$, where

$$H = -\sum_{a \in \mathcal{A}} p(a) \log p(a)$$

is the source entropy. If for each $a \in \mathcal{A}$ there is a unique $b \in \mathcal{B}$ that minimizes $d(a, b)$, and each $b \in \mathcal{B}$ minimizes $d(a, b)$ for at most one $a \in \mathcal{A}$, then $R(D_{\min}) = H$.

Some of these properties were established by Shannon [15], including the essential convexity property (d). For proofs of the others see Gallager [3], Jelinek [16], and Berger [17].

In the special case of a binary equiprobable source and an error-frequency (or Hamming) distortion measure, calculations reveal that

$$R(D) = 1 - h(D) = 1 + D \log_2 D + (1 - D) \log_2 (1 - D), \ 0 \leq D \leq 1/2 = D_{\max}$$

where $h(\cdot)$ is Shannon's binary entropy function,

$$h(x) = -x \log_2 x - (1 - x) \log_2 (1 - x)$$

The desired end-to-end system behavior then becomes that of a binary symmetric channel (BSC) with crossover probability D. It follows that, if one seeks to send a Bernoulli(1/2) source over a BSC that is available once per source letter, then optimum performance with respect to the single-letter fidelity criterion generated by $d(a, b) = 1 - \delta_{a,b}$ can be obtained simply by connecting the source directly to the BSC and using the raw BSC output as the system output. There is need to do any source and/or channel coding. The average distortion will be $D = \varepsilon$, where ε is the crossover probability of the BSC.

This is another instance of a double "coincidence." This time the first coincidence is that a Bernoulli(1/2) source drives every BSC at capacity, and the second coincidence is that BSC(ε) provides precisely the end-to-end system transition probabilities that solve the rate-distortion problem for the Bernoulli(1/2) source at $D = \varepsilon$. Again, their combination represents a precarious saddle point. If the channel were not available precisely once per source symbol, if the Bernoulli source were to have a bias $p \neq 1/2$ if the channel were not perfectly symmetric, or if the distortion measure were not perfectly symmetric (i.e., if $d(0, 1) \neq d(1, 0)$), it would become necessary to employ source and channel codes of long memory and high complexity in order to closely approach performance that is ideal in the sense of achieving equality in the information transmission inequality (5). To enhance appreciation for the fragility of the double-coincidence saddle point, let us replace the Bernoulli(1/2) source with a Bernoulli(p) source, $p \neq \frac{1}{2}$. Calculations (see [17], p. 46–47) reveal that the rate-distortion function then becomes

$$R(D) = h(p) - h(D), \ 0 \leq D \leq \min(p, 1 - p) = D_{\max}$$

Although the optimum backward system transition probabilities $P(a|b)$ remain those of BSC(D), the optimum forward transition probabilities become those of a binary asymmetric channel. Hence, it is no longer possible to obtain an optimum system simply by connecting the source directly to the BSC and using the raw channel output as the system's reconstruction of the source. Not only does the asymmetric source fail to drive the BSC at capacity, but the BSC fails to provide the asymmetric system transition probabilities required in the $R(D)$ problem for $p \neq 1/2$. For example, suppose $p = 0.25$ so that $R(D) = 0.811 - h(D)$ bits/letter, $0 \leq D \leq 0.25 = D_{\max}$. Further suppose that $\varepsilon = 0.15$ so that the channel capacity is $C = 1 - h(0.15) = 0.390$ bits/channel use. Direct connection of the source to the channel yields an error frequency of $D = \varepsilon = 0.15$. However, evaluating the distortion-rate function at C in accordance with Eq. (5) shows that a substantially smaller error frequency of $R^{-1}(0.390) = 0.0855$ can be achieved using optimum source and channel coding. It is noteworthy that, even when treating continuous-amplitude sources and reconstructions, Shannon always employed discrete output random variables. "Consider a finite selection of points $z_i (i = 1, 2, \ldots, l)$ from the B space, and a measurable assignment of transition probabilities $q(z_i|m)$"[15]. One reason for why he did this is the he appreciated that the representation of the source would always have to be stored digitally; indeed, his major motivation for Section V in 1948 had been to overcome the challenge posed by the fact that continuous-amplitude data has infinite entropy. But, an even better explanation is that it turns out that the output random variable $\hat{X}$ that results from solving the rate-distortion problem for a continuous-amplitude source usually is discrete! In retrospect, it seems likely that Shannon knew this all along. (For more about this discreteness, see the excellent article by Rose [18] and also work of Fix [19] dealing with cases in which X has finite support.) Shannon did not state or prove any lossy source coding theorems in his classic 1948 paper. He did, however, state and sketch the proof of an end-to-end information transmission theorem for a communication system, namely his Theorem 21. Since the notation $R(D)$ did not exist in 1948, Shannon's theorem statement has v_1 in place of D and R_1 in place of $R(D)$. It reads:

Theorem 21. If a source has a rate R_1 for a valuation v_1 it is possible to encode the output of the source and

transmit it over a channel of capacity C with fidelity as near v_1 as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$.

In 1959, however, Shannon proved many lossy source coding theorems, information transmission theorems, and their converses in the quite general case of stationary and ergodic sources. These theorems have since been considerably generalized by various authors; see, for example, the work of Gray and Davisson [20], Bucklew [21], and Kieffer [22]. It is not our purpose here to enter into the details of proofs of source coding theorems and information transmission theorems. Suffice it to say that at the heart of most proofs of positive theorems lies a random code selection argument, Shannon's hallmark. In the case of sources with memory, the achievability of average distortion D at coding rate $R_n(D)$ is established by choosing long-code words constructed of concatenations of "super-letters" from $\mathcal{B}^n$. Each super-letter is chosen independently of all the others in its own code word and in the other code words according to the output marginal $q(b)$ of the joint distribution $p(a)Q(b|a)$ associated with the solution of the variational problem that defines $R_n(D)$. Shannon concluded his 1959 paper on rate-distortion theory with some memorable, provocative remarks on the duality of source theory and channel theory. He mentions that, if costs are assigned to the use of its input letters of a channel, then determining its capacity subject to a bound on expected transmission cost amounts to maximizing a mutual information subject to a linear inequality constraint and results in a capacity-cost function for the channel that is concave downward. He says, "Solving this problem corresponds, in a sense, to finding a source that is just right for the channel and the desired cost." He then recapitulates that finding a source's rate-distortion function is tantamount to minimizing a mutual information subject to a linear inequality constraint and results in a function that is convex downward. "Solving this problem," Shannon says, "corresponds to finding a channel that is just right for the source and allowed distortion level." He concludes this landmark paper with the following two provocative sentences:

> This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.

Gallager [3] introduced the following dual to the convex mathematical programming problem that defines $R(D)$: Let λ denote a vector with components $\lambda(a)$ indexed by the letters of the source alphabet, $\mathcal{A}$. Given any real s and any $\lambda 0$ let c denote the vector with components $c(b)$, $b \in \mathcal{B}$ defined by

$$c(b) = \sum_{a \in \mathcal{A}} \lambda(a) p(a) \exp[sd(a,b)]$$

Let

$$\Lambda_s = \{\lambda \geq 0 : c \leq 1\}$$

Gallager proved that

$$R(D) = \max_{s \leq 0, \lambda \in \Lambda_s} \left[sD + \sum_{a \in \mathcal{A}} p(a) \log \lambda(a) \right]$$

Expressing $R(D)$ as a maximum rather than a minimum allows one to generate lower bounds to $R(D)$ readily. Just pick any $s \leq 0$ and any $\lambda \geq 0$. Then evaluate c. If the largest component of c exceeds 1, form a new λ by dividing the original λ by this largest $c(b)$. The new λ then belongs to Λ_s. It follows that the straight line $sD + \Sigma_a p(a) \log \lambda(a)$ in the (D, R)-plane underbounds $R(D)$. Not only are lower bounds to $R(D)$ produced aplenty this way, but we are assured that the upper envelope of all these lines actually *is* $R(D)$. This dual formulation is inspired by and capitalizes on the fact that a convex downward curve always equals the upper envelope of the family of all its tangent lines. It turns out that all known interesting families of lower bounds to $R(D)$ are special cases of this result. In particular, choosing the components of λ such that $\lambda(a)p(a)$ is constant yields Shannon's lower bound [15] for cases in which the distortion measure is balanced (i.e., every row of the distortion matrix is a permutation of the first row and every column is a permutation of the first column) and yields a generalization of the Shannon lower bound when the distortion measure is not balanced. The families of lower bounds introduced by Wyner and Ziv also can be shown to be obtainable via dual problem investigations.

Although it may have appeared in the early 1970s that the then 25-year-old subject of rate-distortion theory was reaching maturity, this has not turned out to be the case. Rate-distortion theory thrived at Stanford under Gray, at Cornell under Berger, who wrote a text devoted entirely to the subject [17], at JPL under Posner, at UCLA under Omura and Yao, and at Bell Labs under Wyner.[4]

Attending a seminar on the mathematics of population genetics and epidemiology somehow inspired Blahut to work on finding a fast numerical algorithm for the computation of rate-distortion functions. He soon thereafter found that the point on an $R(D)$ curve parameterized by s could be determined by the following iterative procedure [23]:[5]

Step 0. Set $r = 0$. Choose any probability distribution $q_0(\cdot)$ over the destination alphabet that has only positive components, for example, the uniform distribution $q_0(b) = 1/|\mathcal{B}|$.

Step 1. Compute $\lambda_r(a) = (\Sigma_b q_r(b) \exp[sd(a,b)])^{-1}$, $a \in \mathcal{A}$.

[4] Centers of excellence in rate-distortion emerged in Budapest under Csiszar, in Tokyo under Amari, in Osaka under Arimoto, in Israel under Ziv and his "descendants," in Illinois under Pursley and at Princeton under Verdu.

[5] Blahut and, independently, Arimoto [24] found an analogous algorithm for computing the capacity of channels. Related algorithms have since been developed for computing other quantities of information-theoretic interest. For a treatment of the general theory of such max-max and min-min alternating optimization algorithms, see Csiszar and Tusnady [25].

Step 2. Compute $c_r(b) = \Sigma_a \lambda_r(a) p(a) \exp[sd(a,b)]$, $b \in \mathcal{B}$. If $\max_b c_r(b) < 1 + \varepsilon$, halt.

Step 3. Compute $q_{r+1}(b) = c_r(b) q_r(b)$. $r \leftarrow r+1$. Return to Step 1.

Blahut proved the following facts.

1. The algorithm terminates for any rate-distortion problem for any $\varepsilon > 0$.
2. At termination, the distance from the point (D_r, I_r) defined by

$$D_r = \sum_{a,b} p(a)\lambda_r(a) q_r(b) \exp[sd(a,b)] d(a,b)$$

and

$$I_r = sD_r + \sum_a p(a) \log \lambda_r(a)$$

to the point D, $R(D)$ parameterized by s (i.e., the point on the $R(D)$-curve at which $R'(D) = s$) goes to zero as $\varepsilon \to 0$. Moreover, Blahut provided upper and lower bounds on the terminal value of $I_r - R(D_r)$ that vanish with ε.

Perhaps the most astonishing thing about Blahut's algorithm is that it does not explicitly compute the gradient of $R + sD$ during the iterations, nor does it compute the average distortion and average mutual information until after termination. In practice, the iterations proceed rapidly even for large alphabets. Convergence is quick initially but slows for large r; Newton-Raphson methods could be used to close the final gap faster, but practitioners usually have not found this to be necessary. The Blahut algorithm can be used to find points on rate-distortion functions of continuous-amplitude sources, too; one needs to use fine-grained discrete approximations to the source and user alphabets. See, however, the so-called "mapping method" recently introduced by Rose [18], which offers certain advantages especially in cases involving continuous alphabets; Rose uses reasoning from statistical mechanics to capitalize on the fact, alluded to earlier, that the support of the optimum distribution over the reproduction alphabet usually is finite even when $\mathcal{B}$ is continuous. Following his seminal work on autoregressive sources and certain generalizations thereof, Gray joined the Stanford faculty. Since rate-distortion is a generalization of the concept of entropy and conditional entropy plays many important roles, Gray sensed the likely fundamentality of a theory of conditional rate-distortion functions and proceeded to develop it [26] in conjunction with his student, Leiner [27,28]. He defined

$$R_{X|Y}(D) = \min I(X; \hat{X}|Y)$$

where the minimum is over all r.v., $\hat{X}$ jointly distributed with (X, Y) in such a manner that $E_{X,Y,\hat{X}} d(X, \hat{X}) \le D$. This not only proved of use per se but also led to new bounding results for classical rate-distortion functions. However, it did not treat what later turned out to be the more challenging problem of how to handle side-information $\{Y_k\}$ that was available to the decoder only and not to the encoder. That had to await groundbreaking research by Wyner and Ziv [29].

Gray also began interactions with the mathematicians Ornstein and Shields during this period. The fruits of those collaborations matured some years later, culminating in a theory of sliding block codes for sources and channels that finally tied information theory and ergodic theory together in mutually beneficial and enlightening ways. Other collaborators of Gray in those efforts included Neuhoff, Omura, and Dobrushin [30–32]. The so-called process definition of the rate-distortion function was introduced and related to the performance achievable with sliding block codes with infinite window width (codes in the sense of ergodic theory). It was shown that the process definition agreed with Shannon's 1959 definition of the rate-distortion function $\liminf_{n\to\infty} R_n(D)$ for sources and/or distortion measures with memory. More importantly, it was proved that one could "back off" the window width from infinity to a large, finite value with only a negligible degradation in the tradeoff of coding rate versus distortion, thereby making the theory of sliding block codes practically significant.

Seeing that Slepian and Wolf [33] had conducted seminal research on lossless multiterminal source coding problems analogous to the multiple access channel models of Ahlswede [34] and Liao [35], Berger and Wyner agreed that research should be done on a lossy source coding analog of the novel Cover-Bergmans [36,37] theory of broadcast channels. Gray and Wyner were the first to collaborate successfully on such an endeavor, authoring what proved to the first of many papers in the burgeoning subject of multiterminal lossy source coding [38]. The seminal piece of research in multiterminal lossy source coding was the paper by Wyner and Ziv [29], who considered lossy source coding with side information at the decoder. Suppose that in addition to the source $\{X_k\}$ that we seek to convey to the user, there is a statistically related source $\{Y_k\}$. If $\{Y_k\}$ can be observed both by the encoder and the decoder, then we get conditional rate-distortion theory a la Gray. The case in which neither the encoder nor the decoder sees $\{Y_k\}$, which perhaps is under the control of an adversary, corresponds to Berger's source coding game [39]. The case in which the encoder sees $\{Y_k\}$ but the decoder does was long known [40] to be no different from the case in which there is no $\{Y_k\}$. But the case in which the decoder is privy to $\{Y_k\}$ but the encoder is not proves to be both challenging and fascinating. For the case of a single-letter fidelity criterion and (X_k, Y_k)-pairs that are i.i.d. over the index k, Wyner and Ziv showed that the rate-distortion function, now widely denoted by $_{WZ}(D)$ in their honor, is given by

$$R_{WZ}(D) = \min_{Z \in cal Z_D} I(X; Z \mid Y) \tag{13}$$

where $\mathcal{Z}_D$ is the set of auxiliary r.v. $Z \in \mathcal{Z}$ jointly distributed with a generic (X, Y) such that:

1. $Y - X - Z$ is a Markov chain; i.e., $p_{Y,X,Z}(y,x,z) = p_Y(y) p_{X|Y}(x|y) p_{Z|X}(z|x)$.
2. There exists $g : \mathcal{Z} \times \mathcal{Y} \to \mathcal{X}$ such that $Ed(X, g(Z, Y)) \le D$.

3. The cardinality of the alphabet $\mathcal{Z}$ may be constrained to satisfy $|\mathcal{Z}| \leq |\chi| + 1$.

In the special case in which $\{X_k\}$ and $\{Y_k\}$ are Bernoulli(1/2) and statistically related as if connected by a BSC of crossover probability $p \leq 1/2$ and $d(a, b) = 1 - \delta_{a,b}$,

$$R_{WZ}(D) = \begin{cases} h(p) - h(p * D), & \text{if } 0 \leq D \leq D_{ci} \\ \text{straight line from } (D_c, h(p) - h(p * D_c) \text{ to } (p, 0) & \text{if } D_c \leq D \leq p \end{cases} \tag{14}$$

where $p * d = p(1 - D) + (1 - p)D$ and D_c is such that the straight-line segment for $D \geq D_c$ is tangent to the curved segment for $D \leq D_c$. Berger had used Bergmans [37] theory of "satellites and clouds" to show that Eq. (17) was a upper bound to $R(D)$ for this binary symmetric case. The major contribution of Wyner and Ziv's paper resided in proving a converse to the unlikely effect that this performance cannot be improved upon, and then generalizing to Eq. (17) for arbitrary (X, Y) and $d(\cdot, \cdot)$.

The advent of Wyner-Ziv theory gave rise to a spate of papers on multiterminal lossy source coding, codified, and summarized by Berger in 1977 [41]. Contributions described therein include works by Korner and Marton, [42–44], Berger and Tung [45,46], Chang [47], Shohara [48], Omura and Housewright [49], Wolfowitz [50], and Sgarro [51]. In succeeding decades further strides have been made on various side-information lossy coding problems [52–58]. Furthermore, challenging new multiterminal rate-distortion problems have been tackled with considerable success, including the multiple descriptions problem [59–68], the successive refinements problem [69], and the *CEO problem* [70–72]. Applications of multiple descriptions to image, voice, audio and video coding are currently in development, and practical schemes based on successive refinement theory are emerging that promise application to progressive transmission of images and other media. In order for rate-distortion theory to be applied to images, video, and other multidimensional media, it is necessary to extend it from random processes to random fields (i.e., collections of random variables indexed by multidimensional parameters or, more generally, by the nodes of a graph). The work of Hayes, Habibi, and Wintz [73] extending the water-table result for Gaussian sources to Gaussian random fields already has been mentioned. A general theory of the information theory of random fields has been propounded [74], but we are more interested in results specific to rate-distortion. Most of these have been concerned with extending the existence of critical distortion to the random field case and then bounding the critical distortion for specific models. The paper of Hajek and Berger [75] founded this subfield. Work inspired thereby included Bassalygo and Dobrushin [76], Newman [77], Newman and Baker[78] in which the critical distortion of the classic Ising model is computed exactly, and several papers by Berger and Ye [79,80]. For a summary and expansion of all work in this arena, see the monograph by Ye and Berger [81].

Work by Fitingof, Lynch, Davisson, and Ziv in the early 1970s showed that lossless coding could be done efficiently without prior knowledge of the statistics of the source being compressed, so-called universal lossless coding. This was followed by development of Lempel-Ziv coding [82,83], arithmetic coding [84–86] and context-tree weighted encoding [87,88], which have made universal lossless coding practical and, indeed, of great commercial value.

Universal lossy coding has proven more elusive as regards both theory and practice. General theories of universal lossy coding based on ensembles of block codes and tree codes were developed [89–95], but these lack sufficient structure and hence require encoder complexity too demanding to be considered as solving the problem in any practical sense. Recent developments are more attractive algorithmically [96–103]. The paper by Yang and Kieffer [100] is particularly intriguing; they show that a lossy source code exists that is universal not only with respect to the source statistics but also with respect to the distortion measure. Though Yang–Kieffer codes can be selected a priori in the absence of any knowledge about the fidelity criterion, the way one actually does the encoding does, of course, depend on which fidelity criterion is appropriate to the situation at hand. All universal lossy coding schemes found to date lack the relative simplicity that imbues Lempel–Ziv coders and arithmetic coders with economic viability. Perhaps as a consequence of the fact that approximate matches abound whereas exact matches are unique, it is inherently much faster to look for an exact match than it is to search a plethora of approximate matches looking for the best, or even nearly the best, among them. The right way to tradeoff search effort in a poorly understood environment against the degree to which the product of the search possesses desired criteria has long been a human enigma. This suggests it is unlikely that the "holy grail" of implementable universal lossy source coding will be discovered soon.

We have several times noted how sources can be matched to given channels and channels can be matched to given sources. Indeed, we even quoted Shannon's 1959 comments about this. Doubly matched situations that require no coding in order to for optimum performance to be achieved have been stressed. Whereas these situations are rarely encountered in the context of man-made communication systems, there is growing evidence that *doubly matched* configurations are more the norm that the exception in sensory information processing by living organisms. A growing community of biologists and information theorists are engaged in active collaboration on mathematical and experiments treatments of information handling within living systems. These bioinformation theorists are finding that chemical pathways and neural networks within organisms not only exhibit double matching but do so robustly over a range of data rates and energy consumption levels. The evolutionist's explanation for how this comes to pass is that, over eons, natural selection evolves channels within successful organisms that are matched to the information sources that constitute the environment. These channels extract just enough information to provide a representation of each facet of the environment that is sufficiently accurate for the organism's purposes. That is, lossy coding that is nearly optimal in the sense of rate-distortion theory is routinely performed by living

organisms. Moreover, these channels evolve so as to reside at saddle points which correspond not only to minimizing information flow subject to fidelity constraints but also simultaneously to maximizing the mutual information between their inputs and outputs subject to constraints on rates at which they consume bodily resources, especially metabolic energy. Further information about lossy coding in living systems can be found in the works of Levy and Baxter [104,105] and especially in the Shannon Lecture recently delivered by Berger [106].

BIOGRAPHY

Toby Berger received the B.E. degree in electrical engineering from Yale in 1962 and the M.S. and Ph.D. degrees in applied mathematics from Harvard in 1964 and 1966. From 1962 to 1968, he was a senior scientist at Raytheon Company. In 1968, he joined the faculty of Cornell University where he is presently the Irwin and Joan Jacobs professor of engineering. His research interests include information theory, random fields, communication networks, wireless communications, video compression, voice and signature compression and verification, biological information theory, quantum information theory, and coherent signal processing. Berger has been a Guggenheim Fellow, a Fulbright Fellow, a Fellow of the Japan Association for Advancement of Science, and a Fellow of the Ministry of Education of the PRC. An IEEE Fellow, he has served as editor-in-chief of the IEEE Transactions on Information Theory and as president of the IEEE Information Theory Society. Berger received the 1982 Frederick E. Terman Award from the American Society for Engineering Education and the 2002 Shannon Award from the IEEE Information Theory Society.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**: 379–423, 623–656, (July and Oct. 1948). (Also in N. J. A. Sloane and A. D. Wyner, eds., *Claude Elwood Shannon: Collected Papers*, IEEE Press, Piscataway, NJ, 1993, 5–83.)
2. T. Berger and J. D. Gibson, Lossy Source Coding, Invited paper for Special 50th Anniversary Issue, *IEEE Trans. Inform. Theory* **44**(6): 2693–2723 (Oct. 1998). (Reprinted in *Information Theory: 50 Years of Discovery* IEEE Press, Piscataway, NJ 1999.)
3. R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.
4. J. L. Holsinger, Digital Communication over Fixed Time-Continuous Channels with Memory — with Special Application to Telephone Channels. Sc. D. dissertation, M.I.T., Cambridge, MA (TR No. 366, Lincoln Labs, Lexington, MA), 1968.
5. A. N. Kolmogorov, A new metric invariant of transitive dynamic systems and automorphisms in Lebesgue spaces, *Dokl. Akad. Nauk. SSSR* **119**: 861–864 (1958).
6. A. N. Kolmogorov, The theory of transmission of information, Plenary Session of the Academy of Sciences of the USSR on the Automization of Production, Moscow, 1956. *Iz. Akad. Nauk SSSR* 66–99 (1957).
7. Ya. G. Sinai, On the concept of entropy of a dynamical system, *Dokl. Akad. Nauk. SSSR* **124**: 768–771 (1959).
8. A. N. Kolmogorov, On the Shannon theory of information transmission in the case of continuous signals, *IRE Transactions on Information Theory* **IT-2**: 102–108 (1956).
9. D. S. Ornstein, Bernoulli shifts with the same entropy are isomorphic, *Advances in Math.* **4**: 337–352 (1970).
10. E. C. Posner and E. R. Rodemich, Epsilon entropy and data compression, *The Annals of Math. Statist.* **42**: 2079–2125 (1971).
11. R. J. McEliece and E. C. Posner, Hiding and covering in a compact metric space, *Annals of Statistics* **1**: 729–739 (1973).
12. C. E. Shannon, Communication in the presence of noise, *Proc. IRE* **37**: 10–21 (1949).
13. M. S. Pinsker, Mutual information between a pair of stationary Gaussian random processes, *Dokl. Akad. Nauk. USSR* **99**(2): 213–216 (1954).
14. M. S. Pinsker, Computation of the message rate of a stationary random process and the capacity of a stationary channel, *Dokl. Akad. Nauk. USSR* **111**(4): 753–756 (1956).
15. C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion. *IRE Convention Record* **7**: 142–163 (1959). (Also in R. E. Machol, ed., *Information and Decision Processes*, McGraw-Hill, Inc. New York, 1960, 93–126, and in N. J. A. Sloane and A. D. Wyner, eds., *Claude Elwood Shannon:Collected Papers*, IEEE Press, Piscataway, NJ, 1993, 325–350.)
16. F. Jelinek, *Probabilistic Information Theory*, McGraw-Hill, New York, 1968.
17. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
18. K. Rose, A mapping approach to rate-distortion computation and analysis, *IEEE Trans. Inform. Theory* **IT-42**: 1939–1952 (1996).
19. S. L. Fix, Rate distortion functions for squared error distortion measures, In *Proc. 16th Annual Allerton Conference on Comm., Contr. and Comput.*, Monticello, IL, (1978).
20. R. M. Gray and L. D. Davisson, Source coding theorems without the ergodic assumption, *IEEE Trans. Inform. Theory* **IT-20**: 625–636 (1974).
21. J. A. Bucklew, The source coding theorem via Sanov's theorem, *IEEE Trans. Inform. Theory* **IT-33**: 907–909 (1987).
22. J. C. Kieffer, A survey of the theory of source coding with a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-39**: 1473–1490 (1993).
23. R. E. Blahut, Computation of channel capacity and rate distortion functions, *IEEE Trans. Inform. Theory* **IT-18**: 460–473 (1972).
24. S. Arimoto, An algorithm for calculating the capacity of an arbitrary discrete memoryless channel, *IEEE Trans. Inform. Theory* **IT-18**: 14–20 (1972).
25. I. Csiszar and G. Tusnady, Information Geometry and Alternating Minimization Procedures, in *Statistics and Decisions / Supplement Issue*, R. Oldenbourg Verlag, Munich, Germany, E. J. Dudewicz, D. Plachky and P. K. Sen, Eds., No. 1, 205–237, 1984. (Formerly entitled On Alternating Minimization Procedures, Preprint of the Mathematical

Institute of the Hungarian Academy of Sciences, No. 35/1981, 1981.)

26. R. M. Gray, A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions, *IEEE Trans. Inform. Theory* **IT-19**: 480–489 (1973).
27. B. M. Leiner, *Rate-Distortion Theory for Sources with Side Information*, Ph.D. dissertation, Stanford University, Calif., August 1973.
28. B. M. Leiner and R. M. Gray, Rate-distortion for ergodic sources with side information, *IEEE Trans. Inform. Theory* **IT-20**: 672–675 (1974).
29. A. D. Wyner and J. Ziv, The rate-distortion function for source coding with side-information at the receiver, *IEEE Trans. Inform. Theory* **IT-22**: 1–11 (1976).
30. R. M. Gray, D. L. Neuhoff and J. K. Omura, Process definitions of distortion rate functions and source coding theorems, *IEEE Trans. Inform. Theory* **IT-21**: 524–532 (1975).
31. R. M. Gray, D. L. Neuhoff and D. S. Ornstein, Nonblock source coding with a fidelity criterion, *Annals of Probability* **3**: 478–491 (1975).
32. R. M. Gray, D. S. Ornstein and R. L. Dobrushin, Block synchronization, sliding-block coding, invulnerable sources and zero error codes for discrete noisy channels, *Annals of Probability* **8**: 639–674 (1975).
33. D. Slepian and J. K. Wolf, Noiseless coding of correlated information sources, *IEEE Trans. Inform. Theory* **IT-19**: 471–480 (1973).
34. R. Ahlswede, Multi-way Communication Channels, In *Proc. 2nd. Int. Symp. Information Theory* (*Tsahkadsor, Armenian SSR*), 23–52, 1971.
35. H. Liao, Multiple Access Channels, Ph.D. dissertation, University of Hawaii, Honolulu, HI, 1972.
36. T. M. Cover, Broadcast channels, *IEEE Trans. Inform. Theory* **IT-18**: 2–14 (1972).
37. P. Bergmans, Random coding theorem for broadcast channels with degraded components, *IEEE Trans. Inform. Theory* **IT-19**: 197–207 (1973).
38. R. M. Gray and A. D. Wyner, Source coding for a simple network, *Bell Syst. Tech. J.* **58**: 1681–1721 (1974).
39. T. Berger, The Source coding game, *IEEE Trans. Inform. Theory* **IT-17**: 71–76 (1971).
40. T. J. Goblick, Jr., Coding for a Discrete Information Source with a Distortion Measure, Ph.D. dissertation, M.I.T., Cambridge, MA, 1962.
41. T. Berger, Multiterminal Source Coding, In *The Information Theory Approach to Communications*, CISM Courses and Lectures No. 229, 171-231, Springer-Verlag, Wien–New York, 1977.
42. J. Korner and K. Marton, The Comparison of Two Noisy Channels, *Trans. Keszthely Colloq. Inform. Theory*, Hungarian National Academy of Sciences, August 8–12, Keszthely, Hungary, 411–423.
43. J. Korner and K. Marton, Images of a set via two channels and their role in multi-user communications, *IEEE Trans. Inform. Theory* **IT-23**: 751–761 (1977).
44. J. Korner and K. Marton, How to encode the modulo-two sum of binary sources, *IEEE Trans. Inform. Theory* **IT-25**: 219–221 (1979).
45. T. Berger and S. Y. Tung, Encoding of Correlated Analog Sources, *Proc. 1975 IEEE-USSR Joint Workshop on Information Theory*, IEEE Press, 7–10, December 1975.
46. S. Y. Tung, Multiterminal Rate-Distortion Theory, Ph.D. dissertation, Cornell University, Ithaca, NY, 1977.
47. M. U. Chang, Rate-Distortion with a Fully Informed Decoder and a Partially Informed Encoder, Ph.D. dissertation, Cornell University, Ithaca, NY, 1978.
48. A. Shohara, Source Coding Theorems for Information Networks, Ph.D. dissertation, University of California at Los Angeles, Tech. Rep. UCLA-ENG-7445, 1974.
49. J. K. Omura and K. B. Housewright, Source Coding Studies for Information Networks, *Proc. IEEE 1977 International Conference on Communications*, IEEE Press, 237–240, Chicago, Ill., June 13–15, 1977.
50. T. Berger et al., An upper bound on the rate-distortion function for source coding with partial side information at the decoder, *IEEE Trans. Inform. Theory* **IT-25**: 664–666 (1979).
51. A. Sgarro, Source coding with side information at several decoders, *IEEE Trans. Inform. Theory* **IT-23**: 179–182 (1977).
52. A. D. Wyner, The rate-distortion function for source coding with side information at the decoder — II: General sources, *Information and Control* **38**: 60–80 (1978).
53. H. Yamamoto, Source coding theory for cascade and branching communication systems, *IEEE Trans. Inform. Theory* **IT-27**: 299–308 (1981).
54. H. Yamamoto, Source coding theory for a triangular communication system, *IEEE Trans. Inform. Theory* **IT-42**: 848–853 (1996).
55. A. H. Kaspi and T. Berger, Rate-distortion for correlated sources with partially separated encoders, *IEEE Trans. Inform. Theory* **IT-28**: 828–840 (1982).
56. C. Heegard and T. Berger, Rate distortion when side information may be absent, *IEEE Trans. Inform. Theory* **IT-31**: 727–734 (1985).
57. T. Berger and R. W. Yueng, Multiterminal source encoding with one distortion criterion, *IEEE Trans. Inform. Theory* **IT-35**: 228–236 (March 1989).
58. T. Berger and R. W. Yueng, Multiterminal source encoding with encoder breakdown, *IEEE Trans. Inform. Theory* **IT-35**: 237–244 (1989).
59. A. Gersho and A. D. Wyner, The Multiple Descriptions Problem, Presented by A. D. Wyner, IEEE Information Theory Workshop, Seven Springs Conference Center, Mt. Kisco, NY, September 1979.
60. H. S. Witsenhausen, On source networks with minimal breakdown degradation, *Bell Syst. Tech. J* **59**: 1083–1087 (1980).
61. J. K. Wolf, A. D. Wyner and J. Ziv, Source coding for multiple descriptions, *Bell Syst. Tech.* **59**: 1417–1426 (1980).
62. L. H. Ozarow, On the source coding problem with two channels and three receivers, *Bell Syst. Tech. J* **59**: 1909–1922 (1980).
63. H. A. Witsenhausen and A. D. Wyner, Source coding for multiple descriptions II: A binary source, *Bell Syst. Tech. J.* **60**: 2281–2292 (1981).

64. A. El Gamal and T. M. Cover, Achievable rates for multiple descriptions, *IEEE Trans. Inform. Theory* **IT-28**: 851–857 (1982).

65. T. Berger and Z. Zhang, Minimum breakdown degradation in binary source encoding, *IEEE Trans. Inform. Theory* **IT-29**: 807–814 (1983).

66. R. Ahlswede, The rate-distortion region for multiple descriptions without excess rate, *IEEE Trans. Inform. Theory* **IT-31**: 721–726 (1985).

67. New results in binary multiple descriptions, *IEEE Trans. Inform. Theory* **IT-33**: 502–521 (1987).

68. Z. Zhang and T. Berger, Multiple description source coding with no excess marginal rate, *IEEE Trans. Inform. Theory* **IT-41**: 349–357 (1995).

69. W. E. Equitz and T. M. Cover, Successive refinement of information, *IEEE Trans. Inform. Theory* **IT-37**: 269–275 (1991). (See also W. E. Equitz and T. M. Cover, Addendum to Successive refinement of information, *IEEE Trans. Inform. Theory* **IT-39**: 1465–1466 (1993).

70. T. Berger, Z. Zhang and H. Viswanathan, The CEO problem, *IEEE Trans. Inform. Theory* **IT-42**: 887–903 (May 1996).

71. H. Viswanathan and T. Berger, The quadratic gaussian CEO problem, *IEEE Trans. Inform. Theory* **43**: 1549–1561 (1997).

72. Y. Oohama, The rate distortion problem for the quadratic gaussian CEO problem, *EEE Trans. Inform. Theory* (1998).

73. J. F. Hayes, A. Habibi, and P. A. Wintz, Rate-distortion function for a gaussian source model of images, *IEEE Trans. Inform. Theory* **IT-16**: 507–509 (1970).

74. T. Berger, S. Y. Shen, and Z. Ye, Some communication problems of random fields, *International Journal of Mathematical and Statistical Sciences* **1**: 47–77 (1992).

75. A decomposition theorem for binary markov random fields, *Annals of Probability* **15**: 1112–1125 (1987).

76. L. A. Bassalygo and R. L. Dobrushin, ε-Entropy of the random field, *Problemy Peredachi Informatsii* **23**: 3–15 (1987).

77. C. M. Newman, Decomposition of binary random fields and zeros of partition functions, *Annals of Probability* **15**: 1126–1130 (1978).

78. C. M. Newman and G. A. Baker, Decomposition of Ising Model and the Mayer Expansion, In S. Albeverio et al., eds., *Ideas and Methods in Mathematics and Physics — In Memory of Raphael Hoegh-Krohn* (*1938–1988*), Cambridge University Press, Cambridge, UK, 1991.

79. T. Berger and Z. Ye, ε-Entropy and critical distortion of random fields, *IEEE Trans. Inform. Theory* **IT-36**: 717–725 (1990).

80. Z. Ye and T. Berger, A new method to estimate the critical distortion of random fields, *IEEE Trans. Inform. Theory* **IT-38**: 152–157 (1992).

81. Z. Ye and T. Berger, *Information Measures for Discrete Random Fields*, Chinese Academy of Sciences, Beijing, 1998.

82. J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Inform. Theory* **IT-23**: 337–343 (1977).

83. J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Trans. Inform. Theory* **IT-24**: 337–343 (1978).

84. R. Pasco, Source Coding Algorithms for Fast Data Compression, Ph.D. dissertation, Stanford University, California, 1976.

85. J. Rissanen, Generalized kraft inequality and arithmetic coding, *IBM J. Res. Devel.* **20**: 198 (1976).

86. J. Rissanen, Universal coding, information, prediction and estimation, *IEEE Trans. Inform. Theory* **IT-30**: 629–636 (1984).

87. F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, The context-tree weighting method: Basic properties, *IEEE Trans. Inform. Theory* **IT-41**: 653–664 (1995).

88. F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, Context weighting for general finite-context sources, *IEEE Trans. Inform. Theory* **IT-42**: 1514–1520 (1996).

89. D. L. Neuhoff, R. M. Gray, and L. D. Davisson, Fixed rate universal block source coding with a fidelity criterion, *IEEE Trans. Inform. Theory* **21**: 511–523 (1975).

90. K. M. Mackenthum, Jr. and M. B. Pursley, Strongly and Weakly Universal Source Coding, In *Proc. 1977 Conference on Information Science and Systems*, 286–291, Johns Hopkins University, 1977.

91. M. B. Pursley and K. M. Mackenthum, Jr., Variable-rate source coding for classes of sources with generalized alphabets, *IEEE Trans. Inform. Theory* **IT-23**: 592–597 (1977).

92. K. M. Mackenthum, Jr. and M. B. Pursley, Variable-rate universal block source coding subject to a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-24**: 349–360 (1978).

93. H. H. Tan, Tree coding of discrete-time abstract alphabet stationary block-ergodic sources with a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-22**: 671–681 (1976).

94. J. Ziv, Coding of sources with unknown statistics — part II: Distortion relative to a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-18**: 389–394 (1972).

95. T. Hashimoto, *Tree Coding of Sources and Channels*, Ph.D. dissertation, Osaka University, Japan, 1981.

96. Y. Steinberg and M. Gutman, An algorithm for source coding subject to a fidelity criterion based on string matching, *IEEE Trans. Inform. Theory* **IT-39**: 877–886 (1993).

97. Z. Zhang and V. K. Wei, An on-line universal lossy data compression algorithm by continuous codebook refinement, *IEEE Trans. Inform. Theory* **IT-42**: 803–821 (1996).

98. Z. Zhang and V. K. Wei, An on-line universal lossy data compression algorithm by continuous codebook refinement, part two: Optimality for ϕ-mixing source models, *IEEE Trans. Inform. Theory* **IT-42**: 822–836 (1996).

99. I. Sadeh, Universal Compression Algorithms Based on Approximate String Matching, *Proc. 1995 IEEE International Symposium on Information Theory*, Whistler, British Columbia, September 17–22, 1995, p. 84.

100. E. H. Yang and J. Kieffer, Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm, *IEEE Trans. Inform. Theory* **IT-42**: 239–245 (1996).

101. E. H. Yang, Z. Zhang, and T. Berger, Fixed-slope universal lossy data compression, *IEEE Trans. Inform. Theory* **IT-43**: 1465–1476 (1997).

102. I. Kontoyiannis, An implementable lossy version of the Lempel-Ziv algorithm — part I: Optimality for memoryless sources, NSF Technical Report no. 99, Department of Statistics, Stanford University, April 1998.

103. I. Kontoyiannis, Asymptotically Optimal Lossy Lempel-Ziv Coding. To be presented at 1998 IEEE International Symposium on Information Theory, MIT, Cambridge MA, August 16–21, 1998.

104. W. B. Levy and R. A. Baxter, Energy efficient neural codes, *Neural Computation* **8**(3): 531–543 (1996).

105. W. B. Levy and R. A. Baxter, Energy efficient neural computation via quantal synaptic failures, *J. Neurosci.* **22**: 4746–4755 (2002).

106. T. Berger, "Living Information Theory," The Shannon Lecture, IEEE International Symposium on Information Theory, Lausanne, Switzerland, July 4, 2002. (Visit *http://www.itsoc.org* to download the slides of this presentation.)

REFLECTOR ANTENNAS

CAREY RAPPAPORT
Northeastern University
Boston, Massachusetts

1. APERTURE ANTENNA FUNDAMENTALS

1.1. Introduction

A critical component in a high-performance wireless telecommunications system is the antenna, which couples the signals carried by wires into and from waves propagating through space. For point-to-point communications links, specific antennas are used to constrain radiated power in a prescribed manner. In satellite and terrestrial applications, the antenna concentrates waves in an angular region of high intensity in desired directions and minimizes the power that would be wasted elsewhere. This key measure of antenna performance, the gain, depends directly on the effective antenna aperture area. Common high-gain antennas with large effective apertures include lenses, reflectors, and phased arrays.

The first two devices increase the effective aperture area by a purely geometric transformation, whereas the phased array electrically transforms the aperture. The lens antenna is a heavy and often complex structure, which, like the phased array, is (in most cases) frequency-dependent. The reflector antenna is the simplest, cheapest, and lightest alternative and has been the primary means of providing high-gain microwave beams for over half a century. With each type of antenna, the input signal flowing on cables is distributed in a prescribed way across the outer radiating surface. The particular phase and amplitude distribution of field (or equivalently, current density) at this aperture governs the radiation pattern shape of the antenna.

1.2. Far-Field Radiation Concepts

Most telecommunications applications involve links with distant stations. When the distance r to these stations is great compared with the antenna size D, $r \gg 2D^2/\lambda$, the stations are in the far-field of the antenna. The wavelength is given by $\lambda = c/f$, for frequency f and speed of light c. In the farfield, or Fraunhofer region, the spatial field distribution pattern is independent of the radial distance from the source. Indeed the electromagnetic field falls off as $\exp(-jkr)/kr$, with wavenumber $k = 2\pi f/c$, while the radiation pattern is a function of polar angle from boresight (direction the antenna is aimed) θ, and circumferential angle φ. The radiation pattern is proportional to the two-dimensional spatial Fourier transform of the aperture current phase and amplitude distribution [1–9]. Thus, the spatial current distribution across a finite aperture $A(x, y)$ is mapped to the angular radiation pattern $P(k_x, k_y)$, with angular wavenumbers, $k_x = k\cos\theta\cos\varphi$, and $k_y = k\cos\theta\sin\varphi$. Clearly, a larger aperture (bigger antenna) produces a higher-intensity beam, with a narrower beamwidth. A rectangular aperture of dimension $2a$ by $2b$ — illuminated by a wave with uniform amplitude and constant phase — produces a beam proportional to $\sin(k_x a)/(k_x a) \bullet \sin(k_y b)/(k_y b)$, while a circular aperture of radius R produces a radiation pattern proportional to $J_1(u)/u$, where $u = k_\rho R$, $k_\rho = \sqrt{(k_x^2 + k_y^2)}$ and J_1 is the first-order Bessel function of first kind. This radiation pattern of radiated power as a function of angle is shown in Fig. 1 for a 20-wavelength-diameter aperture. Figure 2 displays the same pattern in 3D polar format. The antenna pattern is very sensitive to the current distribution at the aperture. Beam direction, maximum gain, power outside the mainlobe, and associated sidelobe levels depend strongly on the aperture phase function [10,11].

The gain of an antenna is the measure of field intensity in a particular direction relative the average intensity over all directions. The peak gain, or directivity, specifies the ratio of the power density at boresight in the middle of the antenna's main beam at a distance r to the average power density (total radiated power divided by $4\pi r^2$). An aperture antenna with area A has maximum peak gain when the current amplitude and phase are uniform across the aperture. Such an aperture is referred to as having 100% illumination efficiency. In this case the peak gain is

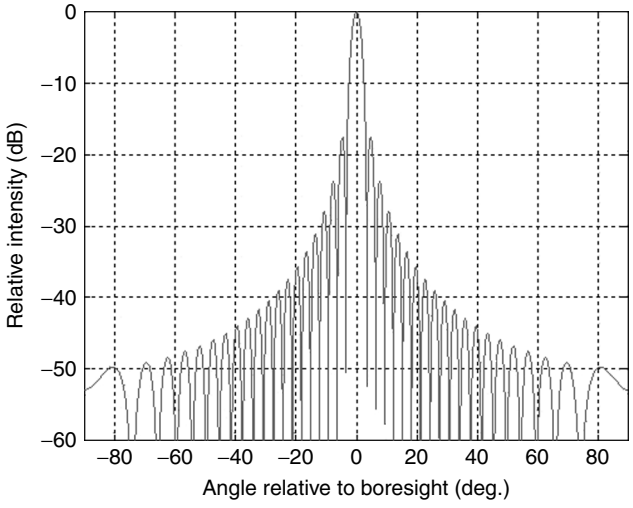


Figure 1. Far-field radiation pattern of an ideal, uniformly illuminated 20-wavelength-diameter circular aperture.

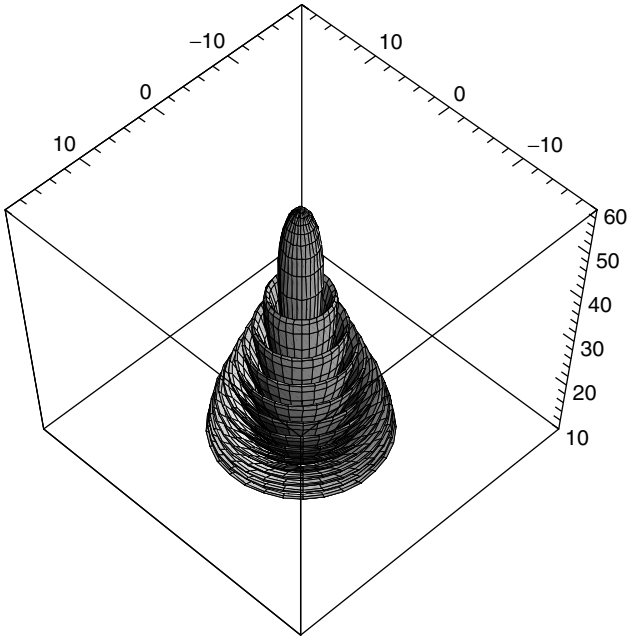


Figure 2. Three-dimensional polar plot of the far-field antenna pattern of field intensity (in dB) produced by a uniformly illuminated 20-wavelength-diameter circular aperture.

given by [2]

$$G = \frac{4\pi A}{\lambda^2}$$

and for circular apertures of diameter D

$$G = \left(\frac{\pi D}{\lambda}\right)^2$$

or in decibels

$$G(dB) = 20 \log \frac{\pi D}{\lambda}$$

Doubling the antenna diameter — or doubling its frequency of operation — leads to a 6-dB increase in antenna gain (as long as no aberrations are introduced in the electrically larger aperture).

The 3-dB, or half-power, antenna beamwidth can be approximated for large apertures as [1]

$$BW = \lambda / D$$

and the beamwidth between first nulls is double this.

The first sidelobe level is another important antenna specification, as it gives a measure of crosstalk intensity outside the main beam. For circular apertures it is −17.3 dB, and for rectangular apertures it is −13.5 dB [9]. By altering the amplitude distribution, it is possible to reduce these first sidelobe levels; but in doing so, the peak gain decreases as well. For example, a same-size rectangular aperture with a triangular amplitude distribution in the x direction could be thought of as the spatial convolution of two half-size uniform amplitude distributions, which would produce a radiation pattern with dependence $\sin^2(k_x a/2)/(k_x a/2)^2 \bullet \sin(k_y b)/(k_y b)$. The first sidelobe level in the x direction is lowered by 6 dB to −19.5 dB, but since the effective area is halved, the gain is also reduced by 3 dB.

Variations in the current phase across the aperture have an even greater effect on the radiation pattern than amplitude variations. A linear phase variation steers or "scans" the beam in the direction perpendicular to a plane of constant phase, without changing the beam pattern shape. Any quadratic or higher-order phase variations across the aperture are aberrations and lead to beam pattern degradation. It is essential to understand the effects of phase aberrations when designing antenna systems [12].

The simplest aperture antennas provide a single high-gain beam in the boresight direction. More sophisticated antennas produce specifically tailored beam shapes, or multiple simultaneous or independently excited beams. To maintain good beam quality, each radiated beam must have only linear phase variation. Although maintaining a flexible linear phase function while minimizing higher-order optical aberration terms is generally difficult, a phased-array antenna can accomplish these goals fairly simply by applying the required phase shifts to successive elements. With enough elements, a linear variation can be approached as closely as desired.

For single fixed-beam antennas, the aperture conditions depend only on the reflector surface placement and geometry. There is only one source input, the antenna feed, with fixed position and orientation. However, a scanning antenna must be able to produce a variable linear phase distribution that depends on varying source conditions.

One important issue with reflector antennas is blockage of the aperture with the feed or subreflector structure. There is no blockage with the transmission-based lens antennas, because the feed structure is on the other side of the antenna from the radiating aperture. However, surface impedance-matching requirements and dissipative transmission losses, as well as weight and bandwidth limitations, render lenses inferior to reflectors in most satellite applications.

2. REFLECTOR DESIGN FUNDAMENTALS

2.1. Geometric Optics Analysis

While the radiation pattern of a reflector antenna is determined by diffraction analysis of the fields across its radiating aperture, the conventional method for reflector shape synthesis and optimization makes use of geometric optics. Geometric optics follows from the eikonal equation [13,6] $|\nabla L(\boldsymbol{r})|^2 = n^2$, which is the lowest order (infinite frequency) approximation of wave equation, in which all field variation is assumed to be contained in the phase $\Phi = \omega/c\ L(r)$. This approximation assumes that all waves in uniform media propagate along straight rays, where each wavefront is perpendicular to these rays, occurring at the same distance from a given source. The rays are reflected by metal reflector surfaces as if they encountered piecewise perfectly conducting planes with the same surface normal at the reflection point [14–17]. Geometric optics cannot be used to find the

far-field radiation pattern of a well-formed beam, because this pattern is entirely determined by diffraction [18,19]. However, since a focused reflector antenna must receive plane waves from distant sources, it can be thought of as a single surface — or collection of multiple surfaces — that converts parallel incoming rays into rays that converge on a feed element. Conversely, when transmitting, reflectors *collimate* rays by ideally converting spherically diverging rays from a point-source feed to a set of parallel rays. Furthermore, the pathlength from the source point (or any starting wavefront) to any given wavefront along each ray must be the same regardless of one or more redirections of the rays by reflector surfaces.

The behavior of rays incident on metal surfaces is governed by Snell's law of reflection, which ensures that incident and reflected rays and the surface normal at the point of intersection are coplanar, and that the angle of incidence equals the angle of reflection. Snell's law and the constant-pathlength condition are sufficient to synthesize a focused reflector system. In general, two additional constraints are applied to the reflector surface synthesis formulation: surface continuity and field amplitude concentration.

Although phased arrays and some lens antennas divide the aperture into separate physical regions [20–22], reflectors tend to be continuous surfaces, with few steps or cusps. Discontinuities in the surface or the surface normal lead to diffraction effects that are unpredictable with geometric optics. Thus, the surface reflection point and its normal are coupled, and specifying the pathlengths of the set of reflected rays along with their directions greatly restricts the family of focusing surfaces. For a single reflector, only the circular paraboloid satisfies the Snell's law, constant-pathlength, and continuous-surface constraints. For multireflector systems, Snell's law must be specified on each reflector, and the pathlength condition applies to the full path from source to aperture plane. In designing reflector antenna systems, usually a two-dimensional profile is generated first; then this cross section is rotated about the central system axis line of symmetry.

The other useful addition to geometric optics analysis is the assumption that each ray represents a constant differential power flow [9]. The power within a bundle of rays must remain constant throughout the optical system. Power density on any two wavefronts is inversely proportional to the ratio of areas on those wavefronts bounded by a given set of rays. This property of conservation of energy is useful in many synthesis applications.

Additional parameters must be specified if the energy-conservation constraint is imposed. This constraint merely describes how the feed phase and amplitude distribution are transformed to an aperture distribution. Thus it is necessary to specify the input and output distributions. In many cases discussed in the literature, the input is assumed to be a tapered feed pattern with amplitude of the form $\cos^n \theta$, and with a spherical phase distribution [4]. In many cases, attempts are made to ensure that the aperture has high illumination efficiency, with the output field as close as possible to a uniformly amplitude plane wave. Occasionally the exact pathlength constraints are relaxed in favor of a particular aperture amplitude distribution.

Many attempts have been made to alter the geometry of dual reflectors to improve their performance [23–32]. Unlike with the single reflector, there are an infinite number of pairs of surfaces that produce a plane-wave output for spherical wave input at the focus. Although it is possible to generate a symmetric dual reflector with various arbitrary phase and amplitude distributions at the aperture [33,34], most efforts have been directed toward simply improving the main reflector illumination efficiency.

2.2. Single Reflectors

The simplest reflector antenna systems are single reflectors with parabolic cross section, as shown in Fig. 3. Often referred to as *prime focus reflectors*, these paraboloidal reflectors are specified entirely by their focal length F, aperture diameter D, and vertex position z_0. The paraboloid shape is given by [35]

$$z = \frac{x^2 + y^2}{4F} + z_0$$

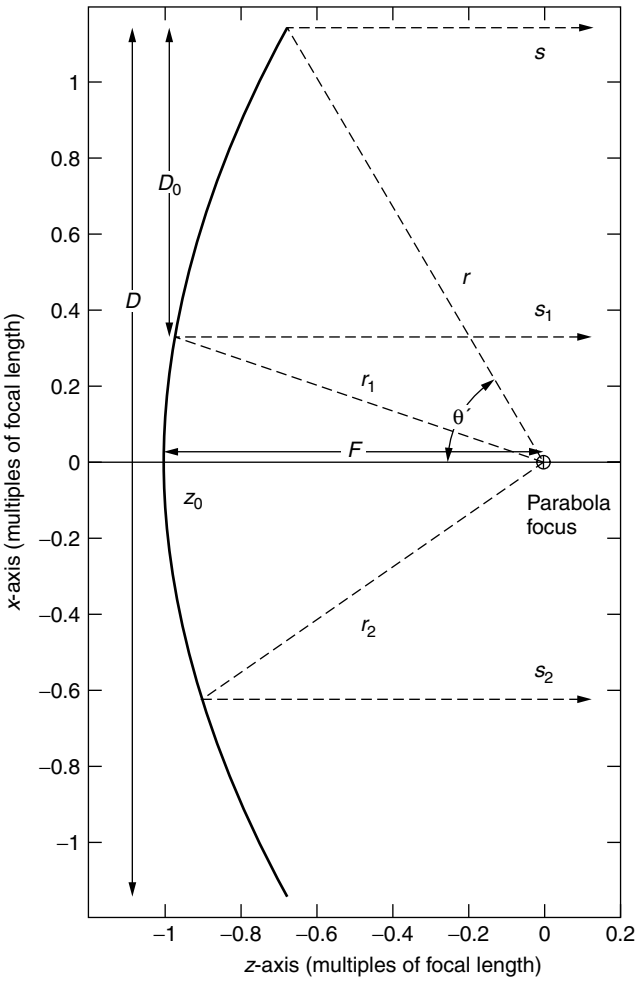


Figure 3. Parabolic single-reflector geometry, with offset section shown.

or equivalently, in terms of distance from the focus to the reflector surface r and angle from the negative axis of symmetry θ' [7,8]:

$$r = \frac{2F}{1+\cos\theta'} = F\sec^2\frac{\theta'}{2}$$

The feed is positioned at the parabola focus, facing the paraboloid at a distance F from its vertex. The pathlength condition demands that for any combination of rays, the total distance to the planar wavefront $r+s$, or r_1+s_1, or r_2+s_2 be the same. Since the feed structure lies in the path of outwardly reflected rays, it tends to block the central portion of the aperture.

An important characteristic of this single-reflector antenna is the ratio of focal length to diameter F/D [also known as the *f number* (or *f stop*; aperture) of camera lenses]. Reflectors with smaller F/D have greater surface curvature, and are more physically compact, with closer feed structures than those with larger F/D ratios, but are more sensitive to the precise feed position and beam pattern.

Offset single reflectors avoid the blockage problem by using only a portion of a larger symmetric paraboloid that reflects unblocked rays [36,37]. In Fig. 3, a possible offset parabolic section would be bounded by rays s and s_1, and have diameter D_0. While offset designs enhance both radiated power and prevent some beam distortion, the lack of perfect symmetry introduces differential effects for waves polarized parallel and perpendicular to the offset direction.

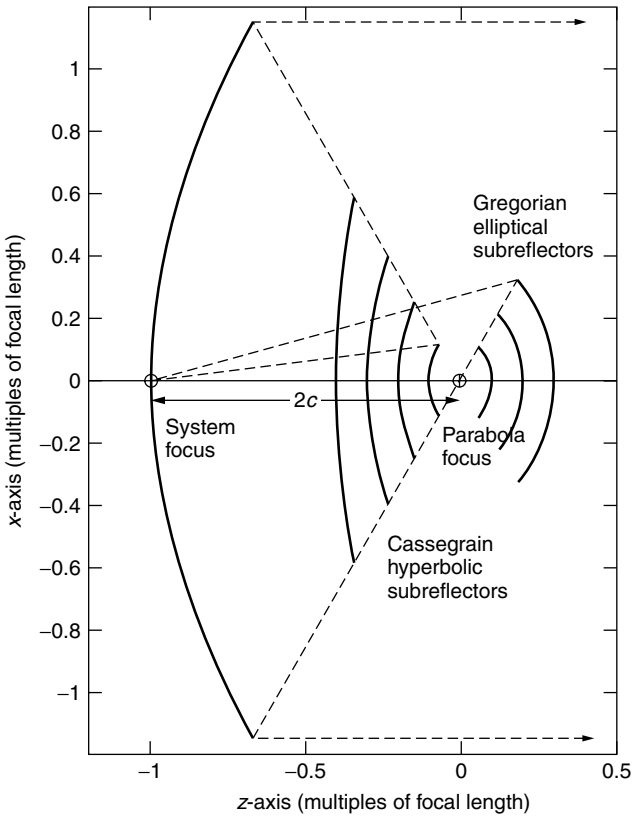


Figure 4. Cassegrain and Gregorian dual-reflector profiles with sample ray paths.

2.3. Multiple Reflectors

Conventional dual reflectors are of two major types: the Cassegrain and the Gregorian. Based on the optical telescope designs first introduced in 1672, they consist of a paraboloidal main reflector and, respectively, either a hyperboloidal or ellipsoidal secondary or subreflector [38]. As long as the paraboloid focus coincides with one hyperboloid (ellipsoid) focus, and the feed is positioned at the other hyperboloid (ellipsoid) focus, the dual reflector will collimate outgoing rays. For a given paraboloidal main reflector, there are infinitely many possible subreflectors with differing size and position relative to the main reflector. A family of possible Cassegrain and Gregorian subreflector profiles for a parabola with unit focal length and system focus at the parabola vertex is shown in Fig. 4. All rays originating at the system focus reflect from a subreflector as if they had originated at the common focal point. Since the difference of segments from each hyperbola focus to a point on the hyperbola is constant and equal to the major hyperbola axis length $2a$, the pathlength from feed to an aperture plane is again constant, $2a$ greater than for the main reflector if it were used as a prime focus single reflector. The same argument applies for the ellipse, with constant sum of segments from each focus equal to $2a$.

The equation for the subreflector profile is given by

$$\frac{(z-z_c)^2}{a^2} + \frac{x^2+y^2}{c^2-a^2} = 1$$

where the central point is found by associating the common foci $z_c = z_0 + F - c$. The equivalent polar form of the subreflector equation is [5]

$$r_s = \frac{c^2-a^2}{a+c\cos\theta'}$$

where the distance between subreflector foci $2c = F$ for the profiles of Fig. 4, with positive values of r_s corresponding to hyperbolas and negative values corresponding to ellipses.

It should be noted that the system focus need not be positioned at the parabola vertex. The subreflector size to chosen so that the extreme ray from feed to the edge of the paraboloid reflects from its edge. Generally, subreflectors are kept as small as possible to minimize blockage without requiring too narrow a taper of the feed radiation pattern.

Dual-reflector systems offer an improvement over paraboloidal single reflectors in both packaging efficiency and design flexibility. Because the subreflector redirects rays to the main reflector, the actual length involved in the system is much less than its equivalent focal length. Also, when a subreflector is used, the feed can be positioned near or behind the main reflector, so that feedlines are shortened, and the placement of the associated electronics is simplified. Servicing and maintenance of the feed is simplified as well. Another mechanical advantage is the ease of supporting a relatively lightweight subreflector.

There are several disadvantages to symmetric (nonoffset) dual reflectors. The subreflector blocks the center

of the aperture, thereby reducing efficiency and increasing the sidelobe levels. The fact that two surfaces must intercept the rays also introduces the requirement for minimizing power missing the surfaces (spillover) while keeping the illumination efficiency high. Alignment is crucial in dual reflectors. Unlike single reflectors, where the feed concentrates power over the wide angle subtended by the main reflector, the dual-reflector feed must illuminate a much smaller subreflector. Thus larger feedhorns are required, and their placement must be precise. Also, the subreflector placement relative to the main reflector is critical [39–42]. Larger feedhorns are less preferable for multibeam antennas, since their size may prevent their phase centers from being close enough to generate adjacent component beams.

With symmetric dual reflectors, subreflector blockage can be regarded as the removal of a circle of uniform, constant-amplitude power from the center of the aperture. Thus one subtracts a lower intensity, much wider $J_1(u')/u'$ pattern from the original far-field amplitude pattern. This effect lowers the mainlobe power level, lowering gain and increasing sidelobe magnitude. Power spilling past the subreflector is sufficiently large and sufficiently close to the axis to also become apparent on the combined pattern. With single reflectors, the spillover is almost always at least 150° away from the boresight axis, so its effect is negligible for communications antennas.

Methods of making the subreflector less obstructive have been proposed [43–47], such as serrating or perforating the surface; or fabricating it from a linearly polarized material, illuminating it with radiation of the orthogonal polarization, and using a twist reflection material for the main reflector. A much more feasible concept is the offset configuration, which consists of a (usually circular) section of the main reflector and the corresponding section of the subreflector of a symmetric dual reflector. Blockage is eliminated, and the VSWR (reflection back into the feed) is reduced, but the focal length : diameter ratio (F/D) increases by the ratio of the parent main reflector diameter to the offset section diameter. The lack of symmetry also tends to increase cross-polarization and make analysis much more difficult.

Offset designs often make use of the Gregorian configuration. In the symmetric case, having a concave subreflector increases the reflector separation without much improvement in performance. For an offset geometry, however, the Gregorian subreflector can be placed entirely below the antenna axis, where it can still illuminate the entire main reflector above the axis. Blockage is avoided while still using a large offset section of the main reflector. By properly tilting the axis of the ellipsoidal subreflector with respect to the main reflector focal axis, cross-polarization can be significantly reduced [48,49].

Several methods are used for analyzing Cassegrain antennas. One very powerful approximate technique is that of defining the equivalent parabola (Fig. 5), which has the same aperture and feed illumination angle as the Cassegrain [50,51]. The equivalent focal length can be found using the maximum subreflector angle θ_m to be

$$F_e = F\frac{c+a}{c-a}$$

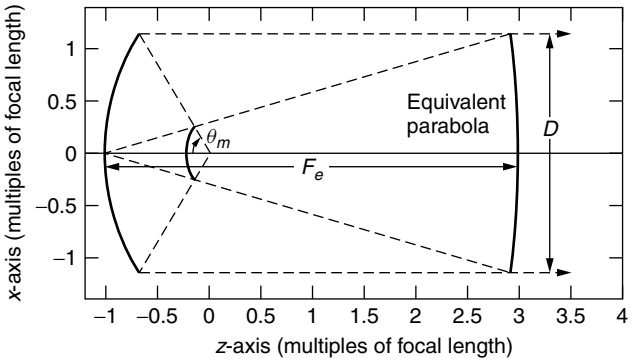


Figure 5. Equivalent parabola for Cassegrain dual reflector.

The factor $m = (c + a)/(c - a)$ is called the *magnification factor*. The F/D ratio of the equivalent parabola is m times larger than the physical Cassegrain system. The equivalent parabola concept can be used to determine the amplitude distribution across the aperture by replacing the real dual-reflector system with a single reflector.

Scattering theory and geometric theory of diffraction have been used for a more detailed analysis of dual reflectors [40–42,52–54]. The mathematics is complicated, and numerical methods are often required. Dual reflectors have less frequency dependence because of their relatively long effective focal length for their given packaging volume. Also, illumination efficiency is directly dependent on the system's geometry, and this tends to have the dominant effect in scanning.

2.4. Dual-Reflector Shaping

Much effort has been spent on dual-reflector designs that keep one of the standard Cassegrain surfaces fixed (usually the main reflector) and shape the other. Green first showed that by shaping the subreflector and then by making small alterations to the main reflector to preserve the planar phase front, uniform illumination could be attained [28]. The resulting beam is not exactly focused, but the improvement in performance is appreciable. Other systems consider subreflectors that correct for main reflector deformations or errors, or generate particular aperture distributions and polarizations [29,32,55,57].

Simultaneous shaping of both surfaces of dual-offset reflectors has been explored [26,27,34,35,43–47]. The problem involves solving a pair of partial nonlinear differential equations. These simplify to ordinary differential equations for symmetric systems. Approximate methods of solving the equations show reasonable numerical results [46], but the processes involve considerable trial and error and significant amounts of numerical computation. However, Westcott et al. [47], claim to have arrived at an exact solution.

3. BEAM SCANNING WITH REFLECTOR ANTENNAS

3.1. Analysis of Reflector Beam Scanning

For reflectors and lenses, beam steering, or "scanning" is accomplished by transversely displacing the feed from

the antenna focus, or unscanned, aberration-free source position [58,59]. For systems with a single perfect focal point, this displacement always produces higher-order phase terms. The absence of higher-order terms at any other position would imply the existence of a second perfect focal point there.

Scanning is of great concern for antennas used in satellite applications. When high resolution is important, as is the case with point-to-point networking and direct broadcast communications, the antenna must provide a large number of high-gain, well-formed spot beams over its entire field of view [60–67]. Well-formed beams are also required in multibeam antennas. High-efficiency shaped coverage contours are useful for illuminating specific geographic regions. Because they are produced by coherently combining component beams, each beam must have low-intensity sidelobes and well-defined nulls to prevent undesirable interference.

The optical aberrations, caused by lateral feed displacement, cause an asymmetric effect in the plane of scan [13–15,68]. The beam broadens, the first null in the scanned direction fills, and the first sidelobe on the axial side, the coma lobe, rises. Several authors have addressed the problem of finding the best focal locus of a paraboloidal antenna [58,59,61,69]. The most general solution appears to be given in Balling [62].

For small scan angles, the equivalent parabola concept is very useful [51] for predicting the performance degradation in dual reflectors. The scanning tradeoff analysis of dual reflectors is much more complex than it is for paraboloids. The effective focal length increases as subreflector curvature increases; however, the subreflector is smaller and spillover is of greater concern. For larger feed displacements and scan angles, spillover and distorted imaging prevents acceptable modeling of the dual reflector as an equivalent paraboloid [70,71]. Choosing the optimum focal point for a particular scan angle is strongly dependent on the particular Cassegrain geometry. Changing the surface curvature of either reflector surface will greatly change the position of the focal surface.

Scanning with an offset antenna system is slightly different from that in the symmetric case. The optimum focal locus is no longer symmetrically disposed about the antenna axis. Instead, the best focal positions for scanned beams, to the first approximation, are on the plane perpendicular to the offset axis [36,37,64,72]. The precise feed position depends on spillover and illumination efficiency as well as pathlength and phase error considerations.

3.2. Reflectors Designed for Scanning

Several antennas have been developed specifically for efficient scanning applications. The spherical cap [73] is the simplest scanning antenna. With a source positioned approximately halfway between the sphere and its center, the surface resembles a paraboloid over a limited angular region. In fact, if the parabola's curvature value at its vertex is inverted and used as the radius of the osculating circle at that vertex, this circle corresponds to the profile of the spherical cap. Several feeds can be positioned on the spherical focal surface (with radius $F/2$), each generating a beam in a different scanned direction.

Because it is not a perfect paraboloid, the spherical cap introduces phase errors in the reflected wave in the form of spherical aberration. This optical aberration increases with larger angular illumination from the feed or with higher frequencies. Since the symmetry of the sphere produces identical patterns for all scan directions, the entire reflector is underilluminated for wide-field-of-view ($>5^\circ$) applications; that is, only a fraction of the reflector is illuminated by each feed for any given beam. A compromise is made frequently between scanning perfection and reflector inefficiency.

Subreflector correctors have been designed to eliminate the spherical aberration in the cap [74]. The Gregorian type corrector makes use of caustics formed by rays reflected by the spherical surface to collimate incoming rays. The corrector applies only for a single scan direction; therefore, it must be moved mechanically if the beam is to be redirected. This characteristic makes scanning arrays impractical and shaped beams impossible for corrected spherical caps.

One improvement to the basic spherical cap is the torus [3,75]. The torus is a double-curved surface with a circular profile in the plane of scan and a parabolic profile in the orthogonal plane. Reflected waves suffer from spherical aberration in the scan plane, but are focused without aberration in the orthogonal plane. The torus can scan in only one plane, so it is well suited for communication with multiple stations situated on a single arc, such as a group of satellites in geostationary orbit. Like the spherical cap, the torus has poor illumination efficiency in the scan direction.

An alternative to the torus is a specifically shaped single-reflector surface that balances the aperture phase errors with illumination efficiency [76,77]. This surface is described by a polynomial whose coefficients minimize the variations across a symmetric pair of specified subapertures of the pathlengths from a symmetric pair of specified focal points to a pair of scanned aperture planes. In addition to minimizing the optical aberrations for these extreme scan angles, a third beam direction — that of boresight — is considered as well. The feed is positioned along the axis of symmetry, at the point that minimizes phase errors across a central subaperture, and the surface polynomial coefficients are adjusted to reduce the errors for the boresight beam without greatly worsening the errors for the scanned beams. It has been shown that for equivalent beam quality across a 60° wide field of view, this type of reflector can be made about 40% smaller than the torus.

Dual offset reflectors have been designed specifically for scanning and multiple beam applications [60,63–67,78–85]. Acceptable multibeam systems are achievable for large F/D conventional Cassegrain systems with oversized main and subreflectors. Unconventional designs include the confocal paraboloid [64,79], which has a feed array in the near field of a small offset parabolic subreflector with focal point coinciding with that of a parabolic main reflector. This type of antenna has been shown to scan effectively up to 3°.

The bifocal dual reflector [82] is a totally redesigned antenna system, with two perfect focal points, one for each extreme scanned beam. Since it consists of two reflectors, each tracing of rays through the system has 2 degrees of freedom. The two Snell's law conditions allow for two independent collimating pathlength conditions. The design principle for the bifocal is based on this idea by insisting that each point on each reflector will lie on the path from one focal point to its corresponding aperture plane, as well as on the path from the second point to its aperture plane. First, the focal points are selected, an initial subreflector point and it normal are chosen, and the total pathlength to the first scanned aperture plane is specified. Ray tracing uniquely determines the corresponding main reflector point and its normal. Next, this main reflector point is used with ray tracing from the second focal point to the second scanned aperture to determine the subsequent subreflector point. This process continues until a full set of reflector profile points are generated. These points are joined by spline fitting, and then the profiles are rotated about the axis of symmetry to generate a pair of reflector surfaces. The focal points are also smeared out into a focal ring, which unfortunately introduces aberrations.

The need for an oversized subreflector to simultaneously redirect both positively and negatively scanned rays to the main reflector causes severe blockage in the symmetric bifocal. The offset bifocal reflector antenna system [83] overcomes this deficiency by choosing asymmetric focal points, limiting the illuminated sections of the main reflector to only an unblocked aperture, and synthesizing each entire reflector surfaces from the ray-tracing procedure rather than by rotating the profiles about the axis of symmetry. The offset bifocal surfaces only approximate the continuous surface condition, but the resulting dual-reflector system produces insignificant phase errors. Performance results indicate that it is possible to design a compact dual-reflector system that can radiate well-formed 1° beams across the 17° field of view subtended by the earth as seen from geostationary orbit.

One last type of scanning dual-reflector antenna system incorporates the ideas of the shaped single reflector [76,77] with a mechanically tilting subreflector [84–86]. This configuration is driven by the need to minimize antenna cost, by using only a single feed along with a single front end (amplifier, filter, polarizer), yet providing the capability to scan a single beam across a wide field of view. Having the smaller subreflector rotate instead of the entire antenna structure provides mechanical and cost advantages for mass-market communication systems such as direct broadcast television. The subreflector can be planar or specially shaped [86] for additional packaging benefits.

BIOGRAPHY

Carey M. Rappaport (M., S.M. 1996) received five degrees from the Massachusetts Institute of Technology: the S.B. in Mathematics, the S.B., S.M., and E.E. in Electrical Engineering in June 1982, and the Ph.D. in Electrical Engineering in June 1987. He is married to Ann W. Morgenthaler and has two children, Sarah and Brian.

Professor Rappaport has worked as a teaching and research assistant at MIT from 1981 until 1987, and during the summers at COMSAT Labs in Clarksburg, Maryland, and The Aerospace Corp. in El Segundo, California. He joined the faculty at Northeastern University in Boston, in 1987. He has been Professor of Electrical and Computer Engineering since July 2000. During fall 1995, he was Visiting Professor of Electrical Engineering at the Electromagnetics Institute of the Technical University of Denmark, Lyngby, as part of the W. Fulbright International Scholar Program. He has consulted for Geo-Centers, Inc., PPG, Inc., and several municipalities on wave propagation and modeling, and microwave heating and safety. He is Principal Investigator of an ARO-sponsored Multidisciplinary University Research Initiative on Demining and Co-Principal Investigator of the NSF-sponsored Center for Subsurface Sensing and Imaging Systems (CenSSIS) Engineering Research Center.

Professor Rappaport has authored over 190 technical journal and conference papers in the areas of microwave antenna design, electromagnetic wave propagation and scattering computation, and bioelectromagnetics, and has received two reflector antenna patents, two biomedical device patents, and three subsurface sensing device patents. He was awarded the IEEE Antenna and Propagation Society's H.A. Wheeler Award for best applications paper, as a student in, 1986. He is a member of Sigma Xi and Eta Kappa Nu professional honorary societies.

BIBLIOGRAPHY

1. S. Silver, ed., *Microwave Antenna Theory and Design*, Dover Publications, New York, 1965.
2. D. Staelin, A. Morgenthaler, and J. Kong, *Electromagnetic Waves*, Prentice-Hall, Englewood Cliffs, NJ, 1994.
3. R. E. Collin and F. J. Zucker, *Antenna Theory*, McGraw-Hill, New York, 1969.
4. J. D. Kraus and R. Marhefka, *Antennas*, McGraw-Hill, New York, 2002.
5. R. S. Elliot, *Antenna Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
6. J. Kong, *Electromagnetic Wave Theory*, Wiley, New York, 1999.
7. W. Stutzman and G. Thiele, *Antenna Theory Design*, Wiley, New York, 1981.
8. C. Balanis, *Antenna Theory Analysis and Design*, Wiley, New York, 1998.
9. R. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985.
10. D. K. Cheng, Effect of arbitrary phase error on the gain and beamwidth characteristics of radiation pattern, *IRE Trans. Antennas Propag.* **AP-3**: 145–147 (July 1965).
11. T. B. Vu, The effect of phase errors on the forward gain, *IEEE Trans. Antennas Propag.* **AP-13**: 981–982 (Nov. 1965).
12. K. S. Kelleher, *Antenna Wavefront Problems*, Naval Research Lab., Washington, DC, Sept. 1949.

13. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, New York, 1970.
14. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill, New York, 1957.
15. A. E. Conrady, *Applied Optics and Optical Design, Part 1*, Dover Publications, New York, 1957.
16. P. S. Holt, *Application of Geometric Optics to the Design and Analysis of Microwave Antennas*, AFCRL, Bedford, MA, AFCRL-67-0501, Sept. 1967.
17. G. A. Fry, *Geometrica1 Optics*, Chilton, Philadelphia, 1969.
18. S. Cornbleet, *Microwave Optics*, Academic Press, New York, 1976.
19. R. S. Longhurst, *Geometrica1 and Physical Optics*, Wiley, New York, 1967.
20. F. S. Holt and A. Mayer, A design procedure for dielectric microwave lenses of large aperture ratio and large scanning angle, *IRE Trans. Antennas Propag.* 25–30 (Jan. 1957).
21. R. M. Brown, Dielectric bifocal lenses, *IRE National Convention Record*, 1956, pp. 180–187.
22. C. Rappaport and A. Zaghloul, Optimized three dimensional lenses for wide-angle scanning, *IEEE Trans. Antennas Propag.* 1227–1236 (Nov. 1985).
23. S. P. Morgan, Some examples of generalized Cassegrainian and Gregorian antennas, *IEEE Trans. Antennas Propag.* **AP-12**: 685–691 (Nov. 1964).
24. P. Brickell and B. S. Westcott, Reflector design as an initial-value problem, *IEEE Trans. Antennas Propag.* (Communication) **AP-24**: 531–533 (July 1976).
25. B. S. Westcott and A. P. Norris, Reflector synthesis for generalized far fields, *J. Phys. A, Math. Nucl. Gen.* **8**: 521–532 (1975).
26. G. W. Collins, Shaping subreflectors in Cassegrainian antennas for maximum aperture efficiency, *IEEE Trans. Antennas Propag.* **AP-21**: 309–313 (May 1973).
27. W. F. Williams, High efficiency antenna reflector, *Microwave J.* **8**: 79–82 (July 1965).
28. K. A. Green, Modified Cassegrain antenna for arbitrary aperture illumination, *IRE Trans. Antennas Propag.* (Communication) **AP-11**: 589–590 (Sept. 1963).
29. T. Kitsuregawa and M. Mizusawa, Design of the shaped reflector Cassegrainian antenna in consideration of the scattering pattern of the subreflector, *IEEE Group; Antennas and Propagation, Int. Symp. Digest*, Sept. 9–11, 1968, pp. 391–396.
30. P. Rouffy, Design of dual reflector antennas, *URSI. 1968 Symp. Digest*, Sept. 10–12, 1968, p. 88.
31. S. Von Hoerner, The design of correcting secondary reflectors, *IEEE Trans. Antennas Propag.* **AP-24**: 336–340 (May 1976).
32. M. O. Millner and R. H. T. Bates, Design of subreflectors to compensate for Cassegrain main reflector deformations, *Proc. IEEE, Microwaves, Optics and Antennas*.
33. V. Galindo, Design of dual reflector antennas with arbitrary phase and amplitude distributions, *IEEE Trans. Antennas Propag.* **AP-12**: 403–408 (July 1964).
34. B. Y. Kinber, On two reflector antennas, *Radio Eng. Electron. Phys.* **6**: 914–921 (June 1962).
35. W. V. T. Rusch and P. O. Potter, *Analysis of Reflector Antennas*, Academic Press, New York, 1970.
36. P. G. Ingerson and W. C. Wong, Focal region characteristics of offset fed reflectors, *1974 Int. IEEE/AP-S Symp. Program and Digest*, June 10–12, 1974, pp. 121–123.
37. A. W. Rudge and N. A. Adatia, Offset-parabolic-reflector antennas: A review, *Proc. IEEE* **66**: 1592–1618 (Dec. 1978).
38. P. W. Hannan, Microwave antenna derived from the Cassegrain telescope, *IRE Trans. Antennas Propag.* **AP-9**: 140–153 (March 1961).
39. A. M. Isber, Obtaining beam pointing accuracy with Cassegrain antennas, *Microwaves* 40–44 (Aug. 1967).
40. W. V. T. Rusch, Scattering from a hyperboloidal reflector in a Cassegrainian feed system, *IEEE Trans. Antennas Propag.* **AP-11**: 414–421 (July 1963).
41. P. D. Potter, Application of spherical wave theory to Cassegrainian-fed paraboloids, *IEEE Trans. Antennas Propag.* **AP-15**: 727–736 (Nov. 1967).
42. P. D. Potter, Aperture illumination and gain of a Cassegrain system, *IEEE Trans. Antennas Propag.* **AP-71**: 373–375 (May 1963).
43. G. Bjontegaard and T. Pettersen, A shaped offset dual reflector antenna with high gain and low sidelobe levels, *IEE 2nd Int. Conf. Antennas and Propagation*, April 1981, York, UK, pp. 163–167.
44. R. Mittra and V. Galindo-Israel, Shaped dual reflector synthesis, *IEEE Antennas Propag. Newsl.* 5–9 (Aug. 1980).
45. V. Galindo-Israel, R. Mittra, and A. Cha, Aperture amplitude and phase control of offset dual reflectors, *IEEE Trans. Antennas Propag.* **AP-27**: 159–164 (March 1979).
46. J. J. Lee, L. I. Parad, and R. S. Chu, Shaped offset-fed dual reflector antenna, *IEEE Trans. Antennas Propag.* **AP-27**: 165–171 (March 1979).
47. B. S. Westcott, F. A. Stevens, and P. Brickell, GO synthesis of offset dual reflectors, *IEEE Proc.* **128**: 11–18 (Feb. 1981).
48. T. Mizuguchi, M. Akagawa, and H. Yokoi, Offset Gregorian antenna, *Electron. Commun. Jpn.* **61-B**(3): 58–66 (1978).
49. M. Tanaka and M. Mizusawa, Elimination of cross polarization in offset dual-reflector antennas, *Electron. Commun. Jpn.* **58-B**(12): 71–78 (1975).
50. W. D. White and L. K., DeSize, Focal length of a Cassegrain reflector, *IRE Trans. Antennas Propag.* **AP-9**: 412 (Jan. 1961).
51. W. C. Wong, On the equivalent parabola technique to predict the performance characteristics of a Cassegrainian system with an offset feed, *IEEE Trans. Antennas Propag.* **AP-21**: 335–339 (May 1973).
52. L. K. DeSize, D. J. Owen, and G. K. Skahill, *Investigation of multibeam antennas and wide-angle optics*, Airborne Instruments Laboratory Report 7358–1, Jan. 1960.
53. O. Sorensen and W. V. T. Rusch, Application of the geometric theory of diffraction to Cassegrain subreflectors with laterally defocused feeds, *IEEE Trans. Antennas Propag.* **AP-73**: 698–701 (Sept. 1975).
54. C. A. Mentzer and L. Peters, A GTD analysis of the far-out sidelobes of Cassegrain antennas, *IEEE Trans. Antennas Propag.* **AP-23**: 702–709 (Sept. 1975).
55. Q. Ji-zen, Equivalent phase center of the sub-reflector in the shaped Cassegrain antenna, *IEE 2nd Int. Conf. Antennas and Propagation*, April 1981, York, UK, pp. 204–206.
56. S. K. Buchmeyer, An electrically small Cassegrain antenna with optically shaped reflectors, *IEEE Trans. Antennas Propag.* **AF-25**: 346–351 (May 1977).

57. P. J. B. Clarricoats, Some recent developments in microwave reflector antennas, **126**(1): 9–25 (Jan. 1979).

58. J. Ruze, Lateral feed displacement in a paraboloid, *IEEE Trans. Antennas Propag.* **AP-13**: 660–665 (Sept. 1965).

59. W. A. Imbriale, P. G. Ingerson, and W. C. Wong, Large lateral feed displacements in a parabolic reflector, *IEEE Trans. Antennas Propag.* **AP-22**: 742–743 (Nov. 1974).

60. E. A. Ohm, A proposed multiple-beam microwave antenna for Earth stations and satellites, *Bell Syst. Tech. J.* **53**: 1657–1665 (Oct. 1974).

61. A. W. Rudge, Multiple-beam antennas: Offset reflectors with offset feeds, *IEEE Trans. Antennas Propag.* **AP-23**: 234–239 (May 1975).

62. P. Balling, R. Jorgensen, and K. Pontoppidan, *Study of techniques for design of high gain antennas with Contoured Beams*, Final Report ESTEC Contract 3371/77/NL/AK, Dec. 1978.

63. T. S. Bird, J. L. Boomars, and P. J. B. Clarricoats, Multiple-beam-dual-offset reflector antenna with an array feed, *Electron. Lett.* **14**(14): 439–440 (July 1978).

64. K. Woo, Array-fed reflector antenna design and applications, *IEE 2nd Int. Conf. Antennas and Propagation*, April 1981, York, UK, pp. 209–213.

65. E. A. Ohm and M. J. Gans, Numerical analysis of multiple-beam offset Cassegrainian antennas, *AIAA/CASI 6th Communications Satellite Conf.*, Apri1 5–8, 1976.

66. E. A. Ohm, System aspects of a multibeam antenna for full U.S. coverage, *Int. Conf. Communications*, June 10–14, 1979, pp. 49.2.1–49.2.5.

67. D. C. Chang and W. V. T. Rusch, Transverse beam scanning for an offset dual reflector system with symmetric main reflector, *IEE 2nd Int. Conf. Antennas and Propagation*, Apri1 1981, York, UK, pp. 207–208.

68. J. R. Cogdell and J. H. Davis, Astigmatism in reflector antennas, *IEEE Trans. Antennas Propag.* **AP-21**: 565–567 (July 1973).

69. W. V. T. Rusch and A. C. Ludwig, Determination of the maximum scan-gain contours of a beam-scanning paraboloid and their relation to the Petzval surface, *IEEE Trans. Antennas Propag.* **AP-21**: 141–147 (March 1973).

70. M. Akagawa and D. P. DiFonzo, Beam scanning characteristics of offset Gregorian antennas, *APS Symp. Digest*, June 1979, pp. 262–265.

71. W. D. White and L. K. DeSize, Scanning characteristics of two-reflector systems, *1962 IRE Convention Record*, Part I, pp. 44–70.

72. N. A. Adatia and A. W. Rudge, Beam squint in circularly polarized offset-reflector antennas, *Electron. Lett.* **11**(21): 513–515 (Oct. 1975).

73. T. Li, A study of spherical reflectors as wide-angle scanning antennas, *IRE Trans. Antennas Propag.* **47** (July 1959).

74. F. S. Holz and E. L. Bouche, A Gregorian corrector for a spherical reflector, *IEEE Trans. Antennas Propag.* **AP-12**: 44–47 (Jan. 1964).

75. C. Sletten, *Reflector and Lens Antennas*, Artech House, Norwood, MA, 1988.

76. C. Rappaport and W. Craig, High aperture efficiency, symmetric reflector antennas with up to 60 degrees field of view, *IEEE Trans. Antennas Propag.* **AP-39**(3): 336–344 (March 1991).

77. W. Craig, C. Rappaport, and J. Mason, A high aperture efficiency, wide-angle scanning offset reflector antenna, *IEEE Trans. Antennas Propag.* **41**(11): 1481–1490 (Nov. 1993).

78. G. Tong, P. J. B. Clarricoats, and G. L. James, Evaluation of beam-scanning dual reflector antennas, *Proc. IEEE* **124**(12): 1111–1113 (Dec. 1977).

79. W. D. Fitzgerald, *Limited Electronic Scanning with an Offset Feed Near-Field Gregorian System*, MIT Lincoln Laboratory, Technical Report 486, (Sept. 1971, DDC AP-736029.

80. M. Kumazawa and M. Karikomi, Multiple-beam antenna for domestic communication satellites, *IEEE Trans. Antennas Propag.* **AP-21**: 876–877 (Nov. 1973).

81. M. Karikomi, A limited steerable dual reflector antenna, *Electron. Commun. Jpn.* **55-B**(10): 62–68 (1972).

82. B. L. J. Rao, Bifocal dual reflector antenna, *IEEE Trans. Antennas Propag.* **AP-22**: 711–714 (Sept. 1974).

83. C. Rappaport, An offset bifocal reflector antenna design for wide angle scanning, *IEEE Trans. Antennas Propag.* 1196–1204 (Nov. 1984).

84. A. Garcia Pino, C. Rappaport, J. Rubinos, and E. Lorenzo, A shaped dual reflector antenna with a tilting flat subreflector for scanning applications, *IEEE Trans. Antennas Propag.* **43**(10): 1022–1028 (Oct. 1995).

85. E. Lorenzo, A. Pino, and C. Rappaport, An inexpensive scanning dual offset reflector antenna with rotating flat subreflector, *1995 Antennas and Propagation Society/URSI Symposium Digest*, June 20, 1995, pp. 1178–1181.

86. E. Lorenzo, C. Rappaport, and A. Pino, Scanning dual reflector antenna with rotating curved subreflector, *Progress Electromagn. Res. Symp.*, July 1998, p. 347.

RADIO RESOURCE MANAGEMENT IN FUTURE WIRELESS NETWORKS

JENS ZANDER
Royal Institute of Technology
Stockholm, Sweden

1. INTRODUCTION

The rapid increase of the size of the wireless mobile community and their demands for high speed, multimedia communications stands in clear contrast to the rather limited spectrum resource that has been allocated in international agreements. Efficient spectrum or Radio Resource Management (RRM) is of paramount importance due to these increasing demands. Figure 1 illustrates the principles of wireless network design. The network consists of a fixed network part and a wireless access system. The fixed network provides connections between base stations or Radio Access Ports (RAP), which in turn provide the wireless "connections" to the mobiles. The RAPs are distributed over the geographical area where we wish to provide the mobile users with communication services. We will refer to this area simply as the service area. The area around a RAP where the transmission conditions are favorable enough to maintain a connection of the required quality between a mobile and the RAP, is denoted the coverage area of the RAP. The transmission quality and thus

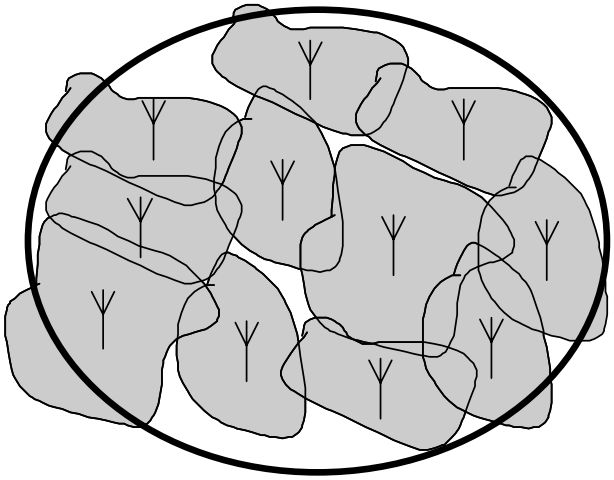


Figure 1. Wireless area communication system.

the shape of these regions will, as we may expect, depend heavily on the propagation conditions and the current interference from other users in the system. The coverage areas are therefore usually of highly irregular shape.

The fraction of the service area where communication with some required quality of service (QoS) is possible is called the coverage or the area availability of the system. In two-way communication systems (such as mobile telephone systems), links have to be established both from the RAP to the mobile (down- or forward link) and between the mobile terminal and the RAP (up- or reverse link). At first glance these two links seem to have very similar properties, but there are some definite differences from a radio communication perspective. The propagation situation is quite different, in particular in wide area cellular phone systems, where the RAP (base station) usually has its antennas at some elevated location free of obstacles. The terminals, however, are usually located amidst buildings and other obstacles creating shadowing and multipath reflections. Also, the interference situation in the up- and downlink will be different since there are many terminals and varying locations and quite few RAPs at fixed locations.

For obvious economical reasons, we would like our wireless network to provide ample coverage with as few RAPs as possible. Clearly this would not only minimize the cost of the RAP hardware and installation, but also limit the extent of the fixed wired part of the infrastructure. Coverage problems due to various propagation effects

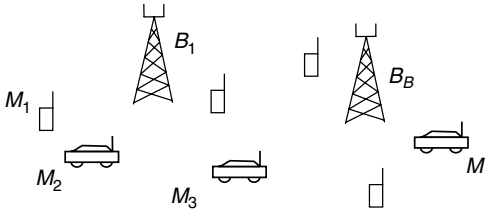


Figure 2. Resource management problem formulation.

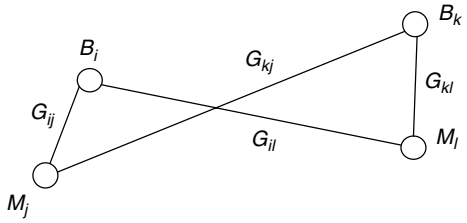


Figure 3. Link gains.

puts a lower limit to the number of RAP that are required. However not quite correct, one could say that the range of the RAPs is too small, compared to the inter-RAP distance. Such a system where this type of problem is dominant is called a range limited system. As the number of transmitters in the system becomes large within some fixed chunk of available RF-spectrum, the number of simultaneous connections (links) will become larger than the number of orthogonal signals that the available bandwidth may provide. In order to provide service for such a large population of users, it is obvious that the bandwidth used by the RAPs and terminals has to be reused in some clever way at the cost of mutual interference. The system is said to be bandwidth or interference-limited. Absolutely vital to the study of any resource management problem is a thorough understanding of the user requirements, that is, the required QoS and the traffic characteristics. All resource management schemes are designed (or optimized) using some model for the traffic. The resulting performance will clearly be a function of not only how well our design has been adapted to the traffic model, but also how accurate the traffic model is. Most wireless systems of today use circuit switched speech as the main design model (e.g., GSM). This does not prevent such systems to carry other types of traffic, but they always do this at a performance penalty. Future wireless access systems are expected to carry both large bandwidths as well as a mixture of services with very different and often conflicting service requirements. Particularly in these scenarios, accurate modeling is imperative for efficient resource utilization. Unfortunately, however, future user applications are little known and most work to derive realistic traffic models for these kinds of applications lies still ahead of the telecommunication community.

In the remainder of the article we present a more rigorous formulation of the radio resource management and review some of the ideas and results from the literature. Finally, we give an outlook on how these results can be applied to future wideband systems and which are the key problems that should be addressed in further studies.

2. RADIO RESOURCE MANAGEMENT — A GENERAL PROBLEM FORMULATION

Assume that M mobiles $(M_1, M_2, \ldots, M_M)$ are served by access ports (base stations), numbered from the set

$$\mathbf{B} = \{1, 2, 3, \ldots, B\}.$$

Now, let us assume that there are C (pairs of) waveforms (in conventional schemes these can be seen as orthogonal channels (channel pairs)) numbered from the set

$$\mathbf{C} = \{1, 2, 3, \ldots, C\}$$

available for establishing links between access ports and mobile terminals. To establish radio links, to each mobile the system has to assign

(a) an access port from the set **B**,
(b) a waveform (channel) from the set **C**,
(c) a transmitter power for the access port and the terminal.

This assignment (of access port, channel, and power) is performed according to the resource allocation algorithm (RAA) of the wireless communication system. The assignment is restricted by the interference caused by the access ports and mobiles as soon as they are assigned a "channel" and when they start using it. Another common restriction is that access ports are in many cases restricted to use only a certain subset of the available channels. Good allocation schemes will aim at assigning links with adequate SIR to as many (possibly all) mobiles as possible. Note that the RAA may well (should) opt for not assigning a channel to an active mobile if this assignment would cause excessive interference to other mobiles.

Let us now study the interference constraints on resource allocations in somewhat more detail. We now may compute the signal and interference power levels in all access ports and mobiles, given the link (power) gains, G_{ij}, between access port i and mobile terminal j. For the sake of simplicity, we will here consider only rather wideband modulation schemes that will make the link gains virtually independent of the frequency. Collecting all link gains in matrix form, we get a BxM rectangular matrix — the link gain matrix $\boldsymbol{G}$. The link gain matrix describes the (instantaneous) propagation conditions in the system. Note that in a mobile system, both the individual elements of the matrix (due to mobile motion) and the dimension of the matrix (due to the traffic pattern) may vary over time.

The task of the resource allocation scheme is to find assignments for the QoS that is sufficient in as many links as possible (preferably all). Providing a stringent definition to the QoS for a practical communication service is a complex and multifaceted problem. In this treatment we will confine ourselves to a simple measure the signal-to-interference ratio (SIR), or actually, to be precise, the signal-to-interference+noise ratio. This measure is strongly connected with performance measures as the bit or message error probability in the communication link. We require the SIR in link i to exceed a given threshold γ_i which is determined by both the QoS requirements for that particular link as well as by the modulation and coding formats of the system. This means that the following inequality must hold for both the up-(mobile-to-access port) and down-(access port-to-mobile) link of the connection:

$$\Gamma_i = \frac{P_j G_{ij}}{\sum_m P_m G_{im}\theta_{jm} + N} \geq \gamma_i \tag{1}$$

where Γ_j denotes the SIR at the receiver and N denotes the receiver (thermal) noise power at the access port. P_j denotes the transmitter power used by terminal j. The quantity θ_{jm} is the normalized crosscorrelation between the signals from mobiles j and m at the access port receiver, that is, the effective fraction of the received signal power from transmitter m contributes to the interference when receiving the signal from access port j. If the waveforms are chosen to be orthogonal (as in FDMA and TDMA) these correlations are either zero or one depending on if the station has been assigned the same frequency (time slot) or not. In nonorthogonal access schemes (e.g., DS-CDMA) the θ_{jm} take real values between zero and one. Note that we may not be certain that it is possible to comply with all the constraints (2) for all the M mobiles, in particular if M happens to be a large number. As system designers, we may have to settle for finding resource allocation schemes that assign channels with adequate quality to as many mobiles as possible.

In the classical single service (e.g., mobile telephony) case, the service requirements are identical in all links, that is, $\gamma_i = \gamma_0$ for all i. In this case the *system capacity* can be measured by the largest number of users that may be successfully handled by the system. Since the number of mobiles is a random quantity and the constraints (2) depend on the link matrix, that is, on the relative position of the mobiles, such a capacity measure is not a well-defined quantity. The classical approach for telephone type of traffic is to use as capacity measure the maximal relative arrival rate of calls ρ for which the blocking probability (the probability that a newly arrived session request is denied) can be kept below some predetermined level. Due to the mobility of the mobiles this is not an entirely satisfying measure. A call or session may be lost due to adverse propagation conditions. To include such phenomena into our capacity would require detailed specification of call-handling procedures (e.g., handling of new vs. old calls, hand-off procedures as a mobile moves from one access port to another, etc.). It may therefore be practical to choose a simpler and more fundamental capacity measure that will reflect the performance of the resource allocation scheme as such. For this purpose, the assignment failure probability ν (or assignment failure rate [14,15]) has been proposed. The instantaneous capacity $\omega * (\nu_o)$ of a wireless

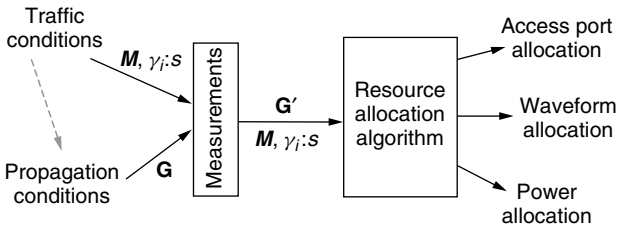


Figure 4. A resource allocation algorithm.

system is the maximum allowed traffic load in order to keep the assignment failure rate below some threshold level ν_o, that is,

$$\omega * (\nu_o) = \{\max \omega: \nu \leq \nu_o\} \quad (2)$$

As we have seen above, finding the optimum resource allocation for each mobile determining

(i) a waveform assignment (determining the θ_{jm})
(ii) an access port assignment (of one or more (!) ports)
(iii) a transmitter power assignment

that maximizes Y for a given link gain matrix, is a formidable problem. No efficient general algorithm that is capable of doing such an optimal assignment for arbitrary link gain matrices and mobile sets is known. Instead, partial solution and a number of more less complex heuristic schemes have been proposed (and are used in the wireless systems of today). These schemes are usually characterized by low complexity and by using simple heuristic design rules. The capacity $\omega*$ achieved by these schemes is, as expected, often considerably lower that can be expected to be achieved by optimum channel assignment.

3. CURRENT APPROACHES TO RESOURCE ALLOCATION STRATEGIES

The subproblem that has attracted most of the interest in the literature so far, is the choice and allocation of waveforms. Orthogonal waveforms such as frequency division multiplexing (FDMA) and time division (TDMA), which provide a "channelization" of the spectrum, have no doubt been the most popular ones, although considerable interest has recently been devoted to nonorthogonal waveforms, such as the IS-95 DS-CDMA waveforms [25]. Given the set of signaling waveforms **C**, the next problem is the allocation of waveforms to the different terminal-access port links. This allocation can be done numerous ways depending on the amount and quality of the information available regarding the matrix **G** and the traffic situation (activity of different terminals). Another important issue is the time scale on which resource (re-)allocation is feasible.

Channel allocation in early FDMA cellular radio systems operates on a long-term basis. Based on average type statistical information regarding **G** (i.e., large scale propagation predictions), frequencies are on a more or less permanent basis assigned to different access ports. Such a "cell plan" provides a sufficient reuse distance between RAPs providing a reasonably low probability of outage (to low SIR) [1]. Inhomogenuities in the traffic load can also be taken care of by adapting the number of channels in each RAP to the expected traffic carried by that access port. To minimize the planning effort, adaptive cell planning strategies (e.g., "channel segregation" [2]) have been devised using long-term average measurements of the interference and traffic to automatically allocate channels to the access port. These "static" (or "quasi-static") channel allocation schemes work quite well when employed in macrocellular systems with high-traffic loads. In short range (microcellular) systems propagation conditions tend to change more abruptly. Since each of the RAPs tend to carry less total traffic in small microcells, the relative traffic variations are also large, particularly in multimedia traffic scenarios. Employing "static" channel allocation schemes in the situations require considerable design margins. Large path loss variations are countered with large reuse distances, unfortunately at a substantial capacity penalty. In the same way microcellular traffic variations are handled by assigning excess capacity to handle traffic peaks. In recent years two principally different methods to approach this problem have been devised: Dynamic channel allocation (DCA) and Random Channel Allocation (RCA).

In dynamic (real-time) channel allocation(DCA), real-time measurements of propagation and/or traffic conditions are used to (re-)allocate spectrum resources. Early graph theoretic schemes, adapted only to traffic variations [4,5] yielding only moderate capacity gains (<50%) compared to static systems in microcellular environments. Other schemes adapt their channel allocation to the received wanted signal strength. One example of the latter type of schemes is the class of reuse-partitioning schemes [6,8]. Here, several overlaid cell plans with different reuse distances are used. Terminals with a high received signal level are tolerant to interference and can be allocated a channel from a dense reuse cell plan, whereas the "weaker" terminals get channels with a large reuse distance and lower interference levels. Capacity gains in the order of up to 100% have been reported for these schemes [7]. Also, schemes directly estimating the C/I and thereby in a distributed way finding channels with adequate quality have been proposed [2,9]. Similar gains as in the reuse partitioning schemes are found in the literature. A comprehensive survey of different DCA-schemes is provided in Ref. 46.

The performance of the DCA schemes is critically dependent on the rate at which allocation or reallocation occurs. Purely traffic adaptive schemes act on incoming user requests and users releasing capacity. Channel reallocation has to occur at these rates to fully utilize the potential of such a DCA scheme. For speech traffic this means that reallocations typically occur at second rates.

Path loss and interference adaptive schemes that "track" (at least slow fading) signal level variations and reallocation rates in the 10's of millisecond range may be required. An alternative class of allocation schemes are the random channel allocation schemes. The principle is most easily explained using Fig. 5. Figure 5a shows a typical set of C/I-trajectories of five terminals in a cellular system. As we can see, 4 of the 5 terminals achieve an adequate C/I, corresponding to an (ensemble) outage rate of 20%. Compare this situation to the one in Fig. 5b exhibiting the same outage rate. In contrast to the situation in Fig. 5a where 20% of the terminals are experiencing too low C/I, here each terminal will experience insufficient quality 20% of the time. In Fig. 5a channel coding is a waste of capacity since four terminals have sufficient quality and the last unlucky terminal is probably "beyond salvage." In Fig. 5b however, there

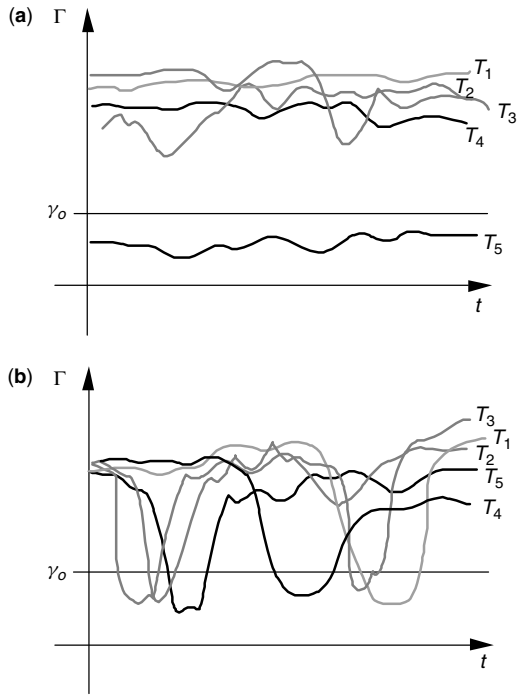


Figure 5. Typical realizations of terminal SIR:s in cellular systems with slowly moving terminals (**a**), and with rapidly moving terminals (**b**).

are probably a sufficient number of reliable channel symbols in all terminals to make reception possible, provided suitable constraint lengths and interleaving are used. The obvious way to achieve the latter situation, regardless of the mobile speed, is to permute channel allocations in a random fashion. The simplest way is to use (orthogonal) frequency hopping which can be seen as a static channel allocation where terminals allocated to a certain access port swap channels with each other [28,33]. Frequency hopping occurs typically 100–1000 times/second. Also nonorthogonal waveforms can be used as in the DS-CDMA based IS-95 scheme [25]. Effectively, a new random waveform is used for every transmitted bit. DS-CDMA schemes require only a very low level of synchronization and no cell planning, which has made them attractive. Regarding capacity the comparison between DS and FH schemes is not obvious although orthogonal schemes seem to have advantages in mixed cell environments [26]. Comparing the performance of (deterministic) DCA to the performance of the random allocation schemes is even more complex and stands out as one of the more fundamental research topics of the near future. Another quite different situation where similar interference conditions as in frequency hopping prevail are certain packet communication systems. Here the "randomness" is mainly induced by the random arrivals of packets triggering transmission events.

The selection of the proper transmitter power in terminals and access ports is another topic that has attracted considerable interest in recent years. There can be several objectives for this: to suppress adjacent channel (cross correlation) interference in nonorthogonal schemes, to minimize power consumption in order to extend terminal battery life, and to control cochannel interference (in schemes with orthogonal waveforms). In the resource allocation problem context, it can be shown that the maximum number of terminals is supported under a power control (PC) regime that balances the C/I of all terminals that can be supported and shuts off the rest [10].

Figure 6, showing the cumulative distribution function (CDF) of the received C/I in an (orthogonal signaling) cellular system under three power control regimes, illustrates why this is so. As we can see, the uncontrolled system exhibits a rather flat CDF with a high outage probability (at the threshold C/I). A received signal strength based algorithm, such as the constant received power scheme, reduces some of the variations in the C/I by limiting the variations in the "C" component. The variations in the interference part ("I") are however now larger than before and the figure shows a typical net result CDF. The outage probability is now slightly lower. In the C/I balancing scheme all stations have the same C/I, here slightly over the threshold, leaving only a small fraction of terminals without support. Finding this optimum set of nonsupported terminals is a problem closely related to the design of DCA schemes. Distributed implementations and different implementational constraints [11,12] have been studied. Results show that very robust near-optimum power control schemes can be devised at very low complexity. Performance results indicate that in static channel allocations substantial (>100%) capacity gains can be achieved using optimum power control. These gains are, of course, not additive with the gains obtained by DCA schemes. However, preliminary results regarding combined DCA/PC schemes show substantial capacity gains [13,15].

For packet communication with short messages or in frequency hopping environments, power control as described above may not work properly due to the fact that the feedback delay in the power control loop may in fact be longer than the time required to transmit the message (or in FH the chip/burst duration). The bursty interference caused by other users compounds to the problem of accurately measuring and predicting the C/I. Several approaches to this problem have been proposed. In systems utilizing mainly forward error correction, C/I-balancing power control strategies involving estimating statistical parameters of the C/I, such as the average C/I (measured over many packet/chip durations), and as

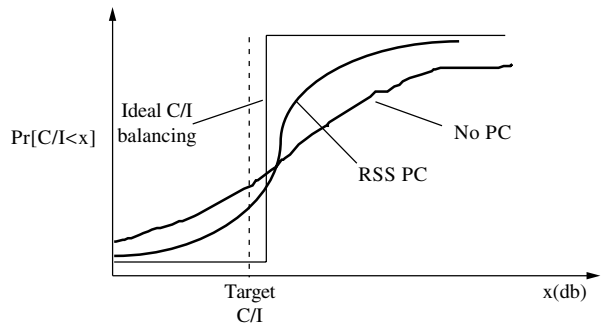


Figure 6. Outage probability minimization-CDF of received C/I.

well as fast C/I estimation/tracking schemes have been proposed [34–37].

In random access systems utilizing retransmissions (ARQ), however, the picture is quite different. Two important differences between these systems and the continuous ("circuit switched") systems can be noted. Providing equal transmission quality to all packet transmission (e.g., letting all packets be received with equal powers), turns out to be a disastrous strategy, since it guarantees that no packet can be received properly when several packets happen to be transmitted simultaneously (a "collision"). It has clearly been demonstrated that a received power spread, such as the one caused by near-far effects or Rayleigh/Shadow fading actually improves the capacity (throughput) of such systems [39]. Results show that power control, which creates an even larger power spread, can produce even better results [40–42].

4. MANAGING CHANGE — THE DYNAMICS OF RESOURCE ALLOCATION

As terminals move about in the service area, the propagation and interference situation may turn such that the terminal cannot be supported by the same access port on any waveform. New terminals may enter the service area requiring services, while others are terminating their communication sessions. As most of the basic resource allocation strategies described above deal mainly with static or quasi-static situations that are encountered on a microscopic, short term, time scale, we have to devise resource management schemes capable of handling these variations.

In the first case, where we have signal quality variations as a result of the terminal moving during a communication session some kind of resource reallocation may become necessary. This may be a waveform reallocation, or an "intra-port" handoff, which in principle involves a reexecution of the basic channel allocation scheme, or an inter-port ("inter-cell") handoff. In early cellular systems which are mainly noise-limited systems these handoffs were basically triggered by too low received signal levels. The handoff mechanism has, for these cases, often been modeled as a selection (macro-) diversity scheme where the terminals are assigned to the access port with the highest received signal level. This situation where hand-offs occur at more or less well-defined "cell-borders" has been extensively studied.

Maximizing the instantaneous received signal level may, however, not be neither very practical nor produce the best results. In high density wireless systems, the coverage areas of the access port overlap to a large extent. Low signal levels is rarely a problem since normally several access ports provide sufficient signal levels. In these cases the variations in the interference and not cell boundary crossings is the most probable cause of a handoff. When several access ports may provide sufficient C/I, the system is also able to handle traffic variations by means of *load sharing*, that is, by letting less loaded access ports support terminals even though they are providing less C/I than the best (often the closest) port [20]. Combinations of power control and access port selection also show promising results [21,22]. Instead of conventional handoff schemes ("switching diversity"), continuous combining schemes ("soft handoff") have been studied quite extensively [19].

Keeping track of the mobile terminals in a large (possibly global) wireless system, the mobility management, is a formidable task. Although this is handled mainly on the fixed network end, there are important implications to the resource management. The tradeoff between the capacity required for the air signaling to monitor the whereabouts of the terminals (the "locating" procedures) and the capacity required for finding, or paging, a terminal when a communication request comes from the network end, has received quite some attention [23,24] in CDMA schemes.

Handling arriving and departing terminals poses a slightly different problem. Whenever a new terminal arrives (a new request for service or an inter-port handoff) the RRM system has to decide if this particular terminal may be allowed into the system. An algorithm making these decisions is called an admission control algorithm. Since the exact terminal population and gain matrix may not be tracked exactly at all times and due to the complexity of the RRM-algorithms, determining the success of an admission decision may not be possible beforehand without physically executing the admission itself. The admission procedure may fail in two ways:

1. ("False admission") A terminal is admitted giving rise to a situation where one or more terminals cannot be supported (not necessarily including the admitted terminal).
2. ("False rejection") A terminal is rejected when successful resource allocation actually was possible.

Traditional approaches involving static channel allocation normally use simple thresholding strategies on the available channels in each cell. Access ports have been assigned a fixed set of channels that "guarantee" to provide a certain low-outage probability. Such a system is what we call "blocking limited," whenever a call arrives, we may check if there are channels available or not. Since we may choose to give priority to an ongoing session experiencing an inter-port handoff, new arriving calls are admitted only up to the point where there is only some small fraction of the resource remaining. This spare capacity is "reserved" for calls entering a cell due to an inter-cell handoff.

In systems using dynamic channel allocation or random allocation there is no clear limit on the number of channels/waveform that can be used. In such "interference-limited" systems, the feasibility of admitting new users will depend on the current interference situation. In particular in systems utilizing C/I-balancing power control this is complicated by the fact that already active terminals will react to the admission of a new terminal by adjusting (raising) their transmitter powers. It is therefore quite possible that the admission of yet another user may cause several of the original users (possibly all!) to no longer be supported at the required C/I-level. Admission control schemes can be grouped into noninteractive schemes and interactive schemes [43]. The noninteractive schemes proposed are mainly using

different types of interference or transmitter power thresholds [43], that is, when the measured interference (or the currently used power) on some channel (cell) is too high, admission is denied. The interactive schemes involve the gradual increase of the power of new terminals until they are finally admitted. Such a procedure to protect the already established procedure connection is referred to as "Soft-and Safe (SAS)" admission [44] or channel probing/active link protection [45].

5. MULTIPLE QUALITY-OF-SERVICE RESOURCE MANAGEMENT

Looking at the more general problem definition, most of the capacity definitions provided above fail if the users have different service requirements. In our problem definition this is reflected in the SIR requirements γ_i. Most of the techniques discussed above, properly generalized, lend themselves readily also for the nonspeech and multiservice cases. It has, however, to be understood that (for nonspeech services) the mapping of user perceived performance on technical parameters, such as the SIRs, in the wireless system is certainly a very complex task. With the proliferation of "best-effort"-type backbone networks, the user experience of the service provided is influenced not only by the shortcomings of the wireless system but also by other factors. As indicated by Fig. 7, such factors include the performance of the wireline backbone and switching, the service providers application software and hardware, and even the user interface provided by the service provider and the terminal manufacturer. In order to allow rational design of telecommunication systems, these overall (end-to-end) requirements are broken down to specific (sub-)service requirements for the individual building blocks. Here the interest is focused on studying the behavior of the radio network part of the "transport system." The "services" provided at this level have been coined bearer services in the UMTS/3G standardization process. These services have been divided into four classes as outlined in Table 1 [47] and are mainly distinguished by their delay requirements ranging from very strict delay requirements in "conversational class" (e.g., voice services) to the very relaxed requirements in the "background-best effort" class. In more technical terms, the services in the different classes are characterized by means of sets of service parameters, forming a QoS profile. Some of the QoS parameters (service attributes) in the 3G systems are found in Table 2.

In principle, an infinite set of QoS-profiles and thus different bearer services could be defined by varying these parameters — possibly one combination for each user. In practice, limitations on the number of modulation waveforms, codes, and so on, restrict the number of service

Table 1. 3G (UMTS) Bearer Service Classes [47]

Service Class	Typical Applications	Service Functional Characteristics
Conversational		
Real Time(RT) Streaming RT	Voice Video/Audio streams	• Preserve time relations between entities • Stringent preservation of conversational patterns (low delay)
Interactive		Preserve time relations between entities
Best effort (BE) Background BE	Web-browsing File transfer, E-mail	• Request-response pattern • Preserve payload (low error rate) • Not time critical • Preserve payload (low error rate)

offerings. Most systems will therefore offer a finite set of bearer services, where each parameter will be allowed to take one out of a few discrete values. Table 2 provides an indication of what ranges these service parameters can take in a 3G wireless system. In addition to these service parameters, the *availability* of the services has to be considered since it may vary over time and over the user services. Clearly, users with QoS-profiles with large "resource consumption" (e.g., high bit rate, poor location) or with low-relative priority will more often experience that the system is not capable of accommodating their service request.

Looking at a popular example to illustrate the problems we face, we take a web browsing session. Such a session consists of an irregular sequence of file transfers (using the TCP/IP protocol stack). Typical very short messages are transmitted in the uplink from the terminals (corresponding to a mouse-click) a random instants to request rather large files (web-pages, pictures, etc.) to be downloaded into the terminal. This can be seen as a service of the "interactive" class. The critical QoS characteristic to the user is the response-time (i.e., delay between request and the complete reception of the requested page). For large requested files the delay is dominated by the transfer delay of the files, that is, the average data rate is in fact the critical QoS parameter that will determine the user delay. The undetected error rate at the user level has to be below 10^{-8} corresponding to 1 error in about 10 MB. The radio bearer may however have a larger undetected error probability since the TCP/IP protocol provides end-to-end error control of its own. Classical models for data traffic of this type are based on Poisson distributions — both for

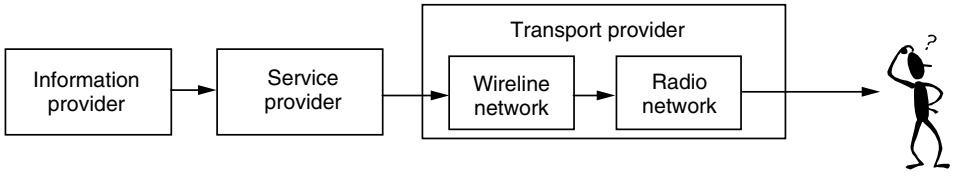


Figure 7. Service provisioning in modern information and communication systems.

Table 2. Some 3G (UMTS) Service Attribute/Parameter Ranges [47]

Traffic Class	Conversational	Streaming	Interactive	Background
Max bit rate(kbps)	<2000	<2000	<2000-overhead	<2000-overhead
Max SDU size(byte)	<1500	<1500	<1500	<1500
Guaranteed bit rate	<2000	<2000		
Transfer delay(ms)	80-max value	500-max value		
Priority	1,2,3	1,2,3	1,2,3	1,2,3
Residual BER	$5*10^{-2}$, 10^{-2}, 10^{-3}, 10^{-6}	$5*10^{-2}$, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}	$4*10^{-3}$,10^{-5}, $6*10^{-8}$	$4*10^{-3}$,10^{-5}, $6*10^{-8}$

the interarrival time of packets as well as the size of the requested files. Recent studies of Internet traffic have however shown that these models tend to underestimate the time between packets and the packet sizes. As an alternative, the Pareto distributions, that is, stochastic variables with density function

$$f(x) = \frac{\beta a^{\beta}}{x^{\beta+1}} a, \beta \geq 0, x \geq a$$

have been proposed [49]. These distributions are "heavy-tailed," they assign higher probability to large values of x compared to the exponential distributions found in Poisson point processes. Another approach for simulation purposes, is to use a more detailed model of the actual TCP/IP protocol [48].

After having discussed the user performance perspective, let us now briefly return to the discussion of the network performance. Defining a capacity performance measure for such a single service system is not obvious. Examples of popular measures (for non–real-time service systems as in the example above) found in the literature are the total throughput, that is, the sum data rate of all connections in the systems (possibly per area unit) and the user Circuit switched equivalent Bit Rate (CBR, i.e., the constant data rate would provide the same user perceived performance). In non–real-time, "best-effort," systems efficient resource management makes use of the extra degree of freedom provided by not fixing the delay. Efficient scheduling and adapting the transmission rate to the current propagation and SIR conditions may be used to maximize throughput. In a fading channel (as illustrated by Fig. 5b), the channel conditions may be bad for a certain user during one time frame. That particular user could thus postpone its transmission, allowing some other users with favorable data rates to transmit — potentially at a much higher data rate. Typically, such schemes maximize the total throughput by favoring terminals with good propagation conditions as they generally have more time slots at their disposal and can use higher data rates [52,53].

When mixing services, the situation becomes even more complex. In order for any capacity definition to be precise in the general (multiservice) case requires a model, not only for the number of users, but also for their behavior. What will be the QoS-profile requested for a certain user (which is then mapped to the individual γ_i), and what is the required service availability? Typically, random models will be used for this purpose. A user will, with some given probability, belong to a certain class of users with an identical QoS-profile. The probability distribution of class membership is usually referred to as the service mix.

Determining which service mix should be used for the capacity definition is indeed difficult. One approach has been to look at the pricing strategy and maximizing the revenue of the operator. This leads to both an optimal service mix and maximal revenue derived from the network. A difficulty is that this type of model disregards the fact that if the demand for the service provided is finite, operators are prone to competition from either other similar operators or alternative technical solutions. In these situations the pricing strategy clearly affects the demand for the services.

Given a certain service mix, several capacity definitions have been tried in the literature, but a generally acceptable single definition has still to be found. Looking at the total throughput or the total operator revenue is one approach. The main difficulty with this is to set the price parameters to reasonable values due to the interdependence of the pricing and the service mix. More promising is to use the (total) number of users satisfied with their respective service, this is one approach that leads to a capacity definition very close to the one discussed above [50,51]. In this approach, the capacity regions, that is, the permissible combinations of numbers of users in the different service classes play an important role in the cases where no price structure can be determined [50].

Another problem caused by mixing traffic, in particular circuit-switched and bursty best-effort traffic, is that it makes link-quality assessment more difficult. The main result from traffic theory tells us to dynamically share all the available spectrum for all types of traffic. A sideeffect of this is, however, that also the interference experienced by different users will exhibit the same wide span in character [38]. In particular if we would like to estimate the link quality for a high quality circuit switched service, the link will be subject to both quasi-constant as well as intermittent interference (from packet service users). Reliably estimating, for instance, using the C/I as a basis for RRM decisions, will be considerably more difficult.

6. DISCUSSION

We have previously presented a formulation of the radio resource problem based on the three basic allocation decisions: waveform, access port, and transmitter power. As the reader may have realized, these are closely related. Most of the recent work indicates that good results are achieved when these decisions are coordinated. Combinations of power control and DCA [13,15], base station assignment and power control [21,22], as well as power control assisted admission schemes, have provided

interesting results. Another area where research is just in its preliminary stages is the combination of detailed modulation waveforms/channel coding and its interaction with DCA and power control. Although the current mobile telephony systems are rather easily modeled in the terms described above, it seems clear that also most of the RRM problems expected in the future can be mapped onto the framework presented here. A key problem in bursty and mixed traffic is the tradeoff between maximizing instantaneous resource utilization (transmit only when data is available) and obtaining reliable quality measurements to facilitate the efficient adaptation of the radio resources to the needs of the users.

Traditionally, we consider the frequency spectrum to be the resource to be shared. Since there, in fact, does not exist any upper limit on the capacity that can be provided (with a dense enough infrastructure), it is important that we widen the resource management perspective. Parameters such as infrastructure density costs and terminal power consumption play important roles. One could easily identify tradeoffs such as where the signal processing load should be put in a wireless system — in the terminal where power is scarce or in the fixed infrastructure. The key question here is: Should the access port infrastructure be very dense (and costly) allowing for dumb, cheap, low-power terminals, or should terminals be more complex allowing for the rapid deployment of a cheap infrastructure at the expense of battery life and terminal cost ?

BIOGRAPHY

Jens Zander (S'82–M'85) received the M.S degree in electrical engineering (Y) and the Ph.D degree (in datatransmission) from Linköping University, Sweden, in 1979 and 1985, respectively.

From 1985 to 1989 he was a partner of SECTRA, a high-tech company in telecommunications systems and applications. In 1989 he was appointed professor and head of the Radio Communication Systems Laboratory at the Royal Institute of Technology, Stockholm, Sweden. Since 1992 he also serves as Senior Scientific Advisor to the Swedish National Defence Research Institute (FOI). He currently is the Scientific Director of the Center for Wireless Systems (Wireless@KTH) at the Royal Institute of Technology, Stockholm.

Dr. Zander has published numerous papers in the field of radio communication, in particular on resource management aspects of personal communication systems. He has also coauthored four textbooks on radio communication systems, including the English textbooks *Principles of Wireless Communications* and *Radio Resource Management for Wireless Networks*. He was the recipient of the IEEE Veh. Tech. Soc. Jack Neubauer Award for best systems paper in 1992.

Dr. Zander is a member of the Royal Academy of Engineering Sciences. He is the chairman of the IEEE VT/COM Swedish chapter. He is associate editor of the ACM *Wireless Networks* Journal and area editor of *Wireless Personal Communications*.

His current research interests include future wireless infrastructures, in particular related resource allocation and economic issues.

BIBLIOGRAPHY

1. W. C. Y. Lee, *Mobile Communication Fundamentals*, Wiley, New York, 1993.
2. Y. Furuya, Y. Akaiwa, Channel segregation — A distributed adaptive channel allocation scheme for mobile communication systems, *Proc DMR-II*, Stockholm, 1987.
3. R. Beck and H. Panzer, Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems, *IEEE Veh Tech Conf VTC89*, May 1989.
4. D. C. Cox and D. O. Reudink, Dynamic channel assignment in high capacity mobile communication systems, *Bell Syst Tech J.* **50**(6): (July–Aug. 1971).
5. D. Everitt and D. Manfield, Performance analysis of cellular communication systems with dynamic channel allocation, *IEEE Trans. Sel. Areas Comm.* **7**(8): (Oct. 1989).
6. S. W. Halpern, Reuse partitioning in cellular systems, *IEEE Veh. Tech. Conf. VTC85*, May 1985.
7. J. Zander and H. Eriksson, Asymptotic bounds on the Performance of a class of dynamic channel Assignment Algorithms, *IEEE Trans. Sel. Areas Comm.* **11**(3): (Aug. 1993).
8. T. Kanai, Autonomous reuse partitioning in cellular systems, *IEEE Veh. Tech. Conf. VTC92*, May 1992.
9. D. J. Goodman, S. A. Grandhi, and Vijayan, Distributed dynamic channel assignment schemes, *IEEE Veh. Tech. Conf. VTC93*, May 1993.
10. J. Zander, Performance of optimum transmitter power control in radio systems, *IEEE Trans. Veh. Tech.* **41**(1): (Feb. 1992).
11. G. J. Foscini and Z. Mijanec, A simple distributed power control algorithm and its convergence, *IEEE Trans. Veh. Tech.* **42**(4): (Nov. 1993).
12. S. A. Grandhi, J. Zander, and R. Yates, Constrained power control, *Wireless Personal Communications*, (Kluwer) **2**(3): (Aug. 1995).
13. G. J. Foscini and Z. Mijanec, Distributed autonomous wireless channel assignment algorithm with power control, *IEEE Trans. Veh. Tech.* **44**(4): (Nov. 1995).
14. M. Frodigh, Bounds on the performance of DCA-algorithms in highway micro cellular radio systems, *IEEE Trans. Veh. Tech.* **43**(3): (Aug. 1994).
15. M. Frodigh, Performance bounds for power control supported DCA-algorithms in highway micro cellular radio systems, *IEEE Trans. Veh. Tech.* **44**(2): (May 1995).
16. S. Tekinay and B. Jabbari, Handover and channel assignment in mobile cellular network, *IEEE Comm Mag.* **29**(11): (Nov. 1991).
17. M. Austin and G. Stüber, Cochannel interference modelling for signal-strength based handoff, *Electronics Letters* **30**: 1914–1915 (Nov. 1994).
18. N. Zhang and J. Holtzman, Analysis of handoff algorithms using both absolute and relative measurements, *IEEE 44th Veh Tech Conf VTC94*, June 1994.
19. N. Zhang and J. Holtzman, Analysis of a CDMA soft handoff algorithm, *Proc PIMRC 95*, Toronto, Sept. 1995.
20. B. Eklundh, Channel utilization and blocking probability in cellular mobile telephone systems with directed retry, *IEEE Trans. Comm.* **34**(4): (Apr. 1986).
21. R. Yates and C-Y. Huang, Integrated power control and base station assignment, *IEEE Trans. Veh. Tech.* **44**(4): (Nov. 1995).

22. S. V. Hanly, An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity, *IEEE Trans. Sel. Areas Comm.* **13**(7): (Sep. 1995).

23. J. G. Markoulidakis and E. D. Sykas, Model for location updating and handover rate estimation in mobile telecommunications, *Electronics Letters* **29**(17): (Aug. 1993).

24. G. Morales-Andes and Villen-Altamirano, An approach to modelling subscriber mobility in cellular radio networks, *Forum Telecom 87*, Geneva, 1987.

25. A. Salmasi and K. S. Gilhousen, On the system design aspects of code division multiple access(CDMA) and personal communication networks, *IEEE 42ndVeh Tech Conf VTC92*, May 1992.

26. H. Eriksson, G. Gudmundson, J. Sköld, and J. K. Ugland et al., Multiple access options for cellular based personal communications, *IEEE 43rdVeh Tech Conf VTC93*, May 1993.

27. E. Anderlind, Resource allocation for heterogenous traffic in a wireless network, *Int. Symp. on Personal, Indoor and Mobile Radio Comm. PIMRC 95*, Toronto, Sept. 1995.

28. G. Gudmundson, J. Sköld, and J. K. Ugland, A comparison between CDMA and TDMA systems, *IEEE 42ndVeh Tech Conf VTC92*, May 1992.

29. A. Acampora, Wireless ATM: A perspective on issues and prospects, *IEEE Personal Comm. Mag.* (Aug. 1996).

30. C. Robool and C. Roobol, Message delay in 1-D indoor packet radio systems with diversity reception, *IEEE First Symposium on Communications and Vehicular Technology in the Benelux*, Delft, The Netherlands Oct. 1993.

31. F. Borgonovo, Zorzi & Acampora, Capture division packet access, *IEEE Comm. Mag.* **34**(9): (Sept. 1996).

32. G.Q. Maguire, B. Ottersten, H. Tenhunen, and J. Zander, Future wireless computing & communication, *Nordisk Radioseminarium*, NRS-94, Linköping, Sweden, Oct. 1994.

33. H. Olofsson, Interference diversity as means for increasing capacity in GSM, *Proc EPMCC'95*, Bologna, Italy, Nov. 1995.

34. Z. Rosberg, Fast power control in cellular networks based on short term correlation of Rayleigh fading, *Proc 6th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ, March 1997.

35. Z. Rosberg and J. Zander, Power control in wireless networks with random interferers, *Internal Report*, Radio Communication lab, Royal Institute of Technology, (Dec. 1995) (to be published, *http://www.s3.kth.se/s3/radio/PUBLICATIONS/documents.html)*.

36. M. Andersin and Z. Rosberg, Time-variant power control, *Proc PIMRC-96*, Taipeh Oct. 1996.

37. D. Mitra and J. A. Morrisson, A distributed power control algorithm for bursty transmissions in cellular CDMA networks, *Proc 5th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ, 1995.

38. D. Mitra and J. A. Morrisson, A novel distributed Power Control Algorithm for classes of service in cellular spread spectrum wireless networks, *Proc 6th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ March 1997.

39. J. A. Arnbak and W. van Blitterswijk, Capacity of slotted ALOHA in Rayleigh fading channels, *IEEE J Sel. Areas Commun.* **SAC-5**(2): (Feb. 1987).

40. J. J. Metzner, On improving utilization in ALOHA Networks, *IEEE Trans. Comm.* **COM-24**: (Apr. 1976).

41. C. Leung and V. Wong, A transmit power control scheme for improving performance in a mobile packet radio system, *IEEE Trans. Veh. Tech.* **VT-43**(1): (Feb 1994).

42. C. Roobol, On the Packet Delay in wireless local area networks with access port diversity and power control, *Proc PIMRC-95*, Toronto Sept. 1995.

43. M. Andersin, *Power Control and Admission Control in Cellular Radio Systems dissertation*, Ph.D., Radio Comm Systems Lab, Royal Institute of Technology, June 1996.

44. M. Andersin, Z. Rosberg, and J. Zander, Soft admission in cellular PCS with constrained power control and noise, *Proc 5th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ, 1995.

45. N. Bambos, S. C. Chen, and D. Mitra, Channel probing for distributed access in wireless communication networks, *Proc GLOBECOM '95*, Singapore Nov. 1995.

46. I. Katzela and M. Naghshneh, Channel allocation schemes for cellular mobile telecommunication systems: A comprehensive survey, *IEEE Personal Commun.* 10–31 (June 1996).

47. 3GPP Specification, TS23.107 Dec. 1999.

48. E. Anderlind and J. Zander, A traffic model for non real-time data users in a wireless radio network, *IEEE Comm. Letters* **1**(2): (Mar. 1997).

49. V. Paxson and S. Floyd, Wide Area Traffic:The Failure of Poisson Modeling, *IEEE/ACM Trans. Networking* **3**(3): (June 1995).

50. A. Furuskär, P. de Bruin, C. Johansson, and A. Simonsson, Managing mixed services with controlled QoS in GERAN—The GSM/EDGE Radio Access Network, *IEE 3G Conference on Mobile Communication Technologies* 147–151, 2001.

51. A. Furuskär, P. de Bruin, C. Johansson, and A. Simonsson, Mixed Service Management with QoS Control for GERAN—The GSM/EDGE Radio Access Network, *IEEE Vehicular Technology Conference 2001* Spring, May 2001.

52. J. Zander, Performance Bounds for Joint Power Control & Link Adaption for NRT bearers in Centralized (Bunched) Wireless Network *PIMRC 99*, Osaka, Japan, Sept. 1999.

53. F. Berggren, S-L Kim, R. Jäntti, and J. Zander, Joint Power Control and Intracell Scheduling of DS-CDMA Nonreal Time Data, *IEEE J. Sel. Areas Commun.* **19**(10): 1860–1870 (2001).

54. J. Zander and S.-L. kim, *Resource Management in Wireless Networks*, Artech House, 2001.

ROUTING AND WAVELENGTH ASSIGNMENT IN OPTICAL WDM NETWORKS

GEORGE N. ROUSKAS
North Carolina State University
Raleigh, North Carolina

1. INTRODUCTION TO OPTICAL WDM NETWORKS

A basic property of a single-mode optical fiber is its enormous low-loss bandwidth of several tens of terahertz. However, because of dispersive effects and limitations in optical device technology, single-channel transmission is limited to only a small fraction of the fiber capacity. To take full advantage of the potential of fiber, the use of wavelength-division multiplexing (WDM) technology

has become the option of choice. With WDM, a number of distinct wavelengths are used to implement separate channels [1]. An optical fiber can carry several channels in parallel, each on a particular wavelength. The number of wavelengths that each fiber can carry simultaneously is limited by the physical characteristics of the fiber and the state of the optical technology used to combine these wavelengths onto the fiber and isolate them off the fiber. With currently available commercial technology, a few tens of wavelengths can be supported within the low-loss window at 1550 nm, but this number is expected to grow rapidly in the near future. Therefore, optical fiber links employing WDM technology have the potential of delivering an aggregate throughput in the order of terabits per second, enough to satisfy the ever-growing demand for more bandwidth per user on a sustained, long-term basis.

Unfortunately, because of the mismatch between aggregate fiber capacity and peak electronic processing speeds, simply upgrading existing point-to-point fiber links to WDM creates the well-known *electrooptic bottleneck* [2]: rather than achieving the multiterabit-per-second throughput of the fiber, one has to settle for the multigigabit-per-second throughput that can be expected of the electronic devices where the optical signals terminate. Overcoming the electrooptic bottleneck, therefore, involves the design of properly structured architectures to interconnect the fiber links. An optical WDM network is a network with optical fiber transmission links and with an architecture that is designed to exploit the unique features of fibers and WDM. Such networks offer the promise of an all-optical information highway capable of supporting a wide range of applications that involve the transport of massive amounts of data and/or require very fast response times. Such applications include video on demand and teleconferencing, telemedicine applications, multimedia document distribution, remote supercomputer visualization, and many more to come. Consequently, optical WDM networks have been a subject of extensive research both theoretically and experimentally [3,4].

The architecture for wide-area WDM networks that is widely expected to form the basis for a future all-optical infrastructure is built on the concept of *wavelength routing*. A wavelength routing network, shown in Fig. 1, consists of two types of nodes: *optical cross-connects* (OXCs), which connect the fibers in the network, and *edge nodes*, which provide the interface between non-optical end systems (such as IP routers, ATM switches, or supercomputers) and the optical core. Access nodes provide the terminating points (sources and destinations) for the optical signal paths; the communication paths may continue outside the optical part of the network in electrical form.

The services that a wavelength routed network offers to end systems attached to edge nodes are in the form of *logical* connections implemented using *lightpaths*. Lightpaths (also referred to as λ-*channels*), are clear optical paths between two edge nodes, and are shown in Fig. 1 as red and green directed lines. Information transmitted on a lightpath does not undergo any conversion to and from electrical form within the optical network, and thus, the architecture of the optical network nodes can be very simple because they do not need to do any signal processing. Furthermore, since a lightpath behaves as a literally transparent "clear channel" between the source and destination edge node, there is nothing in the signal path to limit the throughput of the fibers.

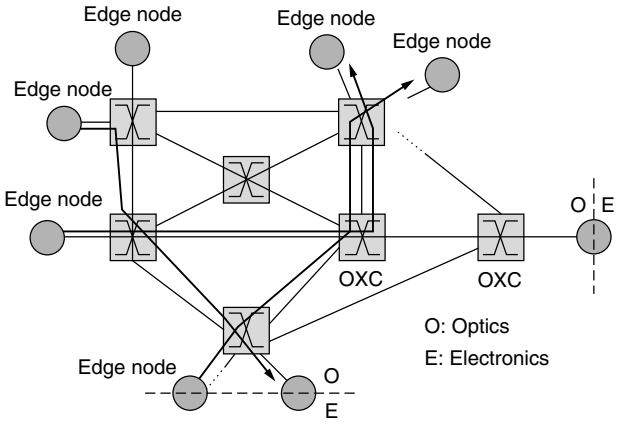


Figure 1. A wavelength-routed WDM network.

The OXCs provide the switching and routing functions for supporting the logical connections between edge nodes. An OXC takes in an optical signal at each wavelength at an input port, and can switch it to a particular output port, independent of the other wavelengths. An OXC with N input and N output ports capable of handling W wavelengths per port can be thought of as W independent $N \times N$ switches. These switches have to be preceded by a wavelength demultiplexer and followed by a wavelength multiplexer to implement an OXC, as shown in Fig. 2. Thus, an OXC can cross-connect the different wavelengths from the input to the output, where the connection pattern of each wavelength is independent of the others. By appropriately configuring the OXCs along the physical path, a logical connection (lightpath) may be established between any pair of edge nodes.

A unique feature of optical WDM networks is the tight coupling between routing and wavelength selection. As can be seen in Fig. 1, a lightpath is implemented by selecting a path of physical links between the source and destination edge nodes, and reserving a particular wavelength on each of these links for the lightpath. Thus, in establishing an optical connection we must deal with both routing (selecting a suitable path) and wavelength assignment (allocating an available wavelength for the connection). The resulting problem is referred to as the *routing and wavelength assignment* (RWA) problem [5], and is significantly more difficult than the routing problem in electronic networks. The additional complexity arises from the fact that routing and wavelength assignment are subject to the following two constraints:

1. *Wavelength continuity constraint* — a lightpath must use the same wavelength on all the links along its path from source to destination edge node. This constraint is illustrated in Fig. 1 by representing

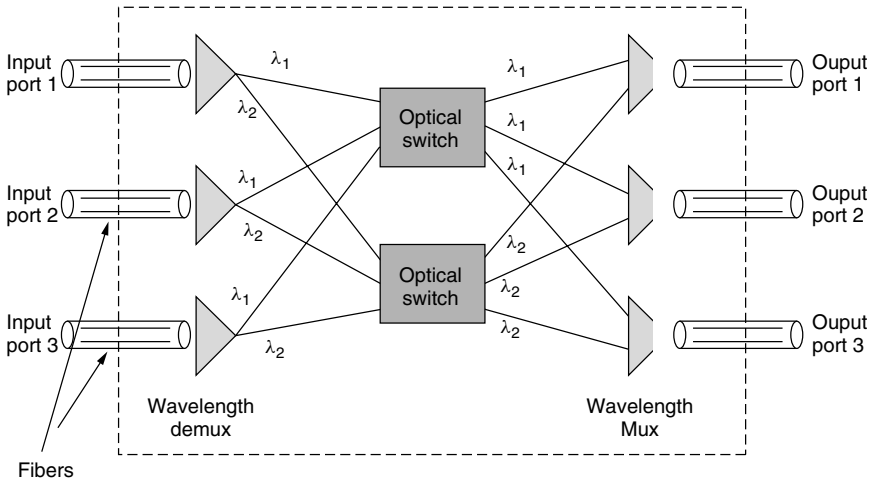


Figure 2. A 3×3 optical cross-connect (OXC) with two wavelengths per fiber.

each lightpath with a single color (wavelength) along all the links in its path.

2. *Distinct wavelength constraint* — all lightpaths using the same link (fiber) must be allocated distinct wavelengths. In Fig. 1 this constraint is satisfied since the two lightpaths sharing a link are shown in different colors (wavelengths).

The RWA problem in optical networks is illustrated in Fig. 3, where it is assumed that each fiber supports two wavelengths. The effect of the wavelength continuity constraint is represented by replicating the network into as many copies as the number of wavelengths (in this case, two). If wavelength i is selected for a lightpath, the source and destination edge node communicate over the ith copy of the network. Thus, finding a path for a connection may potentially involve solving W routing problems for a network with W wavelengths, one for each copy of the network.

The wavelength continuity constraint may be relaxed if the OXCs are equipped with *wavelength converters* [6]. A wavelength converter is a single input/output device that converts the wavelength of an optical signal arriving at its input port to a different wavelength as the signal departs from its output port, but otherwise leaves the optical signal unchanged. In OXCs without a wavelength conversion capability, an incoming signal at port p_i on wavelength λ can be optically switched to any port p_j, but must leave the OXC on the same wavelength λ. With wavelength converters, this signal could be optically switched to any port p_j on some other wavelength λ'. Thus, wavelength conversion allows a lightpath to use different wavelengths along different physical links.

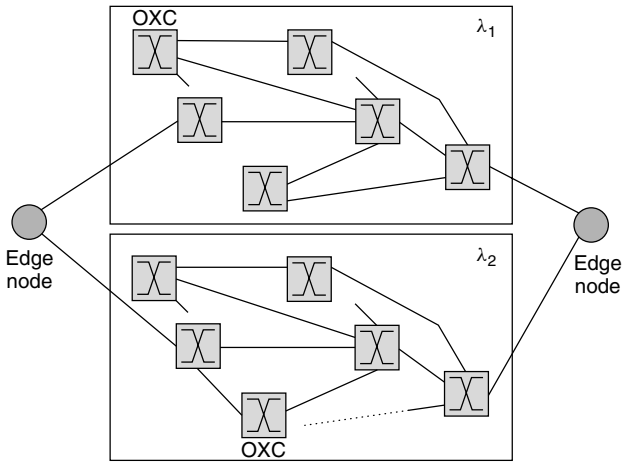


Figure 3. The RWA problem with two wavelengths per fiber.

Different levels of wavelength conversion capability are possible. Figure 4 illustrates the differences for single-input and single-output port situations; the case for multiple ports is more complicated but similar. *Full wavelength conversion* capability implies that any input wavelength may be converted to any other wavelength. *Limited wavelength conversion* [7] denotes that each input wavelength may be converted to any of a specific set of wavelengths, which is not the set of all wavelengths for at least one input wavelength. A special case of this is *fixed wavelength conversion*, where each input wavelength can be converted to exactly one other wavelength. If each wavelength is "converted" only to itself, then we have no conversion.

The advantage of full wavelength conversion is that it removes the wavelength continuity constraint, making it possible to establish a lightpath as long as each link along the path from source to destination has a free wavelength (which could be different for different links). As a result, the RWA problem reduces to the classical routing problem, that is, finding a suitable path for each connection in the

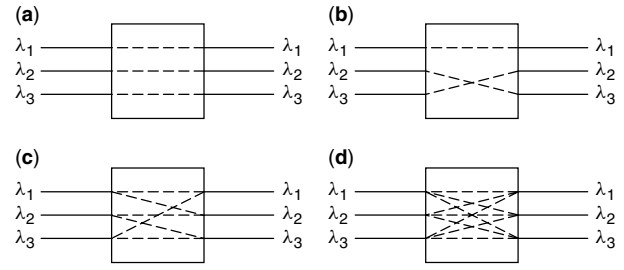


Figure 4. Wavelength conversion: (**a**) no conversion; (**b**) fixed conversion; (**c**) limited conversion; (**d**) full conversion.

network. Referring to Fig. 3, full wavelength conversion collapses the W copies of the network into a single copy on which the routing problem is solved. On the other hand, with limited conversion, the RWA problem becomes more complex than with no conversion. To see why, note that employing limited conversion at the OXCs introduces links between *some* of the network copies of Fig. 3. For example, if wavelength λ_1 can be converted to wavelength λ_2 but not to wavelength λ_3, then links must be introduced from each OXC in copy 1 of the network to the corresponding OXC in copy 2, but not to the corresponding OXC in copy 3. When selecting a path for the connection, at each OXC there is the option of remaining at the same network copy or moving to another one, depending on the conversion capability of the OXC. Since the number of alternatives increases exponentially with the number of OXCs that need to be traversed, the complexity of the RWA problem increases accordingly.

Wavelength conversion (full or limited) increases the routing choices for a given lightpath (i.e., makes more efficient use of wavelengths), resulting in better performance. Since converter devices increase network cost, a possible middle ground is to use *sparse conversion*, that is, to employ converters in some, but not all, OXCs in the network. In this case, a lightpath must use the same wavelength along each link in a segment of its path between OXCs equipped with converters, but it may use a different wavelength along the links of another such segment. It has been shown that implementing full conversion at a relatively small fraction of the OXCs in the network is sufficient to achieve almost all the benefits of conversion [8,9].

Routing and wavelength assignment is the fundamental control problem in optical WDM networks. Since the performance of a network depends not only on its physical resources (OXCs, converters, fibers links, number of wavelengths per fiber, etc.) but also on how it is controlled, the objective of an RWA algorithm is to achieve the best possible performance within the limits of physical constraints. The RWA problem can be cast in numerous forms. The different variants of the problem, however, can be classified under one of two broad versions: a static RWA, whereby the traffic requirements are known in advance, and a dynamic RWA, in which a sequence of lightpath requests arrive in some random fashion. Sections 2 and 3 discuss the static and dynamic versions, respectively, of the RWA problem, and present some algorithms to solve them. Finally, Section 4 presents the multicast RWA problem and algorithms to build *light trees* that connect a source edge node to multiple destinations.

2. STATIC ROUTING AND WAVELENGTH ASSIGNMENT

If the traffic patterns in the network are reasonably well known in advance and any traffic variations take place over long timescales, the most effective technique for establishing optical connections (lightpaths) between edge nodes is by formulating and solving a static RWA problem. For example, static RWA is appropriate for provisioning a set of semipermanent connections. Since these connections are assumed to remain in place for relatively long periods of time, it is worthwhile to attempt to optimize the way in which network resources (e.g., physical links and wavelengths) are assigned to each connection, even though optimization may require a considerable computational effort.

A solution to the static RWA problem consists of a set of long-lived lightpaths that create a *logical* (or *virtual*) topology among the edge nodes. This virtual topology is embedded onto the physical topology of optical fiber links and OXCs. Accordingly, the static RWA problem is often referred to as the *virtual topology design* problem [10]. In the virtual topology, there is a directed link from edge node s to edge node d if a lightpath originating at s and terminating at d is set up (refer also to Fig. 1), and edge node s is said to be "one hop away" from edge node d in the virtual topology, although the two nodes may be separated by a number of physical links. The type of virtual topology that can be created is usually constrained by the underlying physical topology. In particular, it is generally not possible to implement fully connected virtual topologies: for N edge nodes this would require each edge node to maintain $N-1$ lightpaths and the optical network to support a total of $N(N-1)$ lightpaths. Even for modest values of N, this degree of connectivity is beyond the reach of current optical technology, in terms of both the number of wavelengths that can be supported and in the optical hardware (transmitters and receivers) required at each edge node.

In its most general form, the RWA problem is specified by providing the physical topology of the network and the traffic requirements. The physical topology corresponds to the deployment of cables in some existing fiber infrastructure, and is given as a graph $G_p(V, E_p)$, where V is the set of OXCs and E_p is the set of fibers that interconnect them. The traffic requirements are specified in a traffic matrix $\mathbf{T} = [\rho p_{sd}]$, where ρp_{sd} is a measure of the long-term traffic flowing from source edge node s to destination edge node d [11]. Quantity ρ represents the (deterministic) total offered load to the network, while the p_{sd} parameters define the distribution of the offered traffic.

Routing and wavelength assignment are considered together as an optimization problem using mixed-integer programming (MIP) formulations. Usually, the objective of the formulation is to minimize the maximum congestion level in the network subject to network resource constraints [10,12]. While other objective functions are possible, such as minimizing the average weighted number of hops or minimizing the average packet delay, minimizing network congestion is preferable since it can lead to mixed-integer linear programming (MILP) formulations. While we do not present the RWA problem formulation here, the interested reader may refer to the literature [11–12]. These formulations turn out to have extremely large numbers of variables, and are intractable for large networks. This fact has motivated the development of heuristic approaches for finding good solutions efficiently.

Before we describe the various heuristic approaches, we note that the static RWA problem can be logically decomposed into four subproblems. The decomposition is approximate or inexact, in the sense that solving the subproblems in sequence and combining the solutions may

not result in the optimal solution for the fully integrated problem, or some later subproblem may have no solution given the solution obtained for an earlier subproblem, so no solution to the original problem may be obtained. However, the decomposition provides insight into the structure of the RWA problem and is a first step towards the design of effective heuristics. Assuming no wavelength conversion, the subproblems are as follows:

1. *Topology subproblem* — determine the logical topology to be imposed on the physical topology; that is, determine the lightpaths in terms of their source and destination edge nodes.
2. *Lightpath routing subproblem* — determine the physical links which each lightpath consists of; that is, route the lightpaths over the physical topology.
3. *Wavelength assignment subproblem* — determine the wavelength each lightpath uses; that is, assign a wavelength to each lightpath in the logical topology so that wavelength restrictions are obeyed for each physical link.
4. *Traffic routing subproblem* — route packet traffic between source and destination edge nodes over the logical topology obtained.

A large number of heuristic algorithms have been developed in the literature to solve the general static RWA problem discussed here or its many variants. Overall, however, the different heuristics can be classified into three broad categories: (1) algorithms that solve the overall MILP problem suboptimally, (2) algorithms that tackle only a subset of the four subproblems, and (3) algorithms that address the problem of embedding regular logical topologies onto the physical topology.

Suboptimal solutions can be obtained by applying conventional tools developed for complex optimization problems directly to the MILP problem. One technique is to use LP relaxation followed by rounding [13]. In this case, the integer constraints are relaxed creating a nonintegral problem which can be solved by some linear programming method, and then a rounding algorithm is applied to obtain a new solution that obeys the integer constraints. Alternatively, genetic algorithms or simulated annealing [14] can be applied to obtain locally optimal solutions. The main drawback of these approaches is that it is difficult to control the quality of the final solution for large networks: simulated annealing is computationally expensive and thus, it may not be possible to adequately explore the state space, while LP relaxation may lead to solutions from which it is difficult to apply rounding algorithms.

Another class of algorithms tackles the RWA problem by initially solving the first three subproblems listed above; traffic routing is then performed by employing well-known routing algorithms on the logical topology. One approach for solving the three subproblems is to maximize the amount of traffic that is carried on one-hop lightpaths, where traffic that is routed from source to destination edge node directly on a lightpath. A greedy approach taken is to create lightpaths between edge nodes in order of decreasing traffic demands as long as the wavelength continuity and distinct wavelength constraints are satisfied [15]. This algorithm starts with a logical topology with no links (lightpaths) and sequentially adds lightpaths as long as doing so does not violate any of the problem constraints. The reverse approach is also possible [16]; starting with a fully connected logical topology, an algorithm sequentially removes the lightpath carrying the smallest traffic flows until no constraint is violated. At each step (i.e., after removing a lightpath), the traffic routing subproblem is solved in order to find the lightpath with the smallest flow.

The third approach to RWA is to start with a given logical topology, thus avoiding the need to directly solve the first of the four subproblems listed above. Regular topologies are good candidates as logical topologies since they are well understood and results regarding bounds and averages (e.g., for hop lengths) are easier to derive. Algorithms for routing traffic on a regular topology are usually simple, so the traffic routing subproblem can be trivially solved. Also, regular topologies possess inherent load balancing characteristics which are important when the objective is to minimize the maximum congestion.

Once a regular topology is decided on as the one to implement the logical topology, it remains to decide which physical node will realize each given node in the regular topology (this is usually referred to as the *node mapping* subproblem), and which sequence of physical links will be used to realize each given edge (lightpath) in the regular topology (this *path mapping* subproblem is equivalent to the lightpath routing and wavelength assignment subproblems discussed earlier). This procedure is usually referred to as *embedding* a regular topology in the physical topology. Both the node and path mapping subproblems are intractable, and heuristics have been proposed in the literature [16,17]. For instance, a heuristic for mapping the nodes of shuffle topologies based on the gradient algorithm has been developed [17].

Given that all the algorithms for the RWA problem are based on heuristics, it is important to be able to characterize the quality of the solutions obtained. To this end, one must resort to comparing the solutions to known bounds on the optimal solution. A comprehensive discussion of bounds for the RWA problem and the theoretical considerations involved in deriving them can be found in [10]. A simulation-based comparison of the relative performance of the three classes of heuristic for the RWA problem has been presented [12]. The results indicate that the second class of algorithms discussed earlier achieve the best performance.

3. DYNAMIC ROUTING AND WAVELENGTH ASSIGNMENT

Under a dynamic traffic scenario, edge nodes submit to the network requests for lightpaths to be set up as needed. Thus, connection requests are initiated in some random fashion. Depending on the state of the network at the time of a request, the available resources may or may not be sufficient to establish a lightpath between the corresponding source–destination edge node pair. The network state consists of the physical path (route) and

wavelength assignment for all active lightpaths. The state evolves randomly in time as new lightpaths are admitted and existing lightpaths are released. Thus, each time a request is made, an algorithm must be executed in real time to determine whether it is feasible to accommodate the request, and, if so, to perform routing and wavelength assignment. If a request for a lightpath cannot be accepted because of lack of resources, it is blocked.

Because of the real-time nature of the problem, RWA algorithms in a dynamic traffic environment must be very simple. Since combined routing and wavelength assignment is a hard problem, a typical approach to designing efficient algorithms is to decouple the problem into two separate subproblems: the routing problem and the wavelength assignment problem. Consequently, most dynamic RWA algorithms for wavelength routed networks consist of the following general steps:

1. Compute a number of candidate physical paths for each source–destination edge node pair and arrange them in a path list.
2. Order all wavelengths in a wavelength list.
3. Starting with the path and wavelength at the top of the corresponding list, search for a feasible path and wavelength for the requested lightpath.

The specific nature of a dynamic RWA algorithm is determined by the number of candidate paths and how they are computed, the order in which paths and wavelengths are listed, and the order in which the path and wavelength lists are accessed.

Let us first discuss the routing subproblem. If a *static* algorithm is used, the paths are computed and ordered independently of the network state. With an *adaptive* algorithm, on the other hand, the paths computed and their order may vary according to the current state of the network. A static algorithm is executed offline and the computed paths are stored for later use, resulting in low latency during lightpath establishment. Adaptive algorithms are executed at the time when a lightpath request arrives and require network nodes to exchange information regarding the network state. Lightpath setup delay may also increase, but in general adaptive algorithms improve network performance.

The number of path choices for establishing an optical connection is another important parameter. A *fixed* routing algorithm is a static algorithm in which every source–destination edge node pair is assigned a single path. With this scheme, a connection is blocked if there is no wavelength available on the designated path at the time of the request. In *fixed–alternate* routing, a number k, $k > 1$, of paths are computed and ordered off-line for each source–destination edge node pair. When a request arrives, these paths are examined in the specified order, and the first one with a free wavelength is used to establish the lightpath. The request is blocked if no wavelength is available in any of the k paths. Similarly, an adaptive routing algorithm may compute a single path, or a number of alternate paths at the time of the request. A hybrid approach is to compute k paths offline, however, the order in which the paths are considered is determined according to the network state at the time the connection request is made (e.g., least to most congested).

In most practical cases, the candidate paths for a request are considered in increasing order of pathlength. *Pathlength* is typically defined as the sum of the weights assigned to each physical link along the path, and the weights are chosen according to some desirable routing criterion. Since weights can be assigned arbitrarily, they offer a wide range of possibilities for selecting path priorities. For example, in a static (fixed–alternate) routing algorithm, the weight of each link could be set to 1, or to the physical distance of the link. In the former case, the path list consists of the k minimum-hop paths, while in the latter the candidate paths are the k minimum-distance paths (where *distance* is defined as the geographic length). In an adaptive routing algorithm, link weights may reflect the load or "interference" on a link (i.e., the number of active lightpaths sharing the link). By assigning small weights to least loaded links, paths with larger number of free channels on their links rise to the head of the path list, resulting in a *least-loaded* routing algorithm. Paths that are congested become "longer" and are moved further down the list; this tends to avoid heavily loaded bottleneck links. Many other weighting functions are possible.

When pathlengths are sums of link weights, the k shortest path algorithm [18] can be used to compute candidate paths. Each path is checked in order of increasing length, and the first that is feasible is assigned the first free wavelength in the wavelength list. However, the k shortest paths constructed by this algorithm usually share links. Therefore, if one path in the list is not feasible, it is likely that other paths in the list with which it shares a link will also be infeasible. To reduce the risk of blocking, the k shortest paths can be computed so as to be pairwise link-disjoint. This can be accomplished as follows. When computing the ith shortest path, $i = 1, \ldots, k$, the links used by the first $i - 1$ paths are removed from the original network topology and Dijkstra's shortest path algorithm [19] is applied to the resulting topology. This approach increases the chances of finding a feasible path for a connection request.

Let us now discuss the wavelength assignment subproblem, which is concerned with the manner in which the wavelength list is ordered. For a given candidate path, wavelengths are considered in the order in which they appear in the list to find a free wavelength for the connection request. Again, we distinguish between the static and adaptive cases. In the *static* case, the wavelength ordering is fixed (e.g., the list is ordered by wavelength number). The idea behind this scheme, also referred to as *first-fit*, is to pack all the in-use wavelengths toward the top of the list so that wavelengths toward the end of the list will have higher probability of being available over long continuous paths. In the *adaptive* case, the ordering of wavelengths is typically based on usage. Usage can be defined either as the number of links in the network in which a wavelength is currently used, or as the number of active connections using a wavelength. Under the *maximum-reuse* method, the most used wavelengths are considered first (i.e., wavelength are considered in order of decreasing usage). The rationale

behind this method is to reuse active wavelengths as much as possible before trying others, packing connections into fewer wavelengths and conserving the spare capacity of less used wavelengths. This is turn makes it more likely to find wavelengths that satisfy the continuity requirement over long paths. Under the *minimum-reuse* method, wavelengths are tried in the order of increasing usage. This scheme attempts to balance the load as equally as possible among all the available wavelengths. However, minimum-reuse assignment tends to "fragment" the availability of wavelengths, making it less likely that the same wavelength is available throughout the network for connections that traverse longer paths.

Both reuse schemes introduce communication overhead because they require global network information in order to compute the usage of each wavelength. The first-fit scheme, on the other hand, requires no global information, and since it does not need to order wavelengths in real time, it has significantly lower computational requirements than either maximum reuse or minimum reuse. Another adaptive scheme that avoids the communication and computational overhead of maximum reuse and minimum reuse is *random* wavelength assignment. With this scheme, the set of wavelengths that are free on a particular path is first determined. Among the available wavelengths, one is chosen randomly (usually with uniform probability) and assigned to the requested lightpath.

We note that in networks in which all OXCs are capable of wavelength conversion, the wavelength assignment problem is trivial; since a lightpath can be established as long as at least one wavelength is free at each link and different wavelengths can be used in different links, the order in which wavelengths are assigned is not important. On the other hand, when only a fraction of the OXCs employ converters (i.e., a sparse conversion scenario), a wavelength assignment scheme is again required to select a wavelength for each segment of a connection's path that originates and terminates at an OXC with converters. In this case, the same assignment policies discussed above for selecting a wavelength for the end-to-end path can also be used to select a wavelength for each path segment between OXCs with converters.

The performance of a dynamic RWA algorithm is generally measured in terms of the *call blocking probability*, that is, the probability that a lightpath cannot be established in the network because of lack of resources (e.g., link capacity or free wavelengths). Even in the case of simple network topologies (such as rings) or simple routing rules (such as fixed routing), the calculation of blocking probabilities in WDM networks is extremely difficult. In networks with arbitrary mesh topologies, and/or when using alternate or adaptive routing algorithms, the problem is even more complex. These complications arise from both the link load dependencies (due to interfering lightpaths) and the dependencies among the sets of active wavelengths in adjacent links (due to the wavelength continuity constraint). Nevertheless, the problem of computing blocking probabilities in wavelength-routed networks has been extensively studied in the literature, and approximate analytical techniques that capture the effects of link load and wavelength dependencies have been developed in [8,9,20]. A detailed comparison of the performance of various wavelength assignment schemes in terms of call blocking probability can be found in Zhu et al. [21].

Although important, average blocking probability (computed over all connection requests) does not always capture the full effect of a particular dynamic RWA algorithm on other aspects of network behavior, in particular, fairness. In this context, *fairness* refers to the variability in blocking probability experienced by lightpath requests between the various edge node pairs, such that lower variability is associated with a higher degree of fairness. In general, any network has the property that longer paths are likely to experience higher blocking than shorter ones. Consequently, the degree of fairness can be quantified by defining the *unfairness factor* as the ratio of the blocking probability on the longest path to that on the shortest path for a given RWA algorithm. Depending on the network topology and the RWA algorithm, this property may have a cascading effect that can result in an unfair treatment of the connections between more distant edge node pairs — blocking of long lightpaths leaves more resources available for short lightpaths, so that the connections established in the network tend to be short ones. These shorter connections "fragment" the availability of wavelengths, and thus the problem of unfairness is more pronounced in networks without converters since finding long paths that satisfy the wavelength continuity constraint is more difficult than without this constraint.

Several studies [8,9,20] have examined the influence of various parameters on blocking probability and fairness, and some of the general conclusions include the following:

- Wavelength conversion significantly affects fairness. Networks employing converters at all OXCs sometimes exhibit orders of magnitude improvement in fairness (as reflected by the unfairness factor) compared to networks with no conversion capability, despite the fact that the improvement in overall blocking probability is significantly less pronounced. It has also been shown that equipping a relatively small fraction (typically, 20–30%) of all OXCs with converters is sufficient to achieve most of the fairness benefits due to wavelength conversion.
- Alternate routing can significantly improve the network performance in terms of both overall blocking probability and fairness. In fact, having as few as three alternate paths for each connection may in some cases (depending on the network topology) achieve almost all the benefits (in terms of blocking and fairness) of having full wavelength conversion at each OXC with fixed routing.
- Wavelength assignment policies also play an important role, especially in terms of fairness. The random and minimum-reuse schemes tend to "fragment" the wavelength availability, resulting in large unfairness factors (with minimum-reuse having the worst performance). On the other hand, the maximum-reuse assignment policy achieves the best performance in terms of fairness. The first-fit scheme exhibits a

behavior very similar to maximum-reuse in terms of both fairness and overall blocking probability, and has the additional advantage of being easier and less expensive to implement.

4. MULTICAST ROUTING AND WAVELENGTH ASSIGNMENT

In Sections 2 and 3, we considered static and dynamic RWA algorithms, respectively, for establishing lightpaths in optical networks. In Ref. 22, the concept of a lightpath was generalized into that of a *light tree*, which, like a lightpath, is a clear channel originating at given source node and implemented with a single wavelength. But unlike a lightpath, a light tree has multiple destination nodes, hence it is a point-to-multipoint channel. The physical links implementing a light tree form a tree, rooted at the source node, rather than a path in the physical topology, hence the name. That study [22] focused on virtual topology design (i.e., static RWA) for point-to-point traffic and observed that, since a light tree is a more general representation of a lightpath, the set of virtual topologies that can be implemented using light trees is a superset of the virtual topologies that can be implemented only using lightpaths. Thus, for any given virtual topology problem, an optimal solution using light trees is guaranteed to be at least as good and possibly an improvement over the optimal solution obtained using only lightpaths. Furthermore, it was demonstrated that by extending the lightpath concept to a light tree, the network performance (in terms of average packet hops) can be improved while the network cost (in terms of the number of optical transmitters and receivers required) decreases.

Light trees are implemented by employing optical devices known as *power splitters* [2] at the OXCs. A power splitter has the ability to split an incoming signal, arriving at some wavelength λ, into up to m outgoing signals, $m \geq 2$; m is referred to as the *fanout* of the power splitter. Each of these m signals is then independently switched to a different output port of the OXC. Note that, because of the splitting operation and associated losses, the optical signals resulting from the splitting of the original incoming signal must be amplified before leaving the OXC. Also, to ensure the quality of each outgoing signal, the fanout m of the power splitter may have to be limited to a small number. If the OXC is also capable of wavelength conversion, each of the m outgoing signal may be shifted, independently of the others, to a wavelength different than the incoming wavelength λ. Otherwise, all m outgoing signals must be on the same wavelength λ.

While the authors of Ref. 22 considered mainly point-to-point traffic, another attractive feature of light trees is the inherent capability for performing multicasting in the optical domain (as opposed to performing multicasting at a higher layer, e.g., the network layer, which requires electrooptic conversion). Such wavelength-routed light trees are useful for transporting high-bandwidth, real-time applications such as high-definition TV (HDTV). Therefore, OXCs equipped with power splitters will be referred to as *multicast-capable* OXCs (MC-OXCs). Note that, just like with converter devices, incorporating power splitters within an OXC is expected to increase the network cost because of the large amount of power amplification and the difficulty of fabrication.

With the availability of MC-OXCs and the existence of multicast traffic demands, the problem of establishing light trees to satisfy these demands arises. We will call this problem the *multicast routing and wavelength assignment* (MC-RWA) problem. MC-RWA bears many similarities to the RWA problem discussed in Sections 2 and 3. Specifically, the tight coupling between routing and wavelength assignment remains, and even becomes stronger; in the absence of wavelength conversion the same wavelength must be used by the multicast connection not only along the links of a single path but also along the links of the whole light tree. Since the construction of optimal trees for routing multicast connections is by itself a hard problem [23], the combined MC-RWA problem becomes even harder. Depending on the nature of traffic demands, we also distinguish between static and dynamic MC-RWA problems. As we already know, optimal solutions for the point-to-point RWA problems are not practically obtainable, and with a more general construct (the light tree) and hence a much larger search space, this will be even more true for the MC-RWA problems. In general, the approaches to tackling the static and dynamic MC-RWA problems are similar to the ones we described for the static and dynamic RWA problems, respectively. The challenge in this case is to design heuristics that can cope with the increased complexity of the problem and yet produce good solutions. In the following we summarize the most recent work on multicasting in optical networks, but the reader should keep in mind that this is an area of current research.

The benefits of multicasting in wavelength routed optical networks were first demonstrated by Malli et al. [24]. Specifically, it was shown that using light trees (spanning the source and destination nodes) rather than individual parallel lightpaths (each connecting the source to an individual destination) requires fewer wavelengths and consumes a significantly lower amount of bandwidth. In another paper [25] both the static and the dynamic MC-RWA problems were studied. A MILP formulation that maximizes the total number of multicast connections was presented for the static MC-RWA problem. Rather than providing heuristic algorithms for solving the MILP, bounds on the objective function were presented by relaxing the integer constraints. The dynamic MC-RWA problem, on the other hand, was solved by decoupling the routing and wavelength assignment problems. A number of *alternate* trees were constructed for each multicast connection using existing routing algorithms. When a request for a connection arrives, the associated trees are considered in a fixed order. For each tree, wavelengths are also considered in a fixed order (i.e., the first-fit strategy). The connection is blocked if no free wavelength is found in any of the trees associated with the multicast connection.

Finally, the problem of constructing trees for routing multicast connections was studied [26] independently of wavelength assignment, under the assumption that not all OXCs are multicast capable, that is, that there is a limited number of MC-OXCs in the network. Four new algorithms were developed for routing multicast

connections under this *sparse light splitting* scenario. While the algorithms differ slightly from each other, the main idea to accommodate sparse splitting is to start with the assumption that all OXCs in the network are multicast capable and use an existing algorithm to build an initial tree. Such a tree is infeasible if a non-multicast-capable OXC is a branching point. In this case, all but one branches out of this OXC are removed, and destination nodes in the removed branches have to join the tree at a MC-OXC.

BIOGRAPHY

George N. Rouskas received his degree in electrical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1989, and M.S. and Ph.D. degrees in computer science from the College of Computing, Georgia Institute of Technology, Atlanta, Georgia, in 1991 and 1994, respectively. He joined the Department of Computer Science, North Carolina State University, Raleigh, North Carolina, in August 1994, where he is currently an associate professor. During the 2000–2001 academic year he spent a sabbatical term at Vitesse Semiconductor, Morrisville, North Carolina, and in May and June 2000 he was an invited professor at the University of Evry, Evry, France. He is a recipient of a 1997 NSF Faculty Early Career Development (CAREER) Award, and a coauthor of a paper that received the Best Paper Award at the 1998 SPIE conference on all-optical networking. He also received the 1995 Outstanding New Teacher Award from the Department of Computer Science, North Carolina State University, and the 1994 Graduate Research Assistant Award from the College of Computing, Georgia Tech. He was a coguest editor for the *IEEE Journal on Selected Areas in Communications*, Special Issue on Protocols and Architectures for Next Generation Optical WDM Networks, October, 2000, and is on the editorial boards of the *IEEE/ACM Transactions on Networking, Computer Networks, and Optical Networks*. Dr. Rouskas' interests include network architectures and protocols, optical networks, multicast communication, and performance evaluation.

BIBLIOGRAPHY

1. T. E. Stern and K. Bala, *Multiwavelength Optical Networks*, Prentice-Hall, Upper Saddle River, NJ, 2000.
2. B. Mukherjee, *Optical Communication Networking*, McGraw-Hill, New York, 1997.
3. O. Gerstel et al., eds., Special issue on protocols and architectures for next generation optical WDM networks, *IEEE J. Select. Areas Commun.* **18**(10) (Oct. 2000).
4. G.-K. Chung, K.-I. Sato, and D. K. Hunter, eds., Special issue on optical networks, *J. Lightwave Technol.* **18**(12) (Dec. 2000).
5. H. Zang, J. P. Jue, and B. Mukherjee, A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks, *Opt. Networks* **1**(1): 47–60 (Jan. 2000).
6. B. Ramamurty and B. Mukherjee, Wavelength conversion in WDM networking, *IEEE J. Select. Areas Commun.* **16**(7): 1061–1073 (Sept. 1998).
7. V. Sharma and E. A. Varvarigos, Limited wavelength translation in all-optical WDM mesh networks, *Proc. INFOCOM'98 IEEE* 893–901 (March 1999).
8. Y. Zhu, G. N. Rouskas, and H. G. Perros, A path decomposition algorithm for computing blocking probabilities in wavelength routing networks, *IEEE/ACM Trans. Network.* **8**(6): 747–762 (Dec. 2000).
9. S. Subramaniam, M. Azizoglu, and A. Somani, All-optical networks with sparse wavelength conversion, *IEEE/ACM Trans. Network.* **4**(4): 544–557 (Aug. 1996).
10. R. Dutta and G. N. Rouskas, A survey of virtual topology design algorithms for wavelength routed optical networks, *Opt. Networks Mag.* **1**(1): 73–89 (Jan. 2000).
11. R. Ramaswami and K. N. Sivarajan, Design of logical topologies for wavelength-routed optical networks, *IEEE J. Select. Areas Commun.* **14**(5): 840–851 (June 1996).
12. E. Leonardi, M. Mellia, and M. A. Marsan, Algorithms for the logical topology design in WDM all-optical networks, *Opt. Networks* **1**(1): 35–46 (Jan. 2000).
13. D. Banerjee and B. Mukherjee, A practical approach for routing and wavelength assignment in large wavelength-routed optical networks, *IEEE J. Select. Areas Commun.* **14**(5): 903–908 (June 1996).
14. B. Mukherjee et al., Some principles for designing a wide-area WDM optical network, *IEEE/ACM Trans. Network.* **4**(5): 684–696 (Oct. 1996).
15. Z. Zhang and A. Acampora, A heuristic wavelength assignment algorithm for multihop WDM networks with wavelength routing and wavelength reuse, *IEEE/ACM Trans. Network.* **3**(3): 281–288 (June 1995).
16. I. Chlamtac, A. Ganz, and G. Karmi, Lightnets: Topologies for high-speed optical networks, *J. Lightwave Technol.* **11**: 951–961 (May/June 1993).
17. S. Banerjee and B. Mukherjee, Algorithms for optimized node placement in shufflenet-based multihop lightwave networks, *Proc. INFOCOM'93 IEEE*, March 1993.
18. E. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart, Winston, 1976.
19. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
20. E. Karasan and E. Ayanoglu, Effects of wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks, *IEEE/ACM Trans. Network.* **6**(2): 186–196 (April 1998).
21. Y. Zhu, G. N. Rouskas, and H. G. Perros, A comparison of allocation policies in wavelength routing networks, *Photon. Network Commun.* **2**(3): 265–293 (Aug. 2000).
22. L. H. Sahasrabuddhe and B. Mukherje, Light-trees: Optical multicasting for improved performance in wavelength-routed networks, *IEEE Commun.* **37**(2): 67–73 (Feb. 1999).
23. S. L. Hakimi, Steiner's problem in graphs and its implications, *Networks* **1**: 113–133 (1971).
24. R. Malli, X. Zhang, and C. Qiao, Benefit of multicasting in all-optical networks, *Proc. SPIE* **3531**: 209–220 (Nov. 1998).
25. G. Sahin and M. Azizoglu, Multicast routing and wavelength assignment in wide-area networks, *Proc. SPIE* **3531**: 196–208 (Nov. 1998).
26. X. Zhang, J. Y. Wei, and C. Qiao, Constrained multicast routing in WDM networks with sparse light splitting, *J. Lightwave Technol.* **18**(12): 1917–1927 (Dec. 2000).

S

SAMPLING OF ANALOG SIGNALS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Many communication signals that are transmitted from a source to a destination are analog signals. Examples of such signals are speech, images, and video. Since the middle of the twentieth century, the trend has been to convert such analog signals to digital form and to transmit the digital signal using a digital modulation technique. At the receiver, the digital signal can be converted back to an analog signal for the user.

An analog signal is converted to a digital signal through the process of sampling the analog signal periodically and quantizing the samples to obtain a digital signal (a sequence of binary digits), which is then transmitted by a digital communication system. The sampling and quantization processes are generally performed by an analog-to-digital (A/D) converter, whose basic functions are illustrated in Fig. 1. Thus, conceptually, we may view A/D conversion of an analog signal as a three-step process:

1. *Sampling*. This is the conversion of a continuous-time signal $x_a(t)$ into a discrete-time signal $x_a(nT) = x(n)$, where $\{x_a(nT)\}$ are samples of $x_a(t)$ taken at times $t = nT$, and where T is the *sampling interval* and n takes integer values.
2. *Quantization*. This is the conversion of the continuous-valued signal samples $\{x(n)\}$ into discrete-valued signal samples $\{x_q(n)\}$, where the values $\{x_q(n)\}$ are selected from a finite set of possible values. The difference $x(n) - x_q(n)$ is called the *quantization error*.
3. *Coding*. This is the process of representing each quantized value $x_q(n)$ by a b-bit binary sequence.

Although we model the A/D conversion process as shown in Fig. 1, in practice the A/D conversion is performed by a single device whose input is the analog signal $x_a(t)$ and whose output is a sequence of b-bit values representing the quantized samples $x_q(n)$.

2. SAMPLING OF ANALOG SIGNALS

As indicated above, an analog signal $x_a(t)$ that is sampled periodically at $t = nT$ results in the sampled sequence

$$x(n) \equiv x_a(nT), \quad -\infty \le n \le \infty \tag{1}$$

The sampling process is illustrated in Fig. 2. The time interval T between successive samples is called the *sampling interval*, and its reciprocal $1/T$ is called the *sampling frequency*, denoted as

$$F_s = \frac{1}{T} \tag{2}$$

The choice of the sampling frequency is governed by the frequency content of the analog signal $x_a(t)$. To establish this basic relationship, let us consider the sampling of the sinusoidal signal

$$x_a(t) = A \cos 2\pi F t \tag{3}$$

Its sampled values are

$$\begin{aligned} x(n) \equiv x_a(nT) &= A \cos 2\pi F n T \\ &= A \cos 2\pi n \frac{F}{F_s} \end{aligned} \tag{4}$$

We note that $x(n)$ is a discrete-time sinusoidal signal having a normalized frequency $f = F/F_s$. We also observe that discrete-time sinusoids whose normalized frequencies are separated by an integer multiple of unity (or 2π radians) are identical. That is

$$x(n) = A \cos 2\pi n(f + k) = A \cos 2\pi n f \tag{5}$$

where k is any integer. We further observe that the highest rate of oscillation in a discrete-time sinusoid is attained

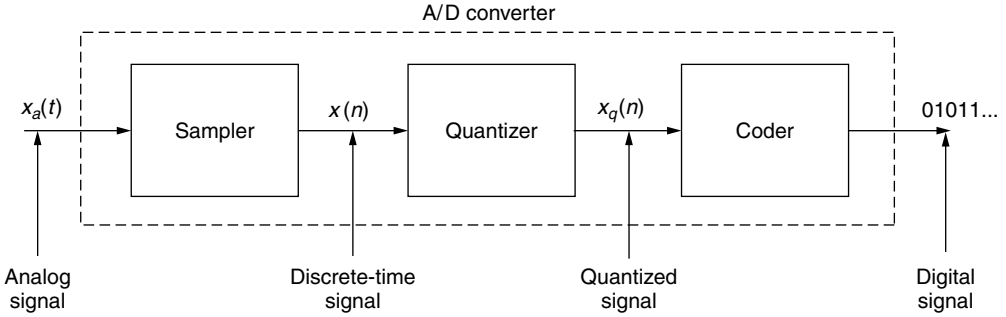


Figure 1. Elements of an analog-to-digital (A/D) converter.

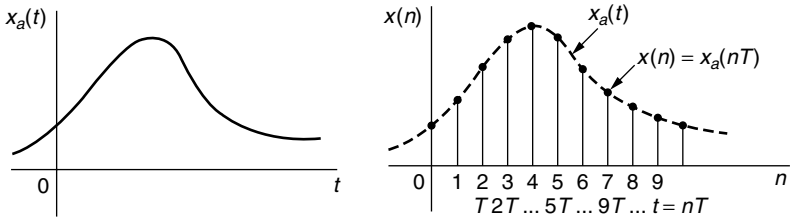


Figure 2. Periodic sampling of an analog signal.

when the normalized frequency $f = \pm\frac{1}{2}$. Any discrete-time sinusoid with frequency $f > \frac{1}{2}$ takes values that are identical to a discrete-time sinusoid whose frequency is contained in the *fundamental range* $-\frac{1}{2} \le f \le \frac{1}{2}$.

The implications of these observations can be fully appreciated by considering the sampling of the two analog sinusoidal signals

$$\begin{aligned} x_1(t) &= \cos 2\pi(10)t \\ x_2(t) &= \cos 2\pi(50)t \end{aligned} \tag{6}$$

at a rate $F_s = 40$ Hz (samples per second). The corresponding discrete-time signals are

$$\begin{aligned} x_1(n) &= \cos 2\pi \left(\frac{10}{40}\right) n = \cos \frac{\pi}{2} n \\ x_2(n) &= \cos 2\pi \left(\frac{50}{40}\right) n = \cos \frac{5\pi}{2} n = \cos \frac{\pi}{2} n \end{aligned} \tag{7}$$

Hence, $x_1(n) = x_2(n)$, so the two sampled signals are identical and, consequently, indistinguishable.

If we are given the sampled values obtained from $\cos \pi n/2$, there is ambiguity as to whether these sampled values correspond to $x_1(t)$ or $x_2(t)$. Since $x_2(t)$ yields exactly the same values as $x_1(t)$ when sampled at $F_s = 40$ samples per second, we say that the frequency $F_2 = 50$ Hz is an alias of the frequency $F_1 = 10$ Hz at the sampling rate of $F_s = 40$ Hz. We also note that F_2 is not the only alias of F_1. At the sampling rate of $F_s = 40$ Hz, the frequencies $F_3 = 90$ Hz, $F_4 = 130$ Hz, and so on, are all aliases of $F_1 = 10$ Hz; that is, the sinusoids $\cos\ 2\pi(F_1 + 40k)t$, $k = 1, 2, \ldots$ sampled at $F_s = 40$ Hz yield identical values.

The preceding observations lead to the conclusion that the highest frequency that can be represented uniquely by sampling an analog signal $x_a(t)$ at a rate F_s is $F_{\max} = F_s/2$. Hence, to avoid aliasing, an analog signal that is to be sampled must be band-limited. This is usually accomplished in practice by prefiltering the analog signal prior to sampling. Thus, the frequency content of the analog signal will be confined to a well-defined frequency band with highest frequency $F_{\max}$. Such a band-limited signal is then sampled at a rate $F_s \ge 2F_{\max}$, so as to avoid aliasing. The critical rate $2F_{\max} = F_N$ is called the *Nyquist sampling rate*.

For example, speech signals that are transmitted over telephone channels are limited to approximately 3200 Hz: $F_{\max} = 3200$ Hz. In this case, the Nyquist sampling rate is $F_N = 6400$ Hz. Such signals are typically sampled at a nominal rate of $F_s = 8000$ Hz. If the speech signal is to be transmitted by pulse code modulation (PCM), for example, each sample is typically (quantized) represented as a 7-bit binary word.

3. FREQUENCY-DOMAIN RELATIONSHIPS

If $x_a(t)$ is an aperiodic signal with finite energy, its (voltage) spectrum $X_a(f)$ is related to $x_a(t)$ by the inverse Fourier transform:

$$x_a(t) = \int_{-\infty}^{\infty} X_a(F) e^{j2\pi Ft} dF \tag{8}$$

The sampled signal sequence is

$$\begin{aligned} x(n) \equiv x_a(nT) &= \int_{-\infty}^{\infty} X_a(F) e^{j2\pi FnT} dF \\ &= \int_{-\infty}^{\infty} X_a(F) e^{j2\pi fF/F_s} dF \end{aligned} \tag{9}$$

The integration range of this integral can be subdivided into an infinite number of intervals of width F_s. Thus, we obtain

$$\begin{aligned} x(n) &= \sum_{k=-\infty}^{\infty} \int_{(k-1/2)F_s}^{(k+1/2)F_s} X_a(F) e^{j2\pi nF/F_s} dF \\ &= \sum_{k=-\infty}^{\infty} \int_{-F_s/2}^{F_s/2} X_a(F - kF_s) e^{j2\pi nF/F_s} dF \\ &= \int_{-F_s/2}^{F_s/2} \left[\sum_{k=-\infty}^{\infty} X_a(F - kF_s) \right] e^{j2\pi nF/F_s} dF \end{aligned} \tag{10}$$

Changing variables in Eq. (10) from F to the normalized frequency $f = F/F_s$ yields

$$x(n) = \left[\int_{-1/2}^{1/2} F_s \sum_{k=-\infty}^{\infty} X_a[(f - k)F_s] \right] e^{j2\pi nf}\, df \tag{11}$$

This equation is simply the inverse Fourier transform that relates the discrete-time signal $x(n)$ to its spectrum $X(f)$. Hence, the spectrum $X(f)$ of the discrete-time signal $x(n)$ is related to the spectrum $X_a(F)$ of the analog signal $x_a(t)$ by the expression

$$X(f) = F_s \sum_{k=-\infty}^{\infty} X_a[f - k)F_s] \tag{12}$$

or, equivalently

$$X\left(\frac{F}{F_s}\right) = F_s \sum_{k=-\infty}^{\infty} X_a(F - kF_s) \tag{13}$$

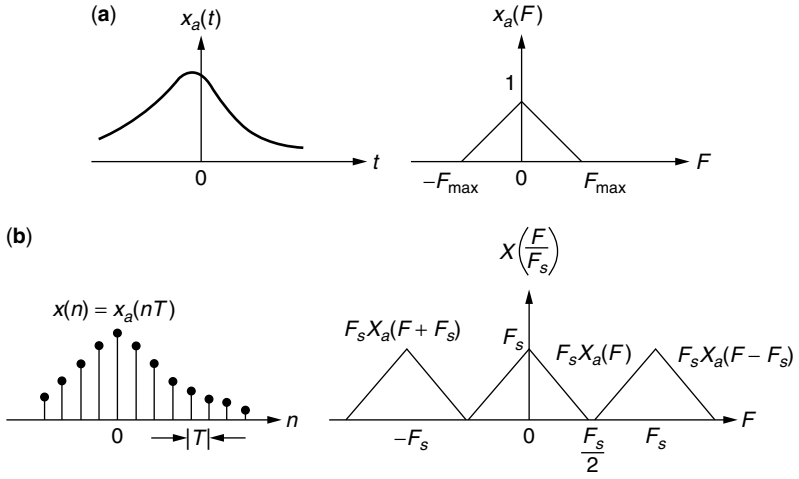


Figure 3. Time-domain and frequency-domain relationships between $x_a(t)$ and $x(n)$.

Figure 3 illustrates the time-domain and frequency-domain relationships between the analog signal $x_a(t)$ and its discrete-time sampled version. We observe that the spectrum of the discrete-time signal $x(n)$ is periodic with period F_s. With $F_s \geq 2F_{\max}$, the spectrum of the discrete-time signal within the fundamental range $|F| \leq F_s/2$ is simply

$$X\left(\frac{F}{F_s}\right) = F_s X_a(F) \tag{14}$$

Hence, the spectrum of the sampled signal is identical (within the scale factor F_s) to the spectrum of the analog signal. This implies that we can reconstruct the analog signal $x_a(t)$ from its samples $x(n)$.

If the sampling frequency F_s is selected such that $F_s < 2F_{\max}$, the periodic continuation of $X_a(f)$, given by Eq. (13) results in spectral overlap. Hence, the spectrum of the sampled signal $x(n)$ contains aliased frequency components of the analog signal spectrum $X_a(F)$. As a consequence, the spectrum of the discrete-time signal $x(n)$ is no longer equal to $F_s X_a(F)$ in the fundamental frequency range $|F| \leq F_s/2$. Therefore, we are unable to reconstruct the analog signal $x_a(t)$ from its samples $x(n)$.

4. THE SAMPLING THEOREM

Given the discrete-time signal sequence $x(n)$ with its spectrum $X(F/F_s)$, as illustrated in Fig. 3, with no aliasing, it is possible to reconstruct the original analog signal $x_a(t)$. This can be accomplished by passing the discrete-time sequence $x(n)$ through an ideal lowpass filter with frequency response

$$G(F) = \begin{cases} T, & |F| \leq \dfrac{F_s}{2} \\ 0, & |F| > \dfrac{F_s}{2} \end{cases} \tag{15}$$

The input to this filter may be expressed as

$$v(t) = \sum_{n=-\infty}^{\infty} x_a(nT)\delta(t - nT) \tag{16}$$

where $T = 1/F_s$ and $\delta(t)$ represents a unit impulse. With $F_s = F_N = 2F_{\max}$, the output of the ideal lowpass filter is

$$x_a(t) = \sum_{n=-\infty}^{\infty} x_a(nT)\frac{\sin(\pi/T)(t - nT)}{(\pi/T)(t - nT)} \tag{17}$$

This equation provides the formula for the ideal reconstruction of the analog signal $x_a(t)$ from its samples $x_a(nT)$. We observe that the ideal reconstruction involves the interpolation function

$$g(t) = \frac{\sin(\pi t/T)}{\pi t/T} \tag{18}$$

and its time-shifted versions (time shifts of nT, $n = \pm 1, \pm 2, \ldots$) which are multiplied by the corresponding samples $x_a(nT)$. We also observe that the interpolation function $g(t - nT)$ is zero at $t = kT$, except at $k = n$. Consequently, $x_a(t)$ evaluated at $t = kT$ is simply the sample $x_a(kT)$. At all other time instants, the weighted sum of the time-shifted versions of the interpolation function combine to yield $x_a(t)$.

The reconstruction formula given by Eq. (17) forms the basis for the *sampling theorem*, which can be stated as follows.

Sampling Theorem. A band-limited continuous-time signal $x_a(t)$ with highest-frequency (bandwidth) $F_{\max}$ can be uniquely recovered from the samples $x_a(nT)$, provided the sampling rate $F_s \geq 2F_{\max}$ samples per second.

5. SAMPLING OF BANDPASS SIGNALS

Suppose that a real-valued analog signal $x(t)$ has a frequency content concentrated in a narrow band of frequencies in the vicinity of a frequency F_c, as shown in Fig. 4. The bandwidth of the signal is defined as $B = B_2 - B_1$ and, usually, $F_c \gg B$. Such a signal is usually called a (narrowband) bandpass signal. Clearly, the highest frequency contained in this signal is $F_{\max} = B_2$. If we blindly apply the sampling principle described in the previous sections, we would sample such a signal at a rate

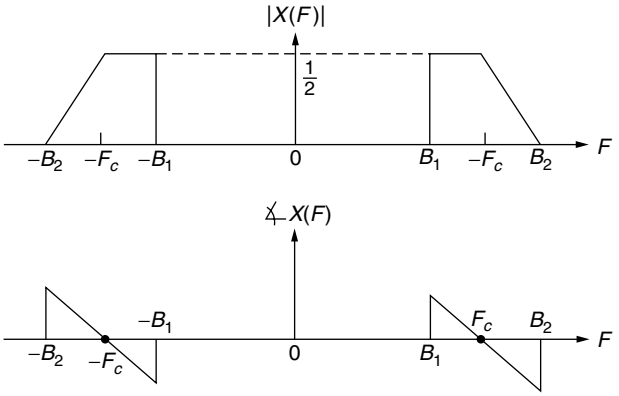


Figure 4. Spectrum of a bandpass signal.

$F_s \geq 2B_2$. However, it is not necessary to sample $x(t)$ at such a high rate.

It is easily shown that any (narrowband) bandpass signal (see, e.g., Ref. 1) can be represented by an equivalent lowpass signal. This representation may be expressed as

$$\begin{aligned} x(t) &= u_c(t)\cos 2\pi f_c t - u_s(t)\sin 2\pi F_c t \\ &= \mathrm{Re}\{[u_c(t) + ju_s(t)]e^{j2\pi F_c t}\} \end{aligned} \tag{19}$$

where $u_c(t)$ and $u_s(t)$ are called the *quadrature components* of the bandpass signal. The complex-valued signal

$$x_l(t) = u_c(t) + ju_s(t) \tag{20}$$

is called the *equivalent lowpass signal*. Its spectrum $X_l(F)$ is illustrated in Fig. 5, and it corresponds to the spectrum obtained by a frequency translation of F_c of the bandpass signal spectrum $X(F)$, shown in Fig. 4. We observe that the spectrum of the bandpass signal is related to the spectrum of the equivalent lowpass signal by the formula

$$X(f) = \tfrac{1}{2}[X_l(f - F_c) + X_l^*(-F - F_c)] \tag{21}$$

Hence, the spectrum of the bandpass signal $x(t)$ can be obtained from the spectrum of the equivalent lowpass signal $x_l(t)$ by a frequency translation.

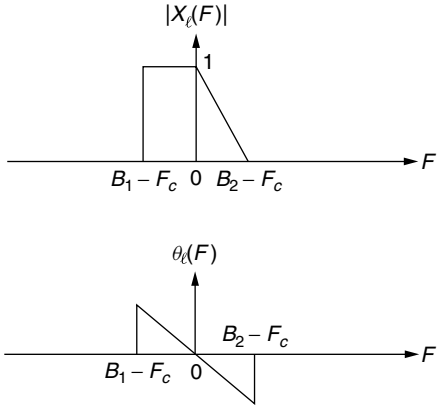


Figure 5. Spectrum of equivalent lowpass signal.

Given the equivalence between the bandpass signal $x(t)$ and the lowpass signal $x_l(t)$, it is advantageous to perform a frequency translation of the bandpass signal by an amount

$$F_c = \frac{B_1 + B_2}{2} \tag{22}$$

and sampling the equivalent lowpass signal. Such a frequency shift can be achieved by multiplying the bandpass signal as given in Eq. (19) by the quadrature carriers $\cos 2\pi F_c t$ and $\sin 2\pi F_c t$ and lowpass-filtering the products to eliminate the signal components at $2F_c$. Clearly, the multiplication and the subsequent filtering are first performed in the analog domain, and then the outputs of the filters are sampled. The resulting equivalent lowpass signal has a bandwidth $B/2$, where $B = B_2 - B_1$. Therefore, it can be represented uniquely by samples taken at the rate of B samples per second for each quadrature component. Thus the sampling can be performed on each lowpass filter output at the rate of B samples per second, as indicated in Fig. 6. Therefore, the resulting rate is $2B$ samples per second.

In view of the fact that frequency conversion to lowpass allows us to reduce the sampling rate to $2B$ samples per second, it should be possible to sample the bandpass signal at a comparable rate. In fact, it is.

Suppose that the upper frequency $F_c + B/2$ is a multiple of the bandwidth B (i.e., $F_c + B/2 = kB$), where k is a positive integer. If we sample $x(t)$ at the rate $2B = 1/T$ samples per second, we have

$$\begin{aligned} x(nT) &= u_c(nT)\cos 2\pi F_c nT - u_s(nT)\sin 2\pi F_c nT \\ &= u_c(nT)\cos\frac{\pi n(2k-1)}{2} - u_s(nT)\sin\frac{\pi n(2k-1)}{2} \end{aligned} \tag{23}$$

where the last step is obtained by substituting $F_c = kB - B/2$ and $T = 1/2B$.

For n even, say, $n = 2m$, Eq. (23) reduces to

$$x(2mT) \equiv x(mT_1)\cos\pi m(2k-1) = (-1)^m u_c(mT_1) \tag{24}$$

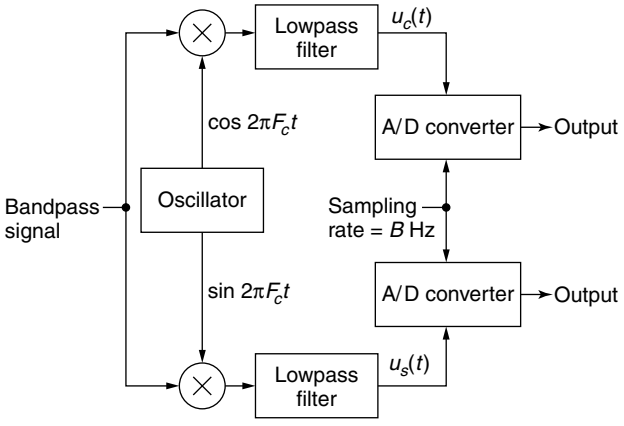


Figure 6. Sampling of a bandpass signal after converting it to a lowpass signal.

where $T_1 = 2T = 1/B$. For n odd, say, $n = 2m - 1$, Eq. (23) reduces to

$$x(2mT - T) \equiv x\left(mT_1 - \frac{T_1}{2}\right) = u_s\left(mT_1 - \frac{T_1}{2}\right)(-1)^{m+k+1} \tag{25}$$

Therefore, the even-numbered samples of $x(t)$, which occur at the rate of B samples per second, produce samples of the lowpass signal component $u_c(t)$. The odd-numbered samples of $x(t)$, which also occur at the rate of B samples per second, produce samples of the lowpass signal component $u_s(t)$.

Now, the samples $\{u_c(mT_1)\}$ and the samples $\{u_s(mT_1 - T_1/2)\}$ can be used to reconstruct the equivalent lowpass signals. Thus, according to the sampling theorem for lowpass signals with $T_1 = 1/B$, we obtain

$$u_c(t) = \sum_{m=-\infty}^{\infty} u_c(mT_1)\frac{\sin(\pi/T_1)(t - mT_1)}{(\pi/T_1)(t - mT_1)} \tag{26}$$

$$u_s(t) = \sum_{m=-\infty}^{\infty} u_s\left(mT_1 - \frac{T_1}{2}\right) \times \frac{\sin(\pi/T_1)(t - mT_1 + T_1/2)}{(\pi T_1)(t - mT_1 + T_1/2)} \tag{27}$$

Furthermore, the relations in Eqs. (24) and (25) allow us to express $u_c(t)$ and $u_s(t)$ directly in terms of samples of $x(t)$. Now, since $x(t)$ is expressed as

$$x(t) = u_c(t)\cos 2\pi F_c t - u_s(t)\sin 2\pi F_c t \tag{28}$$

substitution from Eqs. (27), (26), (25), and (24) into Eq. (28) yields

$$x(t) = \sum_{m=-\infty}^{\infty}\left\{(-1)^m x(2mT)\frac{\sin(\pi/2T)(t - 2mT)}{(\pi/2T)(t - 2mT)} \times \cos 2\pi F_c t + (-1)^{m+k}x((2m-1)T) \times \frac{\sin(\pi/2T)(t - 2mT + T)}{(\pi/2T)(t - 2mT + T)}\sin 2\pi F_c t\right\} \tag{29}$$

But

$$(-1)^m \cos 2\pi F_c t = \cos 2\pi F_c(t - 2mT)$$

and

$$(-1)^{m+k}\sin 2\pi F_c t = \cos 2\pi F_c(t - 2mT + T)$$

With these substitutions, Eq. (29) reduces to

$$x(t) = \sum_{m=-\infty}^{\infty} x(mT)\frac{\sin(\pi/2T)(t - mT)}{(\pi/2T)(t - mT)}\cos 2\pi F_c(t - mT) \tag{30}$$

where $T = 1/2B$. This is the desired reconstruction formula for the bandpass signal $x(t)$, with samples taken at the rate of $2B$ samples per second, for the special case in which the upper-band frequency $F_c + B/2$ is a multiple of the signal bandwidth B.

In the general case, where only the condition $F_c \geq B/2$ is assumed to hold, let us define the integer part of the ratio $F_c + B/2$ to B as

$$r = \left\lfloor \frac{F_c + B/2}{B} \right\rfloor \tag{31}$$

While holding the upper cutoff frequency $F_c + B/2$ constant, we increase the bandwidth from B to B' such that

$$\frac{F_c + B/2}{B'} = r \tag{32}$$

Furthermore, it is convenient to define a new center frequency for the increased bandwidth signal as

$$F_c' = F_c + \frac{B}{2} - \frac{B'}{2} \tag{33}$$

Clearly, the increased signal bandwidth B' includes the original signal spectrum of bandwidth B.

Now the upper cutoff frequency $F_c + B/2$ is a multiple of B'. Consequently, the signal reconstruction formula in Eq. (30) holds with F_c replaced by F_c' and T replaced by T', where $T' = 1/2B'$:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT')\frac{\sin(\pi/2T')(t - mT')}{(\pi/2T')(t - mT')}\cos 2\pi F_c'(t - mT') \tag{34}$$

This proves that $x(t)$ can be represented by samples taken at the uniform rate $1/T' = 2Br'/r$, where r' is the ratio

$$r' = \frac{F_c + B/2}{B} = \frac{F_c}{B} + \frac{1}{2} \tag{35}$$

and $r = \lfloor r' \rfloor$.

We observe that when the upper cutoff frequency $F_c + B/2$ is not an integer multiple of the bandwidth B, the sampling rate for the bandpass signal must be increased by the factor r'/r. However, note that as F_c/B increases, the ratio r'/r tends toward unity. Consequently, the percent increase in sampling rate tends to zero.

The derivation given above also illustrates the fact that the lowpass signal components $u_c(t)$ and $u_s(t)$ can be expressed in terms of samples of the bandpass signal. Indeed, from Eqs. (24)–(27), we obtain the result

$$u_c(t) = \sum_{n=-\infty}^{\infty}(-1)^n x(2nT')\frac{\sin(\pi/2T')(t - 2nT')}{(\pi/2T')(t - 2nT')} \tag{36}$$

and

$$u_s(t) = \sum_{n=-\infty}^{\infty}(-1)^{n+r+1}x(2nT' - T)\frac{\sin(\pi/2T')(t - 2nT' + T)}{(\pi/2T')(t - 2nT' + T)} \tag{37}$$

where $r = \lfloor r' \rfloor$.

In conclusion, we have demonstrated that a bandpass signal can be represented uniquely by samples taken at a rate

$$2B \leq F_s < 4B$$

where B is the bandwidth of the signal. The lower limit applies when the upper frequency $F_c + B/2$ is a multiple

of B. The upper limit of F_s is obtained under worst-case conditions when $r = 1$ and $r' \approx 2$.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.

SATELLITES IN IP NETWORKS

ABBAS JAMALIPOUR
University of Sydney
Sydney, Australia

Satellites have been an important element of telecommunications networks for many years in providing long-distance telephony and television broadcasting. The involvement of satellites in IP networks is a direct result of new trends in global telecommunications where the Internet traffic will have a dominant share of the total network traffic and special features of high-capacity satellite channels to be discussed in this article. The large geographical coverage of the satellite footprint and its unique broadcasting capabilities as well as its high-capacity channel keep the satellite as an irreplaceable part of communications systems, despite the high cost and long development and launching cycle of a satellite system.

In this article, we will review the satellite communications systems and introduce a new era for satellite communications toward broadband satellite systems and satellite-for-the Internet systems. This means that the satellite is changing its traditional role as being simply a relay in space into becoming an active element similar to a switch or a router in terrestrial networks. We will start the article with a short historical overview of satellite communications and then provide up-to-date information on new broadband and Internet satellite systems. We will briefly review third-generation wireless cellular systems, where Internet access is considered, in order to show the role and contribution of satellites in these systems and thus in the mobile Internet. Satellite applications within the third-generation wireless terrestrial systems as well as in the global Internet will be discussed. Several implementation topics, including mobility and location management that are common in satellite and terrestrial mobile networks, will also be discussed. We will then open the topics in satellite transport of Internet traffic and challenging issues that need to be resolved. Finally we will conclude the article with a concise but complete discussion on satellite future perspectives to the global Internet connectivity problem.

1. AN OVERVIEW OF SATELLITE COMMUNICATIONS

A satellite is one of the oldest components in telecommunications systems. For almost half a century, satellite networks have provided long-distance communications services to the public switching telephone network (PSTN) as well as television broadcasting. These services are particularly best justified by the large footprint coverage of the satellite, and to this date, there is no substitute for satellites in this field. In these types of service, a satellite acts as a communications repeater or relay (according to whether the transmitted signal is digital or analog) that communicates with ground stations and solves the problem of transmission of electromagnetic waves between different parts of the world that are not in line of sight of each other. A noteworthy achievement in satellite communications is

the formation of INTELSAT (International Telecommunications Satellite Organization), which in 1964 established a means of fixed-satellite service among nations [1].

In the 1980s, satellites were being deployed for the first time in mobile telecommunications by providing direct communications to maritime vessels and aircrafts. The first major development in this area was the INMARSAT satellite system. INMARSAT started a new era of satellite communications, called *mobile satellite services* (MSSs) in 1982. The INternational MARitime telecommunication SATellite organization used a geostationary satellite system using L-band (1.5–1.6 GHz) to provide telecommunication services mainly to ships. In the first generation of MSS, INMARSAT defined five standards: Standard A (1982), Standard B (1993), Standard C (1991), Standard M (1992/93), and the Aeronautical Standard (1992). Different worldwide telecommunication services, including voice, facsimile, and data were considered in these standards. While INMARSAT A and B are mostly considered the service to ships, INMARASAT C is planned to provide service to small craft, fishing boats, and land mobiles. INMARSAT continues its worldwide services as one of the most reliable satellite communication systems.

Although INMARSAT remains as the most distinguishable satellite system of its kind, there were other MSSs developed during the *first-generation mobile satellite systems*, such as QUALCOMM in North America (1989), ALCATEL QUALCOMM for Europe (1991), and the Japanese NASDA system (1987).

Reduction in size and cost of user terminals was the motive for second-generation MSSs started around 1985. In this generation, interconnection of satellite systems with terrestrial wireless systems has also been considered. INMARSAT defined its mini-M standard in 1995 with worldwide voice, data, facsimile, and telex service at 2.4 kbps (kilobits per second). American Mobile Satellite Corporation (AMSC), NSTAR of Japan, European mobile satellite (EMS), and several others are included in the second-generation MSSs.

Satellite systems have always being faced with unavoidable long propagation delay and large transmission power requirements. Consideration of small-size user terminals and direct radiocommunications between users and satellites (i.e., without using a ground station) led to the idea of using satellites in orbits lower in altitude than the geostationary orbit. Among possible orbit selections, low-earth-orbit (LEO) satellites with altitudes of 500–1500 km and medium-earth-orbit (MEO) satellites of altitude 5000–13000 km were considered [1]. The altitude selection given above assures that the satellites reside outside the two Van Allen belts to avoid the radiation damage to electronic components installed in satellites. The use of these nongeostationary satellite systems for commercial purposes started a new era in mobile satellite communications. Use of spot-beam antennas in these satellites produces a cellular-type structure within coverage areas, and hence a frequency reuse scheme can be applied, making the system a high-capacity cellular-like network on the ground with satellites as the base stations in space.

LEO and MEO satellite systems, because of their shorter distance to the earth, solve the problem of long propagation delay and high power consumption, but they introduced new challenges to the communications industry. Since the satellite is closer to the earth, compared to a geostationary satellite, it is not possible to employ just three satellites to cover all parts of the world as in case of geostationary satellite systems. Therefore, for LEO and MEO, a constellation of satellites in order of tens of satellites is required. This means more complexity and, of course, higher cost to the satellite system, which eventually must be passed to the users. Many LEO and MEO satellite systems were proposed in the early 1990s in North America and around the world and obtained frequency spectrum licenses. Only a few of these systems were completed and became operational, including IRIDIUM (1998) with 66 satellites and GLOBALSTAR (2000) with 48 satellites. However, financial problems associated with the high-cost of LEO systems forced IRIDIUM to cease its operation in 2000. Besides higher network complexity and more expensive control management requirements of nongeostationary satellite systems, a LEO or MEO satellite itself has a shorter lifetime than in geostationary systems. This means more frequent satellite launch requirements and higher maintenance cost to the satellite system.

The operational failure of the advanced but complex IRIDIUM satellite system revealed that although the technology for implementing a mobile satellite phone system is available, it is not possible to compete in the cost and services of such a system with the rapidly growing terrestrial cellular systems and new Internet services using LEO satellites. The roaming capability between different second-generation (2G) cellular networks in different countries and those considered in the third-generation (3G) wireless systems are quite adequate to provide telecommunications services to the majority of world population at lower cost and better quality (e.g., delay) than are those that can be achieved through satellite systems. The new trends in the telecommunications industry in transmitting data traffic and Internet traffic at high speed over wireless channels could not be matched by satellite systems. The IRIDIUM system, for example, could provide short data services at the very low data rate of only 2.4 kbps.

Satellite systems, however, maintain their unique feature of broadcasting. *Satellite broadcasting* has been a success for a long time and continues its dominance for long-distance coverage and service to highly populated telephony networks. If this unique feature of satellite systems can be incorporated into the new trends in telecommunications industry toward high-speed Internet access, then a new era of satellite communications technology will have begun. Broadband satellites are being developed for this market.

Broadband satellites [2–5] are recognized as systems that can provide high data rate transmission in the order of $\geq$1 Mbps. Digital video broadcasting (DVB) systems such as Eutelsat, SES, and INTELSAT; proprietary systems such as Spaceway, Astrolink, and iPSTAR; and proposed systems such as Teledesic, SkyBridge, WEST, and Celestri are among such broadband satellite systems. Standardizations of these satellite systems are

ongoing [6] in order to reduce the cost and increase applicability, similar to the way in which terrestrial cellular systems have developed and become successful. In this standardization, multicasting is also considered as a strong feature. Geostationary satellite systems are becoming of main interest for these services. Some of the applications of broadband satellite systems are shown in Fig. 1, which illustrates how satellites can interconnect geographically distant networks through land gateways. The system shown in this figure is designed to provide access to the Internet contents in one place by users of many other networks. Each network usually includes a caching system for fast local multicast to its Internet users.

Broadband satellite systems can be categorized according to their specifications and capabilities. This could be based on the frequency bands of operation (C band 4–8 GHz, Ku band 10–18 GHz, Ka band 18–31 GHz, and higher bands V and Q); the orbit altitude and hence the satellite lifetime; power requirements and antenna size; usage of bent-pipe or onboard processing (OBP) technologies; global or regional coverage of the system; satellite total capacity and user capacity; use of intersatellite links (ISL); number of supported terminals and required gateways; protocol used in the satellite system such as TCP/IP, DVB, and ATM; use of open or proprietary standards; total number of satellites in the system; and the total cost. Most new designs of satellite systems include onboard processing (OBP) and on-board switching (OBS) facilities so that the satellite node changes its simple role of relaying into being an active element in the network. An example of a functional satellite system that includes onboard switching and an ATM-switch-like satellite is shown in Fig. 2.

There are many regulatory and standard bodies currently involved in development of issues related to the satellite communications industry. Regulatory bodies include WRC, FCC, ITU, ERO, and many national and regional regulatory organizations. Standards are developed mainly in IETF, ETSI, TIA, and ITU.

To conclude this section, we must say that satellite systems have long development cycles compared to terrestrial systems, and usually less funding and fewer engineers are involved in their development. Moreover, they compete over a more limited market than cellular systems. Most satellite systems are still proprietary, and interfaces are not public, which, in turn, prevents competition. The standards for satellites will ensure interoperability and real competition and will be required for a broader consumer market. Therefore, in order to see further development in satellite systems, a widely acceptable standardization is vital.

2. SATELLITES FOR THE GLOBAL INTERNET

New multimedia and Internet services demand more cost-effective high-quality and high-speed telecommunication technologies and architectures. The primary issue is how the current global Internet infrastructure can be expanded so that the quality of service can be improved from current best-effort service and that high-speed access can be achieved. In this context, satellites can play an important role in expanding the Internet infrastructure using the large coverage area feature and in providing high-speed data transmission through a high-bandwidth-capacity channel. Satellites, however, would not perform this task as an isolated network but rather use an efficient integration with current terrestrial networks. So instead of having an IP network in the sky, as suggested in earlier proposals on some satellite IP networks, a combination of terrestrial and satellite networks would be a solution to the future high-speed Internet.

In a global Internet infrastructure, satellites can be used for many purposes. They can be used to connect geographically distant segments of the network or interconnect heterogeneous networks. Satellites can provide direct telecommunication service to aircraft, ships, and isolated local networks on the ground and even to individual users. Flexible and quick deployment of bandwidth by satellite systems, make them easily approachable by densely wired networks when required, as a good backup and supporting network.

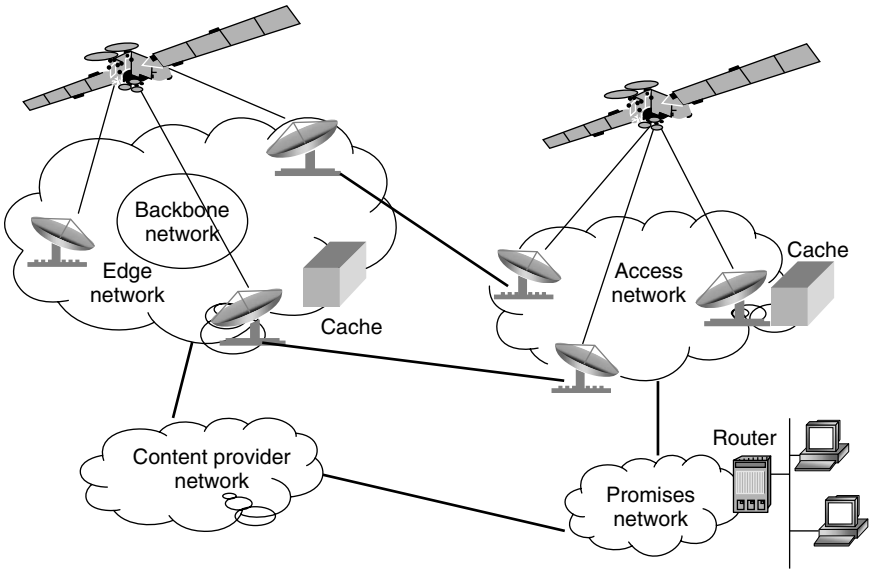


Figure 1. Applications of broadband satellites in interconnecting different networks.

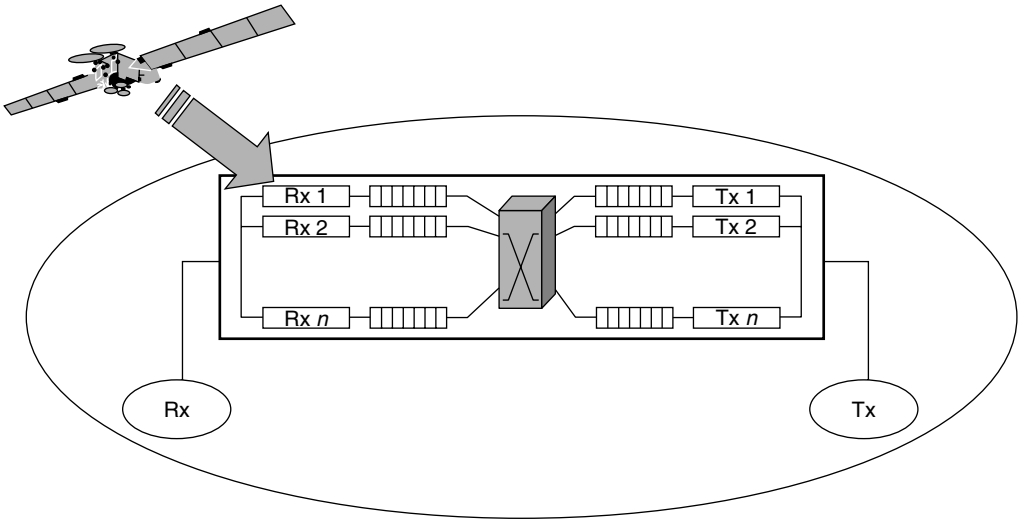


Figure 2. An example of satellite with onboard processing and switching facilities.

Figure 3 shows two different options for the satellite payload that can be used in satellite-based Internet architectures. In Fig. 3a the satellite is used as a reflector in space connecting separate network segments through ground gateways. In Fig. 3b, however, the satellite acts as an active component of the network that can utilize routing and switching processing. The satellites used in Fig. 3 can be on any of the altitudes explained in Section 1; that is, geostationary or nongeostationary (LEO or MEO) or a combination of different altitudes. The satellites shown in Fig. 3b, in addition to having connection to the ground gateways, are also employed in intersatellite links so that network connectivity can be created in the sky independently. This method should be considered as an important option in a future satellite-based Internet architecture. The method requires higher cost for the

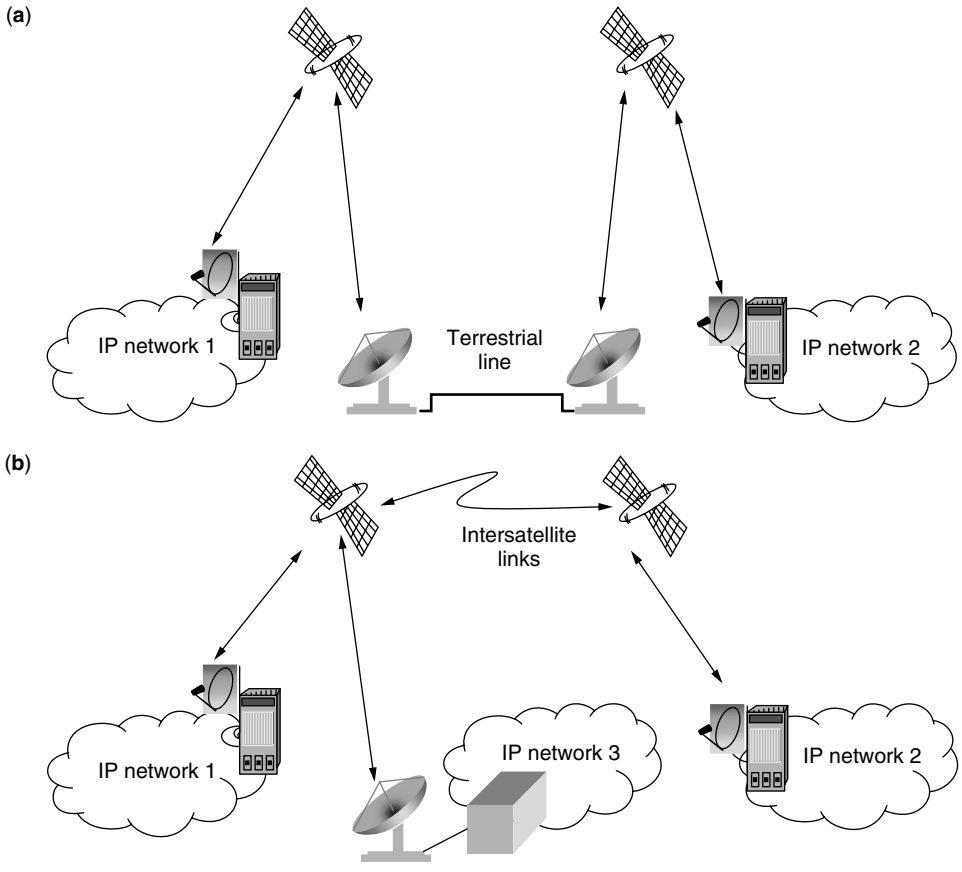


Figure 3. Two different payload options for satellite-based IP architectures: (**a**) bent-pipe architecture; (**b**) onboard processing satellites.

system and more complicated routing management. If special facilities are included in the mobile stations on the earth, both methods can provide direct Internet connectivity to remote users without any other alternative terrestrial telecommunications infrastructure. Hu and Li have summarized the satellite-based Internet architecture described above and current proposals of this kind [7].

A satellite node can also be used as a high-speed downlink for home Internet access. In this method, a home or office user with a satellite receiver, usually used for satellite television, can download the Internet contents at a very high data rate through the satellite downlink channel. A simple architecture of such a satellite–ground high-speed Internet is shown in Fig. 4. In the architecture shown in this figure, the user first connects to its Internet service provider (ISP) using a normal dialup connection. The dialup connection forms a low-speed data communication (e.g., a typical 56-kbps connection) mainly in order to send requests to the Internet servers at the local ISP site. All Internet contents can then be forwarded to the customer through the high-speed satellite downlink on receiving the request. The downlink can send the data to the user at speeds of 1 to a few Mbps using digital videobroadcasting satellites or other types of satellites. This method is especially appropriate for video-on-demand and other type of real-time Internet applications where many users located in the same region want to retrieve the same contents over the Internet. The asymmetry in the Internet traffic that usually results in up to 10 times more data traffic on a typical Internet downlink connection compared to the uplink makes this method of special interest and application. Currently this method is competing with other high-speed Internet access for home users, including cable modems and ADSL technologies. Some prototypes of these satellite Internet systems for home users have already been developed and demonstrated in Europe and other parts of the world [8,9]. With some modifications it is possible to extend the coverage of this type of Internet access to mobile users on the ground and also during long-distance flights and to ships.

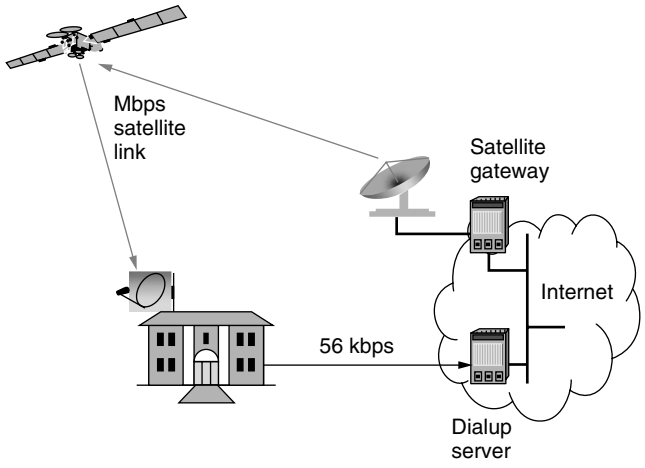


Figure 4. An architecture for satellite high-speed Internet access.

3. SATELLITES IN THIRD-GENERATION WIRELESS NETWORKS

With the increasing popularity of portable computers and the expanding Internet capabilities of mobile phone handsets, a large demand for mobile computing has been generated. Thus, instead of restricting data connections to be maintained always at a fixed position in the network, mobile users will be provided with equivalent multimedia and IP services. There is no doubt that the trend is toward a global mobile networking environment. In such a network, broadband satellites can be considered as an integral part of the network interconnecting the fast-growing terrestrial cellular and wired networks.

Broadband satellite networks for Internet access are the new generation of satellite networks in which Internet-based applications and services will be provided to users regardless of their degree of geographic mobility [2,3]. The main distinction from conventional satellite networks will be that the new satellite networks will support high-data-rate transmission and broadband services and in particular the Internet. The Internet is the most rapidly growing technology, and many new applications such as electronic commerce find their way through the Internet. Therefore, it is not surprising that broadband satellite networks focus on the Internet-based applications for their primary services, although voice and low-bit-rate applications still remain in the list of network services. Asynchronous transfer mode (ATM) will be the envisaged switching mode for future broadband satellites because of its support of a variety of traffic such as constant and variable bit rate and quality of service (QoS) support [2]. Nevertheless, IP routing is considered as another alternative for these satellites due to lower cost and "friendliness" toward the Internet traffic.

In the sphere of terrestrial networks, there are at present two possible approaches for establishing the task of mobile computing: cellular-based and IP-based solutions [3]. Intuitively, while a cellular-based solution enhances the current mobile communications by extending capacity for data and multimedia transmissions, an IP-based solution allows for user mobility by maintaining all ongoing Internet connections even in the presence of frequent handoffs or changes in the network point of attachments. In the forefront of these technologies, third-generation wireless systems are being considered.

Third-generation (3G) wireless communications systems evolve by orienting the integration of three essential domains: broadband, mobile, and Internet (IP). In such a milieu, the increasing feasibility of virtual connections allows mobile users not only to roam freely between heterogeneous networks but also to remain engaged in various forms of multimedia communications. Whether it is geographic coverage, bandwidth, or delay, it would then be up to the users to decide when and how to switch from one access network to another depending on the availability and appropriate cost/performance considerations and, thus, advancing toward an era of all-IP-based communications. Consequently, it will be necessary to implement the 3G system as a universal solution that prompts transparent user roaming (among different wireless networks)

while delivering the widest possible range of cost-effective services [10].

IMT-2000 (International mobile Telecommunications) is a unified 3G mobile system that supports both packet-switched and circuit-switched data transmissions with high spectrum efficiency, making the vision of anywhere, anytime communications a reality. Basically, it is a collection of standards that provides direct mobile access to a range of fixed and wireless networks. Among all, the three most significant developments are UMTS, cdma2000, and UWC-136, which are the 3G successors to the main 2G technologies of GSM, IS95, and AMPS, respectively [11]. The general idea was to make the development of 3G wireless technologies a gradual process from circuit-switched to packet-switched. Take GSM (Global System for Mobile communications) for example. In order to have the system enhanced with improved services (by means of increased capacity, coverage, quality, and data rates), the evolution to 3G was made possible through the incorporation of an intermediate stage called GPRS, the general packet radio services.

Based on the enhanced core network of GPRS, UMTS (Universal Mobile Telecommunications System) is designed to be the backward-compatible 3G standard for GSM. UMTS is the European proposal for a 3G mobile system aiming to support multimedia services with extended intelligent network features and functions. As a first step of the integration, UTRAN (the UMTS terrestrial access radio network) will coexist with GSM access networks. The idea was to develop the UMTS core network by gradually incorporating the desired UMTS features to the GSM/GPRS core network. At this stage, UTRA supports time-division duplex (WCDMA-DSTDD) and frequency-division duplex (WCDMA-DSFDD) modes with the combined operation offering an optimized solution to coverage areas of all sizes. A further multicarrier (MCFDD) mode is to be established at a later date intended mainly for the use in cdmaOne/cdma2000 evolutions [12].

For satellite systems, the situation is somehow different. The most apparent is that the market for satellite systems is much more limited than their cellular counterparts. Therefore, it would be difficult to assume the same approaches for satellite systems. Instead satellite systems can incorporate their global coverage feature for enhancement of the 3G terrestrial networks. Satellites can establish a high-speed backbone network to support the terrestrial networks and also to use their broadcast nature to deliver Internet content at high speed directly to a group a users.

Satellite UMTS (S-UMTS), for example, is considered as a component of 3G networks [13]. The satellite segment of the network connects through appropriate interworking units (IWUs) to the ground segments. An illustration of this incorporation of satellites in providing mobile Internet connectivity is shown in Fig. 5. IWU for the satellite has similar functionality as the gateways used to interconnect 2G and 3G networks for interoperation of these networks during the transition period from 2G to 3G as well as the gateways used for interconnection of different operator networks of the same kind (e.g., GSM). Such a concept is depicted in Fig. 6.

4. TECHNICAL ISSUES FOR SATELLITE-BASED INTERNET IMPLEMENTATION

After the overviews on satellite communications and third-generation wireless networks and introducing the role of satellites within the 3G network architecture in previous sections, in this and the following section we look at some specific but important issues for mobile satellite networks that also apply to terrestrial and cellular networks.

4.1. Mobility Management

It is widely agreed that to allow seamless user mobility, several considerations are necessary to ensure smooth

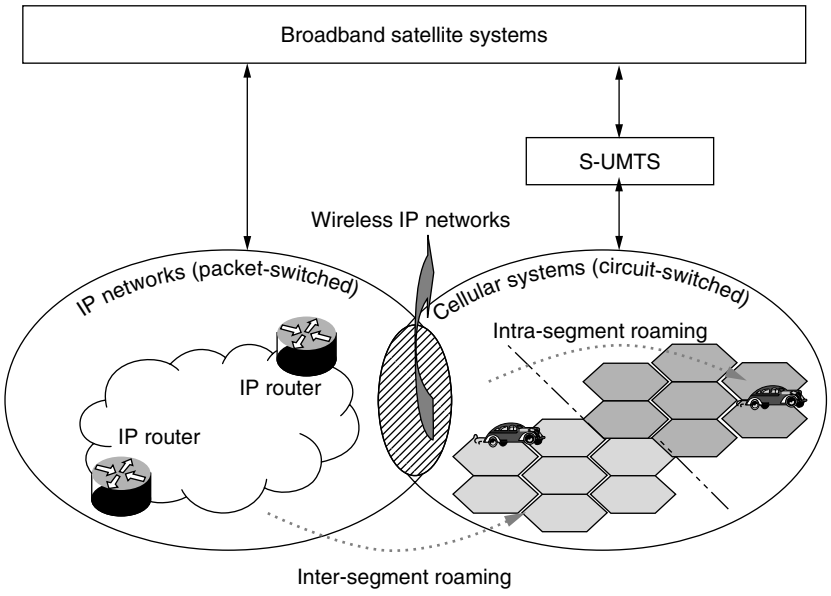


Figure 5. Satellite applications in global communications networks.

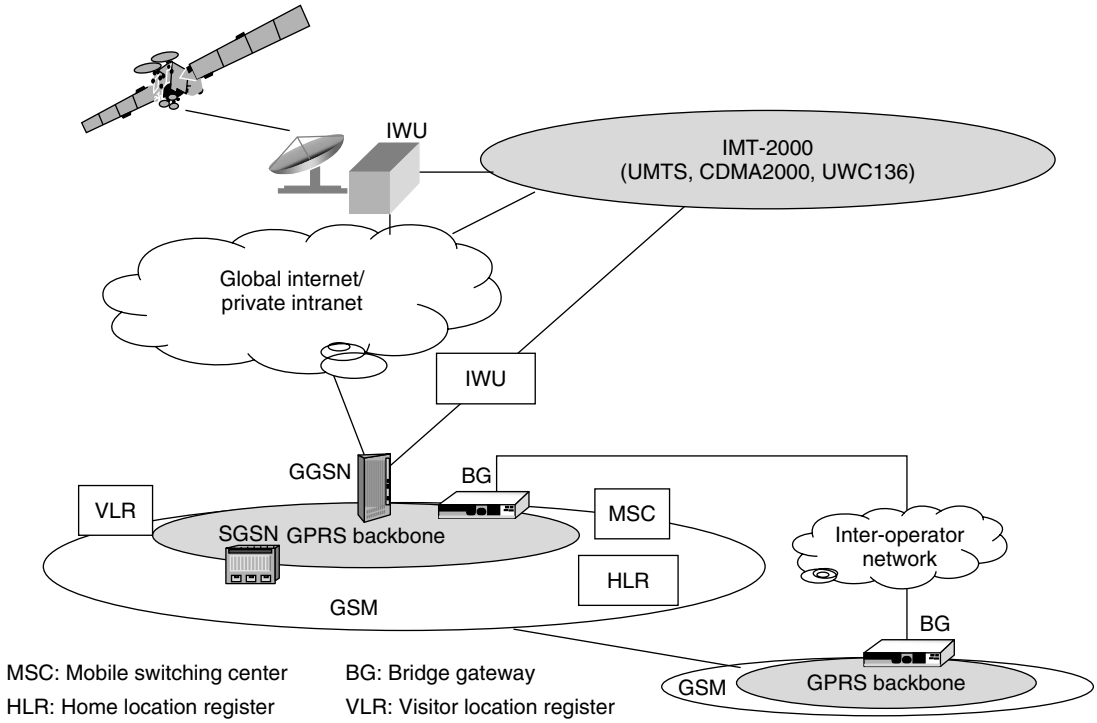


Figure 6. Interconnection of different terrestrial and satellite networks through interworking units.

transitions between different wireless technologies. Ultimately, mobility management is the key in enabling successful convergence between wireless communications and computing. Often, mobility management is interpreted as a process that simply routes packets from one point (the source) to the other (the intended destination). However, such an assumption becomes inadequate as more and more unsolvable issues gradually become apparent. Given its complexity, there seems to be an inevitable need to redefine the previously overlooked issue—*the mobility problem*.

Very briefly, mobility implies adaptability: the capability of maintaining any established network connections by accommodating different system characteristics when a mobile user roams within and/or between networks. More specifically, mobility refers to the initiation of a handoff process, not only when moving between cells (as in a cellular wireless environment) but also when roaming from one wireless network (e.g., satellite) to another (e.g., GPRS). Depending on the level of the network stack from which a movement is considered, mobility can be classified into three categories: air-interface mobility, link-level mobility, and network-level mobility.

Air-interface mobility is perhaps the most common case where a handoff takes place between two adjacent base stations (BSs) or access point (APs) within a radio access network. One can envisage this scenario as a pedestrian walking across microcell boundaries while being engaged in a conversation whether through voice or data transmissions. Link-level mobility goes one level up in the network hierarchy, and is concerned with maintaining a point-to-point protocol (PPP) context across multiple radio access networks. The transitions, however, would still be within the same domain and technology. On the highest network level (among the three categories), network (or IP) mobility provides network level mobility between different access networks (including wireless). Basically, this involves a change in the mobile's domain- (or location-) related IP address due to either (1) a change in radio access technologies or (2) a transition from one network operator to another. Note that in the latter case, the two networks involved might be implemented by the same access technology. Figure 7 attempts to illustrate the differences by listing the hierarchy of concern in the three cases. Note that the overall structure of a subnet can be considered to consist of three distinct stacks, with each stack developed depending on the specific technology (or network architecture) under consideration.

Most issues related to mobility are associated either directly or indirectly with service delivery. By and large, it involves, but is not limited to, the process of routing management, handoff management (including resource management), and also QoS management. Having said that, each operation has its own set of predefined actions. Given that it might be possible to incorporate some of the specific techniques used in satellite systems to enhance the complete operations in a 3G network, the purpose of the brief discussion in the next few paragraphs is to frame this part of the mobility problem.

4.2. Location Management

Obviously, mobility is managed largely through the process of location management. This action involves not only storing and/or retrieving information from the location database but also sending paging signals whenever necessary to locate a roaming user. Although

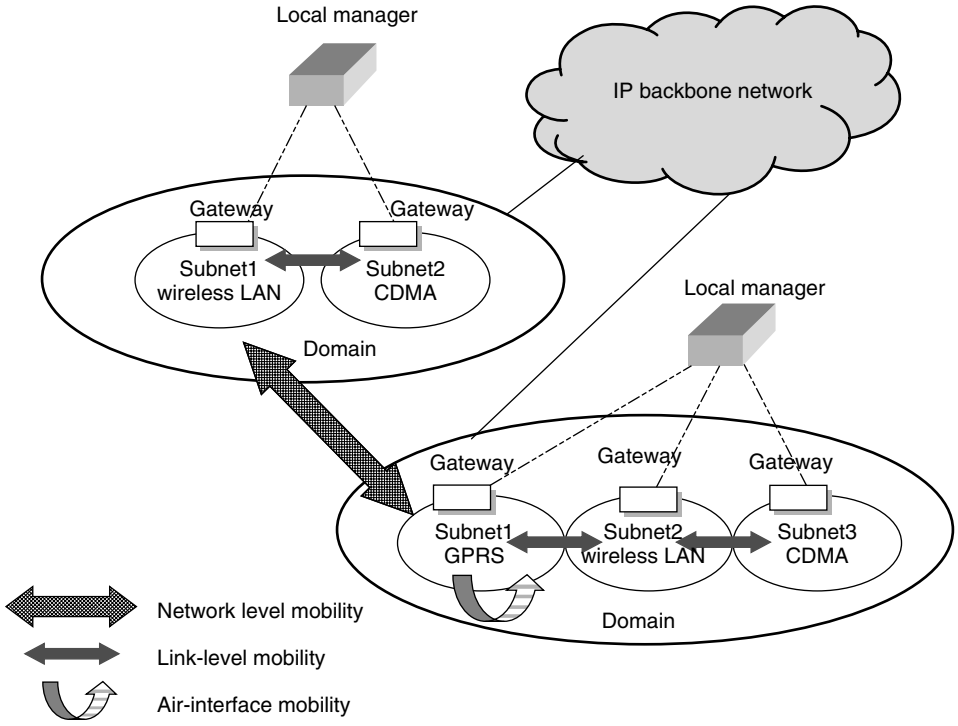


Figure 7. Different levels of mobility.

the need to notify the home network of mobile users' current locations is always present, it is questionable whether accurate location information is essential for mobile nodes (MNs) that are not committed to any data transmissions. Somehow, it seems logical to have separate mobility management techniques for idle and active mobile nodes and thus to allow variations in adjusting the updating frequency at which MNs' movement notifications are sent [14].

4.3. Routing Management

One issue that relates closely to the provision of global roaming is the establishment (or management) of roaming agreements between various networks. As a mobile user roams across various geographic and/or network boundaries, appropriate global roaming agreements between networks (or among ISPs) will also have to be established [15]. In fact, the incorporation of satellites is a particularly significant example of such operations. In scenarios where the limited users population (or terrain conditions) has made it infeasible to implement wireless terrestrial technologies, the availability of satellite systems would instantly form an alternative access medium for a dual-mode handheld terminal.

4.4. Handoff Management

A quality handoff management is particularly important for any real-time transmissions. As continuity plays a crucial part in grading the service quality for such a communicating purpose, it is worth exploring the factors that would have impact on handoff performance.

Mobility implies the necessity of MN handoffs. It is the process of reassociating the roaming mobile with a designated entity (in the new network) to maintain the much needed service continuity. Handoffs are performed at two layers; in the link layer (OSI layer 2, which maintains link connectivity) and also in the IP layer (OSI layer 3, which maintains network access). Physical connection to a base station can often be assumed to be seamless and almost instantaneous. The term network access denotes a MN's capability to exchange traffic in its current subnet without compromising its permanently allocated identity (IP address). A MN is said to have network access when its current subnet location corresponds with its registered network access point [16].

Successful handoffs are also crucial to ensure continuity of ongoing data connections (hence the provision of seamless roaming). However, to appreciate various coexisting network standards, decisions on the necessity of a handoff should be based on underlying network characteristics and specific application scenarios. This is particularly the case for intertechnology roaming (e.g., overlaying networks of Wireless LAN and GPRS), where the system characteristics might be significantly different. Essentially, a tradeoff analysis among the resultant throughput, data rate, latency, and disruption measures would be beneficial in such instances, to determine the actual essentiality of a handoff action. In this respect, the availability of resources has a direct influence on the number of successful handoffs. Thus, specific issues with regard to resource management, such as channel allocation schemes and call admission controls, should all be carefully considered.

4.5. Quality of Service Management

Quality of service management is another significant area that has gained tremendous research interest at both

academic and commercial levels. In particular, with the gradual replacement of voice traffic by data traffic (and multimedia transmissions), a proper implementation of suitable bandwidth allocation mechanisms is crucial to allow successful provision of satisfying customer services. Even apart from this, the possibility of having a network capable of maintaining multiple, concurrent QoS flows (for various applications running simultaneously) would also be advantageous. In addition, the users should be given the opportunity to alter or renegotiate (when desired) the predefined service-level specifications (SLS) with the corresponding providers through a continuous monitoring process of customer requirements [17].

Finally, there is a need to set up necessary SLS between networks. This is important as it allows user mobility to be managed independently of the access and backbone networks (e.g., B-ISDN or GSM), while maintaining a certain quality level that has been specified by any one particular mobile user.

5. MOBILITY MANAGEMENT IN MOBILE SATELLITE NETWORKS

In the previous section we discussed some important aspects of mobility management. Mobile satellite systems using nongeostationary orbits such as LEO and MEO have characteristics very similar to those of cellular networks, and thus those issues are common between both networks. The main similarity is that most of these systems use the cellular concept of increasing the total system capacity through a cellular-like coverage area arrangement introducing similar handoff issues as those involved in cellular systems. There is, however, a major difference between the proposed scheme in satellite systems and what exists in the sphere of cellular networks. The key operational concept that differentiates the specific operations of mobility management in the two systems is the "entity" being considered as the moving object. While the former encounters *mobile* movements within a fixed network architecture, the latter incurs *satellite* movements in reference to fixed mobile nodes. Essentially, this suggests that the mobility management technique will be different.

Furthermore, while the prediction of mobile's deterministic location is relatively easy to obtain (given the traveling characteristics, particularly speed of the LEO satellites), such certainty is not guaranteed in cellular systems. In other words, the difficulties encountered in cellular networks do not seem applicable to its satellite–mobile network counterparts. As a result, appropriate modification of the existing ground based solutions will be necessary before similar operations are suitable for applications on satellite-incorporated third-generation systems. Generally, it would be conceptually a lot easier to anticipate satellite movements than mobile movements. Besides, the rate of call arrival would not be essential in the latter case (i.e., paging seems to be less of a problem for satellite applications).

The use of LEO satellites is most favorable for its high traffic capacity and reduced user power requirements in both satellites and the ground terminal. However, as individual LEO satellites rotate relatively fast along the earth's surface, handoff becomes a particular concern in coping with the nonstationary nature of the coverage area. Depending on the relations of the two satellites involved in the handoff operations, three types of handoff are classified. Basically, *intrasatellite handoffs* are used to describe changes between spot beams under the management of the same satellite. With the increasing involvement of network management, *intersatellite handoffs* indicate handoffs between satellites and link handoffs incur due to changes in the connection pattern of satellite footprints (or satellite network topology). As an example of the latter scenario, such handoffs occur when links to adjacent orbits are turned off when the concerned satellite moves near to the polar region. Thus, the task is not only to utilize the available frequency spectrum efficiently but also to minimize unnecessary forced termination of connections due to handoff failures. In other words, it would be essential to at least attempt to anticipate user motions and to reserve the resources accordingly for the predicted residual time. A brief literature survey indicates that two major prioritization techniques were designed specifically for such purposes: (1) use of guard channels and (2) queuing of handoff requests when the resources were not currently available. Consequently, the operation of call admission control also becomes important as it decides whether sufficient resources would be available to accommodate the newly arrived transmission requests [18].

Cellular networks, on the other hand, focused more on the efficient operations of location management specifically for idle mobile users. Thus, although the basic problem of managing mobility is the same, the actual emphasis on the system developments is different for satellite and cellular systems. In fact, because of the movement of the LEO satellites, definitions of location area (LA) cannot be fixed even for the duration of a connection. Consequently, it is difficult to reapply some of the existing solutions from one system to the other (i.e., from cellular to satellite and vice versa). However, there are certain aspects where the "approaches" might be useful to seek alternative solutions for the open issues identified. For example, in terms of routing [19], a protocol has been developed in an attempt to reduce the frequency rerouting attempts during a link handoff.

Basically, *target probability* is defined to quantify the estimated duration of residency of a mobile terminal on one particular intersatellite link (ISL). During the route establishment of a new call, only links that can demonstrate a lifetime of greater than the target probability will be considered to form segments of the route. Although it might not seem obvious, this idea is similar to the predicting method used in location management operations in cellular networks, specifically, in the sphere of sequential paging where the optimal sequence is selected according to the probability of residence in individual subsections (or subareas). Thus, while acknowledging the fact that the prediction method would have been easier in the satellite systems (because its motion is deterministic and predictable), the goal

of determining efficient operations in both systems is the same.

Consequently, it becomes necessary to more closely identify the differences (or relations) between the operations of handoff and location management. Clearly, handoff management is significant only when the mobile unit is active; specifically, it is about the appropriate reservation of resources (such as bandwidth) along the roaming path of a mobile user while engaging in a call connection. Its efficient operation is important to ensure that the various aspects of the QoS requirements (e.g., throughput versus forced call termination) are satisfactorily complied. Location management, on the other hand, is intended mainly for users who are currently idle but are expected to receive calls (or become active) while they frequently change their point of attachment to the network. In essence, only sufficient location information (about the mobile) is maintained so that the network could loosely track the mobile's movement and subsequently incur a minimal paging (or searching) load when the precise residency is required. Consequently, it would be correct to conclude that the predicted information for handoff needs to be more reliable than that desired for efficient location management. On the basis of this observation, it seems potentially viable to combine (at least to some extent) the operations of the two management processes of handoff and location.

6. SATELLITE TRANSPORT OF INTERNET TRAFFIC

Broadband satellite networks are being developed to transport high-speed multimedia and in particular Internet traffic through high-capacity satellite channels to network segments as well as to individual users. As in the case of any other wireless network designed to deliver Internet traffic, broadband satellite networks need to connect to the backbone wired Internet on the ground. TCP (Transmission Control Protocol) is the most commonly used protocol at the transport layer of the network stack in the Internet, originally developed in wired networks with low bit error rate (BER) in the order of less than 10^{-8}. In this context, any wireless network with Internet service needs to be compatible with the protocol used in the wired network: mainly the TCP/IP protocol. There are, however, some design issues in the TCP/IP protocol, which make it difficult to use it efficiently over the wireless and satellite links. There have been many research activities comparing the performance of TCP in high-BER and high-latency channels and modification proposals to improve its performance in terrestrial and satellite wireless networks [20–24].

TCP has been designed and tuned for networks in which segment losses and corruption of performance are due mainly to network congestion. This assumption might be invalid in many of the emerging networks such as wireless networks. The flow control mechanism used in TCP is based on timeout and window-size adjustment, which can work with high utilization in wired networks with low BER, in the order of 10^{-8}. However, when the wireless channel is used (partially or totally) as the physical layer with a BER as high as 10^{-3}, it may perform inefficiently. The reason is that in the wireless channels the main cause for packet loss is the high BER and not congestion as it is in wired networks. The low efficiency of the TCP in a wireless channel is a result of the fact that the TCP misinterpreted the packet loss because of high error rate and congestion. On the other hand, in high-latency networks (such as satellite networks) adjustment of the window size could take a long time and reduce the system throughput.

TCP has the ability to probe the unused network bandwidth by a mechanism called *slow start* and also to back off the transmission rate on detection of congestion through the *congestion avoidance* mechanism. At the connection startup, TCP initializes a variable called *congestion window* to a value of one segment. This variable determines the transmission rate of TCP. The window size is doubled at every round-trip period until a packet loss is experienced. At this time, the congestion avoidance phase commences, the window size is halved, and the lost packet is retransmitted. During this phase of TCP, the window size is increased only linearly by one segment at each round-trip period and might be halved again on detection of another packet loss. If the retransmitted packet is lost, the timeout mechanism employed in TCP reduces the window size to one. Since all these procedures are performed at the periods equal to round-trip delay of the channel, the system throughput could be degraded significantly where high-latency channels such as geostationary satellites are involved. Therefore, the high-latency satellite channel, combined with the slow increase of the TCP congestion window size, results in underutilization of the satellite high-capacity channel.

Some modifications to the basic TCP can be made to ensure more efficient performance in high-latency satellite networks with Internet services (e.g., see Refs. 21–23 and their reference lists). *Selective acknowledgment* (SACK) TCP (RFC 2018), for example, is a method in which multiple losses in a transmission window can be recovered in one round-trip period instead of two in the basic TCP. *TCP for transactions* (T/TCP) also reduces the user perceived latency to one round-trip delay for short transmissions (RFC 1644). In *TCP spoofing* (cited by partridge and shepard [21]), a router close to the satellite link is considered that sends back acknowledgments for the TCP data. The responsibility of any segment loss in this method belongs to the router. In another method, called *split TCP*, a TCP connection is divided into multiple TCP connections and a special *satellite TCP* connection is employed for the satellite link part.

Another alternative for delivering Internet traffic through broadband satellite networks and simultaneously providing quality of service is to use IP-over-ATM or ATM protocols. In this regard, the IP protocol will provide the availability of various Internet applications, whereas the ATM protocol supports connection between two end-user terminals with a guaranteed end-to-end quality of service. An example of such protocol combination has been proposed for the Astra Return Channel System (ARCS), a geostationary multimedia satellite system using the Ka band on the return channels and the Ku bands on the forward channels [25].

In conclusion, we can say that the use of basic TCP in future broadband satellite networks will impose significant problems, especially in the case of short transmissions (compared with the channel delay–bandwidth product). For the geostationary satellite links the major problem with TCP is the long round-trip time, whereas in the case of non-geostationary satellite networks, the round-trip delay variation or jitter becomes more dominant. In both situations, the burst error nature of the satellite channel and the high BER require more sophisticated flow and congestion control mechanisms that can separate the segment loss because of network congestion or because of high channel error rate.

7. SUMMARY AND CONCLUSIONS

In this article, we summarized the satellite communications from a networking point of view in order to see the role of satellites in future mobile and fixed IP networks. The historical summary provided in the first section of this article revealed that despite the high initial investment and maintenance cost of satellite systems, satellites will remain as an irreplaceable component for long-distance communications and multimedia broadcasting. With the progress in optical communications and increasing number of transoceanic cables, it may be mistakenly thought that cable will replace the satellite for long-distance communications. However, the satellite's easy and quick deployment of additional capacity in any part of the world provides a distinct advantage over the deployment of cable systems. Improvement in cable television also could not replace satellite's broadcasting feature, especially due to the satellite's large footprint and simpler deployment.

When it comes to high-speed Internet access to the home, office, ships, aircraft, and mobile users, again satellite systems show its unique features. The global Internet needs expansion in both the geographic domain and data transport capacity. Satellites would be the main telecommunications component, if not the only one, as in many terrain circumstances, that can promise such expansion. The satellite huge onboard channel capacity and large coverage area are sufficient to provide future deployment of new systems. A very handy example would be the efforts toward realization of in-flight Internet access to passengers using satellite networks by major aircraft companies and airlines [26]. Satellites will soon bring inexpensive Internet access to long-distance flights and using voice-over-IP techniques make a huge reduction in the cost of phonecalls from and to airplanes.

For high-speed Internet access to the home and small-office users, currently ADSL and cable modem are the two leading technologies. With new digital videobroadcasting satellite systems in North America and Europe, however, these technologies found a need to compete with the satellites. The number of subscribers to satellite high-speed Internet access is increasing and close to the other two technologies, and this number is expected to increase even more rapidly by introduction of inexpensive satellite receivers in the near Future.

Satellites have an even larger contribution in IP networks than to the individual access discussed above. A satellite node can be an intelligent ATM switch or IP router in the sky interconnecting segments of the backbone networks on the earth. Similar to the conventional usage of satellites in public switching telephony networks, satellites can play an important role in future packet-switched networks, including the public Internet. The third-generation wireless networks and beyond consider Internet and multimedia traffic to have the dominant share of the network traffic load, and satellites have already shown their role in completion of any terrestrial mobile network. An example of satellite UMTS was given in this article to outline the role of satellites in future mobile communication systems. Satellite ground stations acting as an interworking unit can solve the roaming issue between heterogeneous wired and wireless terrestrial networks, expanding the telecommunications to its ultimate universal stage.

Some important technical implementation issues concerned with a global IP network have been discussed in this article. Mobility management has been revisited and redefined and location, handoff, routing, and quality of service managements have been discussed. All these issues are current research topics in mobile and satellite communications. For the high-latency satellite channel, as well as the error-prone wireless channel (including both satellite and terrestrial), the need for improvement in transport protocols currently employed in the Internet has been discussed and state-of-the art research activities toward improvement of TCP protocols has been reviewed. Note that other researchers are also currently working to improve the error probability of the wireless channel using forward error correction (FEC) schemes and sophisticated coding algorithms. Although these works are of great importance in the establishment of a better-quality wireless channel, we should not forget there are always situations in which the wireless signal-to-noise ratio is too low and no coding scheme can improve it. Therefore, a better solution would lie in the higher layers of the network, including the transport and network layers, where enhanced flow control algorithms speed up the data rate and the throughput of the wireless channel.

BIOGRAPHY

Abbas Jamalipour has been with the School of Electrical and Information Engineering at the University of Sydney, Australia, where he is responsible for teaching and research in wireless data communication networks and satellite systems, since 1998. He holds a Ph.D. degree in Electrical Engineering from Nagoya University, Japan. His current areas of research include wireless broadband data communications and wireless IP networks, mobile and satellite communications, traffic modeling, and congestion control. He is a recipient of a number of technology and paper awards and has authored two technical books and coauthored two others. He has authored numerous publications in these areas, and given short courses and tutorials in major international conferences. He has served on several major conferences technical committees, and organized and chaired many

technical sessions and panels in international conferences, including a symposium on satellite IP in IEEE Globecom 2001. He is the Vice Chair to the Satellite and Space Communications Committee of the IEEE ComSoc and has served as a guest editor to two special issues on 4G networks in IEEE magazines. He is a technical editor to the *IEEE Wireless Communications Magazine* (formerly, *Personal Communications Magazine*) and a Senior Member of IEEE.

BIBLIOGRAPHY

1. A. Jamalipour, *Low Earth Orbital Satellites for Personal Communication Networks*, Artech House, Norwood, MA, 1998.
2. A. Jamalipour, Broadband satellite networks — the global IT bridge, *Proc. IEEE* (special issue on multidimensional broadband wireless technologies and services) **89**(1): 88–104 (Jan. 2001).
3. A. Jamalipour and T. Tung, The role of satellites in global IT: Trends and implications, *IEEE Pers. Commun. Mag.* (special issue on Multimedia Communications over Satellites) **8**(3): 5–11 (June 2001).
4. J. Farserotu and R. Prasad, A survey of future broadband multimedia satellite systems, issues and trends, *IEEE Commun. Mag.* **38**(6): 128–133 (June 2000).
5. P. Chitre and F. Yegenoglu, Next-generation satellite networks: Architectures and implementations, *IEEE Commun. Mag.* **37**(3): 30–36 (March 1999).
6. J. Neale, R. Green, and J. Landovskis, Interactive channel for multimedia satellite networks, *IEEE Commun. Mag.* **39**(3): 192–198 (March 2001).
7. Y. Hu and V. O. K. Li, Satellite-based Internet: A tutorial, *IEEE Commun. Mag.* **39**(3): 154–162 (March 2001).
8. I. Minei and R. Cohen, High-speed Internet access through unidirectional geostationary satellite channels, *IEEE J. Select. Areas Commun.* **17**(2): 345–359 (Feb. 1999).
9. H. D. Clausen, H. Linder, and B. Collini-Nocker, Internet over direct broadcast satellites, *IEEE Commun. Mag.* **37**(6): 146–151 (June 1999).
10. M. Zeng, A. Annamalai, and V. Bhargava, Harmonization of global third-generation mobile systems, *IEEE Commun. Mag.* **38**(12): 94–104 (Dec. 2000).
11. W. Mohr and W. Konhauser, Access network evolution beyond third generation mobile communications, *IEEE Commun. Mag.* **38**(12): 122–133 (Dec. 2000).
12. R. Steele, C. C. Lee, and P. Gould, *GSM, cdmaOne and 3G Systems*, Wiley, Chichester, UK, 2001.
13. F. Priscoli, UMTS architecture for integrating terrestrial and satellite systems, *IEEE Multimedia* **6**(4): 38–44 (Oct.–Dec. 1999).
14. V. Wong and V. Leung, Location management for next-generation personal communications networks, *IEEE Network* **14**(5): 18–24 (Sept./Oct. 2000).
15. J. Solomon, *Mobile IP: The Internet Unplugged*, PTR Prentice-Hall, Englewood Cliffs, NJ, 1998.
16. N. Fikouras, K. Malki, S. Cvetkovic, and C. Smythe, Performance evaluation of TCP over Mobil IP, *Proc. Int. Conf. PIMRC'99*, Osaka, Japan, Sept. 1999.
17. A. Mehrotra and L. Golding, Mobility and security management in the GSM system and some proposed future improvements, *Proc. IEEE* **86**(7): 1480–1497 (July 1998).
18. I. Akyildiz, J. NcNair, J. Ho, H. Uzunalioglu, and W. Wang, Mobility management in next generation wireless systems, *Proc. IEEE* **87**(8) (August 1999).
19. H. Uzunalioglu, Probabilistic routing protocol for low earth orbit satellite networks, *Proc. Int. Conf. ICC'98*, Atlanta, GA, June 1998.
20. R. Goyal et al., Traffic management for TCP/IP over satellite ATM networks, *IEEE Commun. Mag.* **37**(3): 56–61 (March 1999).
21. C. Partridge and T. J. Shepard, TCP/IP performance over satellite links, *IEEE Network* **11**(5): 61–71 (Sept./Oct. 1997).
22. T. R. Henderson and R. H. Katz, Transport protocols for Internet-compatible satellite networks, *IEEE J. Select. Areas Commun.* **17**(2): 326–344 (Feb. 1999).
23. I. Minei and R. Cohen, High-speed Internet access through unidirectional geostationary satellite channels, *IEEE J. Select. Areas Commun.* **17**(2): 345–359 (Feb. 1999).
24. G. Xylomenos, G. C. Polyzos, P. Mahonen, and M. Saaranen, TCP performance issue over wireless links, *IEEE Commun. Mag.* **39**(4): 52–58 (April 2001).
25. *ASTRA Return Channel System, System Description Documentation*, Societe Europeenne des Satellites, Document ARCS.240.DC-Eoo1-0.2, issue 0.2, May 1998.
26. S. Karlin, Take off, plug in, dial up, *IEEE Spectrum* 52–59 (Aug. 2001).

SCALAR AND VECTOR QUANTIZATION

TIMO KAUKORANTA
Turku Centre for Computer Science (TUCS)
University of Turku
Turku, Finland

1. INTRODUCTION

Although the term *quantization* is not commonly used in our everyday life, the phenomenon itself is familiar to everyone. There the word *quantization* is known as *rounding*. We round the ages of the people to exact numbers of years even if the accurate age is a real number, which can have infinite decimal places. We apply rounding to several measurements in our life such as weights, distances, and periods of time. The purpose of rounding, or quantization, is to make these measurements easier to understand and handle. The measurements in these everyday examples are *scalars*, namely, single numbers, which present the amount of some quantity. This technique of rounding scalars is in fact known as *scalar quantization* (SQ).

Quantization can be considered as a technique for analog-to-digital conversion and data compression. By *quantization*, an originally analog signal, for example audio or video, is transformed into a digital form. This is very important operation because of the well-known benefits that digitization offers for signal processing. As

to the compression, speech and image data demand huge amounts of storage space so that many applications, such as in the fields of medicine or data communication, suffer from the lack of efficiency in coding of data.

Analog-to-digital conversion can be performed by SQ. There the original signal is sampled at some frequency f producing a sequence of samples x_i. These samples are then quantized so that they attain values from a set of predetermined *reproduction values* or *points* c_j. While the set of sample values may be unlimited, the size of the reproduction values is small and fixed. A unique binary word (code) w_j, which can simply be the index j of the reproduction value, is assigned to each reproduction value. This codeword is stored or transmitted instead of the original sample x_i.

In many applications we must store and transmit structured collections of data instead of scalars. If these data can be treated as vectors of fixed dimensionality, we can, instead of using the exact values of the vectors, handle their approximations. This is much like in the rounding process for individual scalars, but now we speak about *vector quantization* (VQ). Our hope is again to get advantage in transmission times and the storage space.

Generally, on the basis on the ability to preserve the information, compression techniques can be classified to *lossy* and *lossless*. A lossless compression technique reconstructs information exactly as it was before compression. Instead, a lossy compression technique can produce distortion to information, but the idea is that original information can be reproduced at sufficient accuracy, that is, that no necessary information is missing. The benefit of lossy compression is that it is possible to reach better compression ratios than by using lossless techniques.

Both scalar and vector quantization use fixed number of reproduction levels. Because the number M of these levels is normally much smaller than the number of all different inputs, the quantizer can be interpreted as a compression technique. The number of bits used to represent the binary word w_j may be a fraction of the number of bits needed to represent the corresponding sample of the original signal. On the other hand, quantization clearly loses some information; all possible input values cannot be reproduced. Therefore, quantization is a lossy compression technique.

It is typical that the input signal of a vector quantizer has already been scalar quantized. Single sample values (from the point of view of the vector quantizer) are organized as a sequence, and a fixed number of successive samples are grouped together to form vectors. The order of the samples in the sequence depends on the original source signal; in the case of one-dimensional signals such as audio, it is convenient to use time order. From images, which are two-dimensional inherently, $k \times k$-pixel blocks are often extracted and their pixels are assigned to vector elements in row-major order. In the case of color images, the RGB (red-green-blue) pixels can be interpreted as three-dimensional vectors.

VQ applies the same principles as SQ. In the encoding phase, the *vector quantizer* groups the values of the source signal into K-dimensional input vectors x_i. An appropriate transformation can be applied to the vectors. Then, for a given input vector x_i, the nearest *representative vector* c_j is searched from the *codebook* C of M K-dimensional *codevectors*. We suppose that the codebook, which is known by both the encoder and decoder, has been constructed before quantization as a separate process. The index, or binary word w_j, of the selected codevector c_j is stored or transmitted to the receiver. Compression rate can be improved by applying entropy coding to the indices. In the decoding phase, the receiver selects the representative vector from the codebook C using the transmitted and decoded index. A new reproduced signal is then constructed from the representative vectors. Compression is achieved because the cardinality of the set of the representative vectors is much smaller than the cardinality of the whole vector set. The encoding/decoding process is described as a sequence of successive steps in Fig. 1.

It is impossible to cover all the rich topics in scalar and vector quantization in a single article. We therefore restrict ourselves to topics, which we feel are important to a practitioner or a student in algorithmics who is interested in understanding the basic ideas in quantization methods. Our focus is more biased to the implementation issues whereas the theoretical consideration is largely omitted. This does not mean that the value of theory, including, for example, Shannon's source coding theory or Bennett's approximation on the quantization distortion, could be bypassed when evaluating different quantizers. These are dealt in depth, for example, in textbook of Gersho and Gray [1].

2. SCALAR QUANTIZATION

In scalar quantization we are given a set of numbers X drawn from a known distribution $f(x)$, and our task is to encode the numbers with a lower number of bits in such a way that the average distortion caused by this action is as small as possible. By the *distortion* we mean the difference of the original data x and the corresponding reproduction values $Q(x)$ obtained by the encoding and decoding sequence. When the original numbers are from $\Re$ (the set of real numbers), a natural choice of the distortion is the squared one-dimensional Euclidean distance $d(x, Q(x)) = (x - Q(x))^2$. Now it is clear that the distribution $f(x)$ has a profound influence on the proper choice of the quantization scheme. We should make more accurate quantization at the regions of high density than at the low-density regions.

The simplest way of thinking is to use a fixed number of bits for expressing the indices of the reproduction values $Q(X)$. If this number is M, we need $\log M$ bits for expressing each of them. This kind of method is called *fixed-rate coding*. Now one can fix M and ask for the best way (in distortion sense) to quantize the source data. This corresponds to minimization of the *distortion-rate function*. Another way of looking at the situation is to minimize M when the maximal distortion of the scalar quantizer has been fixed. Nothing restricts the coding to fixed-length codes. We can as well code the indices of the reproduction points by a *variable-length code*, in which case we may spend fewer bits to indices occurring more

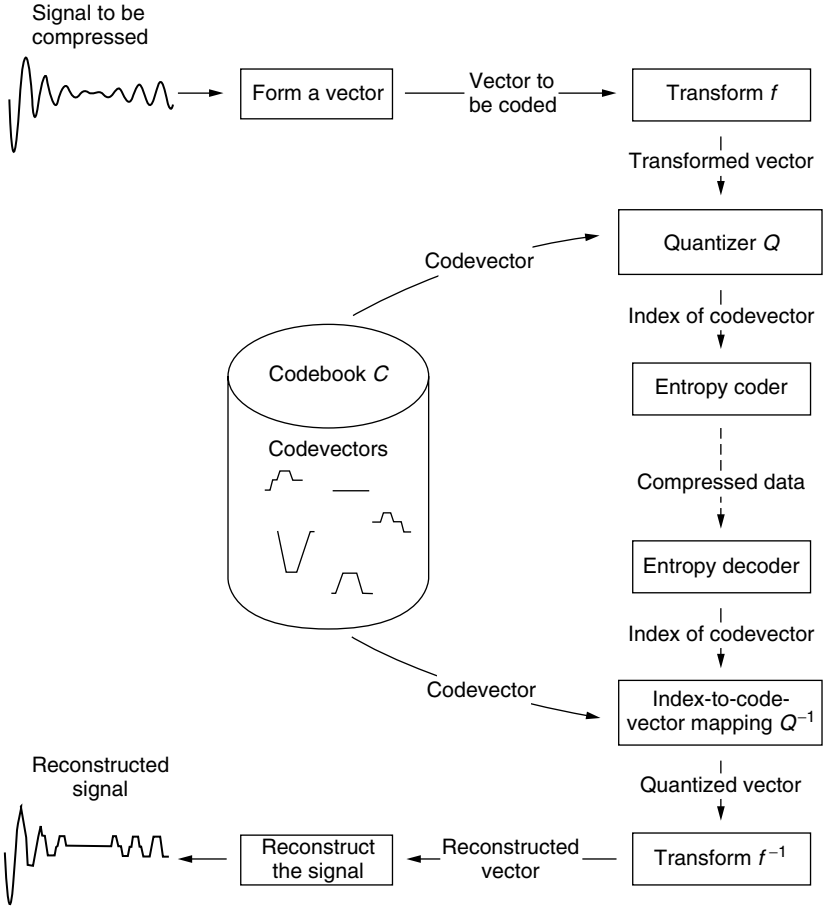


Figure 1. An example of a compression system based on (vector) quantization.

frequently. Note that here the coding must naturally be lossless (like Huffman code); compare with the encoding phase of the scalar quantizer, which is a lossy process. While the basic idea in SQ is the independent quantization of individual scalar data, there is no reason why we could not use entropy coding with some suitable context for the lossless compressor of the indices. By this way, we utilize the knowledge of the dependencies of individual samples on the history of the data when predicting the current quantization index. Then we code (e.g., by arithmetic coding) the difference between the actual and predicted index values.

2.1. Uniform Versus Nonuniform Quantizer

Quantizers can be divided into *uniform* and *nonuniform* depending on whether the reproduction values are at equal distances from their neighbors. A uniform scalar quantizer has an equal space $\Delta \in \Re$ between successive reproduction levels $c_i (i = 1, 2, \ldots, M)$, namely, $c_{i+1} = c_i + \Delta$. Correspondingly, the cell boundaries have the same space between them: $t_{i+1} = t_i + \Delta$. An example of a uniform quantizer is presented in Fig. 2. For this kind of quantizers one can show that the Euclidean squared distortion is of the size [2]

$$D_{un} \cong \frac{\Delta^2}{12} \tag{1}$$

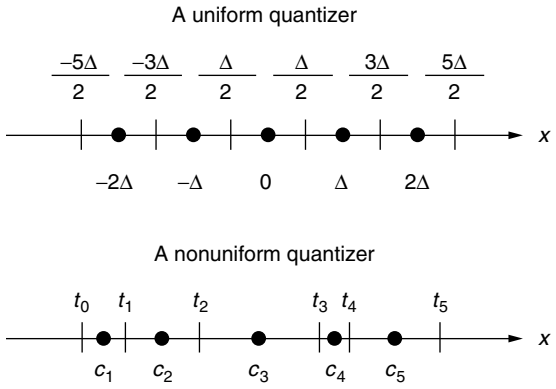


Figure 2. Illustration of a uniform and nonuniform quantizer.

The intervals of a nonuniform quantizer are of different lengths (see Fig. 2). This reflects the varying density of the samples in different regions of $\Re$, giving the overall distortion for samples at the range $[t_0, t_n]$

$$D_n = \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} d(x, c_i) f(x)\, dx \tag{2}$$

Several options for SQ are provided in the JPEG 2000 standard [3].

2.2. Lloyd's Conditions and Algorithm

Lloyd [4] showed two necessary optimality conditions for a scalar quantizer with fixed-length codewords to be optimal: (1) for a fixed setting of reproduction levels the cell boundaries must be selected optimally and (2) for a fixed setting of cell boundaries the reproduction levels should be selected optimally. This important observation can be turned to an algorithm (*Lloyd's algorithm*):

1. Initialize M cells and reproduction levels.
2. Repeat until the distortion converges:
   a. For each sample define the nearest centroid and place the sample to the set of samples mapping to this interval.
   b. Calculate new centroids for each interval by using the sets defined above.

In certain cases (e.g., for Gaussian and Laplacian densities of the samples), Lloyd's algorithm gives a globally optimal quantizer whereas there are applications where a local optimum will be found. Lloyd's algorithm is generalized and considered more closely in Section 4.1.

3. VECTOR QUANTIZATION

The goal and principle of vector quantization (VQ) are the same as in scalar quantization (SQ). The difference is in the nature of the samples. These are for VQ structured collections of scalars organized as vectors. One can separate four main tasks in the design of a vector quantizer: (1) *selection of a training set*, (2) *codebook generation* on the basis of the training set, (3) *encoding* of the source vectors with the aid of the codebook, and (4) *decoding* of the compressed vectors. The decoding commonly is the easiest of these tasks because it can often be performed extremely rapidly and simply by using a lookup table. The two main problems in the design of a vector quantizer are the construction of the codebook and efficient search from the codebook at the encoding phase. These problems are related because fast and accurate search methods are needed in codebook generation also.

A vector quantizer can be defined as a mapping Q of K-dimensional (Euclidean) space $\Re^K$ into a finite subset C of $\Re^K$:

$$Q: \Re^K \rightarrow C \tag{3}$$

where $C = (c_i; i = 1, 2, \ldots, M)$ is the set of *reproduction vectors* (i.e., the *codebook*) and M is the number of codevectors in C (see Fig. 3). Thus, a vector quantizer can be seen as a combination of two functions. The *encoder* produces an index i for input vector x. The *decoder* transforms the index i back to the reproduction vector c_i. As in the case of SQ, we may still have a lossless compressor of the indices. This may again produce a variable-length code, in which case we have a variable-rate VQ. One can also integrate the compression technique to the vector quantizer itself. Entropy-constrained VQ is one example of this kind (see Section 3.4), but other approaches exist also.

VQ has several benefits as a compression technique. The most important of these is the high decoding speed. By changing the number of codevectors in the codebook, the bit rate can be controlled quite easily. A drawback of VQ is that the vector quantizer has to be trained over a representative set of samples of the signal. Thus, a vector quantizer does not perform well for input vectors emitted by a source, which has characteristics very different from those of the original training set. The method suffers also from relatively high computational complexity both in the codebook generation and in the encoding of a signal.

3.1. Transforms

A suitable transform can be applied to vectors before the use of a vector quantizer. The purpose of the transform is to organize data in a vector in such way that important information is concentrated on some elements of the vector. This transformed representation is probably more suitable for quantization and the search for the nearest vector. For example, it may be reasonable to normalize the vectors by subtracting from each element their average. Another way to normalize a vector is to divide its elements by the norm of the vector. Also other transformations, such as the *discrete*

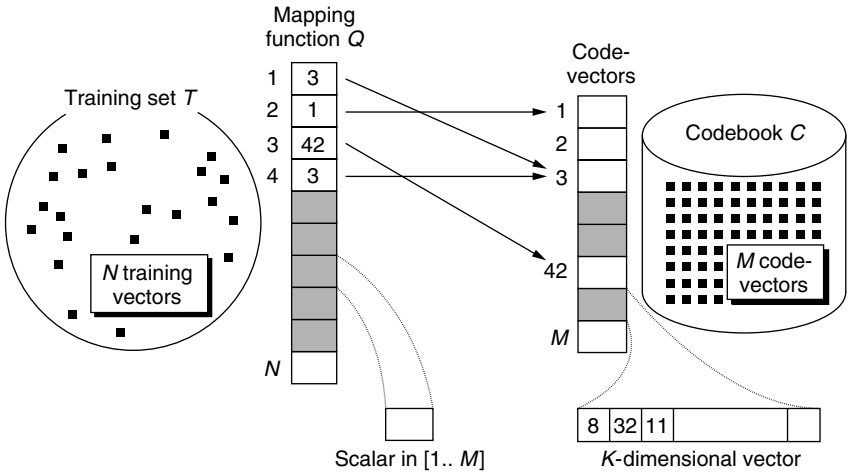


Figure 3. Roles of a training set and a codebook in vector quantization.

cosine transform and the *Walsh–Hadamard transform*, are used in applications. Transforms are often used in *product codevector quantization* (see Section 3.4). For general properties of transform based vector quantizers, see Ding's paper [5].

In some cases, the vectors to be quantized are sequences of the coefficients produced by a signal transformation technique, while in other cases, they may be simply blocks of individual pixels of a digitized image. (Note that this latter case clearly demonstrates the structural dependence of the elements in a vector.) Because of its fast decoding phase, VQ is suitable for compression situations demanding fast reproduction and therefore it has been popular in video codecs.

3.2. Quality and Resolution

The quality of a vector quantizer Q can be measured by the distortion between the original input vector x_i and the reproduction vector $y_j = Q(x_i)$. The smaller the distance, the better the quality. The most widely used distortion measure is the squared error or squared Euclidean distance defined by

$$d(x_i, y_j) = \|x_i - y_j\|^2 = \sum_{k=1}^{K} (x_{i,k} - y_{j,k})^2 \tag{4}$$

where $x_{i,k}$ and $y_{j,k}$ denote the kth element of the vectors x_i and y_j, respectively. Some other distortion measures have been proposed [1,6,7]. The quality of the quantizer is the average distance between the sample vectors and the corresponding reproduction vectors. This average can also be weighted for different components.

The distortion (quality) D of a codebook C can be measured in relation to the training set X for which the codebook has been constructed:

$$D(C, X) = \sum_{i=1}^{N} d(x_i - Q_C(x_i)) \tag{5}$$

Here N is the number of training vectors in the training set X and $Q_C(x_i)$ gives the nearest codevector in the codebook C. In other words, $Q_C(x_i)$ is an optimal mapping that minimizes the distance $\mathrm{d}(x_i, y_j)$, where $y_j \in C$.

There is one question largely omitted in VQ literature, namely, the representativeness of the training set. The VQ codebook is constructed on the basis of a training set of N samples, which are supposed to have the same statistical properties, the density, as the data to which VQ is later applied. A normal way to assert this is to evaluate the operation efficiency of the VQ codebook by an independent test set. Even here caution should be taken to the later changes of the source distribution.

The *resolution* (*code rate* or simply *rate*) of a fixed-rate vector quantizer is

$$r = \frac{\log_2 M}{K} \tag{6}$$

which measures the number of bits per vector element used to represent the K-dimensional input vector. The resolution gives an approximation of the quality of the reproduction of the vector quantizer. One problem of VQ is the *complexity barrier*, which is described as follows. If we limit the resolution r to a fixed value, we can increase the performance of the vector quantizer only by increasing the vector dimensionality K. The reason for this improvement of the performance is that long vectors express the statistical dependencies of the signal more extensively than short vectors. On the other hand, the amount of memory needed to store the codebook and the search complexity (operations per vector element in exhaustive search) are both relative to KM. Thus the space complexity, which is given by

$$KM = K2^{rK} \tag{7}$$

grows exponentially with dimension K. If we suppose that the encoding phase includes an exhaustive search of the codebook, the time complexity of the encoding is exponential on K, too.

3.3. Encoding and Decoding in Unstructured Vector Quantization

In *unstructured vector quantization*, the codebook is just a set of codevectors (reproduction vectors). In *structured vector quantization*, a fast encoding process is pursued by utilizing a special structure of the quantizer. Several structured vector quantizers are discussed in the next section. Here we concentrate on unstructured vector quantizers.

The decoding in unstructured VQ is performed by a table lookup and therefore is very fast and simple. Unfortunately, the encoding is not an easy task. In the encoding, the representative vector for the K-dimensional input vector is searched from the codebook of M codevectors. This *nearest-neighbor problem* is known in other applications, too. Search techniques can be classified into two groups: *exact methods* and *approximate methods*. The former ones always find the nearest codevector, whereas the approximate methods select a vector that is reasonably close and can be found quickly.

Several fast exact search techniques have been introduced to replace the exhaustive search. These techniques typically rely on the properties of the Euclidean vector space. Because the exhaustive search takes $O(MK)$ operations, it is therefore natural to try to reduce the effect of either M or K, or even both. *Partial distortion search* [8], *mean-distance-ordered partial search* [9,10], and *triangular inequality elimination* [11] are widely known methods for the exact search. The importance of efficient search techniques is emphasized by the fact that these techniques are also needed in the codebook generation algorithms, which are discussed later.

In an exact search, the approximate nearest-neighbor search is its own research field also. These techniques often guarantee some property, for example, that the selected codevector is at most at the distance ε from the input vector. One example of approximate techniques is the *tree-structured search*, in which the codebook is organized as a (binary) search tree.

3.4. Vector Quantization Structures

In the case of unstructured VQ, the codebook is just an array of codevectors. In addition, it is typical that the training vectors have been formed from the raw signal data without any special transformations. However, one should keep in mind that there are several alternatives for this basic organization of VQ [1,12,13]. We briefly discuss some of those structures in the following paragraphs.

In several compression methods, *prediction techniques* are utilized. The value of the next sample is predicted by the sample values, which have been coded thus far. The prediction is subtracted from the original pixel value giving a *prediction error* or *residual*, and then the residual is coded by a suitable technique. The same idea is exploited in *predictive vector quantization* (PVQ) [1] by predicting the next vector. The prediction is based on the previously coded vectors so that the decoder is capable of forming the same prediction. The prediction vector is then subtracted from the input vector to form a difference vector, which is finally quantized. Thus, the difference to basic schema of VQ is that here VQ is applied to residual vectors instead of pure vectors in the source domain. The design of a predictive vector quantizer [14–16] includes the design of the predictor in addition to the design of the codebook. These two are dependent on each other, which makes the whole process difficult. Basic approaches are *open-loop*, *closed-loop*, and *semi-closed-loop* designs, [1].

Finite-state vector quantization [1] uses several separate codebooks. The method is based on the assumption that the value of the next input vector depends on the values of the previous vectors. Therefore, a finite-state automaton is constructed to model the dependencies of the vectors. The encoding starts from an initial state of the automaton. In each state, a separate codebook is used for the quantization of the current input vector. The next state is selected on the basis of the reproduction vector. This approach allows us to construct a suitable quantizer for each particular vector context.

In *classified vector quantization* [17], an L-level classifier is used to select a subcodebook for the quantization of the input vectors. First, the encoder transmits the index of the proper subcodebook to the decoder and then the actual index of the selected codevector. The subcodebooks are designed separately and they can be of different sizes. There are several alternatives for the design of the classifier. One possibility is that the classifier is a common vector quantizer, which performs L-level clustering of the vector space. Statistical properties of the input vector can be utilized also. For image compression, the classifier can recognize the pixel blocks as *shade*, *midrange*, *edge*, or *mixed* blocks [17]. This approach has been inspired by the observation that there are too few edge vectors in a single codebook, which has been optimized according to the Euclidean distance. Therefore, the classified vector quantizer tries to guarantee a sufficient number of vectors for those image areas where errors are most annoying (cf. edge areas of the image). Other classified vector quantizers have been discussed [6,14,18].

Operating with high-dimensional vectors slows down both the encoding and the construction of the codebook. *Product code techniques* [1,13] are one way to relieve this problem. The idea is to divide the input vector into a collection of *subvectors*. Each of these describes a certain property of the vector, and therefore one can design a separate codebook for each property. The assumption is that the subvectors are easier to quantize because they take values from a smaller range of the K-dimensional space or have lower dimensionality. In particular, the dimension of a subvector can be one, in which case the property is a scalar. The encoder divides an input vector into a set of subvectors, which are quantized separately, and the corresponding indices are sent to the decoder. The decoder reproduces the subvectors from the indices and constructs a joint reproduction vector from these subvectors. This technique is called *product code vector quantization* because the whole (effective) reproduction codebook is a Cartesian product of all codebooks of the subvectors. Thus, when we have V subcodebooks with M_i codevectors in each, the number of possibilities to construct a reproduction vector is as large as

$$M = \prod_{i=1}^{V} M_i \tag{8}$$

but the storage space and encoding complexities are only of the magnitude

$$\tilde{M} = \sum_{i=1}^{V} M_i \tag{9}$$

The mean-removed vector quantizer [1] is one example of product codevector quantizers. It divides the input vector into two subvectors: a scalar that contains the average of the elements of the *input vector*, and a *difference vector*, where the average has been subtracted from each element of the input vector. *A shape–gain vector quantizer* [1,19] divides the input vector into the *gain* and *shape vectors* similarly. The gain is a scalar, whose value is the Euclidean norm (i.e., length) of the input vector. The shape vector consists of the elements of the input vector divided by the gain.

In *residual* or *multistage vector quantization* [16,20,21], the encoding process is divided into successive applications of separate vector quantizers. In the beginning, the input vector is coded by the first-stage quantizer. A residual vector is formed from the difference between the input vector and its reproduction vector. The second-stage quantizer is then applied to the residual vector. Again, a new residual vector is formed and the same procedure is iterated. By repeating this operation L times, we have an *L-stage vector quantizer*. A separate codebook is generated for each stage. The decoder constructs the whole reproduction vector by summing up the codevector of each quantizer. An upper bound for the number of separate reproduction vectors is the same as for the product codes [see Eq. (8)]. The complexity of the storage and encoding are also the same [see Eq. (9)].

A *lattice vector quantizer* [1,13] is simply a vector quantizer whose codebook is constructed from a lattice. Here the codevectors have a regular arrangement in the K-dimensional vector space. In fact, the lattice vector quantizer can be interpreted as a generalization of a

uniform scalar quantizer. An essential parameter in the design of a lattice vector quantizer is the density of the lattice, which defines how many lattice points per unit volume of the space are contained in the codebook. The higher the density, the higher the bit rate and the smaller the average distortion. The benefit of the lattice vector quantization is that special search methods can be applied because of the regular structure of the codevectors. In addition, the codebook can be expressed by its parameters instead of listing all the codevectors separately.

The output of the vector quantizer can be compressed further by applying entropy coding. Since the entropy codes can reduce the average length of the final code to the entropy of the vector, the vector quantizer should be designed to minimize its output entropy. In the design of an *entropy-constrained vector quantizer* [20,22–24], the task is to generate a codebook, which minimizes the average distortion between the source vectors and reproduction vectors subject to a constraint on the index entropy. Thus the distance between a training vector x_i and a codevector c_j is defined as

$$d(x_i, c_j) = \|x_i - c_j\|^2 + \lambda r_j \tag{10}$$

where r_j is the entropy of the index of codevector c_j, and λ is a weighting parameter for the entropy.

A vector quantizer is typically designed for a particular distribution of source vectors. However, it is possible that the input vectors given to the encoder are from a different distribution of vectors, and therefore the vector quantizer is unable to code these vectors well. In *adaptive vector quantization*, this problem is relieved by modifying the vector quantizer during the coding process. The predictive and finite-state vector quantizers could be considered as adaptive vector quantizers, because they change their encoding rule in time. However, usually, methods that change the codebook in order to match the local properties of input vectors are called *adaptive vector quantizers*.

4. CODEBOOK GENERATION

In the codebook generation problem, the task is to construct a codebook C that minimizes distortion D of Eq. (5) for a given training set X. If the VQ system transfers the codebook to a receiver as a part of compressed data, it is natural to form the training set from the source signal to be compressed. Instead, when the system uses a *static codebook*, which is known by both the encoder and the decoder, the training set is formed from a large set of similar signal samples as the signal source to be compressed. When generating a static codebook, its quality has to be determined against input vectors outside the training set. This process and the construction of training sets have been studied [25,26]. To exclude trivial problem instances, we suppose that the size M of the codebook is smaller than the size N of the training set (in practice we assume that $M \ll N$).

In addition to the distortion D of the produced codebook, the quality of a codebook generation algorithm can be characterized by other attributes. *Robustness* of the algorithm describes how independent the output of the algorithm is from the initial conditions and parameter setups. Another important property of the algorithm is small *running time*; especially in online applications. A *scalable* algorithm is able to improve the quality of the codebook with additional computing resources. *Memory requirements* should be reasonable, so that the algorithm would also work for large problem instances.

The codebook generation problem resembles the *clustering problem* [27], in which the task is to classify N input vectors into M clusters. However, the quality of the clustering is measured both by evaluating the *similarity* of objects (vectors) in the same cluster and *dissimilarity* of the objects in different clusters. Instead, in codebook generation, the task is to find a good set of *representatives* for the input vectors. In addition, the number of codevectors M is usually given in the setup of the codebook construction, whereas in clustering the search for the right number of clusters may be a vital part of the problem. The codebook generation problem also resembles the *P-median problem*, which differs from VQ in that the solution vectors of the P-median problem are limited to be vectors of the training set. A similar approach has been also applied to codebook generation in VQ by modifying the generalized Lloyd algorithm (GLA). Codebook generation for RGB-tuples is known as a *palette generation problem* in color image quantization [28,29].

Construction of an optimal codebook is a combinatorial optimization problem and it is NP-hard [30]. In other words, there is no known polynomial time algorithm for finding the globally optimal solution. However, reasonable suboptimal solutions are typically obtained by *heuristic algorithms* [31,32]. Methods to solve the codebook generation problem can be divided into *problem specific methods* and *general optimization methods*. Problem-specific methods have been developed particularly to solve the codebook generation problem (and the clustering problem). General optimization methods, however, are suitable for all optimization problems on the condition that the problem has been formulated properly for a particular optimization method.

In the following, we concentrate on codebook construction methods for unstructured vector quantizers. However, after minor modifications, these general methods are suitable for design of other vector quantizers also. By *iterative method*, we mean a method that attempts to improve the quality of an existing solution (codebook). Therefore, these methods need an initial codebook, which is feasible for the given problem. Examples of iterative methods are GLA and genetic algorithms, which are discussed later. *Hierarchical methods* are problem-specific and are widely used in clustering problems. Hierarchical methods can be classified as *divisive* and *agglomerative* types, and they repeatedly split or merge clusters until a clustering of the desired size has been reached. The splitting method and the pairwise nearest-neighbor (PNN) algorithm are two examples of hierarchical methods, which are discussed later. In the following three sections we briefly describe problem-specific methods. After that, we also briefly discuss general optimization methods.

4.1. Generalized Lloyd Algorithm

The *generalized Lloyd algorithm* (GLA) [33] is perhaps the most widely known and used method for codebook generation. The algorithm is based on the iterative use of the codebook modification operation generalized from the Lloyd's algorithm for SQ. The GLA is popular also in the context of clustering and pattern recognition, where it is known as a *C-means algorithm* [34]. C refers here to the number of clusters (codevectors).

The codebook modification operation is based on two optimality conditions: *nearest-neighbor condition* and *centroid condition*. These conditions describe how to generate an optimal clustering for a given codebook, and vice versa, how to generate the optimal codebook for a given clustering.

- *Nearest-Neighbor Condition*. For a given codebook, the optimal clustering of the training set is obtained by mapping each training vector to its nearest codevector in the codebook with respect to the distortion function.
- *Centroid Condition*. For a given cluster, the optimal codevector is the *centroid* (average vector) of the vectors within the cluster.

The GLA applies the two optimality conditions in turn. In the *clustering step*, the training set is grouped into clusters according to the existing codebook. The optimal clustering is obtained by mapping each training vector to the nearest codevector. In the *codebook step* a new codebook is constructed by calculating the centroids of the clusters defined in the clustering step. The two optimality conditions guarantee that the new solution is always equal to or better than the previous one. However, it should be noted that although these two steps are locally optimal, the whole process does not necessary produce an optimal codebook. Most of the computational burden of the GLA originates from the clustering step, which can be expedited by several techniques [35].

After the clustering step, it may happen that the codebook contains some codevectors to which no training vectors have been mapped, that is, that do not represent any of the training vectors. This so *empty-cluster problem* is usually solved by replacing the codevectors of empty clusters by some existing training vectors [1].

The GLA can be iterated until the quality of the codebook does not improve anymore. Another *stopping condition* is to require the relative improvement to be higher than a given threshold. The relative improvement can be measured by

$$\Delta D = \frac{D(C^{(t)}, X) - D(C^{(t+1)}, X)}{D(C^{(t)}, X)} \tag{11}$$

where $C^{(t)}$ refers to the codebook on the tth iteration round and $D(C, X)$ is the distortion of coding training set X with C. The process can also be limited to a certain number of iterations.

The GLA is a *descent* algorithm, which means that each iteration decreases (or at least never increases) the distortion. Because of the deterministic nature of the GLA, the process leads to a *local optimum*, which depends on the *initial codebook*. Because of this, the GLA is unable to locate a *globally optimal codebook*. The GLA can be seen as a *fine-tuner* of a codebook, and it has been integrated in many other algorithms.

The GLA tries to improve the quality of an existing codebook, and therefore it needs an initial codebook to start with. Techniques for generating an initial codebook are typically fast and simple; at least faster than the GLA itself. One such technique, which is often satisfactory, is to use M random training vectors as the codevectors. Several other techniques have been proposed in the literature [1,33].

4.2. Splitting Method

The splitting method [36] starts from a *singular codebook* containing a single codevector, which is the centroid of the whole training set. The codebook is then enlarged hierarchically by splitting a cluster (represented by a codevector) into two new clusters. This procedure is repeated until the codebook reaches the desired size. The selection of a codevector to be split is based on the properties of its cluster. For example, this property may be the *variance* of training vectors in the cluster. The problem of selecting the cluster is avoided in *binary splitting* [33], where all current clusters are split. However, this approach may allocate too many codevectors to sparse areas of the vector space.

Splitting of a given cluster is a special codebook generation problem ($M = 2$), where the task is to find two new representative codevectors for a given subcluster. We have two approaches to this problem. The *partition-based* approach divides the training vectors of the cluster into two subclusters and uses the two centroids of the subclusters as new codevectors. In the *codevector-based* approach, two codevectors are selected by some heuristic and the training vectors are mapped to them.

The overall structure of the splitting method is easy to understand. However, internal components of the method can be rather complicated. The selection of these components gives us a versatile way to generate codebooks of varying quality with controllable running time. The top–down process of the splitting method has the benefit that it can be used also for the construction of a *tree-structured vector quantizer* [37].

4.3. Pairwise Nearest Neighbor

The pairwise nearest-neighbor (PNN) algorithm [38] belongs to the class of *agglomerative clustering methods*. In the field of clustering research the PNN is often called *Ward's method* [39]. Because the cluster centroids can be easily used as codevectors of the codebook, the PNN is suitable for codebook construction. The algorithm starts by constructing an initial codebook in which each training vector is considered as its own codevector. Two nearest codevectors, whose merging increases the total distortion least, are merged at each step, and the process is repeated until the desired size of the codebook has been reached.

The algorithm is straightforward to implement in its basic form, and in comparison to the GLA, it gives good-quality results (i.e., codebooks with small distortion). The PNN has also the advantage that the hierarchical approach produces codebooks of differing sizes as a side product. Thus, the PNN can be easily applied to joint minimization of distortion and entropy of codevector indices [23,24,40]. The algorithm can also be used as a part of hybrid methods such as a genetic algorithm [41], or an iterative split-and-merge method [42].

A drawback of the PNN is the relatively high running time in its exact form [43]. There is a large number of steps to be performed by the algorithm, because typically we have $M \ll N$, and at each step, all pairwise distances are calculated for finding the pair of vectors to be merged. This makes the algorithm very slow for large training sets. It should also be noted that although a single merge operation is always performed optimally (i.e., two nearest clusters are merged), the whole process does not guarantee an optimal codebook of the size M.

Approximate PNN variants [44,46] reduce the number of distance calculations by relieving the condition of merging two nearest clusters. In the *K-d tree* variant [38], the nearest cluster is searched from a (small) subset of clusters only, and several cluster pairs are merged at the same iteration round. This approach is clearly faster than the original one, but the quality of the codebooks is worse. Fast *exact PNN variants* have been proposed [45].

4.4. General Optimization Methods

Although the codebook generation problem is NP-hard [30], good solutions can be produced by general heuristic optimization methods [1,47]. In this section, we describe briefly some methods, which have been observed to work well for several other difficult combinatorial problems and have been successful for codebook optimization also. Their application to the generation of unstructured codebooks is straightforward.

Genetic algorithms (GAs) are optimization techniques derived from the model of the natural selection in real-life evolution. The idea is that a *population* of possible solutions (codebooks) called *individuals* is first (randomly) generated. The next generation consists of the survivors and of new individuals generated from the individuals of the former population by genetic operations such as *crossover* and *mutation*. As in the real life, the genes of the best-fitted individuals survive.

A problem of several optimization methods is that the methods may stick in a local minimum. *Tabu search* [48,49] tries to avoid this problem by keeping the previous solutions in a *tabu list*. The method starts from an initial solution by generating a set of candidate solutions. Solutions, that appear in the tabu list are discarded. The best of the remaining candidates is selected as a new solution, and it is inserted in the tabu list. The procedure is repeated until a stopping condition is fulfilled. The purpose of the tabu list is to prevent the search from returning to solutions that have been evaluated recently. This forces the search to proceed into new directions instead of sticking in a local minimum and in its neighborhood.

Stochastic relaxation (SR) is a family of optimization techniques, which add perturbation to the current solution at each iteration. The amount of perturbation decreases with time, making convergence possible. Because the method allows an increase of the distortion, it can continue the search after reaching a local minimum. A popular variant of SR algorithms is known as *simulated annealing* (SA) [50,51]. The solutions, which decrease the distortion, are always accepted in the SA. However, the acceptation of a solution, which increases the distortion, is determined probabilistically. This probability depends on the change of the distortion value and the parameter called *temperature*, whose decrease rate is given by a *cooling schedule*.

Neural network algorithms used in learning applications are proposed for the codebook generation problem in Refs. 52 and 53. The idea is that the neural network learns the properties of the training set and the codebook is the result of this learning process. A popular method is the *self-organization feature maps* (SOM). The method starts from an initial codebook. One training vector is used to modify the codebook at the time. This is done by moving the training vector's nearest codevector and its neighboring codevectors toward the training vector. The size of the neighborhood reduces during the process and finally only the nearest codevector is moved. Because the method is based on the update of the existing codebook, it is also suitable for adaptive vector quantization.

5. SUMMARY

Scalar quantization is a basic technique for analog-to-digital signal transformation. It has an extensive theoretical background in addition to practical usefulness. One can augment the method with variable-rate compression techniques in order to decrease the storage space.

In theory, vector quantization offers very good possibilities for lossy signal compression. The method has the great advantage that its decoding runs extremely fast. Unfortunately, it includes other complexity problems. To obtain good compression results, one has to use codebooks with high dimensionality with respect to both the vector dimension and the number of reproduction vectors. This makes the encoding process slow and the memory consumption of the compression system unpractical.

Acknowledgment

The author would like to thank Professor O. Nevalainen for his helpful comments and support during this work.

BIOGRAPHY

Timo Kaukoranta received his M.Sc. and Ph.D. degrees in computer science from the University of Turku, Finland, in 1994 and 2000, respectively. He has been a researcher at the University of Turku from 1994 to 2000. Since 2001 he has been a postdoctoral researcher at Turku Centre for Computer Science (TUCS). His primary research interests are in distributed interactive simulations, multiplayer computer games, and vector quantization.

BIBLIOGRAPHY

1. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Dordrecht, 1992.
2. W. R. Bennett, Spectra of quantized signals, *Bell Syst. Tech. J.* **27**: 446–472 (1948).
3. M. W. Marcellin et al., An overview of quantization in JPEG 2000, *Signal Process. Image Commun.* **17**: 73–84 (2002).
4. S. P. Lloyd, *Least Squares Quantization in PCM*, unpublished Bell Laboratories Technical Note; portions presented at the Institute of Mathematical Statistics Meeting, Atlantic City, NJ, 1957; published in special issue on quantization, *IEEE Trans. Inform. Theory* **28**: 129–137 (1982).
5. W. Ding, Optimal vector transform for vector quantization, *IEEE Signal Process. Lett.* **1**(7): 110–113 (1994).
6. B. Marangelli, A vector quantizer with minimum visible distortion, *IEEE Trans. Signal Process.* **39**(12): 2718–2721 (1991).
7. V. J. Mathews and P. J. Hahn, Vector quantization using the L_∞ distortion measure, *IEEE Signal Process. Lett.* **4**(2): 33–35 (1997).
8. C.-D. Bei and R. M. Gray, An improvement of the minimum distortion encoding algorithm for vector quantization, *IEEE Trans. Commun.* **33**(10): 1132–1133 (1985).
9. S.-W. Ra and J.-K. Kim, A fast mean-distance-ordered partial codebook search algorithm for image vector quantization, *IEEE Trans. Circuits Syst.-II: Analog Digital Signal Process.* **40**(9): 576–579 (1993).
10. S. Baek, B. Jeon, and K.-M. Sung, A fast encoding algorithm for vector quantization, *IEEE Signal Process. Lett.* **4**(12): 325–327 (1997).
11. S.-H. Chen and W. M. Hsieh, Fast algorithm for VQ codebook design, *IEE Proc.-I* **138**(5): 357–362 (1991).
12. N. M. Nasrabadi and R. A. King, Image coding using vector quantization: A review, *IEEE Trans. Commun.* **36**(8): 957–971 (1988).
13. R. M. Gray and D. L. Neuhoff, Quantization, *IEEE Trans. Inform. Theory* **44**(6): 2325–2384 (1998).
14. K. N. Ngan and H. C. Koh, Predictive classified vector quantization, *IEEE Trans. Image Process.* **1**(3): 269–280 (1992).
15. S. A. Rizvi and N. M. Nasrabadi, Predictive vector quantizer using constrained optimization, *IEEE Signal Process. Lett.* **1**(1): 15–18 (1994).
16. S. A. Rizvi and N. M. Nasrabadi, Predictive residual vector quantization, *Proc. IEEE Int. Conf. Image Processing*, 1994, pp. 608–612.
17. B. Ramamurthi and A. Gersho, Classified vector quantization of images, *IEEE Trans. Commun.* **34**(11): 1105–1115 (1986).
18. K. L. Oehler and R. M. Gray, Combining image compression and classification using vector quantization, *IEEE Trans. Pattern Anal. Mach. Int.* **17**(5): 461–473 (1995).
19. K. L. Oehler and R. M. Gray, Mean-gain-shape vector quantizer, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Minneapolis, MN, 1993, Vol. V, pp. 241–244.
20. F. Kossentini, M. J. T. Smith, and C. F. Barnes, Image coding using entropy-constrained residual vector quantization, *IEEE Trans. Image Process.* **4**(10): 1349–1357 (1995).
21. C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi, Advances in residual vector quantization: A review, *IEEE Trans. Image Process.* **5**(2): 226–262 (1996).
22. P. A. Chou, T. Lookabaugh, and R. M. Gray, Entropy-constrained vector quantization, *IEEE Trans. Acoust. Speech Signal Process.* **37**(1): 31–42 (1989).
23. D. P. de Garrido, W. A. Pearlman, and W. A. Finamore, Vector quantization of image pyramids with the ECPNN algorithm, *SPIE Proc. Visual Commun. Image Process.* **1605**: 221–232 (1991).
24. F. Kossentini and M. J. T. Smith, A fast PNN design algorithm for entropy-constrained residual vector quantization, *IEEE Trans. Image Process.* **7**(7): 1045–1050 (1998).
25. D. Cohn, E. A. Riskin, and R. Ladner, Theory and practice of vector quantizers trained on small training sets, *IEEE Trans. Pattern Anal. Mach. Int.* **16**(1): 54–65 (1994).
26. D. S. Kim, T. Kim, and S. U. Lee, On testing trained vector quantizer codebooks, *IEEE Trans. Image Process.* **6**(3): 398–406 (1997).
27. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
28. M. T. Orchard and C. A. Bouman, Color quantization of images, *IEEE Trans. Signal Process.* **39**(12): 2677–2690 (1991).
29. P. Scheunders, A comparison of clustering algorithms applied to color image quantization, *Pattern Recogn. Lett.* **18**: 1379–1384 (1997).
30. M. R. Garey, D. S. Johnson, and H. S. Witsenhausen, The complexity of the generalized Lloyd-Max problem, *IEEE Trans. Inform. Theory* **28**(2): 255–256 (1982).
31. C.-M. Huang and R. W. Harris, A comparison of several vector quantization codebook generation approaches, *IEEE Trans. Image Process.* **2**(1): 108–112 (1993).
32. N. Akrout, R. Prost, and R. Goutte, Image compression by vector quantization: A review focused on codebook generation, *Image Vision Comput.* **12**(10): 627–637 (1994).
33. Y. Linde, A. Buzo, and R. M. Gray, An algorithm for vector quantizer design, *IEEE Trans. Commun.* **28**(1): 84–95 (1980).
34. J. B. McQueen, Some methods of classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Math. Statist. Probability* 1, Univ. of California, Berkeley, 1967, Vol. 1, pp. 281–296.
35. T. Kaukoranta, P. Fränti, and O. Nevalainen, A fast exact GLA based on code vector activity detection, *IEEE Trans. Image Process.* **9**(8): 1337–1342 (2000).
36. P. Fränti, T. Kaukoranta, and O. Nevalainen, On the splitting method for VQ codebook generation, *Opt. Eng.* **36**(11): 3043–3051 (1997).
37. J. Lin and J. A. Storer, Design and performance of tree-structured vector quantizers, *Inform. Process. Manage.* **30**(6): 851–862 (1994).
38. W. H. Equitz, A new vector quantization clustering algorithm, *IEEE Trans. Acoust. Speech Signal Process.* **37**(10): 1568–1575 (1989).
39. J. H. Ward, Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.* **58**: 236–244 (1963).
40. D. P. de Garrido, W. A. Pearlman, and W. A. Finamore, A clustering algorithm for entropy-constrained vector quantizer design with applications in coding image pyramids, *IEEE Trans. Circuits Syst. Video Technol.* **5**(2): 83–95 (1995).

41. P. Fränti, J. Kivijärvi, T. Kaukoranta, and O. Nevalainen, Genetic algorithms for large scale clustering problem, *Comput. J.* **40**(9): 547–554 (1997).

42. T. Kaukoranta, P. Fränti, and O. Nevalainen, Iterative split-and-merge algorithm for VQ codebook generation, *Opt. Eng.* **37**(10): 2726–2732 (1998).

43. J. Shanbehzadeh and P. O. Ogunbona, On the computational complexity of the LBG and PNN algorithms, *IEEE Trans. Image Process.* **6**(4): 614–616 (1997).

44. T. Kurita, An efficient agglomerative clustering algorithm using a heap, *Pattern Recogn.* **24**(3): 205–209 (1991).

45. P. Fränti, T. Kaukoranta, D.-F. Shen, and K.-S. Chang, Fast and memory efficient implementation of the exact PNN, *IEEE Trans. Image Process.* **9**(5): 773–777 (2000).

46. T. Kaukoranta, P. Fränti, and O. Nevalainen, Vector quantization by lazy pairwise nearest neighbor method, *Opt. Eng.* **38**(11): 1862–1868 (1999).

47. C. R. Reeves, ed., *Modern Heuristic Techniques for Combinatorial Problems*, McGraw-Hill, UK, 1995.

48. K. Al-Sultan, A tabu search approach to the clustering problem, *Pattern Recogn.* **28**(9): 1443–1451 (1995).

49. P. Fränti, J. Kivijärvi, and O. Nevalainen, Tabu search algorithm for codebook generation in vector quantization, *Pattern Recogn.* **31**(8): 1139–1148 (1998).

50. J. K. Flanagan et al., Vector quantization codebook generation using simulated annealing, *Proc. ICASSP*, 1989, pp. 1759–1762.

51. S. Z. Selim and K. Alsultan, A simulated annealing algorithm for the clustering problem, *Pattern Recogn.* **24**(19): 1003–1008 (1991).

52. N. M. Nasrabadi and Y. Feng, Vector quantization of images based upon the Kohonen self-organizing feature maps, *Proc. IEEE Int. Conf. Neural Networks*, 1988, pp. 101–108.

53. N. B. Karayiannis and P.-I. Pai, Fuzzy algorithms for learning vector quantization, *IEEE Trans. Neural Networks* **7**(5): 1196–1211 (1996).

SECURE ULTRAFAST DATA COMMUNICATION AND PROCESSING INTERFACED WITH OPTICAL STORAGE

BAHRAM JAVIDI
University of Connecticut
Storrs, Connecticut

OSAMU MATOBA
University of Tokyo
Tokyo, Japan

1. INTRODUCTION

Pulseshapers of femtosecond laser pulses [1–6] are attractive in wide research fields such as optical communication, information processing, and spectroscopy. In an ultrafast data communication system, spatial data can be sent to remote users at an ultrafast rate, faster than terabit/s, via optical fiber communications. Because there is no device to handle temporal data directly at such an ultrafast rate, the pulseshaper is one of the most promising methods to send or process data. To send a large amount of data, it is also required to use optical storage systems that can readout data in parallel and at an ultrafast rate.

In this article, we present a secure ultrafast data communication system that can link remote users to an encrypted optical database with ultrafast transfer rate [7]. It is well known that holographic memories potentially have a large storage capacity with a fast data transfer rate [8,9]. Security communication is achieved by use of encrypted holographic memory. In the encrypted holographic memory, each data frame to be stored is encoded by optical encryption techniques, such as double-random-phase encryption [10,11] or by an exclusive-OR method. Spatial-temporal converters enable us to send the stored data in the encrypted holographic memory to the receiver at an ultrafast rate. At the remote users, the correct key is required to reconstruct the original data. Without the key information, it is very difficult to decrypt the data because optics can provide a high level of security. In the optical security system more than the one-dimensional nature of light and many physical parameters of the lightwave can be used to encode the data. We show the operation of the secure ultrafast data communication system and present some numerical results of encryption and decryption.

Section 2 describes a secure ultrafast data communication system. Section 3 numerically evaluates the performance of the secure data communication system.

2. SECURE ULTRAFAST DATA COMMUNICATIONS

Figure 1 shows a block diagram and a schematic of a secure data communication system. An encrypted holographic memory is linked to remote users via optical fibers. The secure holographic memory works as an encrypted database. All the data frames to be sent to the remote users were already encrypted by optical encryption techniques and then were recorded in the encrypted holographic memory. The data frame readout from the database is sent to the remote user via an optical fiber by spatio-temporal converters. After the transmission, an original data frame is decrypted at the remote users by using the correct key. The present system consists of four subsystems: the secure holographic memory system, a transmitter based on a space-to-time converter, a receiver based on a time-to-space converter, and a decrypting system to recover the original data. The following subsections describe the operation of each subsystem.

2.1. Secure Holographic Memory

Optical encryption opens new research fields in optical information processing [12–15]. An optical encryption technique, called *double random-phase encryption*, described by Rèfrègier and Javidi in 1995 [10], has been proposed and extensively investigated. The double-random-phase encryption technique can convert an original data frame into a white-noise distribution by using two random phase functions at the input and the Fourier planes. The double random phase encryption technique is easy to introduce into a holographic memory [16–19]. We

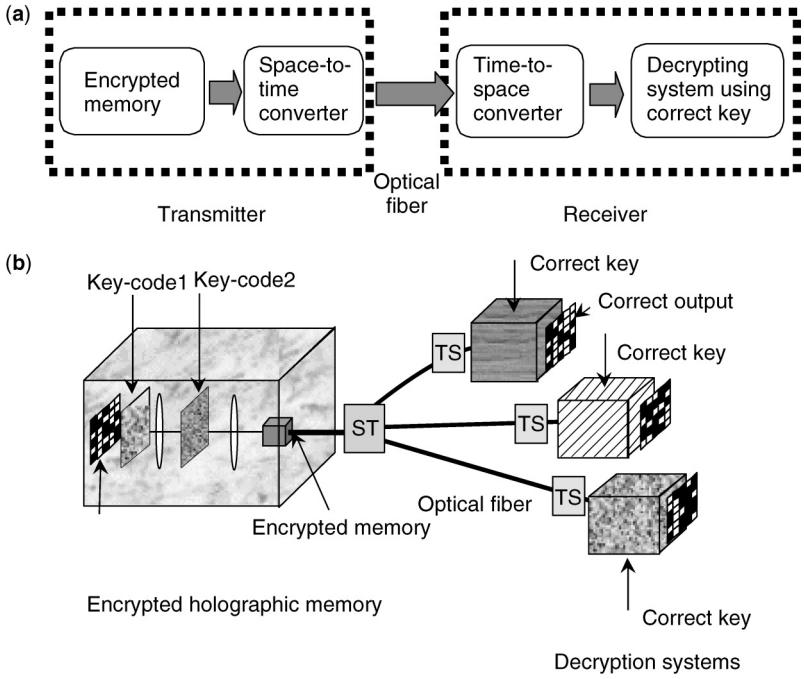


Figure 1. (**a**) Block diagram and (**b**) schematic of secure data communication system. ST and TS denote the space-to-time and time-to-space converters, respectively.

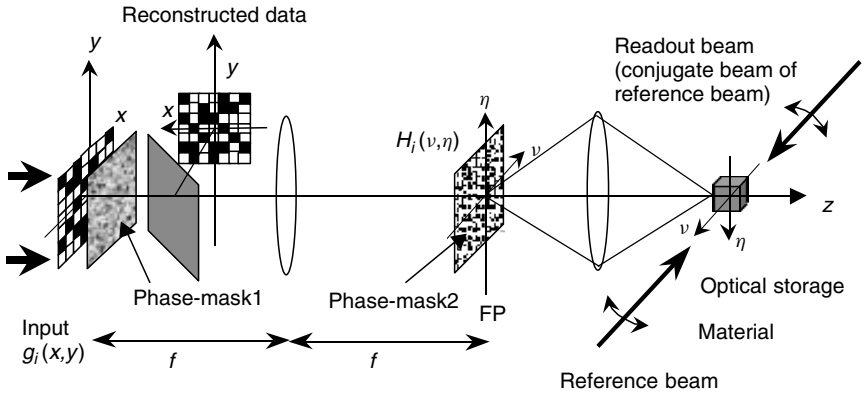


Figure 2. Encrypted holographic memory system.

briefly describe the encrypted holographic storage system, as shown in Fig. 2, using double-random-phase encryption.

Let $g_i(x)$ denote the ith positive and real data to be encrypted. We consider one-dimensional description because the following spatio-temporal converter uses one-dimension signals in the temporal domain. In the encryption process, the ith original data, $g_i(x)$, is multiplied by a random-phase mask (RPM1), $\exp\{-jn_i(x)\}$. This modulated input is Fourier-transformed and then is multiplied by a second random-phase mask (RPM2), $\tilde{H}_i(\nu) = \exp\{-h_i(\nu)\}$. Here x and ν denote the spatial domain and the Fourier domain coordinates, respectively. Two independent white sequences, as $n_i(x)$ and $h_i(\nu)$, are uniformly distributed on the interval $[0, 2\pi]$. After taking another Fourier transform, we obtain the encrypted data, $e_i(x)$:

$$e_i(x) = g_i(x) \exp\{-jn_i(x)\} \otimes F[\exp\{-jh_i(\nu)\}] \qquad (1)$$

where $\otimes$ denotes convolution and $F[\cdot]$ denotes the Fourier transformation. Equation (1) shows that encrypted data are white-noise-like data because of the convolution of two independent white noises.

In a holographic memory, as shown in Fig. 2, the Fourier transformed pattern of the encrypted data described by Eq. (1) is stored holographically together with a reference beam as a *plane wave*. Photorefractive materials are used to record volume holograms. In photorefractive materials, the intensity distribution is recorded as a refractive-index distribution. In order to record many data frames, angular multiplexing is employed. The total intensity distribution, $\phi(\nu)$, stored in the photorefractive material is given by

$$\phi(\nu) = \sum_{i=1}^{M} |\tilde{E}_i(\nu) + \tilde{R}_i(\nu)|^2 \qquad (2)$$

where M is the total number of stored images, $\tilde{E}_i(\nu)$ is the Fourier transform of ith input encrypted data described in Eq. (1), and $\tilde{R}_i(\nu)$ is a reference beam with a specific angle used to record the ith encrypted data.

A sufficient separation angle prevents crosstalk between adjacent stored data in the reconstruction.

Before we present the security of data communication, we show some experimental results in holographic encrypted memory systems as shown in Fig. 2. In the encrypted holographic memory, all stored images are encrypted. In the decryption process, we have to eliminate the phase modulation caused by the two random-phase modulations. Therefore we use the phase conjugate reconstruction in the decryption process. A readout beam is the conjugate beam of the reference beam used in the recording. The conjugate readout can eliminate the phase distortions of the optical field due to the random phase masks and the aberrations of optical elements. The data of the ith stored image can be reconstructed only when the readout beam is incident at a correct angle. The reconstructed data at the photorefractive material is written as $\tilde{E}_i^*(\nu)$, where the asterisk denotes complex conjugate. When we use a phase key, $\tilde{K}_i(\nu) = \exp\{-j2\pi k_i(\nu)\}$ in the Fourier plane, the reconstructed data in the Fourier plane is written as

$$\tilde{S}_i'(\nu) = F[g_i * (x)\exp\{jn_i(x)\}]\tilde{H}_i * (\nu)\tilde{K}_i(\nu) \tag{3}$$

The ith reconstructed image can be obtained by taking another Fourier transform of Eq. (3):

$$s_i'(x) = [g_i * (x)\exp\{jn_i(x)\}] \otimes C_i(x) \tag{4}$$

where

$$C_i(x) = F[\exp\{-jh_i(\nu)\}] \oplus F[\exp\{-jk_i(\nu)\}] \tag{5}$$

In Eq. (5), $\oplus$ denotes correlation. When a correct phase key, $k_i(\nu) = h_i(\nu)$, is used in the Fourier plane, the original data are successfully recovered because Eq. (5) becomes a delta function. The random phase function in the input plane is removed by detecting with an intensity-sensitive device such as a CCD (charge-coupled device) camera. When an incorrect phase key, $k_i(\nu) \neq h_i(\nu)$, is used, the reconstructed data frame is still a white-noise-like image.

Figure 3 shows the experimental setup. An Ar^+ laser at a wavelength of 514.5 nm is used as recording and readout beams. A light beam emitted from the Ar^+ laser was divided into an object beam and a reference beam by a beamsplitter, BS1. The reference beam was again divided into two reference beams by a beamsplitter BS2: one for recording holograms and one for the conjugate readout. All the beams were ordinarily polarized due to the creation of an interference pattern. We use a $10 \times 10 \times 10$ mm^3 $LiNbO_3$ crystal doped with 0.03 mol% Fe as a photorefractive material. The crystal was placed at the Fourier plane and was mounted on a rotary stage for angular multiplexing. The c axis is on the paper and is at 45° with respect to the crystal faces.

An input binary image is displayed on a liquid crystal display controlled by a computer. The input image is illuminated by a collimated beam and is then Fourier-transformed by lens L1. Two random phase-masks, RPM1 and RPM2, are located at the input and the Fourier planes, respectively. A reduced size of the Fourier-transformed image is imaged into the $LiNbO_3$ crystal by lens L2. This Fourier-transformed image and a reference beam interfere with an angle of 90° in the $LiNbO_3$ crystal. In holographic recording, shutters SH1 and SH2 are opened, and SH3 is closed. The encrypted image is observed by a CCD camera (CCD1) after the Fourier transform was taken by lens L3. The focal lengths of L1, L2, and L3 were 400, 58, and 50 mm, respectively.

In the decryption process, the readout beam is the conjugate beam of the reference beam used in the recording. In the experiments, a pair of counterpropagating plane waves is used as the reference and readout beams. Shutters SH1 and SH2 are closed, and SH3 is opened. If the same mask used in the recording is located at the same place, the original image is reconstructed successfully. The intensity information of the reconstructed image is obtained at the CCD camera (CCD2).

We present an example of the experimental results as shown in Fig. 4. In the experiments, angularly multiplexed recording of four binary images was demonstrated. One of the four original binary images is shown in Fig. 4a.

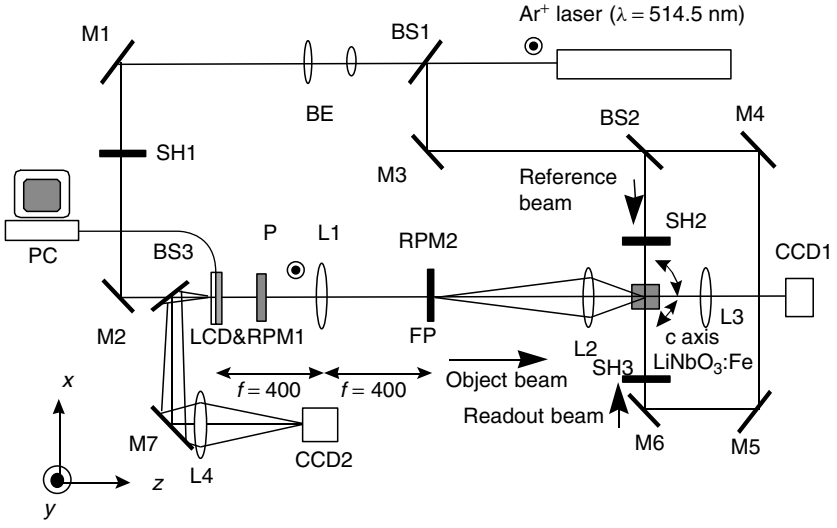


Figure 3. Experimental setup: LCD — liquid crystal display; RPM — random phase masks; BS — beamsplitters; L — lenses; M — mirrors; BE — beam expander; SH — shutters; CCD — CCD cameras; FP — Fourier plane; P — Polarizer; PC — personal computer.

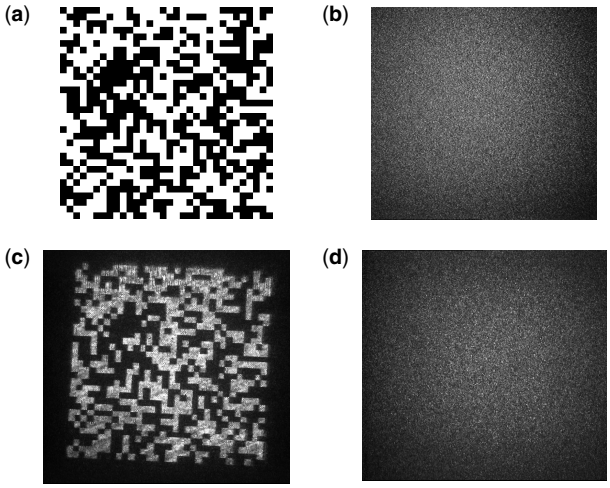


Figure 4. An example of encrypted holographic memory: (**a**) original binary image to be encrypted; (**b**) encrypted image; (**c**) reconstructed image by using correct key; (**d**) reconstructed image by using incorrect key.

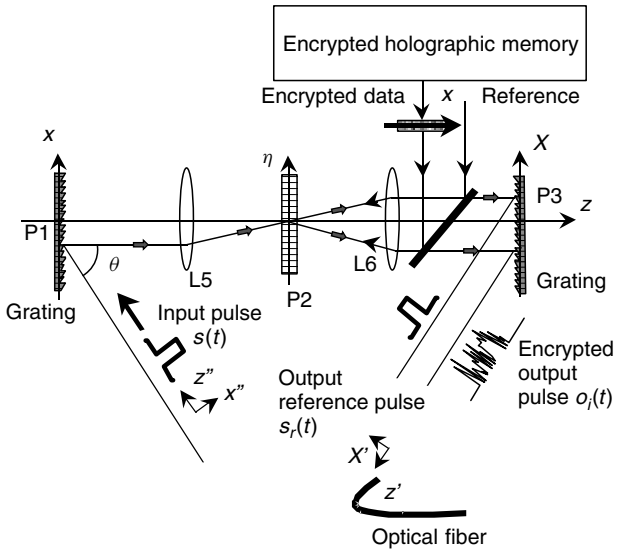


Figure 5. Transmitter based on space-to-time converter.

The image consists of 32×32 pixels and is randomly generated by a personal computer. Two diffusers are used as the random phase-masks, RPM1 and RPM2. Figure 4b shows the intensity distributions of the encrypted images. We can see that random-noise-like image was observed. In the recording process, the optical powers of the object and the reference beams were 78 and 1.4 W/cm^2, respectively. The exposure time was 60 s. Angular multiplexing was achieved by rotating the $LiNbO_3$ crystal in the plane of Fig. 3. The angular separation between adjacent stored images was 0.2°. Figure 4c shows the reconstructed images obtained by using the correct phase keys. These keys are the same as the random-phase masks in the Fourier plane used in the recording. Figure 4c shows that the stored image was reconstructed successfully. After the binarization, we confirmed that there is an error-free reconstruction in the four reconstructed images. Figure 4d shows the reconstructed image when an incorrect key was used. Here the incorrect key was generated by shifting the correct key. The reconstructed image was still a random-noise-like image. The average bit error rate for the reconstructed images with the incorrect keys was 0.502.

In the ultrafast secure communication system as described in Fig. 1, the encrypted data are transmitted to remote users by spatio-temporal converters via optical fibers and then the decryption is implemented at the remote users. To readout the encrypted data from the memory, we use the reference beam as the readout beam. In this case, the reconstructed data are given by Eq. (1) and then are converted into temporal data using the space-to-time converter.

2.2. Transmitter

Figure 5 shows an optical transmitter based on a space-to-time converter with ultrashort pulses. The space-to-time converter converts spatial data to temporal data by controlling the amplitude and phase of the spectra of the input pulse at the Fourier plane, P2 in Fig. 5. Sun and Fainman have extensively investigated the space-to-time converter [5,6]. Using their analysis, we briefly derive an output temporal signal to be sent to the receivers. Because the encrypted data readouts from the secure holographic memory are complex-valued, we introduce here a complex value notation to describe the operation of the space-to-time converter.

An input ultrashort pulse, $s(t)$, is described as

$$s(t) = p(t - t_0)\exp(j\omega_0 t) \tag{6}$$

where $p(t)$ is the envelope of the pulse, t_0 is the time of peak intensity, and ω_0 is the central temporal frequency of the pulse. The temporal frequency distribution of the input pulse, $\tilde{S}(\omega)$, is obtained, by taking the temporal Fourier transform of Eq. (6):

$$\tilde{S}(\omega) = \tilde{P}(\omega - \omega_0)\exp\{-j(\omega - \omega_0)t_0\} \tag{7}$$

where $\tilde{P}(\omega)$ is the temporal Fourier transform of $p(t)$. In the same pulseshaper as described in Fig. 5, each temporal frequency distribution is converted into the spatial distribution at the Fourier plane by using a grating and Fourier-transform lens, L5.

We consider the temporal frequency response of the space-to-time converter in Fig. 5 and then obtain an output temporal pulse. The space-to-time converter consists of two pairs of grating and Fourier transform lens. Suppose that the grating diffracts the light pulse with a temporal frequency of ω_0 into the direction parallel to the optical axis (z axis). When a monochromatic plane wave with a temporal frequency of ω is incident on the grating at an angle of θ, the diffracted optical field at plane P1 is given by

$$\psi_1(x;\omega) = \exp\left\{-j\frac{\omega - \omega_0}{c}\alpha x\right\} w(x) \tag{8}$$

where $\alpha = sin\theta$, w(x) is a pupil function of the grating, and c is the speed of light in vacuum. After taking the Fourier

transform of Eq. (8) by lens L5, the optical field at P2 is given by

$$\tilde{\psi}_2(\eta;\omega) = \tilde{W}\left\{\frac{\omega\eta}{2\pi cf} + \frac{\omega-\omega_0}{2\pi c}\alpha\right\} \tag{9}$$

where $\tilde{W}(\eta)$ is the spatial Fourier transform of $w(x)$, η is the Cartesian coordinate in the plane P2, and f is the focal length of lens L5. Equation (9) shows that the spectral distribution of the temporal delta function is projected into the spatial distribution as a function of the temporal frequency of light. When the pupil function $w(x)$ is infinite [i.e., $\tilde{W}(\eta) = \delta(\eta)$], the relation between η and $(\omega - \omega_0)$ is given by

$$\eta = -f\alpha\frac{\omega-\omega_0}{\omega} \tag{10}$$

The spatially spread spectra described in Eq. (9) are modulated by a hologram that is the interference pattern between the Fourier-transform of the encrypted signal described in Eq. (1) and a reference plane wave.

We suppose that the encrypted signal $e_i(x)$ is spatially sampled at an interval of Δ as follows:

$$\begin{aligned} r_i(x) &= \sum_n e_i(x)\delta(x-n\Delta) \\ &= \sum_n A_i(x)\exp\{j\phi_i(x)\}\delta(x-n\Delta) \\ &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \end{aligned} \tag{11}$$

where $A_i(x) = |e_i(x)|$, $\exp\{j\phi_i(x)\} = e_i(x)/|e_i(x)|$, and Δ is the sampling period. This data sampling is required to avoid the overlap between adjacent data in the reconstructed spatial data at the receiver. In Section 2.3, we show that the spatial data become wide at the receiver after the transmission of spatio-temporal converters. The encrypted data are Fourier-transformed by lens L6. The spatial Fourier transform of the encrypted data at P2 is given by

$$\tilde{R}_i(\eta) = \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}\exp\left(-j\frac{n\Delta\omega'}{cf}\eta\right) \tag{12}$$

where $\omega' = 2\pi c/\lambda'$, f is the focal length of lens L6, and λ' is the wavelength of the light beam used to write the hologram. This signal is recorded as a real-time hologram together with a reference plane wave in an intensity-sensitive medium, such as a multiple-quantum-well photorefractive device with sufficient spectral response. Here we assume that the hologram works as a grating with the transmittance of

$$t_i(\eta) = \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}\exp\left(-j\frac{n\Delta\omega'}{cf}\eta\right) \tag{13}$$

where we neglect the coefficient for normalization of transmittance and the effect of the carrier frequency of the hologram caused by the angle between the encrypted data and the reference beam. This hologram modulates the temporal signal described in Eq. (9). The optical field after the diffraction through the hologram is expressed by

$$\begin{aligned} \tilde{\psi}_3(\eta;\omega) &= \tilde{\psi}_2(\eta;\omega)t_i(\eta) \\ &= \tilde{W}\left\{\frac{\omega\eta}{2\pi cf} + \frac{\omega-\omega_0}{2\pi c}\alpha\right\} \times \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \\ &\quad \times \exp\left(-j\frac{n\Delta\omega'}{cf}\eta\right) \end{aligned} \tag{14}$$

This modulated field is then Fourier-transformed by lens L6:

$$\begin{aligned} \tilde{\psi}_4(X;\omega) &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \\ &\quad \times \exp\left\{j\frac{\omega-\omega_0}{c}\alpha\left(X + n\Delta\frac{\omega'}{\omega}\right)\right\} w\left(-X - n\Delta\frac{\omega'}{\omega}\right) \end{aligned} \tag{15}$$

This optical field is diffracted again by a grating and is given by

$$\begin{aligned} \tilde{\psi}_5(X';\omega) &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \\ &\quad \times \exp\left\{j\frac{\alpha n\Delta}{c}\frac{(\omega-\omega_0)\omega'}{\omega}\right\} w'\left(-X' - n\Delta\frac{\omega'}{\omega}\right) \end{aligned} \tag{16}$$

where X' is the coordinate as shown in Fig. 5, $w'(X')$ is the pupil function of the grating projected onto the X' coordinate. Equation (16) denotes the temporal frequency response of the system. Using Eqs. (7) and (16), we can obtain the output temporal signal by taking an inverse temporal Fourier transform of $\tilde{\psi}_5(X',\omega)\tilde{S}(\omega)$:

$$\begin{aligned} o_i(X',t) &= \int_{-\infty}^{\infty}\tilde{\psi}_5(X';\omega)\tilde{S}(\omega)\exp(-j\omega t)d\omega \\ &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}w'\left(-X' - n\Delta\frac{\omega'}{\omega_0}\right) \\ &\quad \times p(t - t_0 + n\delta t)\exp(j\omega_0 t) \end{aligned} \tag{17}$$

where $\delta t = (\alpha\Delta/c) \times (\omega'/\omega_0)$. To derive Eq. (17) we used the approximation, $1/\omega \approx 1/\omega_0$. This is valid in case of $\Delta\omega = (\omega - \omega_0) \ll \omega_0$ in a few hundreds femtosecond pulse.

Another light pulse, which is passing through P2 without diffraction due to the hologram, is also sent to remote users. This pulse has the same envelope as the input pulse; thus it can be written as

$$s_r(t) = p(t - t_0)\exp(j\omega_0 t) \tag{18}$$

This pulse is used as a reference pulse at the remote users to eliminate the phase distortion in the optical fiber. Both the temporally encrypted signal described in Eq. (17) and the reference pulse in Eq. (18) are sent to the receiver through a single mode fiber. Both pulses travel along the same line with a sufficient temporal interval by using a delay line before the fiber.

2.3. Receiver

At the receiver as shown in Fig. 6, the temporally encrypted data are converted again into spatially

encrypted data using a time-to-space converter. Because a single-mode fiber is used to send the temporally encrypted pulse and the reference pulse, the spatial information of both pulses should be dropped. The time-to-space converter consists of a pair of grating and Fourier transform lens. As shown in Fig. 6, the two light pluses are incident on the grating at an angle of θ and then are diffracted by the grating. These two pulses create the interference pattern after taking the spatial Fourier transform by lens L7. Here the temporal frequency response is considered. Using Eqs. (9), (17), and (18), the intensity distribution of the interference pattern at the Fourier plane P5 is described by

$$\begin{aligned}\tilde{I}_i(\eta,\omega) &= \left|\tilde{O}_i(\omega)\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right. \\ &\quad \left.+\,\tilde{S}_r(\omega)\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &= |\tilde{O}_i(\omega)|^2\times\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &\quad +|\tilde{S}_r(\omega)|^2\times\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &\quad +\tilde{O}_i*(\omega)\tilde{S}_r(\omega)\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &\quad +\tilde{O}_i(\omega)\tilde{S}_r*(\omega)\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \quad (19)\end{aligned}$$

where $W(\eta)$ is the spatial Fourier transform of $w(x)$ and $w(x)$ is a pupil function of the grating at P4. In Eq. (19)

$$\begin{aligned}\tilde{O}_i(\omega) &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}P(\omega-\omega_0) \\ &\quad\times\exp\{-j(\omega-\omega_0)(t_0-n\delta t)\} \quad (20)\end{aligned}$$

and

$$\tilde{S}_r(\omega)=\tilde{P}(\omega-\omega_0)\exp\{-j(\omega-\omega_0)t_0\} \quad (21)$$

Equations (20) and (21) denote the Fourier transforms of Eqs. (17) and (18), respectively.

We use the third term in Eq. (19) to reconstruct the spatially encrypted signal. A continuous-wave (CW) laser beam is incident on the hologram to read out the stored information. The reconstructed optical field is then spatially Fourier transformed by lens L7. When the pupil function of the grating and the beam width are large enough [$\tilde{W}(\eta)=\delta(\eta)$], the reconstructed signal at plane P6 is expressed by

$$\begin{aligned}\xi_i(x') &= F\left[\sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\}|P(\omega-\omega_0)|^2\right. \\ &\quad\left.\times\ \exp\{-j(\omega-\omega_0)n\delta t\}\right] \quad (22)\end{aligned}$$

Note that F[·] denotes the spatial Fourier transform. Equation (22) is calculated as follows:

$$\begin{aligned}\xi_i(x') &= \int_{-\infty}^{\infty}\sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\}\left|P\left(\frac{-\omega_0\eta}{f\alpha}\right)\right|^2 \\ &\quad\times\exp\left\{\frac{j\omega_0\eta n\delta t}{f\alpha}\right\}\exp\left\{\frac{-j2\pi x'\eta}{\lambda'' f}\right\}d\eta \\ &= \sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\} \\ &\quad\times\exp\left\{-\frac{\alpha^2\left(\frac{1}{\lambda''}x'+n\frac{\Delta}{\lambda'}\right)^2}{4\omega_0^2\tau^2}\right\} \quad (23)\end{aligned}$$

To derive Eq. (23), we use a Gaussian-shaped input pulse envelope written as $p(t)=\exp(-t^2/2\tau^2)$, τ is a pulse width, and λ'' is the wavelength of the CW laser beam. Equation (23) shows that each pixel becomes wide by convolving a Gaussian function with a $1/e^2$ width of $w_d=4\sqrt{2}\omega_0\tau\lambda''/\alpha$. This reconstructed signal is used in the following decryption system as shown in Fig. 7 to reconstruct the original data.

2.4. Decryption System

When the sampling interval, Δ, is larger than w_d in Eq. (23) and $\lambda''=\lambda'$, the reconstructed data do not overlap each other. We sample again Eq. (23) at $x=n\Delta$. The sampled data are given by

$$\begin{aligned}\xi_i'(x) &= \sum_n \xi_i(x)\delta(x-n\Delta) \\ &= \sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\} \\ &= \left[\sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}\right]^* \quad (24)\end{aligned}$$

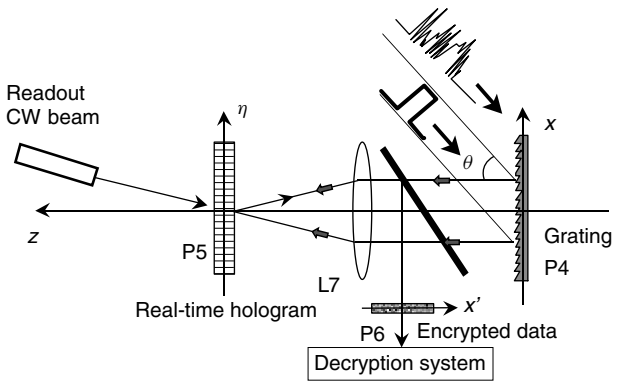


Figure 6. Receiver using time-to-space converter.

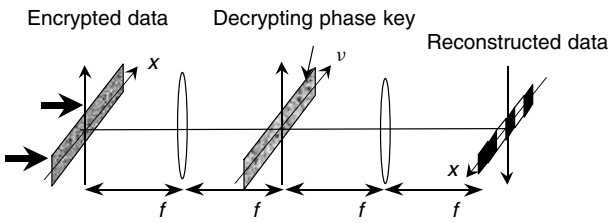


Figure 7. Decrypting system.

where the asterisk denotes complex conjugation. This equation shows that the complex conjugate of the encrypted signal in Eq. (11) is reconstructed. To decrypt the data, the spatial Fourier transform of Eq. (24) is multiplied by the phase key $\tilde{H}_i(\nu) = \exp\{-jh_i(\nu)\}$ as shown in Fig. 7. This phase key is the same random phase mask as that used in the encryption system. In the Fourier plane, the reconstructed data are written by

$$\tilde{\Psi}_i(\nu) = \left\{ \tilde{G}_i^*(\nu) \otimes F[\exp\{-jn_i(x)\}] * \tilde{H}_i^*(\nu) \otimes \exp\left(j\frac{2\pi}{\lambda f} n\Delta\nu\right) \right\} \tilde{H}_i(\nu) \tag{25}$$

where ν denotes the coordinate in the Fourier plane, and $\otimes$ denotes convolution. Finally the reconstructed data are obtained by taking another Fourier transform of Eq. (25):

$$I_{\text{out}}(x) = \sum_n g_i(x) \exp\{-jn_i(x)\} \otimes F * [\exp\{-jh_i(\nu)\}] \cdot \delta(x - n\Delta) \otimes F[\exp\{-jh_i(\nu)\}] \tag{26}$$

This equation shows that the reconstructed data are not exactly the same as the original data, $g_i(x)$ due to the spatial sampling of the encrypted data. In the following section, we numerically evaluate the reconstructed data.

3. NUMERICAL EVALUATIONS

We numerically evaluate the error between the original data and the reconstructed data by use of Eq. (26). When $\omega_0 = 6\pi \times 10^{14}$, $\lambda'' = 1\ \mu\text{m}$, $\alpha = 1/\sqrt{2}$, and $\tau = 50$ fs, the sampling interval, sufficient to avoid the overlap, is 754 μm. If the sampling interval, Δ is smaller than the width of the Gaussian distribution, w_d, the reconstructed spatial data overlap each other. Thus, the original data cannot be reconstructed when the overlap is large, even if we use the correct phase key in the decryption process. To avoid the overlap at the receiver, we have to use a large sampling interval at the transmitter. The sampling results in some loss of the encrypted data and leads to the error in the reconstructed data. In the following calculations, we evaluate the bit error rate in a binary data transmission using undersampled data.

When the encrypted data are undersampled by a factor of 2, half of the encrypted data are lost. This loss of information causes the error in the decrypted data. Using binary data, however, it is possible to reduce the noise in the decrypted data by thresholding. A mean squared error is used as the performance criterion

$$e = E\{|g(x) - m \times g'_\Delta(x)|^2\} \tag{27}$$

where $E\{\cdot\}$ denotes statistical average, $g(x)$ is the original data, m denotes the coefficient to compensate for the loss of total power due to undersampling the encrypted data, and $g'_\Delta(x)$ is the reconstructed data using Eq. (26) when the encrypted data are sampled at the interval of Δ.

The original binary data are 32-bit data, where each bit consists of 64 pixels. Thus the input data have 2048 pixels. This redundancy of the original data is introduced to recover the original binary data. Two random phase masks in the input and Fourier planes consist of 2048 pixels. The original binary data and the two random-phase masks are randomly generated in a computer. The average mean squared error was calculated over 1000 different trials. An example of the original data, encrypted data, and decrypted data is shown in Fig. 8. Here the decrypted data are lowpass-filtered. The lowpass filtering was performed by locally averaging the data with an 11-pixel window. We can see that the encrypted data are random, but the decrypted data have the same structure of the original data. This means that it is possible to recover the original binary data without bit error.

We calculated the bit error rate as a function of the sampling interval, Δ. The bit error rate is defined as a ratio of the number of error bits to the total number of bits. By using the reconstructed analog data as shown in Fig. 8, we calculated the energy of each cell of the reconstructed data. After the calculation of energy, each datum is binarized. To determine the threshold value, all energies are rank ordered. The threshold value is the Nth largest energy of the cells, where N is the number of bright pixels(ones) in the original binary data. Figure 9 shows the bit error rate as a function of the sampling interval Δ. For the binary data used here, the original binary data were reconstructed without error in the case of $\Delta = 1$ and 2. When the sampling interval becomes large, the loss of encrypted data causes a large number of bit errors.

4. CONCLUSION

We have presented a secure data communication system that links remote users to a secure holographic memory system at ultrafast transfer rate. In the encrypted holographic memory each data frame is recorded as a hologram after encrypting the original data by using double-random-phase encryption. The recorded data can be readout in parallel at a fast rate. The spatio-temporal converters based on ultrashort pulse shaper enable us to send the encrypted data at ultrahigh speed via optical fibers. At the remote users, only an authorized user can reconstruct the original data by using the correct keys. We expect that the system can provide a high level of security and ultrafast data communication.

BIOGRAPHIES

Bahram Javidi, distinguished Professor of Electrical and Computer Engineering at University of Connecticut, received the B.S. degree (1980) in electrical engineering from George Washington University, Washington, D.C., and the M.S. (1982) and Ph.D. (1986) degrees in electrical engineering from the Pennsylvania State University, University Park. He is fellow of Institute Of Electrical and Electronics Engineers, fellow of the Optical Society of America, and fellow of the International Society for Optical Engineering (SPIE). In 1990, he was named a Presidential Young Investigator by the National Science Foundation. He has been awarded the University of

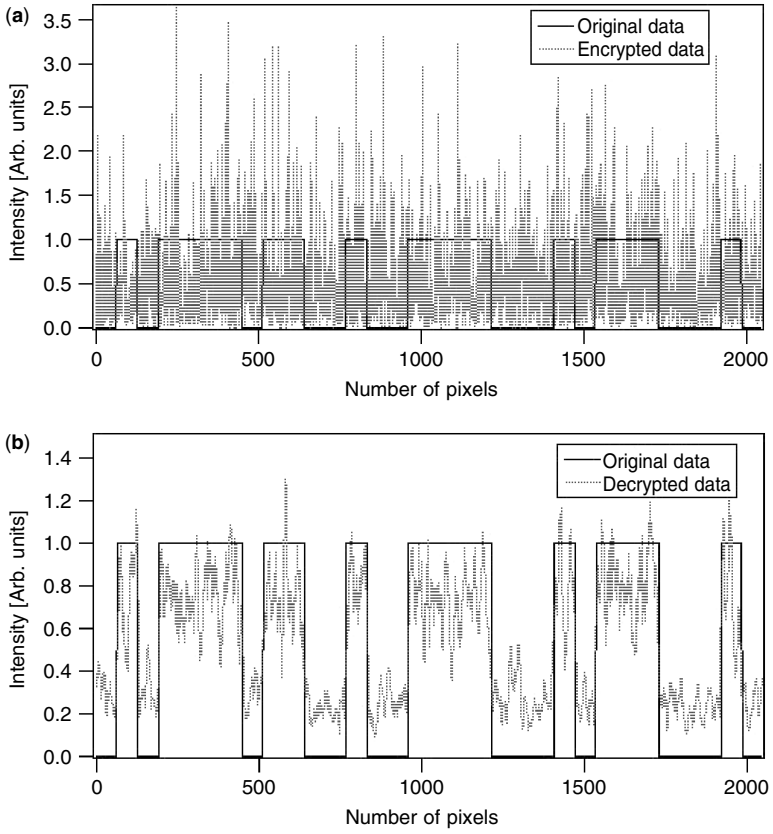


Figure 8. An example of original, encrypted, and decrypted data: (**a**) original and encrypted data; (**b**) original and decrypted data.

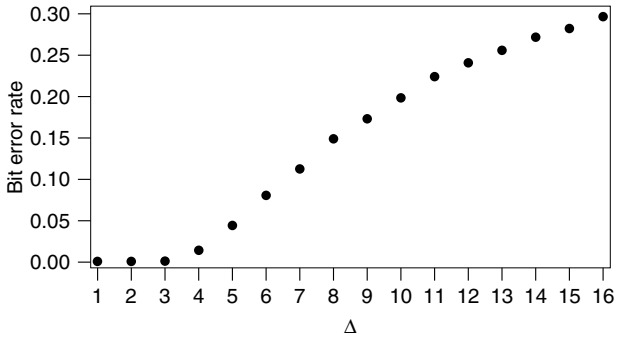


Figure 9. Bit error rate as a function of sampling period, Δ.

Connecticut Alumni Association Research Excellence Award, and the first Electrical and Computer Engineering Department Outstanding Research Award. He is the editor of several books including, "Image Recognition: Algorithms, Systems, and Applications," published by Marcel-Dekker, New York, 2002; "Three Dimensional Television, Video, and Display Technologies," Springer Verlag in 2002; "Smart Imaging Systems," SPIE Press in 2001; and "Real-time Optical Information Processing," Academic Press, 1994. In addition, he has published over 260 technical articles in major journals and conference proceedings, including over 40 invited papers. He has served on the editorial boards for Springer-Verlag, Marcel Dekker, the Optical Engineering Journal, and IEEE/SPIE Press.

His email address is bahram.javidi@uconn.edu

Osamu Matoba received the B.Eng., M.Eng., and D.Eng. degrees in applied physics from Osaka University, Osaka, Japan, in 1991, 1993, and 1996, respectively. Since 1996, he has been a research associate at the Institute of Industrial Science, University of Tokyo, Japan. His current research interests include optical information processing, optical security, three-dimensional display, digital holography, and the development of photorefractive materials. Dr. Matoba is a member of the Japanese Society of Applied Physics, the Optical Society of Japan, the Laser Society of Japan, OSA, SPIE, and IEEE LEOS.

BIBLIOGRAPHY

1. A. M. Weiner, D. E. Leaird, D. H. Reitze, and E. G. Paek, Femtosecond spectral holography, *IEEE J. Quant. Electron.* **QE-28**: 2251–2261 (1992).
2. Y. T. Mazurenko, Holography of wave packets, *Appl. Phys. B* **50**: 101–114 (1990).
3. A. W. Weiner and A. M. Kan'an, Femtosecond pulse shaping for synthesis, processing, and time-to-space conversion of ultrafast optical waveforms, *IEEE J. Select. Topics Quant. Electron.* **4**: 317–331 (1998).
4. Y. Ding, D. D. Nolte, M. R. Melloch, and A. W. Weiner, Time-domain image processing using dynamic holography, *IEEE J. Select. Topics Quant. Electron.* **4**: 332–341 (1998).
5. P. C. Sun et al., All-optical parallel-to-serial conversion by holographic spatial-to-temporal frequency encoding, *Opt. Lett.* **20**: 1728–1730 (1995).

6. D. M. Marom, P. C. Sun, and Y. Fainman, Analysis of spatial-temporal converters for all-optical communication links, *Appl. Opt.* **37**: 2858–2868 (1998).
7. O. Matoba and B. Javidi, Secure ultrafast communication with spatial-temporal converters, *Appl. Opt.* **39**: 2975–2981 (2000).
8. J. F. Heanue, M. C. Bashaw, and L. Hesselink, Volume holographic storage and retrieval of digital data, *Science* **265**: 749–752 (1994).
9. H. J. Coufal, D. Psaltis, and G. T. Sincerbox, eds., *Holographic Data Storage*, Springer, New York, 2000.
10. P. Rèfrègier and B. Javidi, Optical image encryption based on input plane and Fourier plane random encoding, *Opt. Lett.* **20**: 767–769 (1995).
11. B. Javidi, Encrypting information with optical technologies, *Phys. Today* **50**(3): (March 1997).
12. J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, New York, 2000.
13. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
14. B. Javidi, ed., *Smart Imaging Systems*, SPIE Press, Bellingham, WA, 2001.
15. B. Javidi and J. L. Horner, eds., *Real-Time Optical Information Processing*, Academic Press, Boston, 1994.
16. O. Matoba and B. Javidi, Encrypted optical memory system using three-dimensional keys in the Fresnel domain, *Opt. Lett.* **24**: 762–764 (1999).
17. O. Matoba and B. Javidi, Encrypted optical storage with wavelength-key and random phase codes, *Appl. Opt.* **38**: 6785–6790 (1999).
18. O. Matoba and B. Javidi, Encrypted optical storage with angular multiplexing, *Appl. Opt.* **38**: 7288–7293 (1999).
19. O. Matoba and B. Javidi, Encrypted optical memory systems based on multidimensional keys for secure data storage and communications, *IEEE Circuits Devices Mag.* **16**: 8–15 (Sept. 2000).

SEQUENTIAL DECODING OF CONVOLUTIONAL CODES

YUNGHSIANG S. HAN
National Chi Yan University
Taiwan Republic of China

PO-NING CHEN
National Chi Tung University
Taiwan Republic of China

1. INTRODUCTION

The convolutional coding technique is designed to reduce the probability of erroneous transmission over noisy communication channels. The most popular decoding algorithm for convolutional codes is perhaps the Viterbi algorithm. Although widely adopted in practice, the Viterbi algorithm suffers from a high decoding complexity for convolutional codes with long constraint lengths. While the attainable decoding failure probability of convolutional codes generally decays exponentially with the code constraint length, the high complexity of the Viterbi decoder for codes with a long constraint length to some extent limits the achievable system performance. Nowadays, the Viterbi algorithm is usually applied to codes with a constraint length no greater than nine.

In contrast to the limitation of the Viterbi algorithm, sequential decoding is renowned for its computational complexity being independent of the code constraint length [1]. Although simply suboptimal in its performance, sequential decoding can achieve a desired bit error probability when a sufficiently large constraint length is taken for the convolutional code. Unlike the Viterbi algorithm that locates the best codeword by exhausting all possibilities, sequential decoding concentrates only on a certain number of likely codewords. As the sequential selection of these likely codewords is affected by the channel noise, the decoding complexity of a sequential decoder becomes dependent on the noise level [1]. These specific characteristics make sequential decoding useful in particular applications.

Sequential decoding was first introduced by Wozencraft for the decoding of convolutional codes [2,3]. Thereafter, Fano developed the sequential decoding algorithm with a milestone improvement in decoding efficiency [4]. Fano's work subsequently inspired further research on sequential decoding. Later, Zigangirov [5], and independently, Jelinek [6], proposed the *stack algorithm*.

In this article, the sequential decoding will not be introduced chronologically. Rather, Algorithm A [7] — the general sequential search algorithm — will be introduced first because it is conceptually more straightforward. The rest of the article is organized as follows. Sections 2 and 3 provide the necessary background for convolutional codes and typical channel models for performance evaluation. Section 4 introduces Algorithm A, and then defines the general features of sequential decoding. Section 5 explores the Fano metric and its generalization for use to guide the search of sequential decoding. Section 6 presents the stack algorithm and its variants. Section 7 elucidates the well-known Fano algorithm. Section 8 is devoted to the trellis variants of sequential decoding, especially on the proposed maximum-likelihood sequential decoding algorithm (MLSDA). Section 9 examines the decoding performance. Section 10 discusses various practical implementation issues regarding sequential decoding, such as buffer overflow. For completeness, a section on the code construction (Section 11) is included at the end of the article. Section 12 concludes the article.

For clarity, only binary convolutional codes are considered throughout discussions on sequential decoding. Extension to nonbinary convolutional codes can be carried out similarly.

2. CONVOLUTIONAL CODE AND ITS GRAPHICAL REPRESENTATION

Denote a binary convolutional code by a 3-tuple (n, k, m), which corresponds to an encoder for which n output bits are generated whenever k input bits are received, and for which the current n outputs are linear combinations of the present k input bits and the previous $m \times k$ input

bits. Because m designates the number of previous k-bit input blocks that must be memorized in the encoder, m is called the *memory order* of the convolutional code. A binary convolutional encoder is conveniently structured as a mechanism of shift registers and modulo-2 adders, where the output bits are modulo-2 additions of selective shift register contents and present input bits. Then n in the 3-tuple notation is exactly the number of output sequences in the encoder, k is the number of input sequences (and hence, the encoder consists of k shift registers), and m is the maximum length of the k shift registers (i.e., if the number of stages of the jth shift register is K_j, then $m = \max_{1 \le j \le k} K_j$). Figures 1 and 2 exemplify the encoders of binary $(2, 1, 2)$ and $(3, 2, 2)$ convolutional codes, respectively.

During the encoding process, the contents of shift registers in the encoder are initially set to zero. The k input bits from the k input sequences are then fed into the encoder in parallel, generating n output bits according to the shift register framework. To reset the shift register contents at the end of input sequences so that the encoder can be ready for use for another set of input sequences, m zeros are usually padded at the end of each input sequence. Consequently, each k input sequence of length L bits is padded with m zeros, and these k input sequences jointly induce $n(L + m)$ output bits. As illustrated in Fig. 1, the encoder of the $(2, 1, 2)$ convolutional code extracts two output sequences, $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, v_{1,2}, \ldots, v_{1,6}) = (1010011)$ and $\mathbf{v}_2 = (v_{2,0}, v_{2,1}, v_{2,2}, \ldots, v_{2,6}) = (1101001)$, due to the single input sequence $\mathbf{u} = (u_0, u_1, u_2, u_3, u_4) = (11101)$, where u_0 is fed in the encoder first. The encoder then interleaves $\mathbf{v}_1$ and $\mathbf{v}_2$ to yield

$$\begin{aligned}\mathbf{v} &= (v_{1,0}, v_{2,0}, v_{1,1}, v_{2,1}, \ldots, v_{1,6}, v_{2,6}) \\ &= (11\ 01\ 10\ 01\ 00\ 10\ 11)\end{aligned}$$

of which the length is $2(5 + 2) = 14$. Also, the encoder of the $(3, 2, 2)$ convolutional code in Fig. 2 generates the output sequences of $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, v_{1,2}, v_{1,3}) = (1000)$, $\mathbf{v}_2 = (v_{2,0}, v_{2,1}, v_{2,2}, v_{2,3}) = (1100)$, and $\mathbf{v}_3 = (v_{3,0}, v_{3,1}, v_{3,2}, v_{3,3}) = (0001)$, due to the two input sequences $\mathbf{u}_1 = (u_{1,0}, u_{1,1}) = (10)$ and $\mathbf{u}_2 = (u_{2,0}, u_{2,1}) = (11)$, which in turn generate the interleaved output sequence

$$\begin{aligned}\mathbf{v} &= (v_{1,0}, v_{2,0}, v_{3,0}, v_{1,1}, v_{2,1}, v_{3,1}, v_{1,2}, v_{2,2}, v_{3,2}, v_{1,3}, v_{2,3}, v_{3,3}) \\ &= (110\ 010\ 000\ 001)\end{aligned}$$

of length $3(2 + 2) = 12$. Terminologically, the interleaved output $\mathbf{v}$ is called the *convolutional codeword* corresponding to the combined input sequence $\mathbf{u}$.

An important subclass of convolutional codes is the *systematic codes*, in which k out of n output sequences retain the values of the k input sequences. In other words, these outputs are directly connected to the k inputs in the encoder.

A convolutional code encoder can also be viewed as a linear system, in which the relation between its inputs and outputs is characterized by generator polynomials. For example, $g_1(x) = 1 + x + x^2$ and $g_2(x) = 1 + x^2$ can be used to identify $\mathbf{v}_1$ and $\mathbf{v}_2$ induced by $\mathbf{u}$ in Fig. 1, where the appearance of x^i indicates that a physical connection is applied to the $(i + 1)$th dot position, counted from the left. Specifically, putting $\mathbf{u}$ and $\mathbf{v}_i$ in polynomial form as $\mathbf{u}(x) = u_0 + u_1 x + u_2 x^2 + \cdots$ and $\mathbf{v}_i(x) = v_{i,0} + v_{i,1} x + v_{i,2} x^2 + \cdots$ yields that $\mathbf{v}_i(x) = \mathbf{u}(x) g_i(x)$ for $i = 1, 2$, where addition of coefficients is based on modulo-2 operation. With reference to the encoder depicted in Fig. 2, the relation between the input sequences and the output sequences can be formulated through matrix operation as

$$[\mathbf{v}_1(x)\ \mathbf{v}_2(x)\ \mathbf{v}_3(x)] = [\mathbf{u}_1(x)\ \mathbf{u}_2(x)] \times \begin{bmatrix} g_1^{(1)}(x) & g_2^{(1)}(x) & g_3^{(1)}(x) \\ g_1^{(2)}(x) & g_2^{(2)}(x) & g_3^{(2)}(x) \end{bmatrix}$$

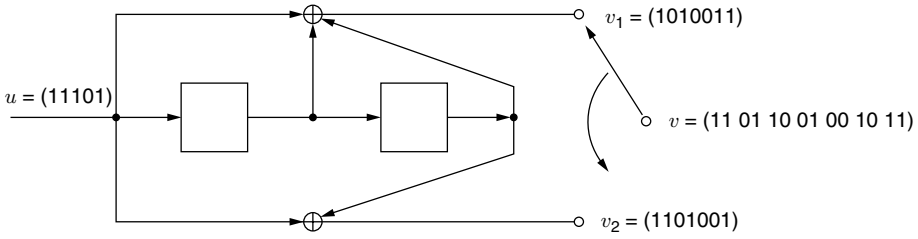


Figure 1. Encoder for the binary $(2, 1, 2)$ convolutional code with generators $g_1 = 7$ (octal) and $g_2 = 5$ (octal), where g_i is the generator polynomial characterizing the ith output.

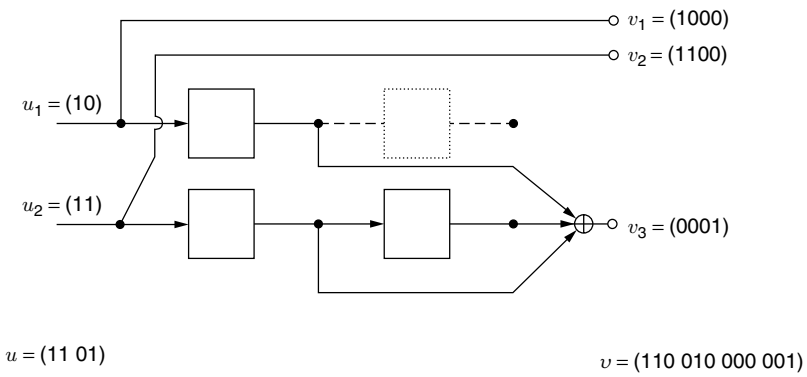


Figure 2. Encoder for the binary $(3, 2, 2)$ systematic convolutional code with generators $g_1^{(1)} = 4$ (octal), $g_1^{(2)} = 0$ (octal), $g_2^{(1)} = 0$ (octal), $g_2^{(2)} = 4$ (octal), $g_3^{(1)} = 2$ (octal), and $g_3^{(2)} = 3$ (octal), where $g_i^{(j)}$ is the generator polynomial characterizing the ith output according to the jth input. The dashed box is redundant and can actually be removed from this encoder; its presence here is only to help demonstrating the derivation of generator polynomials. Thus, as far as the number of stages of the jth shift register is concerned, $K_1 = 1$ and $K_2 = 2$.

where $\mathbf{u}_j(x) = u_{j,0} + u_{j,1}x + u_{j,2}x^2 + \cdots$ and $\mathbf{v}_i(x) = v_{i,0} + v_{i,1}x + v_{i,2}x^2 + \cdots$ define the jth input sequence and the ith output sequence, respectively, and the generator polynomial $g_i^{(j)}(x)$ characterizes the relation between the jth input and the ith output sequences. For simplicity, generator polynomials are sometimes abbreviated by their coefficients in octal number format, led by the least significant one. Continuing the example in Fig. 1 gives $g_1 = 111$ (binary) $= 7$ (octal) and $g_2 = 101$ (binary) $= 5$ (octal). A similar abbreviation can be used for each $g_i^{(j)}$ in Fig. 2.

An (n, k, m) convolutional code can be transformed to an equivalent linear block code with *effective code rate*[1] $R_{\text{effective}} = kL/[n(L+m)]$, where L is the length of the information input sequences. By taking L to infinity, the effective code rate converges to $R = k/n$, which is referred to as the *code rate* of the (n, k, m) convolutional code.

The *constraint length* of an (n, k, m) convolutional code has two different definitions in the literature: $n_A = m + 1$ [8] and $n_A = n(m + 1)$ [1]. In this article, the former definition is adopted, because it is more extensively used in military and industrial publications.

Let $\mathbf{v}_{(a,b)} = (v_a, v_{a+1}, \ldots, v_b)$ denote a portion of codeword $\mathbf{v}$, and abbreviate $\mathbf{v}_{(0,b)}$ by $\mathbf{v}_{(b)}$. The Hamming distance between the first rn bits of codewords $\mathbf{v}$ and $\mathbf{z}$ is given by

$$d_H(\mathbf{v}_{(rn-1)}, \mathbf{z}_{(rn-1)}) = \sum_{i=0}^{rn-1} v_i \oplus z_i$$

where "$\oplus$" denotes modulo-2 addition. The Hamming weight of the first rn bits of codeword $\mathbf{v}$ thus equals $d_H(\mathbf{v}_{(rn-1)}, \mathbf{0}_{(rn-1)})$, where $\mathbf{0}$ represents the all-zero codeword. The *column distance function* (CDF) $d_c(r)$ of a binary (n, k, m) convolutional code is defined as the minimum Hamming distance between the first rn bits of any two codewords whose first n bits are distinct

$$d_c(r) = \min\{d_H(\mathbf{v}_{(rn-1)}, \mathbf{z}_{(rn-1)}) : \mathbf{v}_{(n-1)} \neq \mathbf{z}_{(n-1)} \text{ for } \mathbf{v}, \mathbf{z} \in \mathcal{C}\}$$

where $\mathcal{C}$ is the set of all codewords. Function $d_c(r)$ is clearly nondecreasing in r. Two cases of CDFs are of specific interest: $r = m + 1$ and $r = \infty$. In the latter case, the input sequences are considered infinite in length.[2] Terminologically, $d_c(m + 1)$ and $d_c(\infty)$ (or d_{free} in general) are called the *minimum distance* and the *free distance* of the convolutional code, respectively.

The operational meanings of the minimum distance, the free distance and the CDF of a convolutional code are as follows. When a sufficiently large codeword length is taken, and an optimal (i.e., maximum-likelihood) decoder is employed, the error-correcting capability of a convolutional code [9] is generally characterized by d_{free}. In case a decoder figures the transmitted bits only on the basis of the first $n(m + 1)$ received bits (as in, e.g., the majority-logic decoding [10]), $d_c(m + 1)$ can be used instead to characterize the code error-correcting capability. As for sequential decoding algorithm that requires a rapid initial growth of column distance functions (to be discussed in Section 9), the decoding computational complexity, defined as the number of metric computations performed, is determined by the CDF of the code being applied.

Next, two graphical representations of convolutional codewords are introduced. They are derived from the graphs of *code tree* and *trellis*, respectively. A *code tree* of a binary (n, k, m) convolutional code presents every codeword as a path on a tree. For input sequences of length L bits, the code tree consists of $(L + m + 1)$ levels. The single leftmost node at level 0 is called the *origin node*. At the first L levels, there are exactly 2^k branches leaving each node. For those nodes located at levels L through $(L + m)$, only one branch remains. The 2^{kL} rightmost nodes at level $(L + m)$ are called the *terminal nodes*. As expected, a path from the single origin node to a terminal node represents a codeword; therefore, it is named the *code path* corresponding to the codeword. Figure 3 illustrates the code tree for the encoder in Fig. 1 with a single input sequence of length 5.

In contrast to a code tree, a code *trellis* as described by Forney [11] is a structure obtained from a code tree by merging those nodes in the same *state*. The *state* associated with a node is determined by the associated shift register contents. For a binary (n, k, m) convolutional code, the number of states at levels m through L is 2^K, where $K = \sum_{j=1}^{k} K_j$ and K_j is the length of the jth shift register in the encoder; hence, there are 2^K nodes on these levels. Due to node merging, only one terminal node remains in a trellis. Analogous to a code tree, a path from the single origin node to the single terminal node in a trellis also mirrors a codeword. Figure 4 exemplifies the trellis of the convolutional code presented in Fig. 1.

3. TYPICAL CHANNEL MODELS FOR CODING SYSTEMS

When the $n(L + m)$ convolutional code bits encoded from kL input bits are modulated into respective waveforms (or signals) for transmission over a medium that introduces attenuation, distortion, interference, noise, and other parameters, the received waveforms become "uncertain" in their shapes. A "guess" of the original information sequences therefore has to be made at the receiver end. The "guess" mechanism can be conceptually divided into two parts: demodulator and decoder.

The demodulator transforms the received waveforms into discrete signals for use by the decoder to determine the original information sequences. If the discrete demodulated signal is of two values (i.e., binary), then the demodulator is termed a *hard-decision demodulator*. If the demodulator passes analog (i.e., discrete-in-time but continuous-in-value) or quantized outputs to the decoder, then it is classified as a *soft-decision demodulator*.

[1] The effective code rate is defined as the average number of input bits carried by an output bit [1].

[2] Usually, $d_c(r)$ for an (n, k, m) convolutional code reaches its largest value $d_c(\infty)$ when r is a little beyond $5 \times m$; this property facilitates the determination of $d_c(\infty)$.

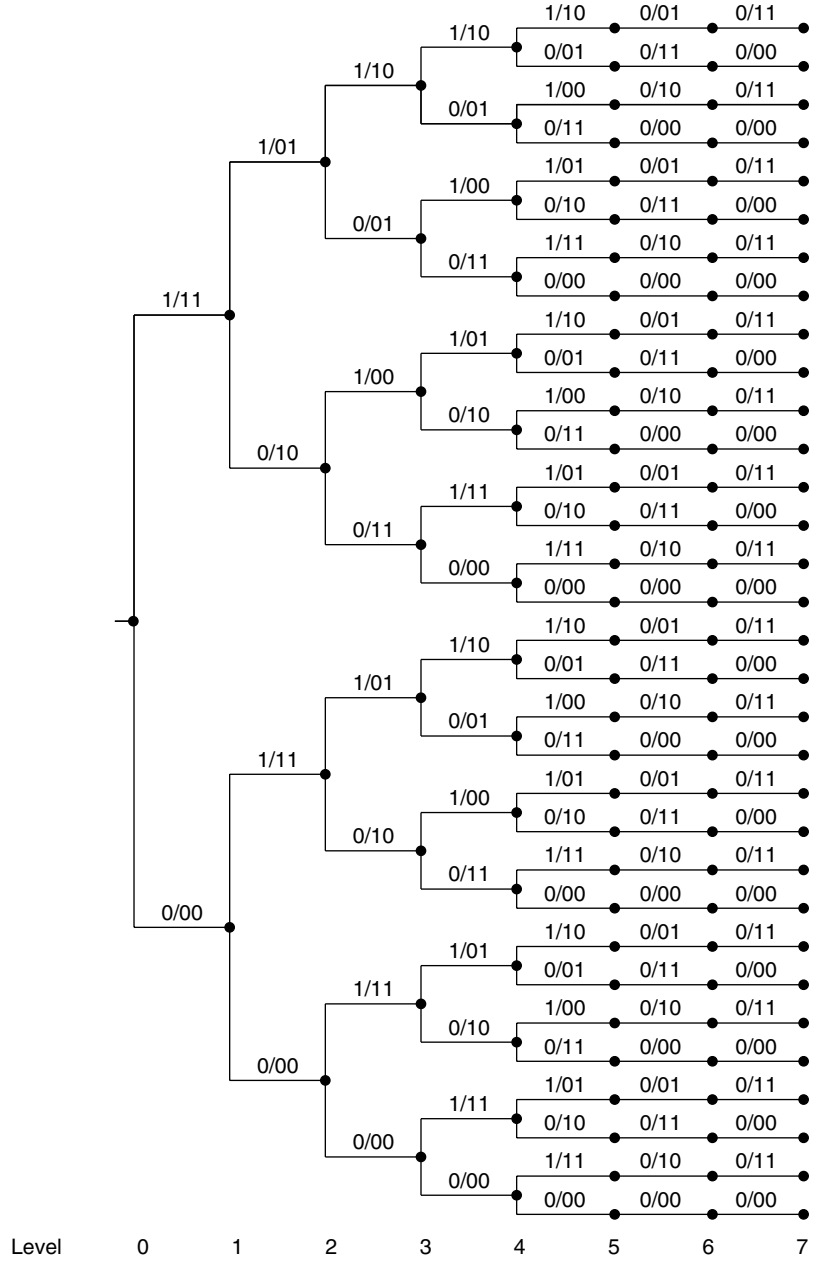


Figure 3. Code tree for the binary $(2, 1, 2)$ convolutional code in Fig. 1 with a single input sequence of length 5. Each branch is labeled by its respective "input bit/output code bits." The code path indicated by the thick line is labeled in sequence by code bits 11, 01, 10, 01, 00, 10 and 11, and its corresponding codeword is $\mathbf{v} = (11\ 01\ 10\ 01\ 00\ 10\ 11)$.

The decoder, on the other hand, estimates the original information sequences on the basis of the $n(L + m)$ demodulator outputs, or equivalently a received vector of $n(L + m)$ dimensions, according to some criterion. One frequently applied criterion is the *maximum-likelihood decoding* (MLD) rule, under which the probability of codeword estimate error is minimized subject to an equiprobable prior on the transmitted codewords. Terminologically, if a soft-decision demodulator is employed, then the subsequent decoder is classified as a *soft-decision decoder*. In a situation in which the decoder receives inputs from a hard-decision demodulator, the decoder is called a *hard-decision decoder* instead.

Perhaps, because of their analytic feasibility, two types of statistics concerning demodulator outputs are of general interest. They are respectively induced from the *binary symmetric channel* (BSC) and the *additive white Gaussian noise* (AWGN) *channel*. The former is a typical channel model for the performance evaluation of hard-decision decoders, while the latter is widely used in examining the error rate of soft-decision decoders. They are introduced after the concept of a channel is elucidated.

For a coding system, a *channel* is a signal passage that mixes all the intermediate effects onto the signal, including modulation, upconversion, medium, downconversion, and demodulation. The demodulator incorporates these aggregated channel effects into a widely adopted additive channel model as $r = s + n$, in which r is the demodulator output, s is the transmitted signal that is a function of encoder outputs, and n represents the aggregated signal distortion, simply termed *noise*. Its extension to multiple

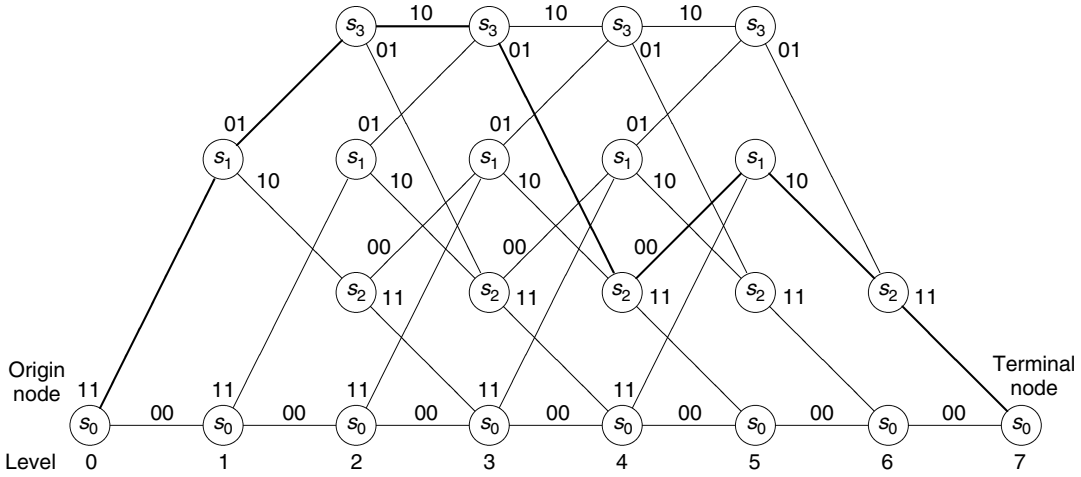


Figure 4. Trellis for the binary $(2, 1, 2)$ convolutional code in Fig. 1 with a single input sequence of length 5. States S_0, S_1, S_2, and S_3 correspond to the states of shift register contents that are 00, 01, 10, and 11 (from right to left in Fig. 1), respectively. The code path indicated by the thick line is labeled in sequence by code bits 11, 01, 10, 01, 00, 10 and 11, and its corresponding codeword is $\mathbf{v} = (11\ 01\ 10\ 01\ 00\ 10\ 11)$.

independent channel usages is given by

$$r_j = s_j + n_j \text{ for } 0 \le j \le N-1$$

which is often referred to as the *time-discrete channel*, since the *time* index j ranges over a *discrete* integer set. For simplicity, independence with common marginal distribution among noise samples $n_0, n_1, \ldots, n_{N-1}$ is often assumed, which is specifically termed *memoryless*. In situation where the power spectrum (i.e., the Fourier transform of the noise autocorrelation function) of the noise samples is a constant, which can be interpreted as the noise contributing equal power at all frequencies and thereby imitating the composition of a white light, the noise is dubbed *white*.

Hence, for a time-discrete coding system, the AWGN channel specifically indicates a memoryless noise sequence with a Gaussian distributed marginal, in which case the demodulator outputs $r_0, r_1, \ldots, r_{N-1}$ are independent and Gaussian distributed with equal variances and means $s_0, s_1, \ldots, s_{N-1}$, respectively. The noise variance exactly equals the constant spectrum value $N_0/2$ of the white noise, where N_0 is *the single-sided noise power per hertz* or $N_0/2$ is *the doubled-sided noise power per hertz*. The means that $s_0, s_1, \ldots, s_{N-1}$ are apparently decided by the choice of mappings from the encoder outputs to the channel inputs. For example, assuming an *antipodal* mapping gives $s_j(c_j) = (-1)^{c_j}\sqrt{E}$, where $c_j \in \{0, 1\}$ is the jth code bit. Under an implicit premise of equal possibilities for $c_j = 0$ and $c_j = 1$, the second moment of s_j is given by

$$E[s_j^2] = \tfrac{1}{2}(\sqrt{E})^2 + \tfrac{1}{2}(-\sqrt{E})^2 = E$$

which is commonly taken to be the average signal energy required for its transmission.

A conventional measure of the noisiness of AWGN channels is the *signal-to-noise ratio* (SNR). For the time-discrete system considered, it is defined as the average signal energy E (the second moment of the transmitted signal) divided by N_0 (the single-sided noise power per hertz). Notably, the SNR ratio is invariable with respect to scaling of the demodulator output; hence, this noisiness index is consistent with the observation that the optimal error rate of guessing c_j based on the knowledge of $(\lambda \cdot r_j)$ through equation

$$\lambda \cdot r_j = \lambda \cdot (-1)^{c_j}\sqrt{E} + \lambda \cdot n_j$$

is indeed independent of the scaling factor λ whenever $\lambda > 0$. Accordingly, the performance of the soft-decision decoding algorithms under AWGN channels is typically given by plotting its error rate against the SNR.[3]

The channel model can be further simplified to that for which the noise sample n and the transmitted signal s (usually the code bit itself in this case) are both elements of $\{0, 1\}$, and their modulo-2 addition yields the hard-decision demodulation output r. Then a binary input/binary output channel between convolutional encoder and decoder is observed. The channel statistics can be defined using the two crossover probabilities: $\Pr(r = 1 \mid s = 0)$ and $\Pr(r = 0 \mid s = 1)$. If the two crossover probabilities are equal, then the binary channel is *symmetric*, and is therefore called the *binary symmetric channel*.[4]

[3] The SNR per information bit, denoted as E_b/N_0, is often used instead of E/N_0 in picturing the code performance in order to account for the code redundancy for different code rates. Their relation can be characterized as $E_b/N_0 = (E/N_0)/R_{\text{effective}} = (E/N_0) \times n(L+m)/(kL)$ because the overall energy of kL uncoded input bits equals that of $n(L+m)$ code bits in principle.

[4] The BSC can be treated as a quantized simplification of the AWGN channel. Hence, the crossover probability p can be derived from $r_j = (-1)^{c_j}\sqrt{E} + n_j$ as $p = (1/2)\text{erfc}(\sqrt{E/N_0})$, where $\text{erfc}(x) = (2/\sqrt{\pi})\int_x^\infty \exp\{-x^2\}\,dx$ is the complementary error function. This convention is adopted here in presenting the performance figures for BSCs.

The correspondence between the transmitted signal s_j and the code bit c_j is often isomorphic. If this is the case, $\Pr(r_j \mid c_j)$ and $\Pr(r_j \mid s_j)$ can be used interchangeably to represent the channel statistics of receiving r_j given that c_j or $s_j = s_j(c_j)$ is transmitted. For convenience, $\Pr(r_j \mid v_j)$ will be used at the decoder end to denote the same probability as $\Pr(r_j \mid c_j)$, where $c_j = v_j$, throughout the article.

4. GENERAL DESCRIPTION OF SEQUENTIAL DECODING ALGORITHM

Following the background introduction to the time-discrete coding system, the optimal criterion that motivates the decoding approaches can now be examined. As a consequence of minimizing the codeword estimate error subject to an equiprobable codeword prior, the MLD rule, on receipt of a received vector $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$, outputs the codeword $\mathbf{c}^* = (c_0^*, c_1^*, \ldots, c_{N-1}^*)$ satisfying

$$\Pr(\mathbf{r} \mid \mathbf{c}^*) \geq \Pr(\mathbf{r} \mid \mathbf{c}) \text{ for all } \mathbf{c} = (c_0, c_1, \ldots, c_{N-1}) \in \mathcal{C}$$

where $\mathcal{C}$ is the set of all possible codewords, and $N = n(L + m)$. When the channel is memoryless, the MLD rule can be reduced to

$$\prod_{j=0}^{N-1} \Pr(r_j \mid c_j^*) \geq \prod_{j=0}^{N-1} \Pr(r_j \mid c_j) \quad \text{for all} \quad \mathbf{c} \in \mathcal{C}$$

which in turn is equivalent to

$$\sum_{j=0}^{N-1} \log_2 \Pr(r_j \mid c_j^*) \geq \sum_{j=0}^{N-1} \log_2 \Pr(r_j \mid c_j) \quad \text{for all} \quad \mathbf{c} \in \mathcal{C} \quad (1)$$

A natural implication of (1) is that by simply letting $\sum_{j=n(\ell-1)}^{n\ell-1} \log_2 \Pr(r_j \mid c_j)$ be the metric associated with a branch labeled by $(c_{n(\ell-1)}, \ldots, c_{n\ell-1})$, the MLD rule becomes a search of the code path with the maximum metric, where the metric of a path is defined as the sum of the individual metrics of the branches of which the path consists. Any suitable graph search algorithm can then be used to perform the search process.

Of the graph search algorithms in artificial intelligence, Algorithm A is one that performs priority-first (or metric-first) searching over a graph [7,12]. In applying the algorithm to the decoding of convolutional codes, the graph G undertaken becomes either a code tree or a trellis. For a graph over which Algorithm A searches, a link between the origin node and any node, either directly connected or indirectly connected through some intermediate nodes, is called a *path*. Suppose that a real-valued function $f(\cdot)$, often referred to as the *evaluation function*, is defined for every path in the graph G. Then Algorithm A can be described as follows:

Algorithm A

Step 1. Compute the associated f-function value of the single-node path that contains only the origin node. Insert the single-node path with its associated f-function value into the stack.

Step 2. Generate all immediate successor paths of the top path in the stack, and compute their f-function values. Delete the top path from the stack.

Step 3. If the graph G is a trellis, check whether these successor paths end at a node that belongs to a path that is already in the stack. Restated, check whether these successor paths merge with a path that is already in the stack. If it does, and the f-function value of the successor path exceeds the f-function value of the subpath that traverses the same nodes as the merged path in the stack but ends at the merged node, redirect the merged path by replacing its subpath with the successor path, and update the f-function value associated with the newly redirected path.[5] Remove those successor paths that merge with some paths in the stack. (Note that if the graph G is a code tree, there is a unique path connecting the origin node to each node on the graph; hence, it is unnecessary to examine the path merging.)

Step 4. Insert the remaining successor paths into the stack, and reorder the stack in descending f-function values.

Step 5. If the top path in the stack ends at a terminal node in the graph G, the algorithm stops; otherwise go to step 2.

In principle, the *evaluation function* $f(\cdot)$ that guides the search of Algorithm A is the sum of two parts, $g(\cdot)$ and $h(\cdot)$; both range over all possible paths in the graph. The first part, $g(\cdot)$, is simply a function of all the branches traversed by the path, while the second part, $h(\cdot)$, called the *heuristic function*, help to predict a future route from the end node of the current path to a terminal node. Conventionally, the heuristic function $h(\cdot)$ equals zero for all paths that end at a terminal node. Additionally, $g(\cdot)$ is usually taken to be zero for the single-node path that contains only the origin node.

A question that follows is how to define $g(\cdot)$ and $h(\cdot)$ so that Algorithm A performs the MLD rule. Here, this question is examined by considering a more general problem of how to define $g(\cdot)$ and $h(\cdot)$ so that Algorithm A locates the code path with maximum metric over a code tree or a trellis. Suppose that a metric $c(n_i, n_j)$ is associated with the branch between nodes n_i and n_j. Define the metric of a path as the sum of the metrics of those branches contained by the path. The g-function value for a path can then be assigned as the sum of all the branch metrics experienced by the path. Let the h-function value of the same path be an estimate of the maximum cumulative metric from the end node of the path to a terminal node. Under such a system setting, if the heuristic function satisfies certain optimality criteria, such as whether it always upper-bounds the maximum cumulative metric

[5] The redirect procedure is sometimes time-consuming, especially when the f-function value of a path is computed based on the branches the path traverses. Section 8 introduces two approaches to reduce the burden of path redirection.

from the end node of the path of interest to any terminal node, then Algorithm A guarantees the finding of the maximum-metric code path.[6]

Following the discussion in the previous paragraph and the observation from Eq. (1), a straightforward definition for function $g(\cdot)$ is

$$g(\mathbf{v}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} \log_2 \Pr(r_j \mid v_j) \tag{2}$$

where $\mathbf{v}_{(\ell n-1)}$ is the label sequence of the concerned path and the branch metric between the end nodes of path $\mathbf{v}_{(\ell(n-1)-1)}$ and path $\mathbf{v}_{(\ell n-1)}$ is given by $\sum_{j=\ell(n-1)}^{\ell n-1} \log_2 \Pr(r_j \mid v_j)$. Various heuristic functions with respect to the g function defined above can then be developed. For example, if the branch metric defined above is nonpositive, which apparently holds when the received demodulator output r_j is discrete for $0 \le j \le N-1$, a heuristic function that equals zero for all paths sufficiently upper-bounds the maximum cumulative metric from the end node of the concerned path to a terminal node, and thereby guarantees the finding of the code path with maximum metric. Another example is the well-known Fano path metric [4], which, by its formula, can be equivalently interpreted as the sum of the g-function defined in (2) and a specific h function. The details regarding the Fano metric will be given in the next section.

Notably, the branch metric used to define (2) depends not only on the labels of the concerned branch (i.e., $v_{\ell(n-1)}, \ldots, v_{\ell n-1}$) but also on the respective demodulator outputs (i.e., $r_{\ell(n-1)}, \ldots, r_{\ell n-1}$). Some researchers also view the received vector $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$ as labels for another (possibly nonbinary) code tree or trellis; hence the term "received branch," which reflects a branch labeled by the respective portion of the received vector $\mathbf{r}$ was introduced. With such a naming convention, this section concludes by quoting the essential attributes of sequential decoding defined in Ref. 15. According to the authors, the very first attribute of sequential decoding is that "the branches are examined sequentially, so that at any node of the tree the decoder's choice among a set of previously unexplored branches does not depend on received branches deeper in the tree." The second attribute is that "the decoder performs at least one computation for each node of every examined path." The authors then remark at the end that "Algorithms which do not have these two properties are not considered to be sequential decoding algorithms." Thus, an easy way to visualize the defined features of sequential decoding is that the received scalars $r_0, r_1, \ldots, r_{N-1}$ are received *sequentially* in time in order of the subindices during the decoding process. The next path to be examined therefore cannot be in any sense related to the received scalars whose subindices are beyond the deepest level of the paths that are momentarily in the stack, because random usage, rather than sequential usage, of the received scalars (such as the usage of r_j, followed by the usage of r_{j+2} instead of r_{j+1}) implicitly indicates that all the received scalars should be ready in a buffer before the decoding process begins.

Based on the two attributes, the sequential decoding is simply Algorithm A with an evaluation function $f(\cdot)$ equal to the sum of the branch metrics of those branches contained by the examined path (i.e., the path metric), where the branch metric is a function of the branch labels and the respective portion of the received vector. Variants of sequential decoding therefore mostly reside on different path metrics adopted. The subsequent sections show that taking a general view of Algorithm A, rather than a restricted view of sequential decoding, promotes the understanding of various later generalizations of sequential decoding.

The next section introduces the most well-known path metric for sequential decoding, which is named after its discover, R. M. Fano.

5. FANO METRIC AND ITS GENERALIZATION

Since its discovery in 1963 [4], the Fano metric has become a typical path metric in sequential decoding. Originally, the Fano metric was discovered through mass simulations, and was first used by Fano in his sequential decoding algorithm on a code tree [4]. For any path $\mathbf{v}_{(\ell n-1)}$ that ends at level ℓ on a code tree, the *Fano metric* is defined as

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} M(v_j \mid r_j)$$

where $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$ is the received vector and

$$M(v_j \mid r_j) = \log_2 \frac{\Pr(r_j \mid v_j)}{\Pr(r_j)} - R$$

is the *bit metric*, and the calculation of $\Pr(r_j)$ follows the convention that the code bits — 0 and 1 — are transmitted with equal probability

$$\begin{aligned} \Pr(r_j) &= \sum_{v_j \in \{0,1\}} \Pr(v_j) \Pr(r_j \mid v_j) \\ &= \tfrac{1}{2} \Pr(r_j \mid v_j = 0) + \tfrac{1}{2} \Pr(r_j \mid v_j = 1) \end{aligned}$$

and $R = k/n$ is the code rate. For example, a hard-decision decoder with $\Pr\{r_j = 0 \mid v_j = 1\} = \Pr\{r_j = 1 \mid v_j = 0\} = p$ for $0 \le j \le N-1$ (i.e., a memoryless BSC channel with crossover probability p), where $0 < p < \frac{1}{2}$, will interpret the Fano metric for path $\mathbf{v}_{(\ell n-1)}$ as

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} \log_2 \Pr(r_j \mid v_j) + \ell n(1-R) \tag{3}$$

where

$$\log_2 \Pr(r_j \mid v_j) = \begin{cases} \log_2(1-p), & \text{for} \quad r_j = v_j \\ \log_2(p), & \text{for} \quad r_j \ne v_j \end{cases}$$

[6] Criteria that guarantee optimal decoding by the Algorithm A are extensively discussed in Refs. 13 and 14.

In terms of the Hamming distance, (3) can be rewritten as

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = -\alpha \cdot d_H(\mathbf{r}_{(\ell n-1)}, \mathbf{v}_{(\ell n-1)}) + \beta \cdot \ell \qquad (4)$$

where $\alpha = -\log_2[p/(1-p)] > 0$, and $\beta = n[1 - R + \log_2(1-p)]$. An immediate observation from (4) is that a larger Hamming distance between the path labels and the respective portion of the received vector corresponds to a smaller path metric. This property guarantees that if no error exists in the received vector (i.e., the bits demodulated are exactly the bits transmitted), and $\beta > 0$ (or equivalently, $R < 1 + \log_2(1-p)$),[7] then the path metric increases along the correct code path, and the path metric along any incorrect path is smaller than that of the equally long correct path. Such a property is essential for a metric to work properly with sequential decoding.

Later, Massey [17] proved that at any decoding stage, extending the path with the largest Fano metric in the stack minimizes the probability that the extending path does not belong to the optimal code path, and the usage of the Fano metric for sequential decoding is thus analytically justified. However, making such a *locally* optimal decision at every decoding stage does not always guarantee the ultimate finding of the *globally* optimal code path in the sense of the MLD rule in (1). Hence, the error performance of sequential decoding with the Fano metric is in general slightly inferior to that of the MLD-rule-based decoding algorithm.

A striking feature of the Fano metric is its dependence on the code rate R. Introducing the code rate into the Fano metric somehow reduces the complexity of the sequential decoding algorithm. Observe from (3) that the first term, $\sum_{j=0}^{\ell n-1} \log_2 \Pr(r_j \mid v_j)$, is the part that reflects the maximum-likelihood decision in (1), and the second term, $\ell n(1-R)$, is introduced as a bias to favor a longer path or specifically a path with larger ℓ, since a longer path is closer to the leaves of a code tree and thus is more likely to be part of the optimal code path. When the code rate increases, the number of incorrect paths for a given output length increases.[8] Hence, the confidence on the currently examined path being part of the correct code path should be weaker. Therefore, the claim that longer paths are part of the optimal code path is weaker at higher code rates. The Fano metric indeed mirrors the above intuitive observation by using a linear bias with respect to the code rate.

[7] The code rate bound below which the Fano-metric-based sequential decoding performs well is the *channel capacity*, which is $1 + p\log_2(p) + (1-p)\log_2(1-p)$ in this case. The alternative larger bound $1 + \log_2(1-p)$, derived from $\beta > 0$, can only justify the subsequent argument, and by no means ensure a good performance for sequential decoding under $1 + p\log_2(p) + (1-p)\log_2(1-p) < R < 1 + \log_2(1-p)$. Channel capacity is beyond the scope of this article. Interested readers can refer to the treatise by Cover and Thomas [16].

[8] Imagine that the number of branches that leave each node is 2^k, and increasing the code rate can be conceptually interpreted as increasing k subject to a fixed n for a (n, k, m) convolutional code.

The effect of taking other bias values has been examined in Refs. 18 and 19. The authors defined a new bit metric for sequential decoding as

$$M_B(r_j \mid v_j) = \log_2 \frac{\Pr(r_j \mid v_j)}{\Pr(r_j)} - B \qquad (5)$$

and found that a tradeoff between computational complexity and error performance of sequential decoding can be observed by varying the bias B.

Researchers began to investigate the effect of a joint bias on $\Pr(r_j)$ and R, providing new generalization of sequential decoding. Han et al. [20] observed that universally adding a constant to the Fano metric of all paths does not change the sorting result at each stage of the sequential decoding algorithm (see step 4 of Algorithm A). They then chose the additive constant $\sum_{j=0}^{N-1} \log_2 \Pr(r_j)$, and found that

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) + \sum_{j=0}^{N-1} \log_2 \Pr(r_j) = \sum_{j=0}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R] + \sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j) \qquad (6)$$

for which the two terms on the right-hand side of (6) can be immediately reinterpreted as the g function and the h function from the perspective of Algorithm A.[9] As the g function is now defined based on the branch metric $\sum_{j=\ell(n-1)}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R]$, Algorithm A, according to the discussion in the previous section, becomes a search to find the code path $\mathbf{v}^*$ that satisfies

$$\sum_{j=0}^{N-1} [\log_2 \Pr(r_j \mid v_j^*) - R] \geq \sum_{j=0}^{N-1} [\log_2 \Pr(r_j \mid v_j) - R] \quad \text{for all code path } \mathbf{v}$$

This criterion is equivalent to the MLD rule in (1). Consequently, the Fano path metric indeed implicitly uses $\sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j)$ as a heuristic estimate of the upcoming metric from the end node of the current path to a terminal node. A question that naturally follows regards the trustworthiness of this estimate. The question can be directly answered by studying the effect of varying weights

[9] Notably, defining a path metric as $\sum_{j=0}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R] + \sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j)$ does not yield a sequential decoding algorithm according to the definition of sequential decoding in Ref. 15, for such a path metric depends on information of "the received branches" beyond level ℓ, i.e., $r_{\ell n}, \ldots, r_{N-1}$. However, a similarly defined evaluation function surely gives Algorithm A.

on the g function (i.e., the cumulative metric sum that is already known) and h function (i.e., the estimate) using

$$f_\omega(\mathbf{v}_{(\ell n-1)}) = \omega \sum_{j=0}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R] + (1-\omega) \sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j) \tag{7}$$

where $0 \le \omega \le 1$. Subtracting a universal constant $(1-\omega)\sum_{j=0}^{N-1} \Pr(r_j)$ from (7) gives the *generalized Fano metric* for sequential decoding as

$$M_\omega(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} (\log_2 \frac{\Pr(r_j \mid v_j)^\omega}{\Pr(r_j)^{1-\omega}} - \omega R) \tag{8}$$

When $\omega = \frac{1}{2}$, the generalized Fano metric reduces to the Fano metric with a multiplicative constant, $\frac{1}{2}$. As ω is slightly below $\frac{1}{2}$, which can be interpreted from (7) as the sequential search is guided more by the estimate on the upcoming metrics than by the known cumulative metric sum, the number of metric computations reduces but the decoding failure probability grows. When ω is closer to one, the decoding failure probability of sequential decoding tends to be lower; however, the computational complexity increases. In the extreme case, taking $\omega = 1$ makes the generalized Fano metric completely mirror the MLD metric in (1), and the sequential decoding becomes a maximum-likelihood (hence, optimal in decoding failure probability) decoding algorithm. The work in Ref. 20 thereby led to the conclusion that an (implicit) heuristic estimate can be elaborately defined to reduce fairly the complexity of sequential decoding with a slight degradation in error performance. Notably, for discrete symmetric channels, the generalized Fano metric is equivalent to the metric defined in Eq. (5) [21]. However, the generalized Fano metric and the metric of (5) are by no means equal for other types of channels such as the AWGN channel.

6. STACK ALGORITHM AND ITS VARIANTS

The stack algorithm or the ZJ algorithm was discovered by Zigangirov [5] and later independently by Jelinek [6] to search a code tree for the optimal codeword. It is exactly the Algorithm A with g-function equal to the Fano metric and zero h function. Because a stack is involved in searching for the optimal codeword, the algorithm is called the *stack algorithm*. An example is provided blow to clarify the flow of the stack algorithm.

Example 1. For a BSC with crossover probability $p = .045$, the Fano bit metric for a convolutional code with code rate $R = \frac{1}{2}$ can be obtained from (3) as

$$M(v_j \mid r_j) = \begin{cases} \log_2(1-p) + (1-R) = 0.434, & \text{for } r_j = v_j \\ \log_2(p) + (1-R) = -3.974, & \text{for } r_j \ne v_j \end{cases}$$

Consequently, only two Fano bit metric values are possible, 0.434 and −3.974. These two Fano bit metric values can be "scaled" to equivalent "integers" to facilitate the simulation and implementation of the system. Taking the multiplicative scaling factor of 2.30415 yields

$$M_{\text{scaled}}(v_j \mid r_j) = \begin{cases} 0.434 \times 2.30415 \approx 1, & \text{for } r_j = v_j \\ -3.974 \times 2.30415 \approx -9, & \text{for } r_j \ne v_j \end{cases}$$

Now, the convolutional code in Fig. 1 is decoded over its code tree (cf. Fig. 3) using the stack algorithm with the scaled Fano metric. Assume that the received vector is **r**=(11 01 00 01 10 10 11). Figure 5 presents the contents

Loop 1	Loop 2	Loop 3	Loop 4	Loop 5
1 (1 + 1 = 2)	11 (2 + 1 + 1 = 4)	111 (4 – 9 + 1 = –4)	1110 (–4 + 1 + 1 = –2)	110 (–4)
0 (–9 –9 = – 18)	10 (2 –9 –9 = –16)	110 (4 + 1 – 9 = –4)	110 (–4)	11100 (–2 + 1 – 9 = –10)
	0 (–18)	10 (–16)	10 (–16)	11101 (–2 –9 + 1 = –10)
		0 (–18)	0 (–18)	10 (–16)
			1111 (–4 –9 –9 = –22)	0 (–18)
				1111 (–22)

Loop 6	Loop 7	Loop 8	Loop 9
11100 (–10)	11101 (–10)	111010 (–10 + 1 + 1 = –8)	1110100 (–8 + 1 + 1 = –6)
11101 (–10)	1100 (–12)	1100 (–12)	1100 (–12)
1100 (–4 –9 + 1 = –12)	1101 (–12)	1101 (–12)	1101 (–12)
1101 (–4 + 1 – 9 = –12)	10 (–16)	10 (–16)	10 (–16)
10 (–16)	111000 (–10 – 9 + 1 = –18)	111000 (–18)	111000 (–18)
0 (–18)	0 (–18)	0 (–18)	0 (–18)
1111 (–22)	1111 (–22)	1111 (–22)	1111 (–22)

Figure 5. Stack contents after each path metric reordering in Example 1. Here, different from that used in the Fano metric computation, the input bit labels rather than the code bit labels are used for each recorded path. The associated Fano metric follows each path label sequence (inside parentheses).

of the stack after each path metric reordering. Each path in the stack is marked by its corresponding input bit labels rather than by the code bit labels. Notably, while both types of labels can uniquely determine a path, the input bit labels are more frequently recorded in the stack in practical implementation since the input bit labels are the desired estimates of the transmitted information sequences. Code bit labels are used more often in metric computation and in characterizing the code, because the code characteristic, such as error-correcting capability, can only be determined from the code bit labels (codewords).[10] The path metric associated with each path is also stored. The algorithm is terminated at the ninth loop, yielding an ultimate decoding result of $\mathbf{u} = (11101)$.

Maintaining the stack is a significant implementation issue of the stack algorithm. In a straightforward implementation of the stack algorithm, the paths are stored in the stack in order of descending f-function values; hence, a sorting mechanism is required. Without a proper design, the sorting of the paths within the stack may be time-consuming, limiting the speed of the stack algorithm.

Another implementation issue of the stack algorithm is that the stack size in practice is often insufficient to accommodate the potentially large number of paths examined during the search process. The stack can therefore overflow. A common way of compensating for a stack overflow is to simply discard the path with the smallest f-function value [1], since it is least likely to be the optimal code path. However, when the discarded path happens to be an early predecessor of the optimal code path, performance is degraded.

Jelinek proposed the so-called *stack-bucket technique* to reduce the sorting burden of the stack algorithm [6]. In his proposal, the range of possible path metric values is divided into a number of intervals with prespecified, fixed spacing. For each interval, a separate storage space, a *bucket*, is allocated. The buckets are then placed in order of descending interval endpoint values. During decoding, the next path to be extended is always the top path of the first nonempty bucket, and every newly generated path is directly placed on top of the bucket in which interval the respective path metric lies. Some data structure can be used to reduce the maintenance burden and storage demand of stacked buckets, since some buckets corresponding to a certain metric range may occasionally (or even always) be empty during decoding. The sorting burden is therefore removed by introducing the stacked buckets. The time taken to locate the next path no longer depends on the size of the stack, rather on the number of buckets, considerably reducing the time required by decoding. Consequently, the stacked bucket technique was used extensively in the software implementation of the stack algorithm for applications in which the decoding time is precisely restricted [1,22,23]

The drawback of the stacked bucket technique is that the path with the best path metric may not be selected, resulting in degradation in performance. A *metric-first stacked bucket* implementation overcomes the drawback by sorting the top bucket when it is being accessed. However, Anderson and Mohan [24] indicated that the access time of the metric-first stacked buckets will increase at least to the order of $S^{1/3}$, where S is the total number of the paths ever generated.

Another software implementation technique for establishing a sorted stack was discussed by Mohan and Anderson [25], who suggested the adoption of a *balanced binary tree* data structure, such as an AVL tree [26], to implement the stack, offering the benefit that the access time of the stack becomes of order $\log_2(S)$, where S represents the momentary stack size. Briefly, a balanced binary tree is a sorted structure with node insertion and deletion schemes such that its depth is maintained equal to the logarithm of the total number of nodes in the tree, whenever possible. As a result, inserting or deleting a path (which is now a node in the data structure of a balanced binary tree) in a stack of size S requires at most $\log_2(S)$ comparisons (i.e., the number of times the memory is accessed). The balanced binary tree technique is indeed superior to the metric-first stacked bucket implementation, when the stack grows beyond certain size. Detailed comparisons of time and space consumption of various implementation techniques of sequential decoding, including the Fano algorithm to be introduced in the next section, can be found in another article [24].

In 1994, a novel systolic priority queue, called the *parallel entry systolic priority queue* (PESPQ), was proposed to replace the stacked buckets [27]. Although it does not arrange the paths in the queue in strict order, the systolic priority queue technique can identify the path with the largest path metric within a constant time. This constant time was shown to be comparable to the time required to compute the metric of a new path. Experiments revealed that the PESPQ stack algorithm is several times faster than its stacked bucket counterpart. Most importantly, the invention of the PESPQ technique has given a seemingly promising future to hardware implementation of the stack algorithm.

As stated at the end of Section 4, one of the two essential features of sequential decoding is that the next visited path cannot be selected based on the basis of the information deeper in the tree [15]. The above feature is subsequently interpreted as the information that is deeper in the tree is supposed to be received in some future time and hence is perhaps not available at the current decoding stage. This conservative interpretation apparently arises from the aspect of an online decoder. A more general reinterpretation, simply following the wording, is that any codeword search algorithm that decides the next visited path in sequence without using the information deeper in its own search tree is considered to be a sequential decoding algorithm. This new interpretation precisely suits the bidirectional sequential decoding algorithm

[10] For a code tree, a path can be also uniquely determined by its end node in addition to the two types of path labels, so putting the "end node" rather than the path labels of a path in the stack suffices to fulfill the need for the tree-based stack algorithm. Nevertheless, such an approach, while easing the stack maintenance load, introduces an extra conversion load from the path end node to its respective input bit labels (as the latter is the desired estimates of the transmitted information sequence).

proposed by Forney [11], in which each of the two decoders still performs the defined sequential search in its own search tree. Specifically, Forney suggested that the sequential decoding could also start its decoding from the end of the received vector, and proposed a bidirectional stack algorithm in which two decoders simultaneously search the optimal code path from both ends of the code tree. The bidirectional decoding algorithm stops whenever either decoder reaches the end of its search tree. This idea has been greatly improved by Kallel and Li by stopping the algorithm whenever two stack algorithms with two separate stacks meet at a common node in their respective search trees [28]. Forney also claimed that the same idea can be applied to the Fano algorithm introduced in the next section.

7. FANO ALGORITHM

The Fano algorithm is a sequential decoding algorithm that does not require a stack [4]. The Fano algorithm can only operate over a code tree because it cannot examine path merging.

At each decoding stage, the Fano algorithm retains the information regarding three paths: the current path, its immediate predecessor path, and one of its successor paths. On the basis of this information, the Fano algorithm can move from the current path to either its immediate predecessor path or the selected successor path; hence, no stack is required for queuing all examined paths.

The movement of the Fano algorithm is guided by a dynamic threshold T that is an integer multiple of a fixed step size Δ. Only the path whose path metric is no less than T can be next visited. According to the algorithm, the process of codeword search continues to move forward along a code path, as long as the Fano metric along the code path remains nondecreasing. Once all the successor path metrics are smaller than T, the algorithm moves backward to the predecessor path if the predecessor path metric beats T; thereafter, threshold examination will be subsequently performed on another successor path of this revisited predecessor. In case the predecessor path metric is also less than T, the threshold T is one step lowered so that the algorithm is not trapped on the current path. For the Fano algorithm, if a path is revisited, the presently examined dynamic threshold is always lower than the momentary dynamic threshold at the previous visit, guaranteeing that looping in the algorithm does not occur, and that the algorithm can ultimately reach a terminal node of the code tree, and stop.

Figure 6 displays a flowchart of the Fano algorithm, in which $\mathbf{v}_p$, $\mathbf{v}_c$, and $\mathbf{v}_s$ represent the path label sequences of the predecessor path, the current path and the successor path, respectively. Their Fano path metrics are denoted by M_p, M_c, and M_s, respectively. The algorithm begins with the path that contains only the origin node. The label sequence of its predecessor path is initially set to "dummy," and the path metric of such a dummy path is assumed to be negative infinity. The initialization value of the dynamic threshold is zero.

The algorithm then proceeds to find, among the 2^k candidates, the successor path $\mathbf{v}_s$ with the largest path metric M_s. Thereafter, it examines whether $M_s \geq T$. If so, the algorithm moves forward to the successor path and updates the necessary information. Then, whether the new current path is a code path is determined, and a positive result immediately terminates the algorithm. A delicate part of the Fano algorithm is "threshold tightening." Whenever a path is first visited, the dynamic threshold T must be "tightened" such that it is adjusted to the largest possible value below the current path metric, namely, $T \leq M_c < T + \Delta$. Notably, the algorithm can determine whether a path is first visited by simply examining $\min\{M_p, M_c\} < T + \Delta$. If $\min\{M_p, M_c\} < T + \Delta$ holds because of the validity of $M_c < T + \Delta$, the threshold is automatically tightened; hence, only the condition $M_p < T + \Delta$ is required in the tightening test. The above mentioned procedures are repeated until a code path is finally reached.

Along the other route of the flowchart, following a negative answer to the examination of $M_s \geq T$ (which implicitly implies that the path metrics of all the successor paths of the current path are less than T), the algorithm must lower the threshold if M_p is also less than T; otherwise, a deadlock on the current path arises. Using the lowered threshold, the algorithm repeats the finding of the best successor path whose path metric exceeds the new threshold, and proceeds the followup steps.

The rightmost loop of the flowchart considers the case for $M_s < T$ and $M_p \geq T$. In this case, the algorithm can only move backward, since the predecessor path is the only one whose path metric is no smaller than the dynamic threshold. The information regarding the current path and the successor path should be subsequently updated. Yet, the predecessor path, as well as its associated path metric, should be recalculated from the current $\mathbf{v}_p$, because information about the predecessor's predecessor is not recorded. Afterward, the Fano algorithm checks for the existence of a new successor path $\mathbf{v}_t$ that is not the current successor path from which the algorithm has just moved, and whose associated path metric exceeds the current M_s. Restated, this step finds the best successor path other than those that have already been examined. If such a new successor path does not exist, then the algorithm seeks either to reduce the dynamic threshold or to move backward again, depending on whether $M_p \geq T$. In case such a new successor path $\mathbf{v}_t$ with metric M_t is located, the algorithm refocuses on the new successor path by updating $\mathbf{v}_s = \mathbf{v}_t$ and $M_s = M_t$, and repeats the entire process.

A specific example is provided below to help in understanding of the Fano algorithm.

Example 2. Assume the same convolutional code and received vector as in Example 1. Let the step size Δ be four. Figure 7 presents the traces of the Fano algorithm during its decoding.

In this figure, each path is again represented by its input bit labels. S and D denote the paths that contains only the origin node and the dummy path, respectively. According to the Fano algorithm, the possible actions taken include MFTT = "move forward and tighten the threshold," MF = "move forward only," MBS = "move

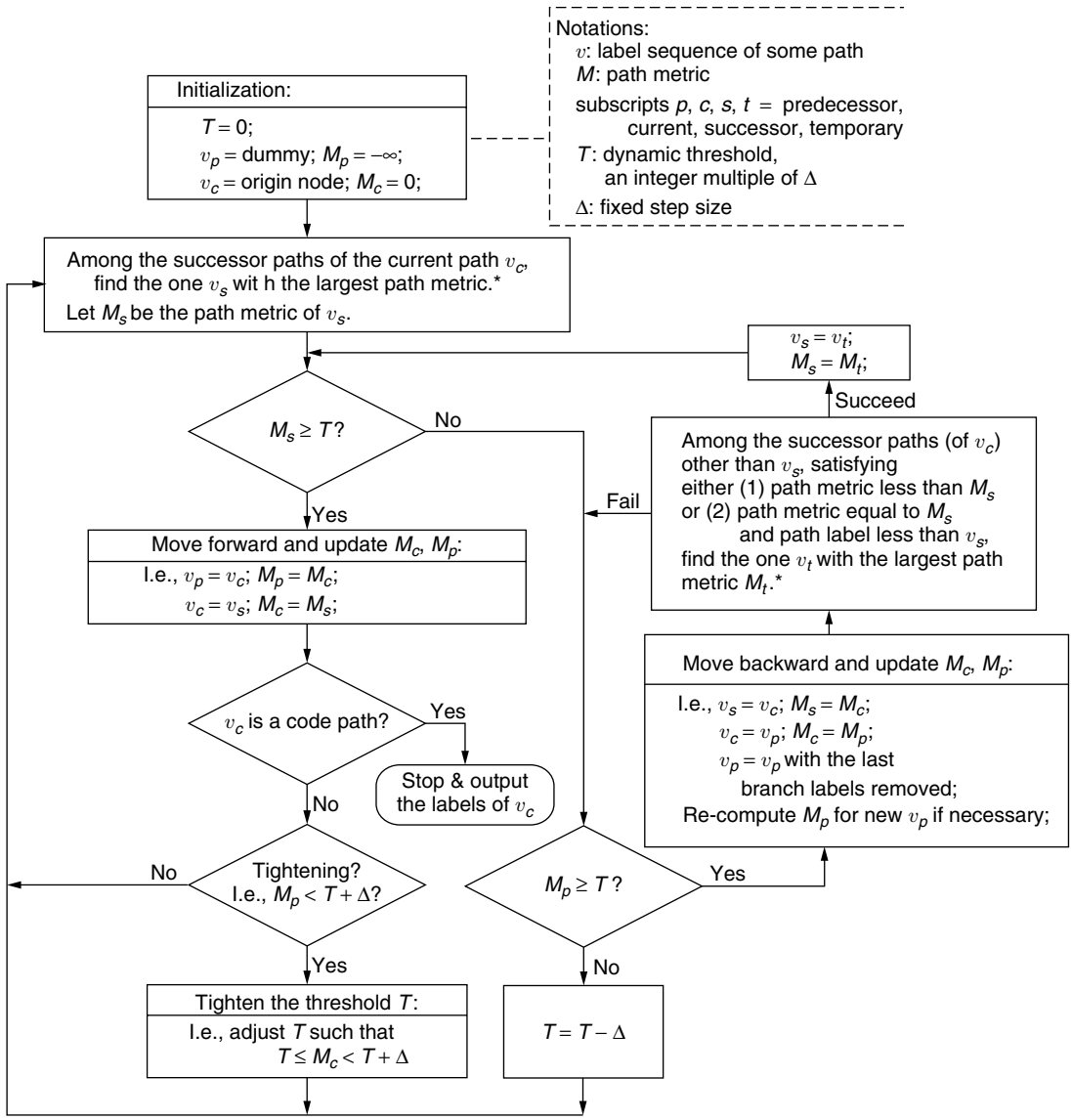


Figure 6. Flowchart of the Fano algorithm.

backward and successfully find the second best successor," MBF = "move backward but fail to find the second best successor," and LT = "lower threshold by one step." The algorithm stops after 36 iterations, and decodes the received vector to the same code path as that obtained in Example 1. This example clearly shows that the Fano algorithm revisits several paths more than once, such as path 11 (eight visits) and path 111 (five visits), and returns to path S three times.

As described in the previous example, the Fano algorithm may move backward to the path that contains only the origin node, and discover all its successors with path metrics less than T. In this case, the only action the algorithm can take is to keep decreasing the dynamic threshold until the algorithm can move forward again because the path metric of the predecessor path of the single-node path is set to $-\infty$. The impact of varying Δ should also be clarified. As stated by Lin and Costello [1], the load of branch metric computations executed during the finding of a qualified successor path becomes heavy when Δ is too small; however, when Δ is too large, the dynamic threshold T might be lowered too much in a single adjustment, forcing an increase in both the decoding error and the computational effort. Experiments suggest [1] that a better performance can be obtained by taking Δ within 2 and 8 for the unscaled Fano metric (Δ should be analogously scaled if a scaled Fano metric is used).

The Fano algorithm, perhaps surprisingly, while quite different in its design from the stack algorithm, exhibits broadly the same searching behavior as the stack algorithm. In fact, with a slight modification to the update procedure of the dynamic threshold (e.g., setting $\Delta = 0$, and substituting the "tightening test" and subsequent "tightening procedure" by "$M_c \neq T$?" and "$T = M_c$," respectively), both algorithms have been proved to visit almost the same set of paths during the decoding process [29]. Their only dissimilarity is that unlike the

Iteration	v_p	v_c	v_s	M_p	M_c	M_s	T	Action
0	*D*	*S*	1	$-\infty$	0	2	0	MFTT
1	*S*	1	11	0	2	4	0	MFTT
2	1	11	111	2	4	−4	4	LT
3	1	11	111	2	4	−4	0	MBS
4	*S*	1	10	0	2	−16	0	MBS
5	*D*	*S*	0	$-\infty$	0	−18	0	LT
6	*D*	*S*	1	$-\infty$	0	2	−4	MF
7	*S*	1	11	0	2	4	−4	MF
8	1	11	111	2	4	−4	−4	MF
9	11	111	1110	4	−4	−2	−4	MFTT
10	111	1110	11100	−4	−2	−10	−4	MBS
11	11	111	1111	4	−4	−22	−4	MBS
12	1	11	110	2	4	−4	−4	MF
13	11	110	1100	4	−4	−12	−4	MBF
14	1	11	110	2	4	−4	−4	MBS
15	*S*	1	10	0	2	−16	−4	MBS
16	*D*	*S*	0	$-\infty$	0	−18	−4	LT
17	*D*	*S*	1	$-\infty$	0	2	−8	MF
18	*S*	1	11	0	2	4	−8	MF
19	1	11	111	2	4	−4	−8	MF
20	11	111	1110	4	−4	−2	−8	MF
21	111	1110	11100	−4	−2	−10	−8	MBS
22	11	111	1111	4	−4	−22	−8	MBS
23	1	11	110	2	4	−4	−8	MF
24	11	110	1100	4	−4	−12	−8	MBF
25	1	11	110	2	4	−4	−8	MBS
26	*S*	1	10	0	2	−16	−8	MBS
27	*D*	*S*	0	$-\infty$	0	−18	−8	LT
28	*D*	*S*	1	$-\infty$	0	2	−12	MF
29	*S*	1	11	0	2	4	−12	MF
30	1	11	111	2	4	−4	−12	MF
31	11	111	1110	4	−4	−2	−12	MF
32	111	1110	11100	−4	−2	−10	−12	MF
33	1110	11100	111000	−2	−10	18	−12	MBS
34	111	1110	11101	−4	−2	−10	−12	MF
35	1110	11101	111010	−2	−10	−8	−12	MFTT
36	11101	111010	1110100	−10	−8	−6	−8	Stop

Figure 7. Decoding traces of the Fano algorithm for Example 2.

stack algorithm which visits each path only once, the Fano algorithm may revisit a path several times, and thus has a higher computational complexity. From the simulations over the binary symmetric channel as illustrated in Fig. 8, the stack algorithm with stacked bucket modification is apparently faster than the Fano algorithm at crossover probability $p = .057$, when their software implementations are concerned. The stack algorithm's superiority in computing time gradually disappears as p becomes smaller (or as the channel becomes less noisy) [22].

In practice, the time taken to decode an input sequence often has an upper limit. If the decoding process is not completed before the time limit, the undecoded part of the input sequence must be aborted or erased; hence, the probability of input erasure is another system requirement for sequential decoding. Figure 9 confirms that the stack algorithm with stacked bucket modification remains faster than the Fano algorithm except when either admitting a high erasure probability (e.g., erasure probability > 0.5 for $p = .045$, and erasure probability > 0.9 for $p = .057$) or experiencing a less noisier channel (e.g., $p = .033$) [22].

An evident drawback of the stack algorithm in comparison with the Fano algorithm is its demand for an extra stack space. However, with recent advances in computer technology, a large memory requirement is no longer a restriction for software implementation.

Hardware implementation is now considered. In hardware implementation, stack maintenance normally

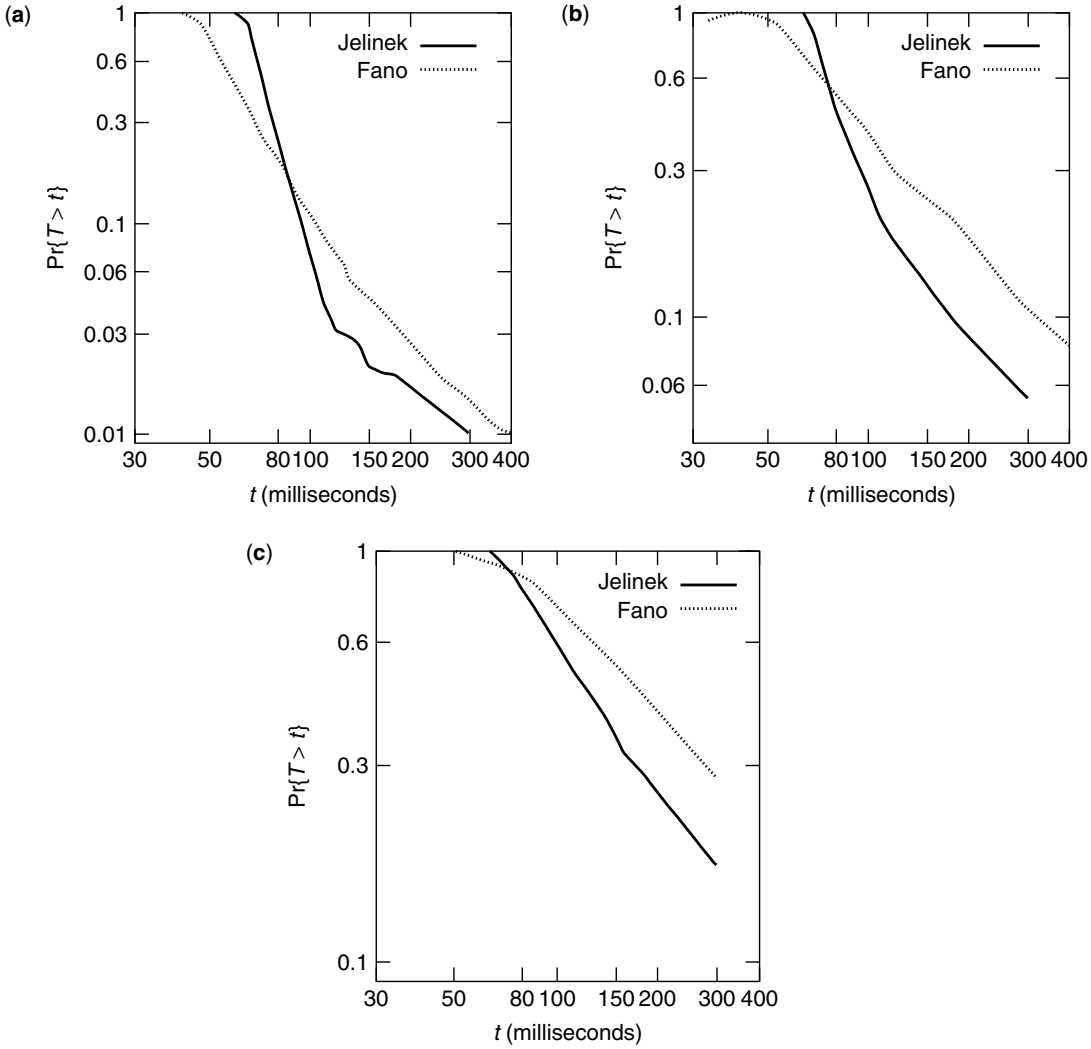


Figure 8. Comparisons of computational complexities of the Fano algorithm and the stack algorithm (with stacked bucket modification) based on the $(2, 1, 35)$ convolutional code with generator polynomials $g_1 = 53, 533, 676, 737$ and $g_2 = 733, 533, 676, 737$ and a single input sequence of length 256; simulations are performed over the binary symmetric channel with crossover probability p: (**a**) $p = .033$; (**b**) $p = .045$; (**c**) $p = .057$. $\Pr\{T \geq t\}$ is the empirical complement cumulative distribution function for the software computation time T. In simulations, $(\log_2[2(1-p)] - \frac{1}{2},\ \log_2(2p) - \frac{1}{2})$, which is derived from the Fano metric formula, is scaled to $(2, -18)$, $(2, -16)$ and $(4, -35)$ for $p = 0.033$, $p = 0.045$ and $p = 0.057$, respectively. In subfigures (a), (b) and (c), the parameters for the Fano algorithm are $\Delta = 16$, $\Delta = 16$ and $\Delta = 32$, and the bucket spacings taken for the stack algorithm are 4, 4, and 8, respectively. (Reproduced from Figs. 1–3 in Ref. 22).

requires accessing external memory a certain number of times, which usually bottlenecks the system performance. Furthermore, the hardware is renowned for its efficient adaptation to a big number of computations. These hardware implementation features apparently favor the no-stack Fano algorithm, even when the number of its computations required is larger than the stack algorithm. In fact, a hard-decision version of the Fano algorithm has been hardware-implemented, and can operate at a data rate of 5 Mbps (megabits per second) [30]. The prototype employs a systematic convolutional code to compensate for the input erasures so that whenever the prespecified decoding time expires, the remaining undecoded binary demodulator outputs that directly correspond to the input sequences are immediately outputted. Other features of this prototype include the following:

- The Fano metric is fixedly scaled to either $(1, -11)$ or $(1, -9)$, rather than adaptable to the channel noise level for convenient hardware implementation.
- The length (or depth) of the backward movement of the Fano algorithm is limited by a technique called *backsearch limiting*, in which the decoder is not allowed to move backward more than some maximum number J levels back from its furthest penetration into the tree. Whenever the backward

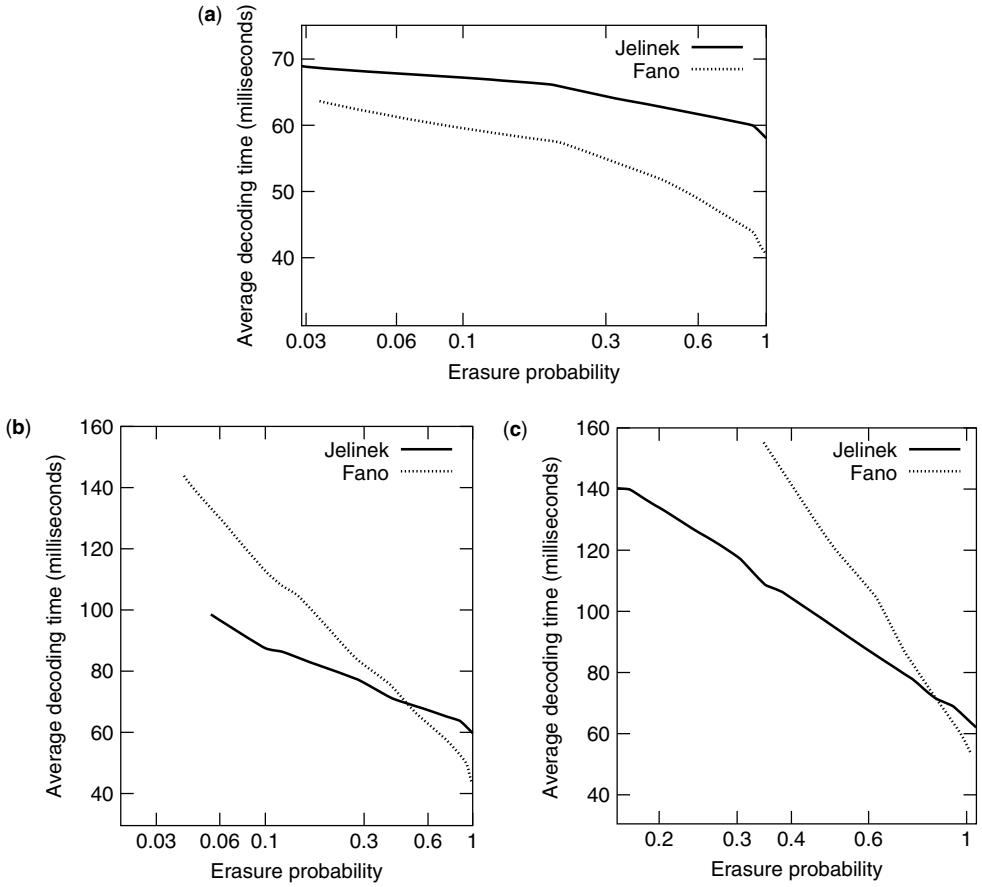


Figure 9. Comparisons of erasure probabilities of the Fano algorithm and the stack algorithm with stacked-bucket modification; all simulation parameters are taken to be the same as those in Fig. 8: (**a**) $p = .033$; (**b**) $p = .045$; (**c**) $p = .057$. (Reproduced from Figs. 5–7 in Ref. 22).

limit is reached, the decoder is forced forward by lowering the threshold until it falls below the metric of the best successor path.

- When the hardware decoder maintains the received bits that correspond to the backsearch J branches for use of forward and backward moves, a separate input buffer must, at the same time, actively receive the upcoming received bits. Once an input buffer overflow is encountered, the decoder must be resynchronized by directly jumping to the most recently received branches in the input buffer, and those information bits corresponding to J levels back are forcefully decoded. The forcefully decoded outputs are exactly the undecoded binary demodulator outputs that directly correspond to the respective input sequences. Again, this design explains why the prototype must use a systematic code.

Figure 10 shows the resultant bit error performances for this hardware decoder for BSCs [30]. An anticipated observation from Fig. 10 is that a larger input buffer, which can be interpreted as a larger decoding time limit, gives a better performance.

A soft-decision version of the hardware Fano algorithm was used for space and military applications in the late 1960s [31,32]. The decoder built by the Jet Propulsion Laboratory [32] operated at a data rate of 1 Mbps, and successfully decoded the telemetry data from the Pioneer Nine spacecraft. Another soft-decision variable-rate hardware implementation of the Fano algorithm was reported in Ref. 33, wherein decoding was accelerated to 1.2 Mbps.

A modified Fano algorithm, named *creeper algorithm*, was proposed in 1999 [34]. This algorithm is indeed a compromise between the stack algorithm and the Fano algorithm. Instead of placing all visited paths in the stack, it selectively stores a fixed number of the paths that are more likely to be part of the optimal code path. As anticipated, the next path to be visited is no longer restricted to the immediate predecessor path and the successor paths but is extended to these selected likely paths. The number of likely paths is usually set less than 2^k times the code tree depth. The simulations given in Ref. 34 indicated that in computational complexity, the creeper algorithm considerably improves the Fano algorithm, and can be made only slightly inferior to the stack algorithm.

8. TRELLIS-BASED SEQUENTIAL DECODING ALGORITHM

Sequential decoding was mostly operated over a code tree in early publications, although some early published

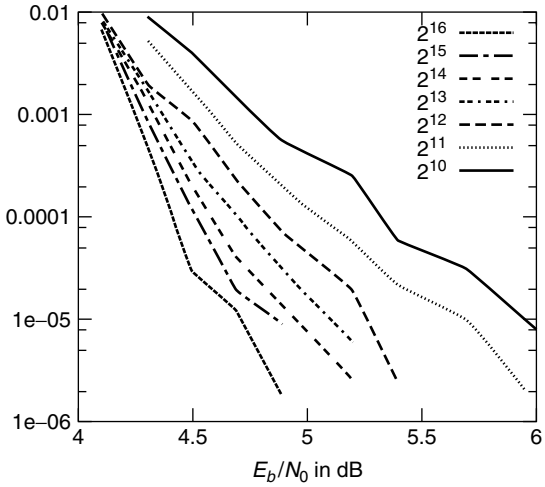


Figure 10. Bit error rate of the hardware Fano algorithm based on systematic $(2, 1, 46)$ convolutional code with generator polynomials $g_1 = 4,000,000,000,000,000$ and $g_2 = 7,154,737,013,174,652$. The legend indicates that the input buffer size tested ranges from $2^{10} = 1024$ to $2^{16} = 65,536$ branches. Experiments are conducted with 1 Mbps data rate over the binary symmetric channel with crossover probability $p = \frac{1}{2}\text{erfc}(\sqrt{E_b/N_0})$, where $\text{erfc}(x) = (2/\sqrt{\pi}) \int_x^\infty \exp\{-x^2\}\, dx$ is the complementary error function. The step size, the backsearch limit, and the Fano metric are set to $\Delta = 6$, $J = 240$, and $(1, -11)$, respectively. (Reproduced from Fig. 18 in Ref. 30).

work already hinted at the possibility of conducting sequential decoding over a trellis [35,36]. The first feasible algorithm that sequentially searches a trellis for the optimal codeword is the *generalized stack algorithm* [23]. The generalized stack algorithm simultaneously extends the top M paths in the stack. It then determines, according to the trellis structure, whether any of the extended paths merge with a path that is already in the stack. If so, the algorithm deletes the newly generated path after ensuring that its path metric is smaller than the cumulative path metric of the merged path up to the merged node. No redirection on the merged path is performed, even if the path metric of the newly generated path exceeds the path metric of the subpath that traverses along the merged path, and ends at the merged node. Thus, the newly generated path and the merged path may coexist in the stack. The generalized stack algorithm, although it generally yields a larger average computational complexity than the stack algorithm, has lower variability in computational complexity and a smaller probability of decoding error [23].

The main obstacle in implementing the generalized stack algorithm by hardware is the maintenance of the stack for the simultaneously extended M paths. One feasible solution is to employ M independent stacks, each of which is separately maintained by a processor [37]. In such a multiprocessor architecture, only one path extraction and two path insertions are sufficient for each stack in a decoding cycle of a $(2, 1, m)$ convolutional code [37]. Simulations have shown that this multiprocessor counterpart not only retained the low variability in computational complexity as the original generalized stack algorithm but also had a smaller average decoding time.

When the trellis-based generalized stack algorithm simultaneously extends 2^K most likely paths in the stack (i.e., $M = 2^K$), where $K = \sum_{j=1}^{k} K_j$ and K_j is the length of the jth shift register in the convolutional code encoder, the algorithm becomes the maximum-likelihood Viterbi decoding algorithm. The optimal codeword is thereby sought by exhausting all possibilities, and no computational complexity gain can be obtained at a lower noise level. Han et al. [38] proposed a true noise-level-adaptable trellis-based maximum-likelihood sequential decoder, called *maximum-likelihood soft-decision decoding algorithm* (MLSDA). The MLSDA adopts a new metric, other than the Fano metric, to guide its sequential search over a trellis for the optimal code path, which is now the code path with the minimum path metric. Derived from a variation of the Wagner rule [39], the new path metric associated with a path $\mathbf{v}_{(\ell n-1)}$ is given by

$$M_{\text{ML}}(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} M_{\text{ML}}(v_j \mid r_j) \tag{9}$$

where $M_{\text{ML}}(v_j \mid r_j) = (y_j \oplus v_j) \times |\phi_j|$ is the jth-bit metric, $\mathbf{r}$ is the received vector, $\phi_j = \ln[\Pr(r_j \mid 0)/\Pr(r_j \mid 1)]$ is the jth loglikelihood ratio, and

$$y_j = \begin{cases} 1, & \text{if } \phi_j < 0 \\ 0, & \text{otherwise} \end{cases}$$

is the hard-decision output due to ϕ_j. For AWGN channels, the ML-bit metric can be simplified to $M_{\text{ML}}(v_j \mid r_j) = (y_j \oplus v_j) \times |r_j|$, where

$$y_j = \begin{cases} 1, & \text{if } r_j < 0 \\ 0, & \text{otherwise} \end{cases}$$

As described previously, the generalized stack algorithm, while examining the path merging according to a trellis structure, does not redirect the merged paths. The MLSDA, however, genuinely redirects and merges any two paths that share a common node, resulting in a stack without coexistence of crossed paths. A remarkable feature of the new ML path metric is that when a newly extended path merges with an existing path of longer length, the ML path metric of the newly extended path is always greater than or equal to the cumulative ML metric of the existing path up to the merged node. Therefore, a newly generated path that is shorter than its merged path can be immediately deleted, reducing the redirection overhead of the MLSDA only to the case in which the newly generated path and the merged existing path are equally long.[11] Thus, they merged at their end node. In such case, the

[11] Notably, for the new ML path metric, the path that survives is always the one with smaller path metric, contrary to the sequential decoding algorithm in terms of the Fano metric, in which the path with larger Fano metric survives.

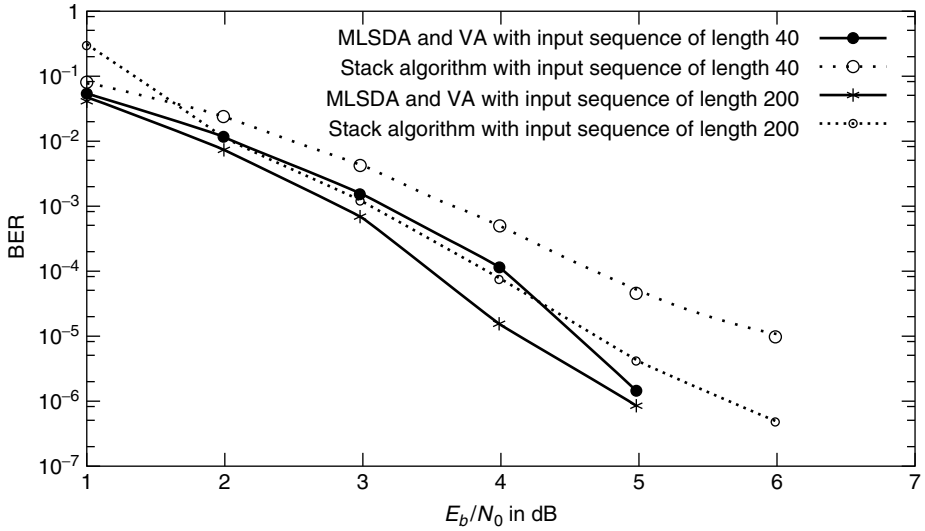


Figure 11. Bit error rates (BERs) of the MLSDA, the Viterbi algorithm (VA), and the stack algorithm for binary $(2, 1, 6)$ convolutional code with generators $g_1 = 634$ (octal), $g_2 = 564$ (octal), and input sequences of lengths 40 and 200. *Source*: Fig. 1 in Ref. 38).

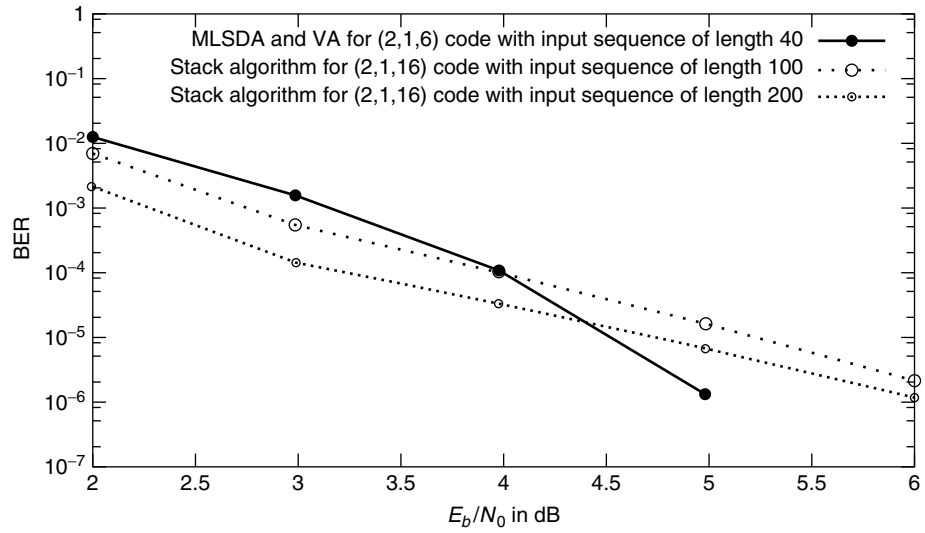


Figure 12. Bit error rates (BERs) of the MLSDA and Viterbi algorithm for binary $(2, 1, 6)$ convolutional code with generators $g_1 = 634$ (octal), $g_2 = 564$ (octal), and an input sequence of length 40. Also, BERs of the stack algorithm for binary $(2, 1, 16)$ convolutional code with generators $g_1 = 1632044, g_2 = 1145734$, and input sequences of lengths 100 and 200. (*Source*: Fig. 2 in Ref. 38).

redirection is merely a deletion of the path with larger path metric.

Figures 11 and 12 show the performances of the MLSDA for $(2, 1, 6)$ and $(2, 1, 16)$ convolutional codes transmitted over the AWGN channel. Specifically, Fig. 11 compares the bit error rate (BER) of the MLSDA with those obtained by the Viterbi and the stack algorithms. Both the MLSDA and the Viterbi algorithm yield the same BER since they are both maximum-likelihood decoders. Figure 11 also shows that the MLSDA provides around 1.5 dB advantage over the stack algorithm at BER $= 10^{-5}$, when both algorithms employ the same input sequence of length 40. Even when the length of the input sequence of the stack algorithm is extended to 200, the MLSDA with input sequence of length 40 still offers an advantage of $\sim$0.5 dB at BER $= 10^{-6}$. Figure 12 collects the empirical results concerning the MLSDA with an input sequence of length 40 for $(2, 1, 6)$ code, and the stack algorithm with input sequences of lengths 100 and 200 for $(2, 1, 16)$ code. The three curves indicate that the MLSDA with an input sequence of smaller length 40 for $(2, 1, 6)$ code provides an advantage of 1.0 dB over the stack algorithm with much longer input sequences and larger constraint length at BER $= 10^{-6}$.

These simulations lead to the conclusion that the stack algorithm normally requires a sufficiently long input sequence to converge to a low BER, which necessarily results in a long decoding delay and high demand for stack space. By adopting a new sequential decoding metric, the MLSDA can achieve the same performance using a much shorter input sequence; hence, the decoding delay and the demand for stack space can be significantly reduced. Furthermore, unlike the Fano metric, the new ML metric adopted in the MLSDA does not depend on the knowledge of the channel, such as SNR, for codes transmitted over the AWGN channel. Consequently, the MLSDA and the Viterbi algorithm share a common nature that their performance is insensitive to the accuracy of the channel SNR estimate for AWGN channels.

9. PERFORMANCE CHARACTERISTICS OF SEQUENTIAL DECODING

An important feature of sequential decoding is that the decoding time varies with the received vector, because the number of paths examined during the decoding process differs for different received vectors. The received vector, in turn, varies according to the noise statistics. The decoding complexity can therefore be legitimately viewed as a random variable whose probability distribution is defined by the statistics of the received vector.

The statistics of sequential decoding complexity have been extensively investigated using the *random coding technique* [15,36,40–44]. Instead of analyzing the complexity distribution with respect to a specific deterministic code, the average complexity distribution for a random code was analyzed. In practical applications, the convolutional codes are always deterministic in their generator polynomials. Nevertheless, taking the aspect of a random convolutional code, in which the coefficients of the generator polynomials are random, facilitates the analysis of sequential decoding complexity. The resultant average decoding complexity (where the decoding complexity is directly averaged over all possible generator polynomials) can nonetheless serve as a quantitative guide to the decoding complexity of a practical deterministic code.

In analyzing the average decoding complexity, a correct code path that corresponds to the transmitted codeword over the code tree or trellis always exists, even if the convolutional encoder is now random. Extra computation time is introduced whenever the search process of the decoder deviates from the correct code path due to a noise-distorted received vector. The incorrect paths that deviate from the correct code path can be classified according to the first node at which the incorrect and the correct code paths depart from each other. Denote by S_j the subtree that contains all incorrect paths that branch from the jth node on the correct path, where $0 \le j \le L-1$ and L is the length of the code input sequences. Then, an upper probability bound[12] on the average computational complexity C_j defined as the number of branch metric computations due to the examination of those incorrect paths in S_j can be established as

$$\Pr\{C_j \ge \mathcal{N}\} \le A\mathcal{N}^{-\rho} \tag{10}$$

for some $0 < \rho < \infty$ and any $0 \le j \le L-1$, where A is a constant that varies for different sequential decoding algorithms. The bound is independent of j because, during its derivation, L is taken to infinity such that all incorrect subtrees become identical in principle. The distribution characterized by the right-hand side of (10) is a *Pareto distribution*, and ρ is therefore named the *Pareto exponent*.

[12] The bound in (10) was first established by Savage for random *tree codes* for some integer value of ρ [41], where random tree codes constitute a super set of convolutional codes. Later, Jelinek [43] extended its validity for random tree codes to real-valued $\rho \ge 1$, satisfying (11). The validity of (10) for random convolutional codes was substantiated by Falconer [42] for $0 < \rho < 1$, and by Hashimoto and Arimoto [44] for $\rho \ge 1$.

Experimental studies indicate that the constant A usually lies between 1 and 10 [1]. The Pareto exponent ρ is uniquely determined by the code rate R using the formula

$$R = \frac{E_0(\rho)}{\rho} \tag{11}$$

for $0 < R < C$, where $E_0(\rho)$ is the *Gallager function* [45 Eq. (5.6.14)] and C is the channel capacity (cf. footnote 7). For example, the Gallager function and the channel capacity for the binary symmetric channel with crossover probability p are respectively given by

$$E_0(\rho) = \rho - (1+\rho)\log_2[p^{1/(1+\rho)} + (1-p)^{1/(1+\rho)}] \tag{12}$$

and

$$C = 1 + p\log_2(p) + (1-p)\log_2(1-p)$$

Equations (11) and (12) together imply that ρ goes to infinity as $R \downarrow 0$, and ρ approaches zero when $R \uparrow C$. Figure 13 gives the Pareto exponents for code rates $R = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$.

An argument converse to (10), due to Jacobs and Berlekamp [15], states that no sequential decoding algorithm can achieve a computational distribution better than the Pareto distribution of (10), given that no decoding error results for convolutional codes. Specifically, they showed that for a Pareto exponent that satisfies (11)

$$\Pr\{C_j \ge \mathcal{N} \mid \text{correct decoding}\} > [1 - o(\mathcal{N})]\mathcal{N}^{-\rho} \tag{13}$$

where $o(\cdot)$ is the little-o function, satisfying $o(x) \to 0$ as $x \to \infty$. The two bounds in (10) and (13) coincide only when $\mathcal{N}$ is sufficiently large.

On the basis of the multiple branching process technique [46], closed form expressions of the average computational complexity of sequential decoding were derived [47–49]. However, these closed form expressions were suited only for small $\mathcal{N}$. Inequalities (10) and (13)

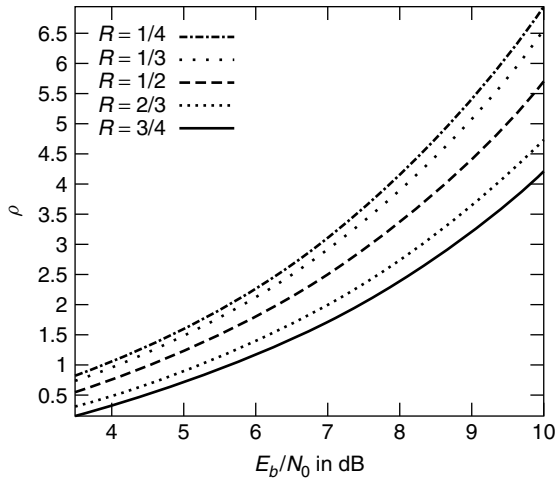


Figure 13. Pareto exponent as a function of E_b/N_0 for a BSC with crossover probability $p = \frac{1}{2}\text{erfc}(\sqrt{E_b/N_0})$, where $\text{erfc}(x) = (2/\sqrt{\pi})\int_x^\infty \exp\{-x^2\}\,dx$ is the complementary error function.

also show that the two bounds are independent of the code constraint length. This observation confirms the claim made in Section 1 that the average computational effort for sequential decoding is in principle independent of the code constraint length.

Inequalities (10) and (13) and the observation that $C_j \geq 1$ jointly yield

$$E[C_j] = \int_1^\infty \Pr\{C_j \geq \mathcal{N}\}\, d\mathcal{N} \leq \int_1^\infty A\mathcal{N}^{-\rho}\, d\mathcal{N}$$

and

$$\begin{aligned} &E[C_j \mid \text{correct decoding}] \\ &\quad = \int_1^\infty \Pr\{C_j \geq \mathcal{N} \mid \text{correct decoding}\}\, d\mathcal{N} \\ &\quad \geq \int_1^\infty [1 - o(\mathcal{N})]\mathcal{N}^{-\rho}\, d\mathcal{N} \end{aligned}$$

Therefore, if the Pareto exponent ρ is greater than unity, $E[C_j]$ is bounded from above. Conversely, if $E[C_j \mid \text{correct decoding}] < \infty$, then $\rho > 1$. Since the probability of correct decoding is very close to unity in most cases of interest, $\rho > 1$ is widely accepted as a sufficient and necessary condition for $E[C_j]$ to be bounded. This result gives rise to the term *computational cutoff rate* $R_0 = E_0(1)$, for sequential decoding.

From (11), $\rho > 1$ if, and only if, $R < R_0 = E_0(1)$, meaning that the cutoff rate R_0 is the largest code rate under which the average complexity of sequential decoding is finite. Thus, sequential decoding becomes computationally implausible once the code rate exceeds the cutoff rate R_0. This theoretical conclusion can be similarly observed from simulations.

Can the computational cutoff rate be improved? The question was answered by a proposal made by Flaconer [42] concerning the use of a hybrid coding scheme. The proposed communication system consists of an $(n_{\text{out}}, k_{\text{out}})$ outer Reed–Solomon encoder, n_{out} parallel $(n_{\text{in}}, 1, m)$ inner convolutional encoders, n_{out} parallel noisy channels, n_{out} sequential decoders for inner convolutional codes, and an algebraic decoder for the outer Reed–Solomon code. These five modules work together in the following fashion. The outer Reed–Solomon encoder encodes k_{out} input symbols into n_{out} output symbols, each of which is b bits long with the last m bits equal to zero. Then, each of n_{out} output symbol is fed into its respective binary $(n_{\text{in}}, 1, m)$ convolutional encoder in parallel, and induces $n_{\text{in}} \times b$ output code bits. Thereafter, these n_{out} $(n_{\text{in}}b)$-bit streams are simultaneously transmitted over n_{out} independent channels, where the n_{out} independent channels may be created by time-multiplexing over a single channel. On receiving n_{out} noise-distorted received vectors of dimension $n_{\text{in}}b$, the n_{out} sequential decoders reproduce the n_{out} output symbols through a sequential codeword search. If any sequential codeword search is not finished before a prespecified time, its output will be treated as an *erasure*. The final step is to use an algebraic Reed–Solomon decoder to regenerate the k_{out} input symbols, based on the n_{out} output symbols obtained from the n_{out} parallel sequential decoders.

The effective code rate of this hybrid system is

$$R_{\text{effective}} = \frac{k_{\text{out}}(b-m)}{n_{\text{out}}n_{\text{in}}b}$$

The largest effective code rate under which the hybrid system is computationally practical has been proved to improve over $E_0(1)$ [42]. Further improvement along the line of code concatenation can be found in Refs. 50 and 51.

The basis of the performance analysis for the aforementioned sequential decoding is random coding, and has nothing to do with any specific properties of the applied code, except the code rate R. Zigangirov [34] proposed to analyze the statistics of C_j for deterministic convolutional codes with an infinite memory order (i.e., $m = \infty$) in terms of recursive equations, and determined that for codes transmitted over BSCs and decoded by the tree-based stack algorithm

$$E[C_j] \leq \frac{\rho}{\rho - 1} 2^{-(-n\alpha+\beta)/(1+\rho)-k} \tag{14}$$

for $R < R_0 = E_0(1)$, where ρ is the Pareto exponent that satisfies (11) and α and β are as defined in the sentence following Eq. (4). Zigangirov's result again suggested that $E[C_j]$ is bounded for $R < R_0$, even when the deterministic convolutional codes are considered. A similar result was also established for the Fano algorithm [34].

Another code-specific estimate of sequential decoding complexity was due to Chevillat and Costello [52,53]. From simulations, they ingeniously deduced that the computational complexity of a sequential decoder is indeed related to the column distance function (CDF) of the applied code. They then established that for a convolutional code transmitted over a BSC with crossover probability p

$$\Pr\{C_j \geq \mathcal{N}\} < A N_d \exp\{-\lambda_1 d_c(\ell) + \lambda_2 \ell\} \tag{15}$$

for $R < 1 + 2p\log_2(p) + (1-2p)\log_2(1-p)$, where A, λ_1, and λ_2 are factors determined by p and code rate R; N_d is the number of length-$[n(\ell+1)]$ paths with Hamming weight equal to $d_c(\ell)$; ℓ is the integer part of $\log_{2^k}\mathcal{N}$; and $d_c(r)$ is the CDF of the applied code. They concluded that a convolutional code with a rapidly increasing CDF can yield a markedly smaller sequential decoding complexity. The outstanding issue is thus how to construct similar convolutional codes for use in sequential decoding. The issue is further explored in Section 11.

Next, the upper bounds on the bit error rate of sequential decoding are introduced. Let P_{S_j} be the probability that a path belonging to the incorrect subtree S_j is decoded as the ultimate output. Then, Chevillat and Costello [53] show that for a specific convolutional code transmitted over a BSC with crossover probability p

$$P_{S_j} < B N_f \exp\{-\gamma d_{\text{free}}\} \tag{16}$$

where B and γ are factors determined by p and code rate R, N_f is the number of code paths in S_j with Hamming weight equal to d_{free}, and d_{free} is the free distance of the applied code. The parameter γ is positive for all convolutional

codes whose free distance exceeds a lower limit determined by p and R. This result indicates that a small error probability for sequential decoding can be obtained by selecting a convolutional code with a large free distance and a small number of codewords with Hamming weight d_{free}. The free distance of a convolutional code generally grows with its constraint length. The bit error rate can therefore be made desirably low when a convolutional code with a sufficiently large constraint length is employed, as the computational complexity of sequential decoding is independent of the code constraint length. However, when a code with a large constraint length is used, the length of the input sequences must also be extended such that the effective code rate $kL/[n(L+m)]$ is closely approximated by code rate $R = k/n$. More discussion of the bit error rates of sequential decoding can be found in the literature [36,40,54].

10. BUFFER OVERFLOW AND SYSTEM CONSIDERATIONS

As already demonstrated by the hardware implementation of the Fano algorithm in Section 7, the *input buffer* at the decoder end for the temporary storage of the received vector is finite. The online decoder must therefore catch up to the input rate of the received vector such that the storage space for obsolete components of the received vector can be freed to store upcoming received components. Whenever an input buffer overflow is encountered, some of the still-in-use content in the input buffer must be forcefully written over by the new input, and the decoder must resynchronize to the new contents of the input buffer; hence input erasure occurs. The overall codeword error P_s of a sequential decoder thus becomes

$$P_s \simeq P_e + P_{\text{erasure}}$$

where P_e is the *undetected word error* under the infinite input buffer assumption and P_{erasure} is the *erasure probability*. For a code with a long constraint length and a practically sized input buffer, P_e is markedly smaller than P_{erasure}, so the overall word error is dominated by the probability of input buffer overflow. In this case, effort is reasonably focused on reducing P_{erasure}. When the code constraint length is only moderately large, a tradeoff between P_e and P_{erasure} must be made. For example, reducing the bucket spacing for the stacked bucket-enabled stack algorithm or lowering the step size for the Fano algorithm results in a smaller P_e, but increases $E[C_j]$ and hence P_{erasure}. The choice of path metrics, as indicated in Section 5, also yields a similar tradeoff between P_e and P_{erasure}. Accordingly, a balance between these two error probabilities must be maintained in practical system design.

The probability of buffer overflow can be analyzed as follows. Let B be the size of the input buffer measured in units of branches; hence the input buffer can store nB bits for an (n, k, m) convolutional code. Also let $1/T$ (bits per second) be the input rate of the received vector. Suppose that the decoder can perform μ branch metric computations in $n \times T$ seconds. Then if over μB branch computations are performed for paths in the jth incorrect subtree, the jth received branch in the input buffer must be written over by the new received branch. To simplify the analysis, assume that the entire buffer is simply reset when a buffer overflow occurs. In other words, the decoder aborts the present codeword search, and immediately begins a new search according to the new received vector. From Eq. (10), the probability of performing more than μB branch computations for paths in S_j is upper-bounded by $A(\mu B)^{-\rho}$. Hence, the erasure probability [41,55] for input sequences of length L is upper-bounded by

$$P_{\text{erasure}} \le L A (\mu B)^{-\rho} \tag{17}$$

Taking $L = 1000, A = 5, \mu = 10, B = 10^5$, and $R = \frac{1}{2}$ yields $\rho = 1.00457$ and $P_{\text{erasure}} \le 4.694 \times 10^{-3}$.

Three actions can be taken to avoid eliminating the entire received vector when the input buffer overflow occurs: (1) just inform the outer mechanism that an input erasure occurs, and let the outer mechanism take care of the decoding of the respective input sequences; (2) estimate the respective input sequences by a function mapping from the received vector to the input sequence; and (3) output the tentative decoding results obtained thus far. The previous section already demonstrated an example of the first action using the hybrid coding system. The second action can be taken whenever an input sequence to a convolutional encoder can be recovered from its codewords through a function mapping. Two typical convolutional codes whose input sequence can be directly mapped from the codewords are the systematic code and the quick-lookin code (to be introduced in the next section). A decoder can also choose to output a tentative decoded input sequence if the third action is taken. A specific example for the third action, named the *multiple stack algorithm*, is introduced in the next paragraph.

In 1977, Chevillat and Costello [56] proposed a *multiple stack algorithm* (MSA), which eliminated entirely the possibility of erasure in the sense that the decoded output is always based on the codeword search. The MSA, as its name implies, acts exactly like the stack algorithm except that it accommodates multiple stacks. During decoding, the MSA first behaves as the stack algorithm by using only its main stack of size z_{main}. When the main stack reaches its limit, the best t paths in the main stack are transferred to a smaller second stack with size $z \ll z_{\text{main}}$. Then the decoding process proceeds just like the stack algorithm, but now using the second stack. If a path reaches the end of the search tree before the second stack is filled, then the path is stored as a tentative decision, and the second stack is eliminated. The MSA then returns to the main stack that has t vacancy spaces for new paths because the top t paths have been removed. If another path reaches the end of the search tree before the main stack is filled again, the decoder compares its path metric with that of the current tentative decision, and outputs the one with larger metric, and stops. Now in case the second stack is filled before a code path is located, a third stack with the same size z is created such that the top t paths in the second stack are transferred to it. The codeword search process then proceeds over the

third stack until either a tentative decision can be made or a new stack needs to be created. Additional stacks of size z are formed whenever necessary. The decoder always compares the newly located code path with the previous tentative decision, and retains the better one. With some properly selected system parameters including z_{main}, z, t, and input buffer size, the MSA guarantees that whenever an input erasure occurs, a tentative decision is always ready for output [56]. Simulation results show that even though the stack maintenance of the MSA is more complex than the stack algorithm, the bit error rate of the former is much lower than that of the latter (see. Fig. 14). Further improvements of the MSA can be found in Refs. 57 and 58.

11. CODE CONSTRUCTION FOR SEQUENTIAL DECODING

A rapid column distance growth in CDF has been conjectured to help the early rejection of incorrect paths for sequential decoding algorithms [59]; this conjecture was later substantiated by Chevillat and Costello both empirically [52] and analytically [53]. In an effort to construct a good convolutional code for sequential decoding, Johannesson proposed [60] that the *code distance profile*, defined as $\{d_c(1), d_c(2), \ldots, d_c(m+1)\}$, can be used, instead of the entire code distance function $d_c(\cdot)$, as a "criterion" for good code construction. His suggestion greatly reduced the number of possible code designs that must be investigated.

A code $\mathcal{C}$ is said to have a better distance profile than another code $\mathcal{C}'$ with the same code rate and memory order, if there exists ℓ with $1 \le \ell \le m+1$ such that $d_c(j) = d'_c(j)$ for $1 \le j \le \ell - 1$ and $d_c(\ell) > d'_c(\ell)$, where $d_c(\cdot)$ and $d'_c(\cdot)$ are the CDFs of codes $\mathcal{C}$ and $\mathcal{C}'$, respectively. In other words, a code with a better distance profile exhibits a faster initial column distance growth in its CDF. A code is said to have an *optimal distance profile*, and is called an ODP code, if its distance profile is superior to that of any other code with the same code rate and memory order.

The searching of ODP codes was extensively studied by Johannesson and Passke [60–63]. They found that the ODP condition could be imposed on a half-rate ($R = \frac{1}{2}$) short constraint length code without penalty in the code free distance [60]; That is, the half-rate ODP code with a short constraint length can also be the code with the largest free distance of all codes with the same code rate and memory order. Tables 1 and 2 list the half-rate ODP codes for systematic codes and nonsystematic codes, respectively. Comprehensive information on ODP codes can be found in Ref. 34. These tables notably reveal that the free distance of a systematic ODP code is always inferior to that of a nonsystematic ODP code with the same memory order.

Employing ODP codes, while notably reduces the number of metric computations for sequential decoding, does not ensure erasure-free performance in practical implementation. If an erasure-free sequential decoding algorithm such as the MSA cannot be adopted due to certain practical considerations, the decoder must still force an immediate decision by just taking a *quick look* at the received vector, once input erasure occurs.

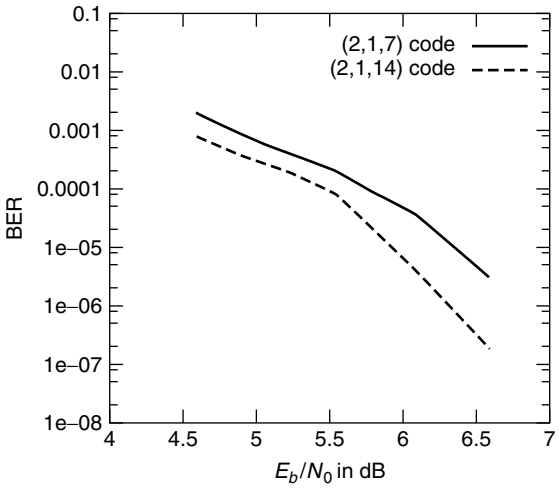


Figure 14. The MSA bit error rates (BER) for $(2, 1, 7)$ and $(2, 1, 14)$ convolutional codes with an input sequence of length 64. The system parameters for $(2, 1, 7)$ and $(2, 1, 14)$ convolutional codes are $(z_{\text{main}}, z, t) = (1024, 11, 3)$ and $(z_{\text{main}}, z, t) = (2900, 11, 3)$, respectively. The input erasure is emulated by an upper limit on branch metric computations C_{lim}, which is 1700 and 3300 for $(2, 1, 7)$ and $(2, 1, 14)$ convolutional codes, respectively. (Reproduced in part from Fig. 3 in Ref. 57).

Table 1. List of Code Rate $R = \frac{1}{2}$ Systematic Codes with Optimal Distance Profile

m	g_2	d_{free}
1	6	3
2	7	4
3	64	4
4	66	5
5	73	6
6	674	6
7	714	6
8	671	7
9	7,154	8
10	7,152	8
11	7,153	9
12	67,114	9
13	67,116	10
14	71,447	10
15	671,174	10
16	671,166	12
17	671,166	12
18	6,711,454	12
19	7,144,616	12
20	7,144,761	12
21	71,447,614	12
22	71,446,166	14
23	67,115,143	14
24	714,461,654	15
25	671,145,536	15
26	714,476,053	16
27	7,144,760,524	16
28	7,144,616,566	16
29	7,144,760,535	18
30	67,114,543,064	16
31	67,114,543,066	18

Source: Ref. 34.

Table 2. List of Code Rate $R = \frac{1}{2}$ Nonsystematic Codes with Optimal Distance Profile

m	g_1	g_2	d_{free}
1	6	4	3
2	7	5	5
3	74	54	6
4	62	56	7
5	77	45	8
6	634	564	10
7	626	572	10
8	751	557	12
9	7,664	5,714	12
10	7,512	5,562	14
11	6,643	5,175	14
12	63,374	47,244	15
13	45,332	77,136	16
14	65,231	43,677	17
15	727,144	424,374	18
16	717,066	522,702	19
17	745,705	546,153	20
18	6,302,164	5,634,554	21
19	5,122,642	7,315,626	22
20	7,375,407	4,313,045	22
21	67,520,654	50,371,444	24
22	64,553,062	42,533,736	24
23	55,076,157	75,501,351	26
24	744,537,344	472,606,614	26
25	665,041,116	516,260,772	27

Source: Ref. 34.

This seems to suggest that a systematic ODP code is preferred, even if it has a smaller free distance than its nonsystematic ODP counterpart. In such case, the deficiency on the free distance of the systematic ODP codes can be compensated for by selecting a larger memory order m. However, when a convolutional code with large m is used, the length of the input sequences must be proportionally extended; otherwise the effective code rate cannot be well approximated by the convolutional code rate, and the performance to some extent degrades. This effect motivates the attempt to construct a class of nonsystematic ODP codes with the "quick look" property and a free distance superior to that of systematic codes.

Such a class of nonsystematic codes has been developed by Massey and Costello, called the *quick-lookin* (QLI) convolutional codes [59]. The generator polynomials of these half-rate QLI convolutional codes differ only in the second coefficient. Specifically, their generator polynomials satisfy $g_1(x) = g_2(x) + x$, where addition of coefficients is based on modulo-2 operation, allowing the decoder to recover the input sequence $\mathbf{u}(x)$ by summing the two output sequences $\mathbf{v}_1(x)$ and $\mathbf{v}_2(x)$ as

$$x \cdot \mathbf{u}(x) = \mathbf{v}_1(x) + \mathbf{v}_2(x) \tag{18}$$

If p is the individual bit error probability in the codeword $\mathbf{v}$, then the bit error probability due to recovering information sequence $\mathbf{u}$ from $\mathbf{v}$ through (18) is shown to be approximately $2p$ [59]. Table 3 lists the QLI ODP convolutional codes [34].

Table 3. List of Code Rate $R = \frac{1}{2}$ QLI Codes with Optimal Distance Profile

m	g_1	d_{free}
2	7	5
3	74	6
4	76	6
5	75	8
6	714	8
7	742	9
8	743	9
9	7,434	10
10	7,422	11
11	7,435	12
12	74,044	11
13	74,046	13
14	74,047	14
15	740,464	14
16	740,462	15
17	740,463	16
18	7,404,634	16
19	7,404,242	15
20	7,404,155	18
21	74,041,544	18
22	74,042,436	19
23	74,041,567	19
24	740,415,664	20
25	740,424,366	20
26	740,424,175	22
27	7,404,155,634	22
28	7,404,241,726	23
29	7,404,154,035	24
30	74,041,567,514	23
31	74,041,567,512	25

Source: Ref. 34.

12. CONCLUSIONS

Although sequential decoding has a longer history than maximum-likelihood decoding based on the Viterbi algorithm, its practical applications are not as popular, because the highly repetitive "pipeline" nature of the Viterbi decoder makes it very suitable for hardware implementation. Furthermore, a sequential decoder usually requires a longer decoding delay (defined as the time between the receipt of a received branch and the output of its respective decoding decision) than a Viterbi decoder. Generally, the decoding delay of a sequential decoder for an (n, k, m) convolutional code is around $n \times B$, where B is the number of received branches that an input buffer can accommodate. Yet, the decoding delay of a Viterbi decoder can be made a small multiple, often ranging from 5 to 10, of $n \times m$. On the other hand, Refs. 64 and 65 showed that sequential decoding is highly sensitive to the channel parameters such as an inaccurate estimate of channel SNR and an incomplete compensation of phase noise. The Viterbi algorithm, however, proved to be robust for imperfect channel identification, again securing the superiority of the Viterbi decoder in practical applications.

Nevertheless, there are certain situations that the sequential decoding fits well, especially in decoding convolutional codes having a large constraint length.

In addition, the sequential decoder can send a timely retransmission request by detecting the occurrence of an input buffer overflow [66]. Sequential decoding has attracted some attention in the field of mobile communications [67] in which a demand of low bit error rate is required. Such applications are beyond the scope of this article, and interested readers can refer to the literature [1,68,69].

Acknowledgments

Professor Marc P. C. Fossorier of the University of Hawaii and Professor John G. Proakis are appreciated for their careful reviews and valuable comments. Mr. Tsung-Ju Wu and Mr. Tsung-Chi Lin are also appreciated for preparing the figures and checking the examples in this manuscript.

BIOGRAPHIES

Yunghsiang S. Han was born in Taipei, Taiwan, on April 24, 1962. He received the B.S. and M.S. degrees in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 1984 and 1986, respectively, and the Ph.D. degree from the school of Computer and Information Science, Syracuse University, Syracuse, New York, in 1993. From 1986 to 1988 he was a Lecture at Ming-Hsin Engineering College, Hsinchu, Taiwan. He was a Teaching Assistant from 1989 to 1992 and from 1992 to 1993 a Research Assistant in the School of Computer and Information Science, Syracuse University. He is a recipient of the 1994 Syracuse University Doctoral Prize. From 1993 to 1997 he was an Associate Professor in the Department of Electronic Engineering at Hua Fan College of Humanities and Technology, Taipei Hsien, Taiwan. He is now with the Department of Computer Science and Information Engineering at National Chi Nan University, Nantou, Taiwan. He was promoted to full Professor in 1998. His research interests are in error-control coding and fault-tolerant computing.

Po-Ning Chen received the B.S. and M.S. degrees in electrical engineering from the National Tsing-Hua University in Taiwan in 1985 and 1987, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland at College Park (USA) in 1994. From 1985 to 1987, he was with Image Processing Laboratory in the National Tsing-Hua University, where he worked on the recognition of Chinese characters. During 1989, he was with StarTech Engineering Inc., where he focused on the development of fingerprint recognition systems. After receiving the Ph.D. degree in 1994, he joined Wan Ta Technology Inc. as a Vice General Manager, conducting several projects on Point-of-Sale systems. In 1995, he joined the research staff at the Advanced Technology Center, Computer and Communication Laboratory, Industrial Technology Research Institute in Taiwan, where he led a project on Java-based Network Managements. Since 1996, he has been an Associate Professor in the Department of Communications Engineering at the National Chiao-Tung University, Taiwan, and became a full Professor in 2001. Dr. Chen received the 2000 Young Scholar Paper Award from Academia Sinica, Taiwan. His areas of interests are information and coding theory, large deviation theory, and distributed detection.

BIBLIOGRAPHY

1. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
2. J. M. Wozencraft, Sequential decoding for reliable communications, *IRE Nat. Conv. Rec.* **5**(Pt. 2): 11–25 (1957).
3. J. M. Wozencraft and B. Reiffen, *Sequential Decoding*, MIT Press, Cambridge, MA, 1961.
4. R. M. Fano, A heuristic discussion of probabilistic decoding, *IEEE Trans. Inform. Theory* **IT-9**(2): 64–73 (April 1963).
5. K. Sh. Zigangirov, Some sequential decoding procedures, *Probl. Peredachi Inform.* **2**: 13–25 (1966).
6. F. Jelinek, A fast sequential decoding algorithm using a stack, *IBM J. Res. Dev.* **13**: 675–685 (Nov. 1969).
7. N. J. Nilsson, *Principle of Artificial Intelligence*, Tioga, Palo Alto, CA, 1980.
8. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
9. A. J. Viterbi, Error bound for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**(2): 260–269 (April 1967).
10. J. L. Massey, *Threshold Decoding*, MIT Press, Cambridge, MA, 1963.
11. G. D. Forney, Jr., Review of random tree codes, in *Appendix A. Study of Coding Systems Design for Advanced Solar Missions*, NASA Contract NAS2-3637, Codex Corp., Dec. 1967.
12. J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, Addison-Wesley, Reading, MA, 1984.
13. Y. S. Han, C. R. P. Hartmann, and C.-C. Chen, Efficient priority-first search maximum-likelihood soft-decision decoding of linear block codes, *IEEE Trans. Inform. Theory* **39**(5): 1514–1523 (Sept. 1993).
14. Y. S. Han, A new treatment of priority-first search maximum-likelihood soft-decision decoding of linear block codes, *IEEE Trans. Inform. Theory* **44**(7): 3091–3096 (Nov. 1998).
15. I. M. Jacobs and E. R. Berlekamp, A lower bound to the distribution of computation for sequential decoding, *IEEE Trans. Inform. Theory* **IT-13**(2): 167–174 (April 1967).
16. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
17. J. L. Massey, Variable-length codes and the fano metric, *IEEE Trans. Inform. Theory* **IT-18**(1): 196–198 (Jan. 1972).
18. E. A. Bucher, Sequential decoding of systematic and nonsystematic convolutional codes with arbitrary decoder bias, *IEEE Trans. Inform. Theory* **IT-16**(5): 611–624 (Sept. 1970).
19. F. Jelinek, Upper bound on sequential decoding performance parameters, *IEEE Trans. Inform. Theory* **IT-20**(2): 227–239 (March 1974).
20. Y. S. Han, P.-N. Chen, and M. P. C. Fossorier, A generalization of the fano metric and its effect on sequential decoding using a stack, *IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, 2002.
21. K. Sh. Zigangirov, *private communication*, Feb. 2002.

22. J. M. Geist, An empirical comparison of two sequential decoding algorithms, *IEEE Trans. Commun. Technol.* **COM-19**(4): 415–419 (Aug. 1971).

23. D. Haccoun and M. J. Ferguson, Generalized stack algorithms for decoding convolutional codes, *IEEE Trans. Inform. Theory* **IT-21**(6): 638–651 (Nov. 1975).

24. J. B. Anderson and S. Mohan, Sequential coding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**(2): 169–176 (Feb. 1984).

25. S. Mohan and J. B. Anderson, Computationally optimal metric-first code tree search algorithms, *IEEE Trans. Commun.* **COM-32**(6): 710–717 (June 1984).

26. D. E. Knuth, *The Art of Computer Programming*, Vol. III: Sorting and Searching, Addison-Wesley, Reading, MA, 1973.

27. P. Lavoie, D. Haccoun, and Y. Savaria, A systolic architecture for fast stack sequential decoders, *IEEE Trans. Commun.* **42**(5): 324–335 (May 1994).

28. S. Kallel and K. Li, Bidirectional sequential decoding, *IEEE Trans. Inform. Theory* **43**(4): 1319–1326 (July 1997).

29. J. M. Geist, Search properties of some sequential decoding algorithms, *IEEE Trans. Inform. Theory* **IT-19**(4): 519–526 (July 1973).

30. G. D. Forney, Jr. and E. K. Bower, A high-speed sequential decoder: prototype design and test, *IEEE Trans. Commun. Technol.* **COM-19**(5): 821–835 (Oct. 1971).

31. I. M. Jacobs, Sequential decoding for efficient communication from deep space, *IEEE Trans. Commun. Technol.* **COM-15**(4): 492–501 (August 1967).

32. J. W. Layland and W. A. Lushbaugh, A flexible high-speed sequential decoder for deep space channels, *IEEE Trans. Commun. Technol.* **COM-19**(5): 813–820 (Oct. 1978).

33. M. Shimada, T. Todoroki, and K. Nakamura, Development of variable-rate sequential decoder LSI, *IEEE Int. Conf. Communications*, 1989, pp. 1241–1245.

34. R. Johannesson and K. Sh. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, Piscataway, NJ, 1999.

35. J. Geist, *Algorithmic Aspects of Sequential Decoding*, Ph.D. thesis, Dept. Electrical Engineering, Univ. Notre Dame, Notre Dame, IN, 1970.

36. G. D. Forney, Jr., Convolutional codes III: Sequential decoding, *Inform. Control* **25**: 267–269 (July 1974).

37. N. Bélanger, D. Haccoun, and Y. Savaria, A multiprocessor architecture for multiple path stack sequential decoders, *IEEE Trans. Commun.* **42**(2–4): 951–957 (Feb.–April 1994).

38. Y. S. Han, P.-N. Chen, and H.-B. Wu, A maximum-likelihood soft-decision sequential decoding algorithm for binary convolutional codes, *IEEE Trans. Commun.* **50**(2): 173–178 (Feb. 2002).

39. J. Snyders and Y. Be'ery, Maximum likelihood soft decoding of binary block codes and decoders for the golay codes, *IEEE Trans. Inform. Theory* 963–975 (Sept. 1989).

40. H. L. Yudkin, *Channel State Testing in Information Decoding*, Ph.D. thesis, MIT, Cambridge, Mass, 1964.

41. J. E. Savage, Sequential decoding-the computation problem, *Bell Syst. Tech. J.* **45**: 149–175 (Jan. 1966).

42. D. D. Falconer, A hybrid decoding scheme for discrete memoryless channels, *Bell Syst. Tech. J.* **48**: 691–728 (March 1969).

43. F. Jelinek, An upper bound on moments of sequential decoding effort, *IEEE Trans. Inform. Theory* **IT-15**(1): 140–149 (Jan. 1969).

44. T. Hashimoto and S. Arimoto, Computational moments for sequential decoding of convolutional codes, *IEEE Trans. Inform. Theory* **IT-25**(5): 584–591 (1979).

45. R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

46. W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley, New York, 1970.

47. R. Johannesson, On the distribution of computation for sequential decoding using the stack algorithm, *IEEE Trans. Inform. Theory* **IT-25**(3): 323–332 (May 1979).

48. D. Haccoun, A branching process analysis of the average number of computations of the stack algorithm, *IEEE Trans. Inform. Theory* **IT-30**(3): 497–508 (May 1984).

49. R. Johannesson and K. Sh. Zigangirov, On the distribution of the number of computations in any finite number of subtrees for the stack algorithm, *IEEE Trans. Inform. Theory* **IT-31**(1): 100–102 (Jan. 1985).

50. F. Jelinek and J. Cocke, Bootstrap hybrid decoding for symmetrical binary input channel, *Inform. Control* **18**: 261–298 (April 1971).

51. O. R. Jensen and E. Paaske, Forced sequence sequential decoding: A concatenated coding system with iterated sequential inner decoding, *IEEE Trans. Commun.* **46**(10): 1280–1291 (Oct. 1998).

52. P. R. Chevillat and D. J. Costello, Jr., Distance and computation in sequential decoding, *IEEE Trans. Commun.* **COM-24**(4): 440–447 (April 1978).

53. P. R. Chevillat and D. J. Costello, Jr., An analysis of sequential decoding for specific time-invariant convolutional codes, *IEEE Trans. Inform. Theory* **IT-24**(4): 443–451 (July 1978).

54. A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, New York, 1979.

55. J. E. Savage, The distribution of the sequential decoding computation time, *IEEE Trans. Inform. Theory* **IT-12**(2): 143–147 (April 1966).

56. P. R. Chevillat and D. J. Costello, Jr., A multiple stack algorithm for erasurefree decoding of convolutional codes, *IEEE Trans. Commun.* **COM-25**(12): 1460–1470 (Dec. 1977).

57. H. H. Ma, The multiple stack algorithm implemented on a zilog z-80 microcomputer, *IEEE Trans. Commun.* **COM-28**(11): 1876–1882 (Nov. 1980).

58. K. Li and S. Kallel, A bidirectional multiple stack algorithm, *IEEE Trans. Commun.* **47**(1): 6–9 (Jan. 1999).

59. J. L. Massey and D. J. Costello, Jr. Nonsystematic convolutional codes for sequential decoding in space applications, *IEEE Trans. Commun. Technol.* **COM-19**(5): 806–813 (Oct. 1971).

60. R. Johannesson, Robustly optimal rate one-half binary convolutional codes, *IEEE Trans. Inform. Theory* **IT-21**(4): 464–468 (July 1975).

61. R. Johannesson, Some long rate one-half binary convolutional codes with an optimal distance profile, *IEEE Trans. Inform. Theory* **IT-22**(5): 629–631 (Sept. 1976).

62. R. Johannesson, Some rate 1/3 and 1/4 binary convolutional codes with an optimal distance profile, *IEEE Trans. Inform. Theory* **IT-23**(2): 281–283 (March 1977).

63. R. Johannesson and E. Paaske, Further results on binary convolutional codes with an optimal distance profile, *IEEE Trans. Inform. Theory* **IT-24**(2): 264–268 (March 1978).

64. J. A. Heller and I. W. Jacobs, Viterbi decoding for satellite and space communication, *IEEE Trans. Commun. Technol.* **COM-19**(5): 835–848 (Oct. 1971).

65. I. M. Jacobs, Practical applications of coding, *IEEE Trans. Inform. Theory* **IT-20**(3): 305–310 (May 1974).

66. A. Drukarev and Jr. D. J. Costello, Hybrid ARQ error control using sequential decoding, *IEEE Trans. Inform. Theory* **IT-29**(4): 521–535 (July 1983).

67. P. Orten and A. Svensson, Sequential decoding in future mobile communications, *Proc. PIMRC '97*, 1997 Vol. 3, pp. 1186–1190.

68. S. Kallel, Sequential decoding with an efficient incremental redundancy ARQ strategy, *IEEE Trans. Commun.* **40**(10): 1588–1593 (Oct. 1992).

69. P. Orten, Sequential decoding of tailbiting convolutional codes for hybrid ARQ on wireless channels, *Proc. IEEE Vehicular Technology Conf.*, 1999, Vol. 1, 279–284.

SERIALLY CONCATENATED CODES AND ITERATIVE ALGORITHMS

S. Benedetto
G. Montorsi
Politecnico di Torino
Torino (Turin), Italy

1. INTRODUCTION

In his goal to find a class of codes whose probability of error decreased exponentially at rates less than capacity, while decoding complexity increased only algebraically, Dave Forney [1] arrived at a solution consisting of the coding structure known as *concatenated code*. It consists of the cascade of an *inner* code and an *outer* code, which, in Forney's approach, would be a relatively short inner code (typically, a convolutional code) admitting simple maximum-likelihood soft decoding, and a long high-rate algebraic outer code (for most applications a nonbinary Reed–Solomon code) equipped with a powerful algebraic error correction decoding algorithm, possibly using reliability information from the inner decoder.

Initially motivated only by theoretical research interests, concatenated codes have since then evolved as a standard for those applications where very high coding gains are needed, such as (deep-)space applications, digital television broadcasting, compact disk players, and many others. Alternative solutions for concatenation have also been studied, such as using a trellis-coded modulation scheme as inner code [2], or concatenating two convolutional codes [3]. In the latter case, the inner Viterbi decoder employs a soft-output decoding algorithm to provide soft-input decisions to the outer Viterbi decoder. An interleaver was also proposed between the two encoders to separate bursts of errors produced by the inner decoder.

We find then, in a "classical" concatenated coding scheme, the main ingredients that formed the basis for the invention of "Turbo codes" [4], namely two, or more, *constituent* codes (CCs) and an *interleaver*. The novelty of Turbo codes, however, consists of the way they use the interleaver, which is embedded into the code structure to form an overall concatenated code with very large block length, and in the proposal of a parallel concatenation to achieve a higher rate for given rates of CCs. The latter advantage is obtained using systematic CCs and not transmitting the information bits entering the second encoder. In the following, we will refer to turbo codes as *parallel concatenated codes* (PCCs). The so-obtained codes have been shown to yield very high coding gains at bit error probabilities in the range 10^{-5}–10^{-7}; in particular, low bit error probabilities can be obtained at rates well beyond the channel cutoff rate, which had been regarded for long time as the "practical" capacity. Quite remarkably, this performance can be achieved by a relatively simple iterative decoding technique whose computational complexity is comparable to that needed to decode the two CCs.

After the invention of PCCs, other forms of code concatenations with interleaver have been proposed, like the serial concatenation [5] and hybrid concatenations [6,7]. This article deals with serial concatenations with interleaver in a wider sense; that is, we apply the iterative algorithm introduced in 1998 [5] for decoding serially concatenated codes (SCCs) also to systems where the two concatenated blocks represent different functions, such as intersymbol interference channel, a modulator, or a multiuser combiner.

In the first part we consider the serial concatenation of interleaved codes or *serially concatenated codes* (SCCs), called SCBC or SCCC according to the nature of CCs, that can be block (SCBC) or convolutional codes (SCCC). For this class of codes, we give analytical upper bounds to the performance of a maximum-likelihood (ML) decoder, present design guidelines leading to the optimal choice of CCs that maximize the so-called *interleaver gain* and the asymptotic code performance, and present the iterative decoding algorithm yielding results close to capacity limits with limited decoding complexity.

In the second part of the paper we extend the concept of serial concatenation, beyond the case of two codes, to systems in which two functional blocks with memory are cascaded and can be separated by an interleaver. This extension will prove that the "Turbo" principle applied to serial concatenation can have a wide variety of applications in the field of digital transmission.

Throughout the paper, a semi-tutorial approach has been adopted, and all but strictly necessary algebra and mathematical subtleties avoided. In this, we tried to stick to Einstein motto: *Everything should be made as simple as possible, but not simpler*, so as to highlight the meaning of the main properties/results while addressing the interested reader to the appropriate references.

2. SERIAL CONCATENATIONS

Figure 1 is a block diagram of a fairly general serial concatenation of three modules, M1, M2, and M3, with two interleavers I1 and I2 separating each pair of modules.

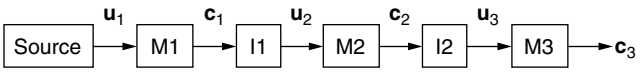


Figure 1. A general serially concatenated structure with three modules and two interleavers.

The symbols **u** and **c** may denote semiinfinite sequences, when the transmitted information flow does not possess a frame structure, or vectors containing a finite number of symbols when the information flow is delivered in a frame-by-frame basis. In the framed case, the interleavers must be *block* interleavers [8]. As we will see in Section 6, a structure like the one depicted in Fig. 1 can represent numerous systems and applications, beyond the basic one in which all modules represent encoders. In the next two sections, however, we will show how to analyze and design a serial concatenation in the particular, yet important, case of two encoders separated by one interleaver. This restriction will simplify the presentation, yet it will lead to conclusions that are applicable to a broader set of systems.

3. ANALYSIS OF SERIALLY CONCATENATED CODES WITH INTERLEAVERS

Consider the serial concatenation shown in Fig. 2. It is formed by the *outer* encoder C_o with rate $R_c^o = k/p$, and the *inner* encoder C_i with rate $R_c^i = p/n$, joined by an interleaver of size N bits, generating a serially concatenated code (SCC) C_S with rate $R_c^S = R_c^o \times R_c^i = k/n$. N will be assumed to be an integer multiple[1] of p. The outer and inner encoders are the constituent codes (CCs) of the SCC. We consider here block constituent codes, in which each code word is formed in a memoryless fashion by the encoder. This is the case when both CCs are block encoders, or when they are *terminated* [9, Chap. 11] convolutional encoders.

For large interleaver sizes, an SCC cannot be decoded as a single code using maximum-likelihood (ML) decoding [8]. Instead, as we will see later, the decoder will use a simpler, suboptimum iterative decoding algorithm based on two soft CC decoders. On the other hand, the analysis seems only be possible for ML decoding, using upper bounds that yield a great insight into the code behavior and provide valuable tools for the code design.

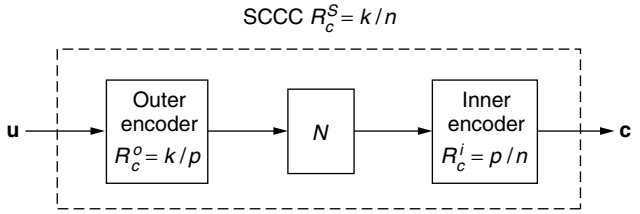


Figure 2. The serial concatenation of two encoders and one interleaver.

[1] Actually, this constraint is not necessary. We can choose in fact inner and outer codes of any rates $R_c^i = k_i/n_i$ and $R_c^o = k_o/n_o$, constraining the interleaver to be an integer multiple of the minimum common multiple of n_o and k_i, that is, $N = K \cdot mcm(n_o, k_i)$. This generalization, though, leads to more complicated expressions and is not considered in the following.

3.1. A Union Bound to the Code Error Probabilities

Upper bounds to the *word*[2] P_w and *bit*[3] P_b error probabilities for an (n, k) linear block code based on the union bound [9] under maximum-likelihood soft decoding for binary PSK (or binary PAM) transmission over an additive white Gaussian noise channel with two-sided noise power spectral density $N_0/2$ are given by

$$P_w \le \frac{1}{2} \sum_{d=d_{\min}}^{n} \sum_{w=1}^{k} A_{w,d} \operatorname{erfc}\left(\sqrt{\frac{dR_cE_b}{N_0}}\right)$$
$$P_b \le \frac{1}{2} \sum_{d=d_{\min}}^{n} \sum_{w=1}^{k} \frac{w}{k} A_{w,d} \operatorname{erfc}\left(\sqrt{\frac{dR_cE_b}{N_0}}\right) \quad (1)$$

where R_c is the code rate, E_b is the energy per *information* bit, and $A_{w,d}$ are the *input–output* coefficients of the encoder, representing the number of codewords with weight d generated by information words of weight w.

Knowledge of the coefficients $A_{w,d}$ is sufficient to evaluate the upper bound to both word and bit error probabilities, and thus we need to evaluate them for the SCC C_S, assuming that we know those of the CCs.

If N in Fig. 2 is low, we can compute the coefficients $A_{w,d}$ by letting each individual information word with weight w be first encoded by the outer encoder C_o and then, after the p bits of the outer codeword have been permuted by the interleaver, be encoded by the inner encoder C_i originating an inner codeword with a certain weight. After repeating this procedure for all the information words with weight w, we should count the inner codewords with weight d, and their number would be the value of $A_{w,d}$.

When N is large the previous operation becomes too complex, and we must resort to a different approach. As thoroughly described elsewhere [5,10], a crucial step in the analysis consists in replacing the actual interleaver that performs a permutation of the N input bits with an abstract interleaver called *uniform interleaver*, defined as a probabilistic device that maps a given input word of weight l into all distinct $\binom{N}{l}$ permutations of it with equal probability $P = 1/\binom{N}{l}$. Use of the uniform interleaver permits the computation of the "average" performance of the SCC, intended as the expectation of the performance of SCCs using the same CCs, taken over the set of all interleavers of a given size. A theorem proved in Ref. 10 guarantees the meaning fullness of the average performance, in the sense that there will always be, for each value of the signal-to-noise ratio, at least one particular interleaver yielding performance better than or equal to that of the uniform interleaver.

[2] The word error probability is defined as the probability that the decoder chooses an incorrect code word, *i.e.*, a code word different from the transmitted one.

[3] The bit error probability is defined as the probability that the decoder delivers an incorrect information bit to the user.

Define as $A_{w,d}^{C_S}$, $A_{w,l}^{C_o}$ and $A_{w,l}^{C_i}$ the input–output coefficients of the SCC C_S, outer and inner encoders, respectively; exploiting the properties of the uniform interleaver, which transforms a codeword of weight l at the output of the outer encoder into all its distinct $\binom{N}{l}$ permutations, we obtain [5]

$$A_{w,d}^{C_s} = \sum_{l=0}^{N} \frac{A_{w,l}^{C_o} \times A_{l,h}^{C_i}}{\binom{N}{l}} \tag{2}$$

This equation (2) permits an easy computation of the input-output coefficients of the SCC, and, together with (1), yields the upper bounds to word and bit error probabilities.

Example 1. Consider the rate-$\frac{1}{3}$ (nominal rate) SCC of Fig. 3 obtained concatenating a terminated rate-$\frac{1}{2}$, 4-state, recursive systematic encoders with generator matrix

$$G_o(Z) = \left[1, \frac{1+Z^2}{1+Z+Z^2}\right] \tag{3}$$

and as inner encoder a terminated 4-state, rate-$\frac{2}{3}$, recursive convolutional encoder with generator matrix

$$G_i(Z) = \begin{bmatrix} 1, 0, \dfrac{1+Z^2}{1+Z+Z^2} \\ 0, 1, \dfrac{1+Z}{1+Z+Z^2} \end{bmatrix} \tag{4}$$

Using the uniform interleaver concept, we have obtained the union bounds to the word and bit error probabilities shown in Figs. 4 and 5. The curves refer to five values of the interleaver size, corresponding to information block sizes $k = 10, 20, 40, 80, 160$. The curves show that increasing the interleaver size yields an increasing coding gain, which is more pronounced for the bit error probability. This phenomenon is known as *interleaving gain*.

Previous analysis can be generalized to n cascaded encoders separated by $n-1$ interleavers. Indeed, the ML performance of the overall encoder improves uniformly when increasing the number of CCs. On the other hand, the suboptimum iterative decoding algorithm that must be used to decode this kind of codes for complexity reasons suffers from a *bootstrap* effect that becomes more and more relevant as we increase the number of CCs. The case of three encoders, which has been proposed and analyzed [11] under the name of double serially concatenated code, seems to be the farthest one can go in practice.

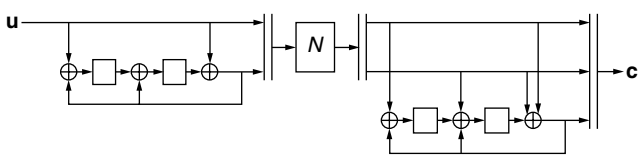


Figure 3. SCCC encoder of Example 1.

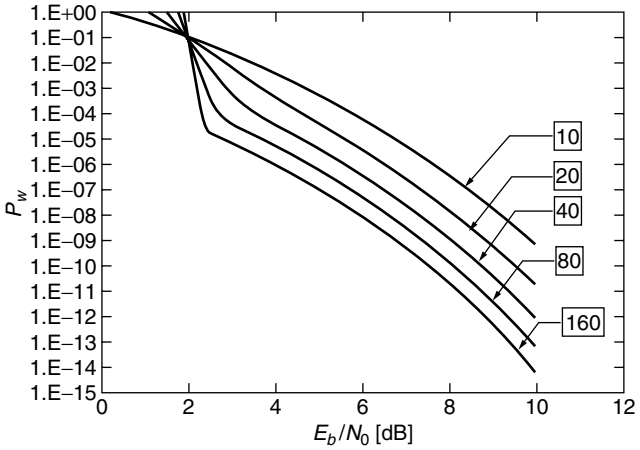


Figure 4. Word error probability bounds for the SCC of Example 1 employing as constituent encoders block codes obtained from 4-state convolutional encoders through trellis termination. The interleaver is uniform and corresponds to information words with size $k = 10, 20, 40, 80, 160$.

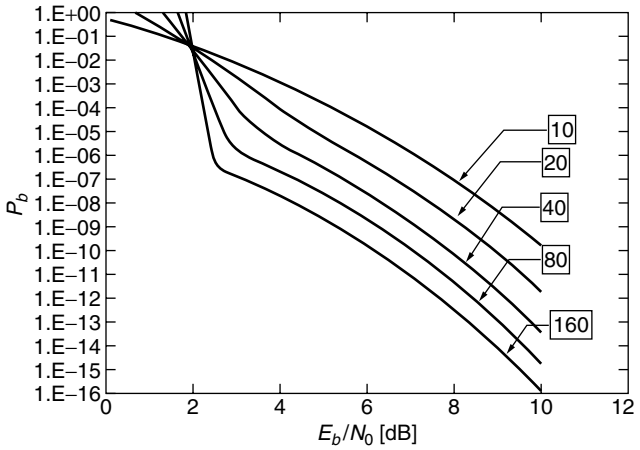


Figure 5. Bit error probability bounds for the SCC of Example 1 employing as constituent encoders block codes obtained from 4-state convolutional encoders through trellis termination. The interleaver is uniform and corresponds to information words with size $K = 10, 20, 40, 80, 160$.

4. DESIGN OF SERIALLY CONCATENATED CODES WITH INTERLEAVER

The error probability performance of concatenated codes with interleaver (and, in particular, of serially concatenated codes) under iterative decoding are invariably represented by curves like the ones depicted in Fig. 6. We can identify three different regions, for increasing values of the signal-to-noise ratio. The first one is the *non-convergence* region, where the error probability keeps high, nearly constant values. At a certain point, the *convergence abscissa*, the curves start a rather steep descent down to medium–low values of the error probability (the *waterfall region*). Finally, in the third region (the *error floor* region), the slope of the curves decreases significantly, and performance improvement are paid with

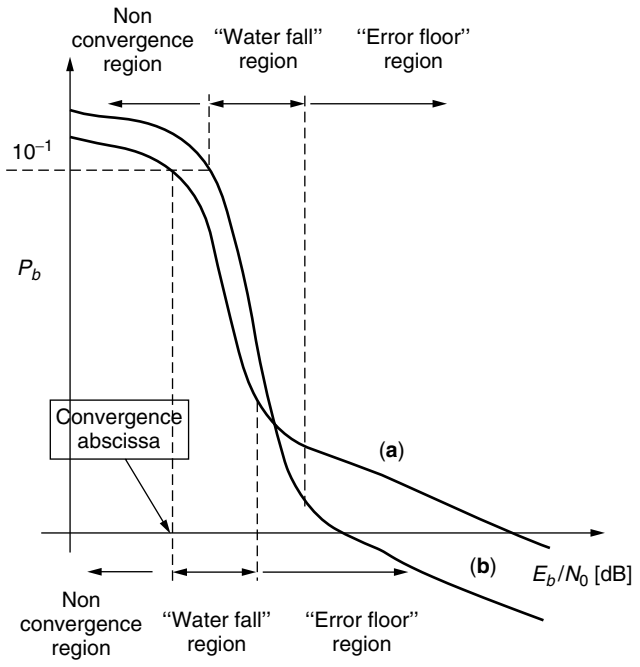


Figure 6. Qualitative behavior of P_b versus E_b/N_0 for concatenated codes with interleavers under iterative decoding.

significant energy expenses. The waterfall region is dominated by the interleaver gain, whereas the error floor region is dictated by the minimum distance of the code. In the first region, the interleaver acts through its size, whereas in the second the kind of interleaver plays a dominant role.

Any attempt to design codes like those in Fig. 2 as a whole leads to discouraging results because of its complexity. So far, only two successful design techniques have been proposed in the literature: the first relies on the previously introduced concept of uniform interleaver, and leads to design rules for the CCs that work nicely for medium–low error probabilities. After designing the CCs, one can improve the code performance by a cut-and-try approach to the interleaver design.

While the first approach is a purely analytical one, the second requires the separate simulation of the behavior of each CC, and is based on the *probability density evolution* technique ([12–14]) or on the *EXIT charts* [15]. This technique leads to codes that exhibit good performance in terms of convergence abscissa, and are suited for medium-high values of the error probability. Unfortunately, the two techniques lead to different code designs; typically, the codes designed (or, better, found) through the second technique show faster convergence of the iterative decoding algorithm, but reach sooner the error floor [see curve (a) in Fig. 6]. The opposite is true for codes designed using the first technique [see curve (b) in Fig. 6]. As a consequence, the choice depends on the quality of service requirements of the system at hand. In the following, we report the main results stemming from the first design technique, addressing the readers to the appropriate references for the second.

4.1. The ML Design

A lengthy analysis [5] shows that, for large interleavers, the union upper bound to the ML error probability of a general concatenated code with uniform interleaver can be written as

$$P \leq \frac{1}{2} \sum_{d=d_{\min}} K_d N^{\alpha(d)} \operatorname{erfc}\left(\sqrt{\frac{dR_c E_b}{N_0}}\right) \tag{5}$$

where K_d does not depend on the interleaver size N. Asymptotically, for large interleavers where $N \to \infty$, the dominant term in the summation of (5) is the one for which the exponent of N is maximum. Denoting that exponent by $\alpha_M \triangleq \max_d \alpha(d)$, and neglecting the other summation terms, yields

$$P \overset{N}{\asymp} \frac{1}{2} K_{d_M} N^{\alpha_M} \operatorname{erfc}\left(\sqrt{\frac{d_M R_c E_b}{N_0}}\right) \tag{6}$$

where $d_M = \arg\max_d [\alpha(d)]$. Negative values of α_M lead to interleaving gains, which are more pronounced for larger magnitudes of α_M. The value of α_M is different in the cases of word and bit error probabilities, and depends on the kind of code concatenation. In the following, we show its values for the bit error probability. Adding one to those values yields the α'_M values pertaining to the word error probability.

For parallel concatenated codes[4] (PCCs) [10] a necessary and sufficient condition to have $\alpha_M < 0$ is that both constituent encoders be *recursive*. In that case, we obtain

$$\begin{aligned} \alpha_M &= -1 \\ d_M &= d_{\text{free,eff}} \end{aligned} \tag{7}$$

where $d_{\text{free,eff}}$ is the *effective* free distance of the PCCC, *i.e.*, the minimum weight of codewords associated to weight 2 information words. The effective free distance of the PCCC is equal to the sum of the effective free distances of the constituent encoders, and thus the optimization of them consists in searching for recursive convolutional encoders with the largest effective free distance, and, possibly, the lowest number of codewords with weight equal to the effective free distance (number of "nearest" neighbors).

For SCCCs, a necessary and sufficient condition to have $\alpha_M < 0$ is that the inner constituent encoder be *recursive* [5]. If that is the case, we obtain

$$\alpha_M = -\left\lfloor \frac{d^o_{\text{free}} + 1}{2} \right\rfloor \tag{8}$$

$$d_M = \begin{cases} \dfrac{d^o_{\text{free}} d^i_{\text{free,eff}}}{2}, & \text{for } d^o_{\text{free}} \text{ odd} \\ \dfrac{(d^o_{\text{free}} - 3) d^i_{\text{free,eff}}}{2} + d_3, & \text{for } d^o_{\text{free}} \text{ even} \end{cases}$$

[4] For comparison purposes, we recall here the main results for the case of parallel concatenation.

where d_3 is the minimum weight of inner code words associated to input words with weight 3.

Example 2. Consider two rate-$\frac{1}{3}$ concatenated encoders. The first is a parallel concatenated convolutional encoder made up with two 4-state systematic recursive convolutional encoders derived from the generator matrix G_0 of (3). The second, instead, is the rate-$\frac{1}{3}$ SCCC of Example 1. The two concatenations use interleaver sizes such that the corresponding information block sizes coincide. Applying the union bound leads to the results reported in Fig. 7, where we plot the bit error probability versus E_b/N_0 for information block sizes equal to 100 and 1000.

The curves make apparent the difference in the interleaver gain. In particular, the parallel concatenation shows an interleaver gain going as $\alpha_M = -1$, whereas the interleaver gain of the SCCC goes as $\alpha_M = -\dfrac{d_{of}+1}{2} = -3$, being the free distance of the outer code equal to 5. This means, for $P_b = 10^{-11}$, a gain of more than 2 dB in favor of the SCCC.

For SCCs, then, the design criteria require the inner code to be recursive, and chosen so as to maximize its effective free distance. The outer encoder can be either recursive or not, and its optimization is based on the usual criterion of maximizing the free distance.

The design criterion based on the *effective* free distance (rather than simply the free distance), and the requirement for the inner constituent encoder to be recursive, is a complete novelty in the panorama of known optimum convolutional encoders, whose search had been based on feedforward encoders with the greatest free distance. The effective free distance of a convolutional encoder can be significantly greater than the free distance, and this, with the interleaver gain, explains the exceptionally good performance of SCCs in the waterfall region. In Refs. 16 and 17 upper bounds to the effective free distance and tables of good recursive convolutional encoders are reported. The findings of the more complete search for good codes matching the design criteria have also been described [18].

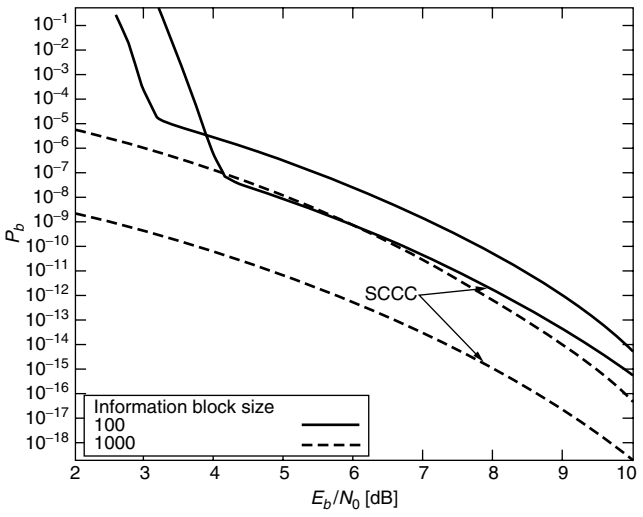


Figure 7. Comparison of serially and parallel rate-$\frac{1}{3}$ concatenated convolutional codes based on 4-state CCs with two interleaver sizes, yielding the same information block size for the two concatenations.

5. ITERATIVE DECODING OF SERIALLY CONCATENATED CODES

In this section, we present the iterative algorithm for decoding serially concatenated codes, with complexity not significantly higher than that needed to separately decode the two CCs. We assume, without loss of generality, that the constituent encoders admit a trellis representation [9].

The core of the decoding procedure consists of a block called *soft-input–soft-output* (SISO). It is a four-port device, which accepts as inputs the likelihood functions (or the corresponding likelihood ratios) of the information and code symbols labeling the edges of the code trellis, and forms as outputs an update of those likelihood functions based on the code constraints. The block SISO is used within the iterative decoding algorithm as shown in Fig. 8, where we also show the block diagram of the encoder to clarify the notations.

We will first explain in words how the algorithm works, according to the blocks of Fig. 8. Subsequently we will give the input-output relationships of the block SISO.

The symbols $\lambda(\cdot;I)$ and $\lambda(\cdot;O)$ at the input and output ports of SISO refer to the logarithmic likelihood ratios (LLRs),[5] unconstrained when the second argument is I, and modified according to the code constraints when it is O. The first argument u refers to the information symbols of the encoder, whereas c refers to code symbols. Finally, the superscript o refers to the outer encoder, and i to the inner encoder. The LLRs are defined as

$$\lambda(x;\cdot) \triangleq \log\left[\frac{P(x;\cdot)}{P(x_{\text{ref}};\cdot)}\right] \tag{9}$$

When x is a binary symbol, "0" or "1," x_{ref} is generally assumed to be the "0." When x belongs to an L-ary alphabet, we can choose as x_{ref} each one of the L symbols.

In contrast to the iterative decoding algorithm employed for PCCC decoding [19], in which only the LLRs

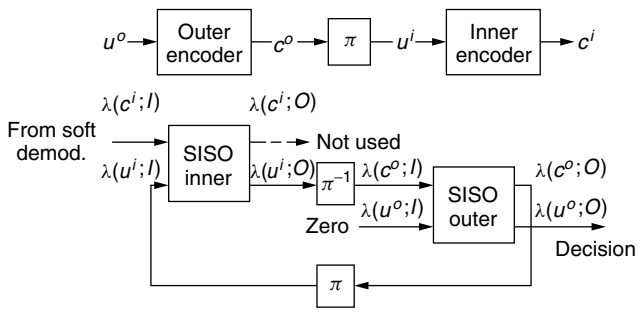


Figure 8. Block diagrams of the encoder and iterative decoder for serially concatenated convolutional codes.

[5] When the symbols are binary, only one LLR is needed; when the symbols belong to an L-ary alphabet, $L-1$ LLRs are required.

of information symbols are updated, here we must update the LLRs of both information and code symbols based on the code constraints.

During the first iteration of the SCC algorithm,[6] the block "SISO inner" is fed with the demodulator soft outputs, consisting of the LLRs of symbols received from the channels, namely, of the code symbols of the inner encoder. The second input $\lambda(u^i;I)$ of the SISO inner is set to zero during the first iteration, since no a- priori information is available on the input symbols u^i of the inner encoder.

The LLRs $\lambda(c^i;I)$ are processed by the SISO algorithm, which computes the *extrinsic* [19] LLRs of the information symbols of the inner encoder $\lambda(u^i;O)$ conditioned on the inner code constraints. The extrinsic LLRs are passed through the inverse interleaver (block labeled "π^{-1}"), whose outputs correspond to the LLRs of the code symbols of the outer code:

$$\lambda(c^o;I) = \pi^{-1}[\lambda(u^i;O)]$$

These LLRs are then fed to the block "SISO outer" in its upper entry, which corresponds to code symbols. The SISO outer, in turn, processes the LLRs $\lambda(c^o;I)$ of its unconstrained code symbols, and computes the LLRs of both code and information symbols based on the code constraints. The input $\lambda(u^o;I)$ of the SISO Outer is always set to zero, which implies assuming equally likely transmitted source information symbols. The output LLRs of information symbols (which yield the a posteriori LLRs of the SCCC information symbols) will be used in the final iteration to recover the information bits. On the other hand, the LLRs of outer code symbols, after interleaving, are fed back to the lower entry (corresponding to information symbols of the inner code) of the block SISO inner to start the second iteration. In fact we have

$$\lambda(u^i;I) = \pi[\lambda(c^o;O)]$$

5.1. Input–Output Relationships for the Block SISO

The block SISO has been described [19]. It represents a slight generalization of the BCJR algorithm [20–22]. Here, we will only recall for completeness its input–output relationships. We will refer, for notations, to the trellis section of the trellis encoder, assumed to be time invariant as we deal with convolutional codes, shown in Fig. 9, where the symbol e denotes the trellis edge, and where we have identified the information and code symbols associated to the edge e as $u(e)$, $c(e)$, and the starting and ending states of the edge e as $s^S(e)$, $s^E(e)$, respectively.

The block SISO works at *symbol* level; thus, for an (n,p) convolutional code, it operates on information symbols u belonging to an alphabet with size 2^p and on code symbols belonging to an alphabet with size 2^n. We will give the general input–output relationships, valid for both outer and inner SISOs, assuming that the information and code symbols are defined over a finite time index set $[1,\ldots,K]$.

[6] To simplify the description, we assume for now that the interleaver acts on symbols instead of bits. In practice, one often uses bit LLRs and bit interleaver, as it will be seen later.

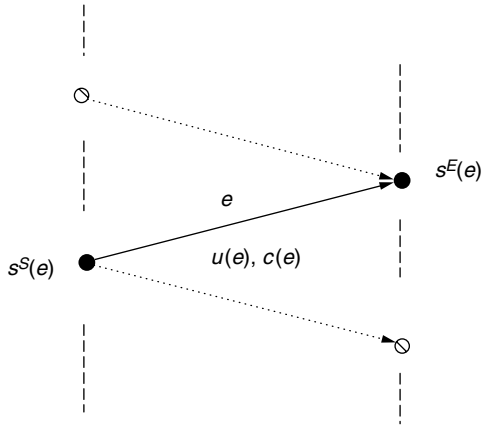


Figure 9. Trellis section defining the notations used for the description of the SISO algorithm.

At time k, $k=1,\ldots,K$, the output extrinsic LLRs are computed as

$$\lambda_k(c;O) = \max_{e:c(e)=c}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[u(e);I] + \beta_k[s^E(e)]\} - \max_{e:c(e)=c_{\text{ref}}}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[u(e);I] + \beta_k[s^E(e)]\} \tag{10}$$

$$\lambda_k(u;O) = \max_{e:u(e)=u}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[c(e);I] + \beta_k[s^E(e)]\} - \max_{e:u(e)=u_{\text{ref}}}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[c(e);I] + \beta_k[s^E(e)]\} \tag{11}$$

The name *extrinsic* given to the LLRs computed according to (10) and (11) derives from the fact that the evaluation of $\lambda_k(c;O)$ [and of $\lambda_k(u;O)$] does not depend on the corresponding simultaneous input $\lambda_k(c;I)$ [and $\lambda_k(u;I)$], so that it can be considered as an update of the input LLR based on informations coming from all homologous symbols in the sequence, except the one corresponding to the same symbol interval.

The quantities $\alpha_k(\cdot)$ and $\beta_k(\cdot)$ in (10) and (11) are obtained through the *forward* and *backward* recursions, respectively, as

$$\alpha_k(s) = \max_{c:s^E(e)=s}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[u(e);I] + \lambda_k[c(e);I]\}, \quad k=1,\ldots,K-1 \tag{12}$$

$$\beta_k(s) = \max_{e:s^S(e)=s}{}^*\{\beta_{k+1}[s^E(e)] + \lambda_{k+1}[u;I] + \lambda_{k+1}[c(e);I]\}, \quad k=K-1,\ldots,1 \tag{13}$$

with initial values

$$\alpha_0(s) = \begin{cases} 0 & s=S_0 \\ -\infty & \text{otherwise} \end{cases}$$

$$\beta_K(S_i) = \begin{cases} 0 & s=S_K \\ -\infty & \text{otherwise} \end{cases}$$

The operator max* performs the following operation:

$$\max_j{}^*(a_j) \triangleq \log\left[\sum_{j=1}^{J} e^{a_j}\right] \tag{14}$$

This operation, a crucial one in affecting the computational complexity of the SISO algorithm, can be performed in practice as

$$\max_j{}^*(a_j) = \max_j(a_j) + \delta(a_1, a_2, \ldots, a_J) \tag{15}$$

where $\delta(a_1, a_2, \ldots, a_J)$ is a correction term that can be computed recursively using a single-entry lookup table [19,23].

The previous description of the iterative decoder assumed that all operations were performed at *symbol* level. Quite often, however, the interleaver operates at *bit* level to be more effective. This is the case, for example, of all results presented in Sections 3–5.

Thus, to perform bit interleaving, we need to transform the symbol extrinsic LLRs obtained at the output of the first SISO into extrinsic bit LLRs, before they enter the deinterleaver. After deinterleaving, the bit LLRs need to be compacted into symbol LLRs before entering the second SISO block, and so on. These operations are performed under the assumption that the LLRs of bits forming a symbol are independent.

Assuming an (n, p) code, and denoting with $\mathbf{u} = [u_1, \ldots, u_p]$ the information symbol formed by p information bits, then the extrinsic LLR λ_i of the ith bit u_i within the symbol $\mathbf{u}$ is obtained as

$$\lambda_{kp+i}(O) = \max_{\mathbf{u}:u_i=1}{}^*[\lambda_k(\mathbf{u}; O) + \lambda_k(\mathbf{u}; I)] - \max_{\mathbf{u}:u_i=0}{}^*[\lambda_k(\mathbf{u}; O) + \lambda_k(\mathbf{u}; I)] - \lambda_{kp+i}(I) \tag{16}$$

Conversely, the extrinsic LLR of the symbol $\mathbf{u}$ is obtained from the extrinsic LLRs of its component bits u_i as

$$\lambda(\mathbf{u}) = \sum_{i=1}^{p} u_i \lambda_i \tag{17}$$

As previous description should have made clear, the SISO algorithm requires that the whole sequence had been received before starting. The reason is due to the backward recursion that starts from the (supposed known) final trellis state. As a consequence, its practical application is limited to the case where the duration of the transmission is short (K small), or, for K long, when the received sequence can be segmented into independent consecutive blocks, like for block codes or convolutional codes with trellis termination [24]. It cannot be used for continuous decoding. This constraint leads to a frame rigidity imposed to the system, and also reduces the overall code rate, because of trellis termination.

A more flexible decoding strategy is offered by modifying the algorithm in such a way that the SISO module operates on a fixed memory span, and outputs the updated LLRs after a given delay D. This algorithm, which we have called the *sliding-window soft-input/soft-output* (SW-SISO) algorithm, is fully described in Ref. 19. In all simulation results the SW-SISO algorithm has been applied.

5.2. Applications of the Decoding Algorithm

We will now use the decoding algorithm to confirm the design rules presented before, and to show the behavior of SCCC in the region of low signal-to-noise ratios (below cutoff rate). Since analytic bounds fail to give significant results in this region, no meaningful quantitative comparisons can be performed between simulated and analytic performance. However, we will show that the hierarchy of the simulation results agrees with the design considerations that had been based on the analysis.

The following aspects will be considered:

- The behavior of the decoding algorithm versus the number of decoding iterations (Section 5.2.1)
- The behavior of the decoding algorithm versus the interleaver length (Subsection 5.2.2)
- The effect of choosing a nonrecursive inner code (Section 5.2.2)
- The SCCC behavior for very low signal-to-noise ratios, to see how close serial concatenation can get to theoretical Shannon bound (Section 5.2.3)
- The comparison between SCCCs and PCCCs (Turbo codes) for the same value of the decoding delay imposed by the two schemes on the input bits (Section 5.2.4).

For all simulated SCCCs, we have used purely random interleavers.

5.2.1. Simulated Coding Gain Versus Number of Iterations. Consider the rate-$\frac{1}{3}$ SCCC1 of Example 1 employing an interleaver of length $N = 2048$. Since the interleaver operates on coded sequences produced by the outer rate-$\frac{1}{2}$ encoder, its length of 2048 bits corresponds to a delay of 1024 information bits. The simulation results are shown in Fig. 10 in terms of bit error probability versus E_b/N_0 for a number of iterations N_I ranging from 1 to 7. The nice convergence of the decoding algorithm is manifest.

5.2.2. The Effect of a Nonrecursive Inner Encoder. In Section 4 we concluded that a nonrecursive inner encoder should yield little interleaver gains. To confirm this theoretical prediction by simulation results, in Fig. 11 we plot the bit error probability versus the input decoding delay obtained by simulating a rate-$\frac{1}{3}$ SCCC that concatenates the outer encoder of Example 1 with the 4-state, nonrecursive inner encoder with generator matrix

$$G_i(Z) = \begin{bmatrix} 1+Z, & Z, & 1 \\ 1+Z, & 1, & 1+Z \end{bmatrix}$$

The curves refer to a signal-to-noise ratio $E_b/N_0 = 1.5$ dB, and to a number of iterations N_I ranging from 1 to 10. It is evident that the bit error probability reaches the floor of

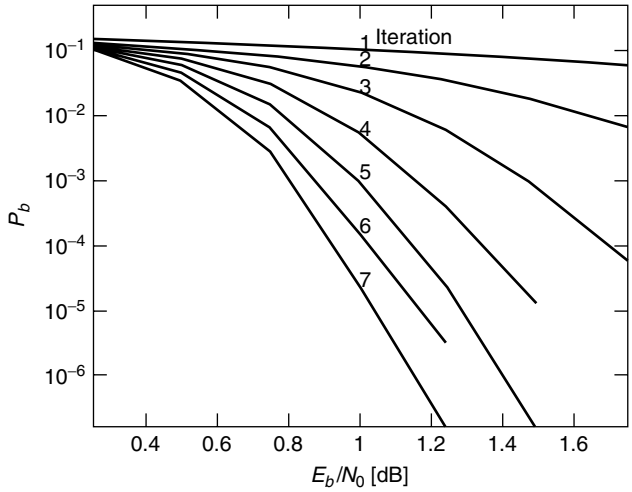


Figure 10. Simulated bit error probability versus the number of iterations for the rate-$\frac{1}{3}$ SCCC of Example 1. The decoding delay in terms of input bits due to the interleaver is 1024.

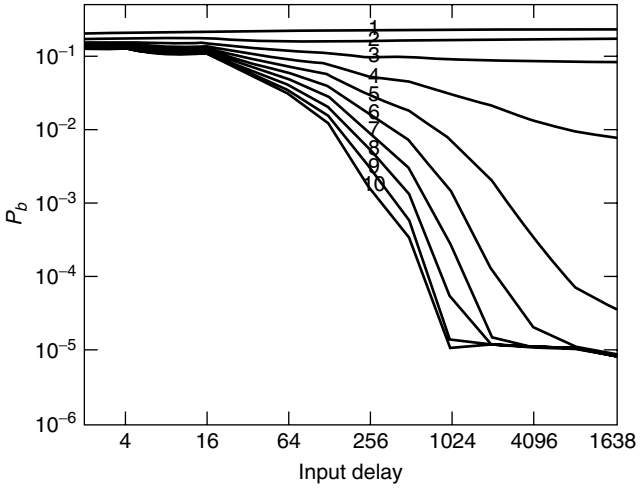


Figure 11. Simulated performance of the SCCC employing a nonrecursive inner encoder described in Section 5.2.3. The bit error probability is plotted versus input decoding delay for different number of iterations. The signal-to-noise ratio is 1.5 dB.

10^{-5} for a decoding delay greater than or equal to 1024, so that no interleaver gain takes place beyond this point. For comparison, in Fig. 12 we show the results obtained for the SCCC concatenating the 4-state, rate-$\frac{1}{2}$ nonrecursive outer encoder with generator matrix

$$G_o(Z) = [1 + Z + Z^2, 1 + Z^2]$$

with the rate-$\frac{2}{3}$ inner encoder of Example 1. The curves refer to a signal-to-noise ratio of 0.75 dB, and show the interleaver gain predicted by the analysis.

5.2.3. Approaching the Theoretical Shannon Limit. We show here the capabilities of SCCCs of yielding results close to the Shannon capacity limit. To this purpose, we have chosen a rate-$\frac{1}{4}$ concatenated scheme with very long interleaver, corresponding to an input decoding delay

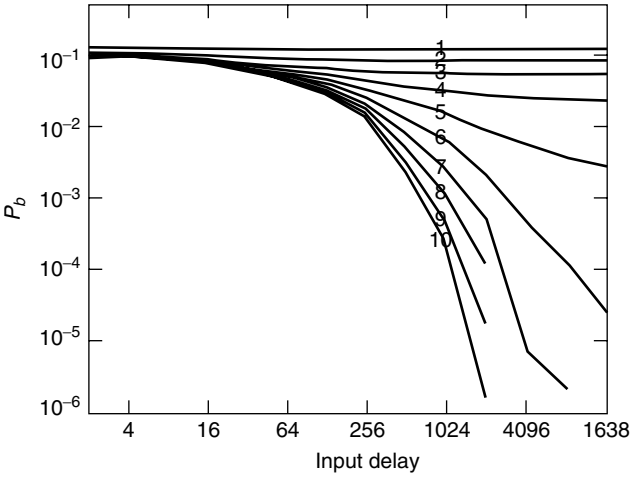


Figure 12. Simulated performance of the SCCC employing a nonrecursive outer encoder described in Section 5.2.2. The bit error probability is plotted versus input decoding delay for different number of iterations. The signal-to-noise ratio is 0.75 dB.

of 16,384. The constituent codes are 8-state codes: the outer encoder is nonrecursive, and the inner encoder is a recursive encoder. Their generating matrices are

$$G_o(Z) = [1 + Z, 1 + Z + Z^3]$$

$$G_i(Z) = \left[1, \frac{1 + Z + Z^3}{1 + Z}\right]$$

respectively. Note the feedback polynomial $(1 + Z)$ in the generator matrix of the inner encoder, which eliminates error events with odd input weights. The results in terms of bit error probability versus signal-to-noise ratio for different number of iterations are presented in Fig. 13 showing that the decoding algorithm works at $E_b/N_0 = -0.05$ dB, at 0.76 dB from the Shannon capacity limit (which is in this case equal to -0.817 dB), with very limited complexity (remember that we are using two rate-$\frac{1}{2}$ codes with 8 states).

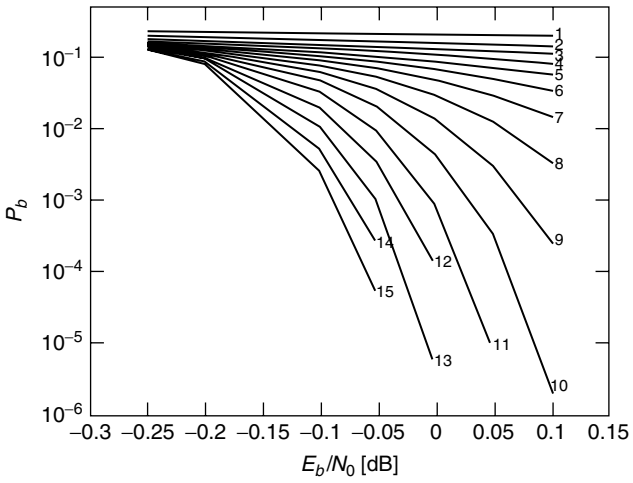


Figure 13. Simulated performance of the rate-$\frac{1}{4}$ SCCC obtained with two 8-state constituent codes and an interleaver yielding an input decoding delay equal to 16,384.

5.2.4. Comparison Between Serially and Parallel Concatenated Codes. Previous analytic results showed that serial concatenation can yield significantly higher interleaver gains and steeper asymptotic slope of the error probability curves. To check whether these advantages are retained when the codes are iteratively decoded at very low signal-to-noise ratios, we have simulated the behavior of two SCCCs and PCCCs in equal system conditions: the concatenated code rate is $\frac{1}{3}$, the CCs (same as Example 2) are 4-state recursive encoders (rates $\frac{1}{2} + \frac{1}{1}$ for PCCCs, and rates $\frac{1}{2} + \frac{2}{3}$ for the SCCCs), and the decoding delays in terms of input bits are 1024 and 16,384, respectively. In Fig. 14 we report the results, in terms of bit error probability versus signal-to-noise ratio, for the case of a decoding delay equal to 1024, after three and seven decoding iterations. As it can be seen from the curves, the PCCC outperforms the SCCC for high values of the bit error probabilities. For bit error probabilities lower than $3 \cdot 10^{-3}$ (for seven iterations), the SCCC outperform PCCC, and does not present the error floor. At 10^{-4}, SCCC has an advantage of 0.7 dB with seven iterations. Finally, in Fig. 15, we report the results for an input decoding delay of 16,384 and six and nine decoding iterations. In this case, the crossover between PCCC and SCCC happens around 10^{-5}. The advantage of SCCC at 10^{-6} is 0.5 dB with nine iterations.

As a conclusion, we can say that the advantages obtained for signal-to-noise ratios above the cutoff rate, where the union bounds can be safely applied, are retained also in the region between channel capacity and cutoff rate. Only when the system quality of service focuses on high values of bit error probability (the threshold depending on the interleaver size) the PCCC are to be preferred. PCCCs, however, present a floor to the bit error probability, which, in the most favorable case seen above, lies around 10^{-6}. This floor is much lower in the case of SCCC.

Finally, it must be recognized that the constituent codes design rules presented in Sec. 4 are based on union bound considerations, and thus yield optimum SCCCs above the cutoff rate. For system applications aiming at very low signal-to-noise ratios, close to the channel capacity (as, for example, in deep-space communications), a general statement is that complex CCs should be avoided, and CCs with low number of states (4–16) should be used. The code design in this signal-to-noise ratio region, a different design approach based on the separate simulation of the behavior of each CC, and based on the probability density evolution technique [12–14] or on the *EXIT charts* [15] must be adopted.

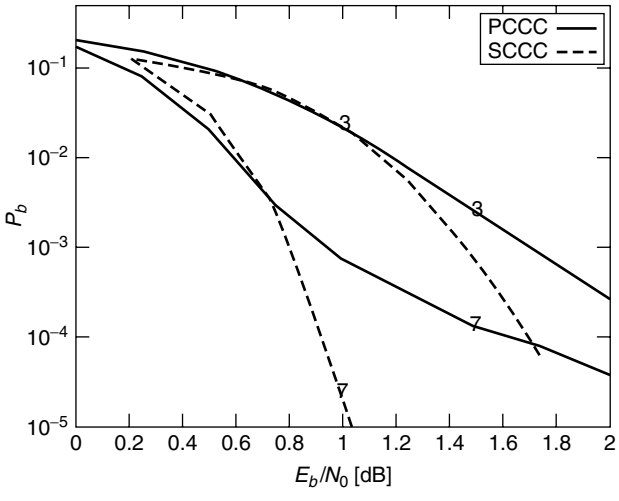


Figure 14. Comparison of the two rate-$\frac{1}{3}$ PCCCs and SCCCs of Example 2. The curves refer to three and seven iterations of the decoding algorithm and to an equal input decoding delay of 1024.

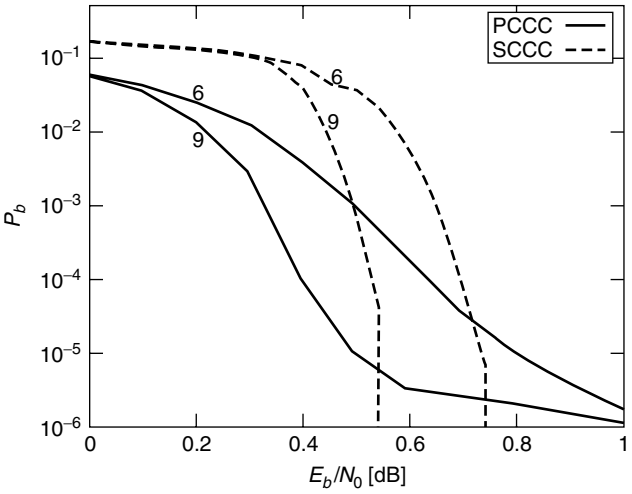


Figure 15. Comparison of the two rate-$\frac{1}{3}$ PCCCs and SCCCs of Example 2. The curves refer to three and seven iterations of the decoding algorithm and to an equal input decoding delay of 16384.

6. EXAMPLES OF SERIAL CONCATENATIONS

In this section we present a few significant examples of digital transmission systems that can be interpreted, and thus be utilized, as serial concatenations of individual modules. We will show that the serial structure fits to a set of systems well beyond the classical one constituted by two concatenated encoders with an interleaver in between.

6.1. Serial Concatenation of an Outer Encoder and an Inner Modulator

The serial concatenation of an outer encoder with rate R_c^o with an inner modulator characterized by an alphabet of $M = 2^m$ channel symbols through an interleaver is shown in Fig. 16 [25].

The binary symbols $\mathbf{u}$ emitted by the source are encoded by the outer binary encoder. The coded bits $\mathbf{c}$ are interleaved first, and then serial-to-parallel converted to m parallel bit streams $\mathbf{c_i}, \mathbf{i} = \mathbf{1}, \ldots, \mathbf{m}$. A mapper associates

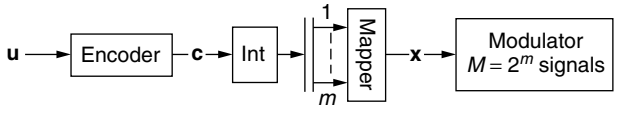


Figure 16. Block diagram of the serial concatenation of an outer encoder with an inner modulator through a bit interleaver.

each bit m-tuple to a modulator symbol x, which is then sent to the channel. The *spectral* efficiency of the scheme is related to the number of information bits per channel symbol $n_m \triangleq mR_c^o$.

A scheme like the one depicted in Fig. 16 can be detected or decoded iteratively, by considering it as a serial concatenation with interleaver. The condition to achieve improved performance through the iterative algorithm is that the mapping of m bits to a modulated signal be an operation with memory

$$\mathrm{P}[u_1, u_2, \ldots, u_m \mid r] \neq \mathrm{P}[u_1 \mid r]\mathrm{P}[u_2 \mid r] \cdots \mathrm{P}[u_m \mid r]$$

where r is the received signal.

Example 3. Consider an outer 8-state, rate-$\frac{1}{2}$ systematic recursive convolutional encoder punctured to rate $\frac{2}{3}$, followed by an interleaver with length 12,003, a mapper performing three different kind of mappings — natural, Gray mapping, or "anti-Gray" mapping — and an 8-PSK modulator. The three mappings are shown in Fig. 17. Note that the anti-Gray mapping has been chosen so as to maximize the sequence d_1, d_2, d_3, d_4, where d_w is the minimum Euclidean distance between pair of signals whose binary labels have a Hamming distance equal to w. This mapping choice is in perfect agreement with the design rule for inner encoders found in Section 4, replacing Hamming with Euclidean distances, and is the opposite with respect to Gray mapping.

The whole system is shown in Fig. 18; its receiver consists of the LLR computation on 8-PSK symbols, followed by a soft demapper that projects the symbol LLRs onto bit LLRs, the interleaver/deinterleaver and the puncturer/depuncturer pairs, and, finally, the outer SISO working on the rate-$\frac{1}{2}$ trellis. In Fig. 19 we report the bit error probability of such a scheme as a function of E_b/N_0, obtained by simulation with one to five iterations of the detection/decoding algorithm. The solid black curves refer to natural mapping, and show a 2.5-dB gain at 10^{-5}. The dashed line refers to Gray mapping, and shows that the iterations do not bring any performance improvement, although yielding better results than the first iteration with natural mapping. Finally, the solid gray curves refer to the anti-Gray mapping. With 1 iteration, this is the worst mapping; as the iterations evolve, however, the gain is very significant, and a gain of almost 4 dB over the Gray mapping is obtained at 10^{-5}, which further increases for higher E_b/N_0 because of the largest slope of the curves.

To further improve the performance, a rate-1, 2-state differential encoder could be added before the modulator, leading to the scheme of Fig. 20, in which the inner "encoder" is now recursive, as prescribed by the design rules of Section 4. Its performance has been reported elsewhere [26], and an extension to the case of intersymbol interference channels and Turbo equalization has been proposed [27].

6.2. The Encoded Continuous-Phase Modulation

Continuous-phase modulation (CPM) is a class of *constant-envelope*, *continuous-phase* modulation schemes obtained through a modulator with finite memory that can be described as a finite-state machine, or, equivalently, a trellis. Because of its attractive properties of constant envelope and bandwidth compactness, also maintained through nonlinear devices, CPM has been a subject of extensive studies and publications in the late 1970s and 1980s, and has then been applied in various radio systems, such as for example in GSM. Reference 28 contains an in-depth treatment of this subject.

A general CPM modulator can be split [29] into the cascade of a finite-state machine, the continuous-phase encoder (CPE), and a memoryless modulator (MM). Adding an outer convolutional encoder and an interleaver leads then to the system depicted in Fig. 21, where the information bits are first encoded by a convolutional encoder with rate k/n, followed by a bit interleaver with length N, and by an M-ary CPM modulator. Such a structure is capable of transmitting $(k/n)\log_2 M$

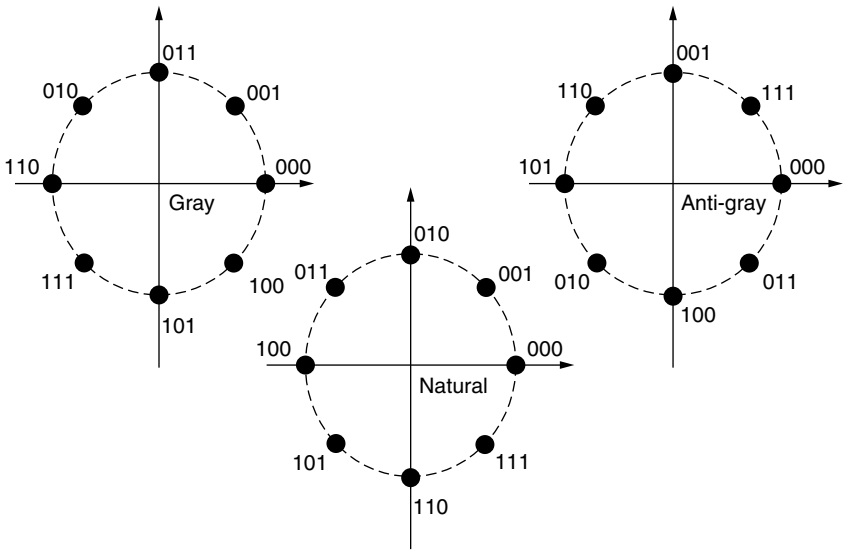


Figure 17. Three different mappings from triplets of bits to 8-PSK signals: Gray, natural, and anti-Gray.

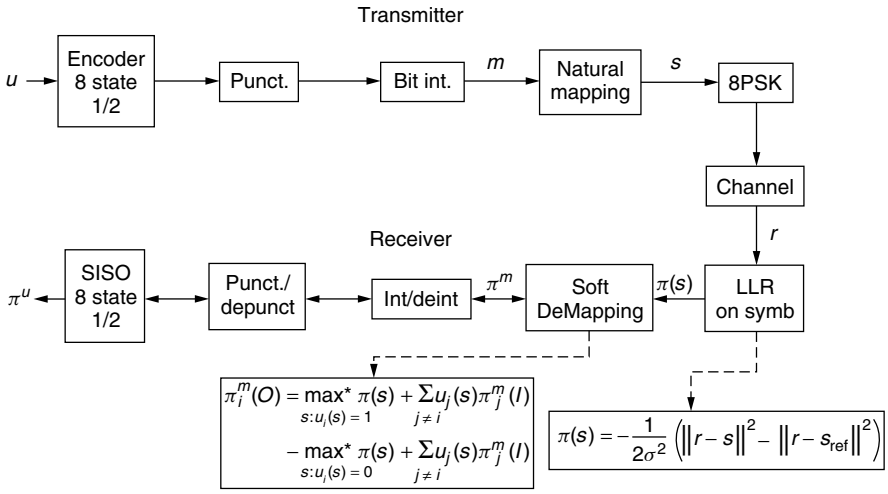


Figure 18. Block diagram of the transmitter and receiver of the serial concatenation of an outer encoder with an inner modulator through a bit interleaver.

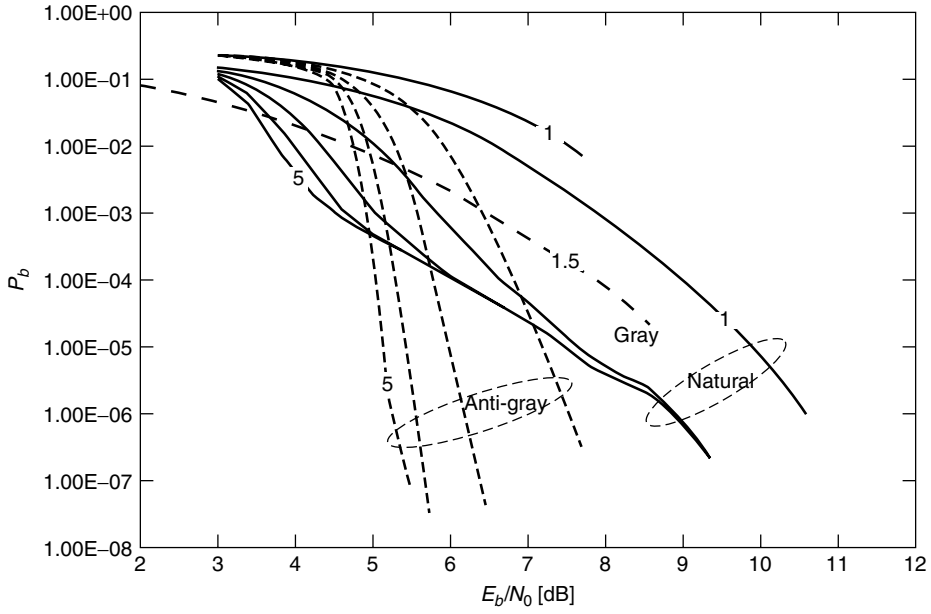


Figure 19. Bit error probability versus E_b/N_0 for the serial concatenation of Example 3. The solid black curves refer to natural mapping, the dashed one to Gray mapping, and the solid gray curves to anti-Gray mapping, with one to five iterations of the decoding algorithm.

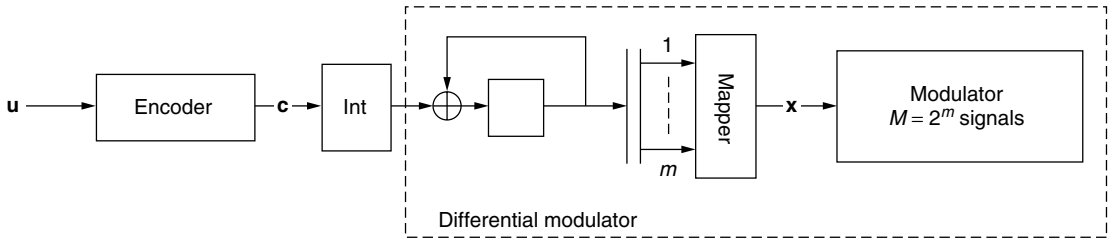


Figure 20. Block diagram of the serial concatenation of an outer encoder with an inner differential encoder and a modulator through a bit interleaver.

information bits per CPM signal, achieving a bandwidth efficiency that depends on the frequency pulse and modulation index of the CPM scheme.

According to Fig. 21, the cascade of a convolutional encoder, an interleaver, and a CPM modulator, a configuration that we will call *serially concatenated convolutional CPM* (SCC-CPM), can be seen as a particular form of serially concatenated convolutional codes with interleaver, and thus demodulated-decoded with an iterative scheme yielding high coding gains.

Example 4. Consider a coded CPM scheme (like the one depicted in Fig. 21), in which the outer encoder is a 4-state, rate-$\frac{2}{3}$ recursive systematic convolutional encoder,

and the inner module is a quarternary partial response CPM scheme [28] employing a rectangular frequency pulse with duration equal to 2 symbol intervals (the so-called 2-REC pulse) and modulation index $h = \frac{1}{3}$. Using the iterative receiver of Fig. 21, one obtains the performance shown in Fig. 22. The simulated scheme, owing to the low modulation index, has a high bandwidth efficiency, and achieves very good performance. Also, the performance improve very significantly with iterations.

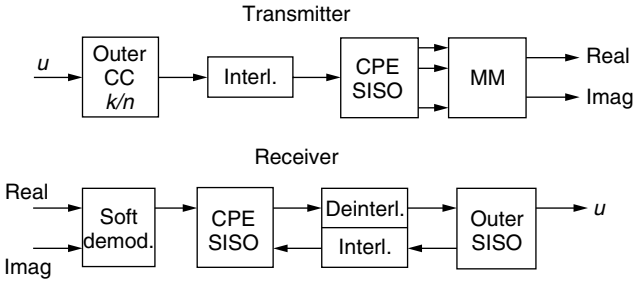


Figure 21. Block diagram of the encoded CPM transmitter and receiver.

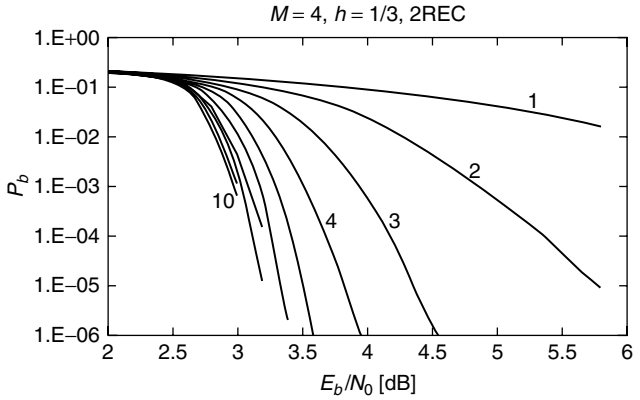


Figure 22. Bit error probability curves versus E_b/N_0 for the quaternary CPM scheme employing the 2-REC pulse with modulation index $h = \frac{1}{3}$. The outer code rate is $\frac{2}{3}$, and the interleaver size corresponds to 8190 information bits.

6.3. Serial Concatenation of an Outer Encoder with the Magnetic Recording Channel

In Fig. 23 we show the block diagram of a coded magnetic recording system based on the serial concatenation paradigm. It uses as outer code a high-rate (like $\frac{8}{9}$, $\frac{15}{16}$; this is a mandatory constraint of this applications that requires very high recording densities) terminated convolutional encoder, and as inner module the magnetic recording channel preceded by a precoder that makes it recursive. An interleaver separates the two blocks. In the same figure, the lower scheme represents the precoder/channel model. Using the Lorentzian model with an EPR4 partial response target [30], the overall magnetic channel transfer function in the transform domain is

$$G_m(Z) = 1 + Z - Z^2 - Z^3$$

It can be seen as an intersymbol interference channel with non binary (five-level) outputs, so this example becomes similar to the case of Turbo equalization [31]. The precoder is characterized by a $1/(1 \oplus Z^2)$ transfer function.

The iterative decoder (see Fig. 23, upper portion) is made by two SISOs. The inner SISO (SISO channel in the figure) is matched to the precoder/channel response, and works on non binary symbols, whereas the outer SISO works for the very high-rate $(n, n-1)$ convolutional encoder. Two solutions are possible: the first makes use of a heavily punctured rate-$\frac{1}{2}$ "mother" encoder, and the second employs unpunctured encoders. In the latter case, the complexity of the SISO, which depends on the number of edges per trellis section per decoded bit is too high, and thus a different approach must be followed, like designing a SISO that works on the $(n, 1)$ *dual* code [32]. This can be made coincident with a standard SISO, provided that its inputs be converted from LLRs to the so-called *log-reflection coefficients* (LRC in the figure).

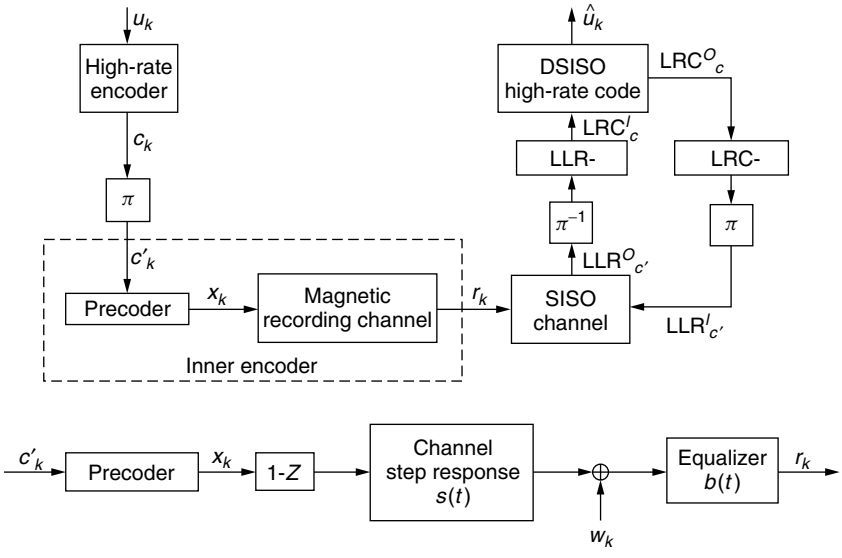


Figure 23. Block diagram of a coded magnetic recording system (upper diagram). Lorentzian model of precoder/magnetic channel (lower diagram).

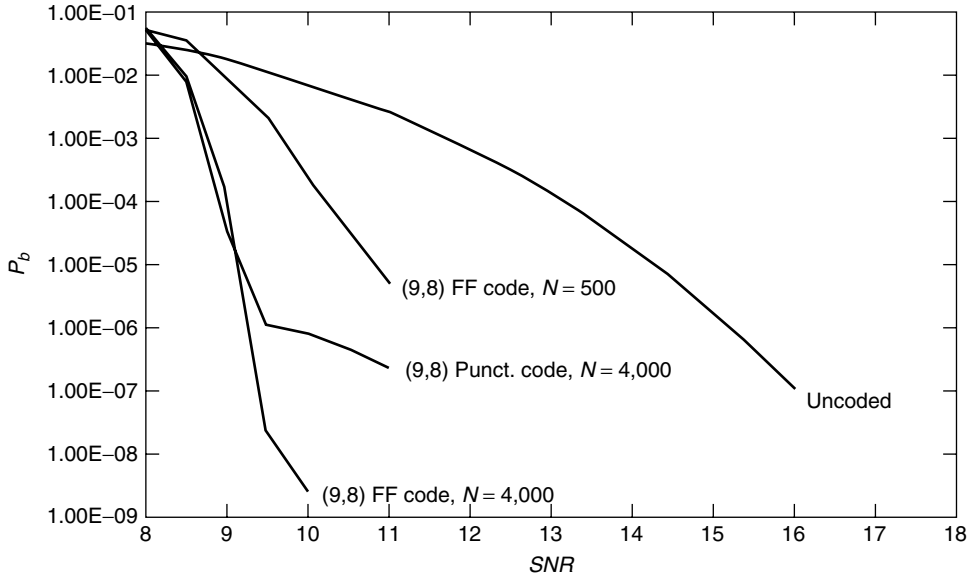


Figure 24. Bit error probability versus signal-to-noise ratio for the magnetic recording system of Fig. 23. The curves refer to the uncoded case, to the coded case employing a punctured rate-$\frac{8}{9}$ outer encoder with input block size 4000, and to the coded case using an unpunctured, optimized rate-$\frac{8}{9}$ outer encoder with block sizes 500 and 4000.

Using the iterative decoding algorithm, we have obtained the results shown in Fig. 24. The curves refer to the uncoded case, to the coded case employing a punctured rate-$\frac{8}{9}$ outer encoder with input block size 4000, and to the coded case using an unpunctured, optimized [33] rate $\frac{8}{9}$ outer encoder with block sizes 500 and 4000. A very large coding gain (6 dB for block size 4000 at 10^{-6}) can be obtained, without any significant error floor down to 10^{-9} for the unpunctured case. The punctured code, on the other hand, behaves similarly down to 10^{-6}, but shows the error floor just below. Notice that the different behavior of the two codes in the error floor region is due to the different free distance of the outer encoder, which is 2 for the punctured code (interleaver gain N^{-1}), and 3 for the unpunctured one (interleaver gain N^{-2}), in perfect agreement with the design rules of Section 4.

6.4. The Multiuser Interfered Channel

In coded code-division multiple access (CDMA) systems, K independent users first encode their information bit streams, then, after interleaving, they transmit them onto the same channel, as shown in Fig. 25. The cascade

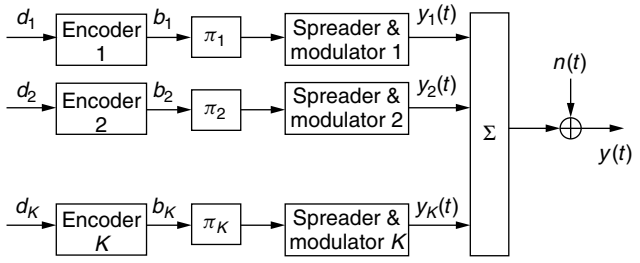


Figure 25. The CDMA transmitter with K users.

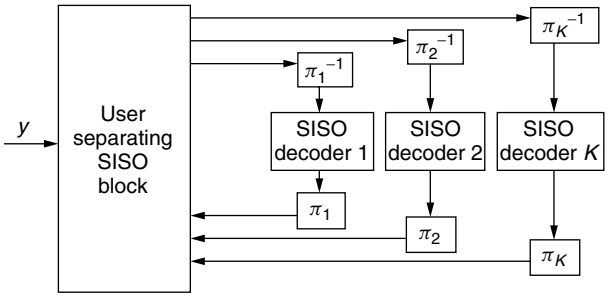


Figure 26. Turbo receiver. At every iteration, the user separator exchanges information with the decoders.

of the channel encoder, interleaver, and multiaccess channel can be viewed as a serial concatenation. As a consequence, at the receiver side, an iterative receiver can be employed, as shown in Fig. 26. A user separator, which is a soft-output version of a multiuser detector, attenuates for each user the multiaccess interference and sends K streams of soft output values to K channel decoders after deinterleaving. The decoders, in their turn, feed back with updated extrinsic LLRs the user separator, which will use the feedback information in the following iteration, to improve the user separation. In the last iteration, the decoders provide the final estimate of the K information bit streams. Figure 27 shows an example of performance along the iterations. It refers to a system with four equipower synchronous users, with correlation among any pairs equal to 0.7, and outer rate $\frac{1}{2}$, equal 16-state convolutional encoders with generator matrix

$$G_o(Z) = [1 + Z + Z^4, 1 + Z^2 + Z^3 + Z^4]$$

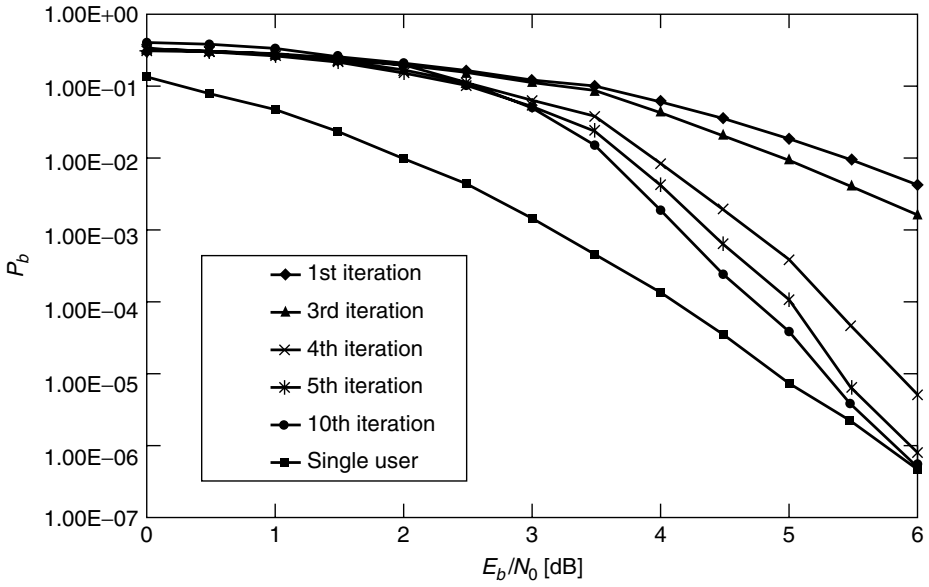


Figure 27. Results of simulation. $K = 4$ synchronous users, all with the same power. The correlation between any pair of users is 0.7. Interleaver size is 256, and the outer encoders are rate-$\frac{1}{2}$, 16-state feedforward convolutional encoders.

The interleavers are purely random interleavers with size 256. Finally, the user separator employs the algorithm described elsewhere [34].

As the curves show, the receiver tends to the single-user performance for sufficiently high signal-to-noise ratios.

7. CONCLUSIONS

In a semitutorial way avoiding all except strictly necessary mathematical developments, in this article we have described the main characteristics of serially concatenated codes with interleavers. They consist of the cascade of an outer encoder, an interleaver permuting the outer codeword bits, and an inner encoder whose input words are the permuted outer code words. Decoding is performed by an iterative technique based on two soft-input/soft-output algorithms tailored to the trellis of the outer and inner encoders. The iterative algorithm can be easily extended to other forms of concatenation, where one or both encoders are replaced with system modules with some form of memory performing different functions, like a modulator, a channel with intersymbol interference, a multiuser combiner. First, upper bounds to the *average* maximum-likelihood error probability of serially concatenated block and convolutional coding schemes have been described. Then, design guidelines for the outer and inner encoders that maximize the *interleaver* gain and the asymptotic slope of the error probability curves have been derived, together with the iterative decoding algorithm. Finally, extensions to different systems that can be seen as serial concatenations and detected or decoded accordingly have been proposed.

Acknowledgments

The authors gratefully acknowledge the contributions of Alex Graell i Amat and Alberto Tarable for providing the magnetic recording and multiuser detection results in Section VII.

BIOGRAPHIES

Sergio Benedetto is a Full Professor of Digital Communications at Politecnico di Torino, Italy since 1981. He has been a Visiting Professor at University of California, Los Angeles (UCLA), at University of Canterbury, New Zealand, and is an Adjoint Professor at Ecole Nationale Superieure de Telecommunications in Paris. In 1998 he received the Italgas Prize for Scientific Research and Innovation. He has also been awarded the title of Distinguished Lecturer by the IEEE Communications Society. He has co-authored two books on probability and signal theory (in Italian), the books *Digital Transmission Theory* (Prentice-Hall, 1987), *Optical Fiber Communications* (Artech House, 1996), and *Principles of Digital Communications with Wireless Applications* (Plenum-Kluwer, 1999), and over 250 papers in leading journals and conferences. He has taught several continuing-education courses on the subject of channel coding for the UCLA Extension Program and for the CEI organization. He was the Chairman of the Communications Theory Symposium of ICC 2001, and is the Area Editor for the *IEEE Transactions on Communications for Modulation and Signal Design*. Professor Benedetto is the Chairman of the Communication Theory Committee of IEEE and a Fellow of the IEEE.

Guido Montorsi was born in Turin, Italy, on January 1, 1965. He received the Laurea in Ingegneria Elettronica in 1990 from Politecnico di Torino, Turin, Italy, with a master thesis, concerning the study and design of coding schemes for high-definition television (HDTV), developed at the RAI Research Center, Turin. In 1992 he spent the year as visiting scholar in the Department of Electrical Engineering at the Rensselaer Polytechnic Institute, Troy, NY. In 1994 he received the Ph.D. degree in telecommunications from the Dipartimento di Elettronica of Politecnico di Torino.

In December 1997 he became assistant professor at the Politecnico di Torino. In July 2001 he became Associate Professor. In 2001–2002 he spent one year in the startup company Sequoia Communications for the innovative design and implementation of a third-generation WCDMA receiver.

He is author of more than 100 papers published in international journals and conference proceedings. His interests are in the area of channel coding and wireless communications, particularly the analysis and design of concatenated coding schemes and study of iterative decoding strategies.

BIBLIOGRAPHY

1. G. D. Forney, Jr., *Concatenated Codes*, MIT Press, Cambridge, MA, 1966.
2. R. H. Deng and D. J. Costello, High rate concatenated coding systems using bandwidth efficient trellis inner codes, *IEEE Trans. Commun.* **COM-37**(5): 420–427 (May 1989).
3. J. Hagenauer and P. Hoeher, Concatenated Viterbi decoding, *Proc. 4th Joint Swedish-Soviet Int. Workshop on Information Theory*, Gotland, Sweden, Studenlitteratur, Lund, Aug. 1989, 29–33.
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. ICC'93*, Geneve, Switzerland, May 1993.
5. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (May 1998).
6. D. Divsalar and F. Pollara, Serial and hybrid concatenated codes with applications, *Proc. Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 1997.
7. S. Benedetto and G. Montorsi, Generalized concatenated codes with interleavers, *Proc. Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 1997.
8. R. Garello, G. Montorsi, S. Benedetto, and G. Cancellieri, Interleaver properties and their applications to the trellis complexity analysis of turbo codes, *IEEE Trans. Commun.* **49**: 793–807 (May 2001).
9. S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, Kluwer Academic/Plenum, New York, 1999.
10. S. Benedetto and G. Montorsi, Unveiling turbo-codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**(2): 409–429 (March 1996).
11. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Analysis, design and iterative decoding of double serially concatenated codes with interleavers, *IEEE J. Select. Areas Commun.* **16**: 231–244 (Feb. 1998).
12. S. Y. Chung, T. J. Richardson, and R. L. Urbanke, Analysis of sum-product decoding of low-density parity-check codes using a Gaussian approximation, *IEEE Trans. Inform. Theory* **47**: 657–670 (Feb. 2001).
13. H. El Gamal and A. R. Hammons, Jr., Analyzing the turbo decoder using the Gaussian approximation, *IEEE Trans. Inform. Theory* **47**: 671–686 (Feb. 2001).
14. D. Divsalar, S. Dolinar, and F. Pollara, Iterative turbo decoder analysis based on density evolution, *IEEE J. Select. Areas Commun.* **19**: 891–907 (May 2001).
15. S. ten Brink, Convergence behavior of iteratively decoded parallel concatenated codes, *IEEE Trans. Commun.* **49**: 1727–1737 (Oct. 2001).
16. S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* **44**: 591–600 (May 1996).
17. D. Divsalar and R. J. McEliece, Effective free distance of turbo codes, *Electron. Lett.* **32**: 445–446 (Feb. 1996).
18. S. Benedetto, R. Garello, and G. Montorsi, A search for good convolutional codes to be used in the construction of turbo codes, *IEEE Trans. Commun.* **46**: 1101–1105 (Sept. 1998).
19. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Soft-input soft-output modules for the construction and distributed iterative decoding of code networks, *Eur. Trans. Telecommun.* **9**: 155–172 (March 1998).
20. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (March 1974).
21. R. J. McEliece, On the BCJR trellis for linear block codes, *IEEE Trans. Inform. Theory* **42**: 1072–1091 (July 1996).
22. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**(2): 429–445 (March 1996).
23. P. Robertson, E. Villebrun, and P. Hoeher, A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain, *Proc. ICC'95*, Seattle, June 1995.
24. D. Divsalar and F. Pollara, Turbo codes for PCS applications, *Proc. ICC'95*, Seattle, June 1995.
25. X. Li and J. A. Ritcey, Trellis-coded modulation with bit interleaving and iterative decoding, *IEEE J. Select. Areas Commun.* **17**: 715–724 (April 1999).
26. P. Hoeher and J. Lodge, Turbo DPSK: Iterative differential PSK demodulation and channel decoding, *IEEE Trans. Commun.* **47**: 837–843 (June 1999).
27. A. Dejonghe and L. Vandendorpe, Low-complexity turbo-equalization for coded multilevel modulations, *Proc. SCVT2001*, Delft (The Nederlands), Oct. 2001.
28. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
29. B. E. Rimoldi, A decomposition approach to CPM, *IEEE Trans. Inform. Theory* **34**: 260–270 (March 1988).
30. H. N. Bertram, *Theory of Magnetic Recording*, Cambridge Univ. Press, 1994.
31. C. Laot, A. Glavieux, and J. Labat, Turbo equalization: adaptive equalization and channel decoding jointly optimized, *IEEE J. Select. Areas Commun.* **19**: 1744–1752 (Sept. 2001).
32. G. Montorsi and S. Benedetto, An additive version of the SISO algorithm for the dual code, *Proc. IEEE Int. Symp. Inform. Theory*, ISIT'2001, 2001.
33. A. Graell i Amat, G. Montorsi, and S. Benedetto, New high-rate convolutional codes for concatenated schemes, *Proc. IEEE Int. Conf. Communications*, ICC'2002, New York, May 2002.
34. A. Tarable, G. Montorsi, and S. Benedetto, A linear front end for iterative soft interference cancellation and decoding in coded CDMA, *Proc. IEEE Int. Conf. Communications*, ICC'2001, Vol. 1, 2001.

SERIALLY CONCATENATED CONTINUOUS-PHASE MODULATION WITH ITERATIVE DECODING

PÄR MOQVIST
TOR AULIN
Chalmers University of Technology
Göteborg, Sweden

1. INTRODUCTION

Digital radio communications has evolved from a tiny research area to mass-market products used by millions of people worldwide. Especially the 1990s saw a tremendous development of mobile telephony. Compared to earlier analog radio systems, digital communications enables cheaper, smaller, and more complex devices through the progress in integrated digital circuit technology. Naturally, this draws increasing attention to research in digital communications.

In this article, we focus on the transmission of digital messages of any kind; thus, we do not cover how these messages are produced. They may stem from an analog source that is digitized (such as speech, audio, or video), or they may be digital in their original nature (such as text messages and computer files). It is, however, essential that the message does not contain any redundancy, that is, that it consists of digital symbols that can be considered independent and identically distributed from a statistical point of view. This normally requires some pre-processing such as speech or video coding, or data compression.

Even in digital communications, messages are converted to analog signals before transmission on a physical channel. This process, known as digital modulation, is adapted to the specific transmission medium in question, such as coaxial cables, optical fibers, or radio channels. Common digital modulation formats include binary and quadrature phase shift keying (BPSK/QPSK), quadrature amplitude modulation (QAM), frequency shift keying (FSK), and digital phase modulation (PM).

A radio transmitter normally consists of a modulator producing radiofrequency (RF) signals, a signal amplifier and an antenna. In many applications, it is both easier and cheaper to use a nonlinear amplifier than a linear one, meaning that only phase-modulated signals can be amplified without considerable distortion. Thus, modulation formats which also vary the signal amplitude (e.g., BPSK, QPSK, QAM) are not suitable in conjunction with nonlinear transmitter amplifiers. It is here that the broad class of continuous-phase modulation (CPM) schemes find their major applications [1,2]. Examples include the GSM mobile telephony standard, microwave radio links, and ground-to-satellite communications.

Reliable digital transmission can be obtained by channel coding, as discussed by Shannon [3] in his 1948 classic paper. Block codes and convolutional codes are basic channel codes that are able to lower the bit error rate (BER) in the receiver at the cost of increased complexity. More advanced codes can be constructed from these basic elements by concatenation. One such technique is Turbo coding or parallel concatenated codes (PCCs), introduced by Berrou et al. [4–7], in which two convolutional codes separated by a random interleaver (permuter) are concatenated in parallel. The two codes are decoded separately in an iterative manner (see Fig. 1). Turbo codes have been shown to perform very close to Shannon's limit, while maintaining a reasonable decoder complexity. The same is true for serially concatenated codes (SCCs) with random interleaving, as shown in Fig. 2 - as examined by Benedetto et al. [8,9]. Turbo codes are currently being introduced in the third-generation mobile system standards UMTS (Universal Mobile Telecommunications System) [10] and cdma2000 [11].

Channel coding and modulation can also be combined, as in convolutionally coded CPM. By a joint design of the code and the CPM system, improved BER performance can be obtained. This leads us to the technique of combining CPM and SCCs with random interleaving, namely, serially concatenated CPM (SCCPM) investigated by several authors [12–14]. Here, the inner (second) code in an SCC is substituted with a CPM modulator, producing a coded and interleaved CPM scheme which can be decoded iteratively (see Fig. 3). As for Turbo codes and SCCs, the use of a random interleaver leads to substantial performance gains. In this article, we describe SCCPM in detail and show some interesting performance examples.

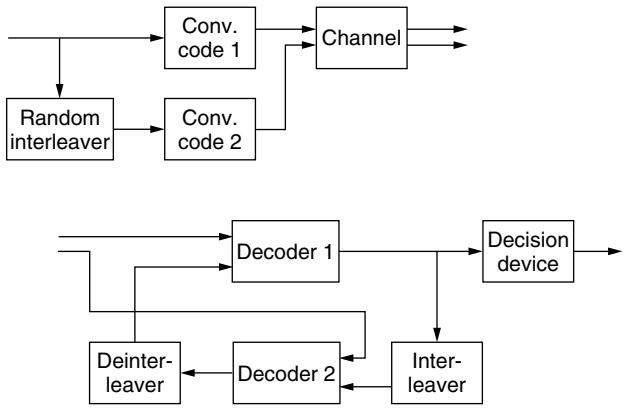


Figure 1. A typical Turbo code (PCC) with encoder (upper) and iterative decoder (lower).

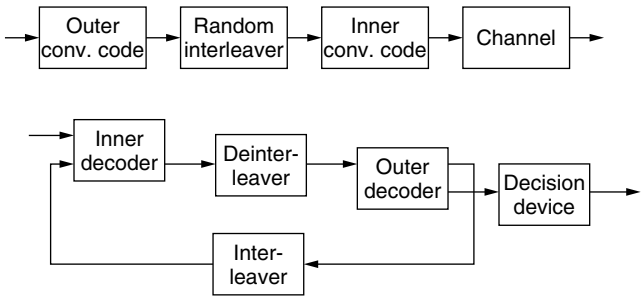


Figure 2. Serially concatenated code (SCC) with encoder (upper) and iterative decoder (lower).

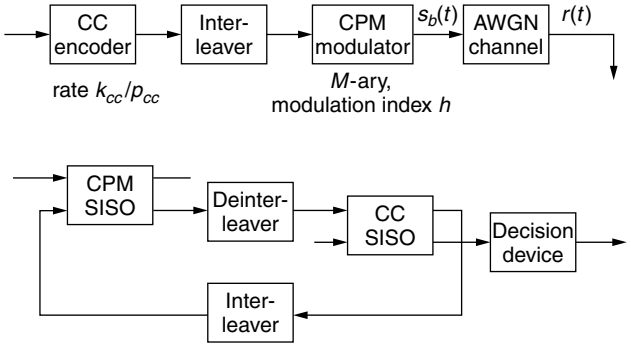


Figure 3. Serially concatenated continuous-phase modulation (SCCPM) with iterative decoding.

2. BACKGROUND

The introduction of the Turbo codes (PCCs) in 1993 [4] led to a vast research interest in concatenated convolutional codes separated by a pseudorandom bit interleaver and decoded iteratively. This technique of achieving large and nearly random codes while maintaining a simple decoder structure has found many applications, primarily where a large coding gain is desired (such as in deep-space applications or terrestrial links). The Turbo codes have shown to yield BERs around 10^{-5} at rates well beyond the channel cutoff rate, [7].

Shortly after the introduction of the turbo codes, it turned out that serially concatenated codes (SCCs) with a pseudo-random bit interleaver were equally suitable for iterative decoding [8]. It is generally understood that SCCs based on convolutional codes can give even better performance than turbo codes for low BERs (typically $<10^{-6}$) [9], while PCCs are better suited for medium-to-high BERs. Still, many concatenated coding schemes assume a simple carrier frequency modulation such as BPSK, meaning that a rate-k/n SCC transmits k/n bits per channel symbol. This bandwidth efficiency can be improved by more advanced modulation techniques such as trellis-coded modulation (TCM) [15] and continuous-phase modulation (CPM) [1,2]. Both TCM and CPM introduce memory into the transmitted signal, in a way that allows them to be described by finite-state machines similar to a convolutional code. Benedetto et al. employed TCM in an SCC with iterative decoding, such that the inner code is a recursive trellis code designed for the chosen signal set [16]. This way, it is possible to use, for example, a rate-$\frac{2}{3}$ SCC combined with an 8-PSK signal set yielding a bandwidth efficiency of 2 bits per symbol.

On the other hand, CPM is the natural choice of modulation when a constant envelope of the transmitted signal is required, such as when the transmitter amplifier is not perfectly linear. CPM was subject to vast research interest in the late 1970s and early 1980s [1], and the reader is referred to Ref. 2 for a comprehensive summary of much of that work. Later, convolutionally coded CPM was investigated as a way of improving performance while maintaining the constant envelope [17–20]. Matched codes [21] were introduced as yielding the minimum number of states in the optimum receiver when combined with CPM. The inherent modulo-2π property of the information-carrying phase in CPM has been exploited [22], such that convolutional codes over rings were combined with CPM. Since these codes showed better performance than did previous ones, much effort was put into code searches for various CPM systems [22–24].

In this article we study SCCPM, *i.e.* coded and interleaved CPM with iterative decoding. First, a brief description of CPM and how it is employed in a serially concatenated system is given (Section 3). Then, we study a transfer function bound to the BER performance, which is based on known concepts from SCCs (Section 4). This gives theoretical insights to the behavior of SCCPM in general. In Section 5, these observations are compared with numerous computer simulation results, demonstrating that the excellent performance of Turbo codes and SCCs indeed can be generalized to a trellis-coded modulation like CPM. Finally, in Section 6, bandwidth considerations are discussed and a bandwidth–performance comparison of some selected systems is presented.

3. SYSTEM DESCRIPTION

3.1. Block Diagram

Figure 3 shows a block diagram of SCCPM on an additive white Gaussian noise (AWGN) channel with iterative decoding. Independent and equiprobable information bits are encoded by a convolutional code (CC) with rate $R_{cc} = k_{cc}/p_{cc}$. For convenience, let p_{cc} be an integer multiple of $\log_2 M$, where M is the symbol alphabet size (cardinality) of the CPM system. As in Turbo codes and SCCs, bitwise block interleaving with block size N bits is employed, yielding groups of $\log_2 M$ bits that are mapped to CPM symbols using some mapping rule, such as natural mapping or gray mapping. The CPM modulator produces a constant envelope signal $s_b(t)$ that is transmitted on the AWGN channel. In this theoretical treatment, continuous transmission is assumed, thus the encoders are not reset to the zero state at the beginning of an interleaver block. The resulting concatenated code is a block code with information words of length $N \cdot R_{cc}$. (In a more practical system, blockwise decoding would be facilitated by trellis termination of both the outer CC and the inner CPM system [25], such that the encoders are enforced to the zero state at both the beginning and the end of a block. This normally leads to a small performance degradation.) In complex baseband representation, the received continuous-time signal is given by $r(t) = s_b(t) + n(t)$, where $n(t)$ is complex white Gaussian noise with double-sided power spectral density $N_0/2$.

3.2. Iterative Decoder

As in Turbo codes and SCCs, the iterative decoder consists of two *a posteriori* probability (APP) algorithms, one for the inner CPM system and one for the outer CC; more specifically, we use the bitwise sliding-window SISO algorithm proposed by Benedetto et al. [26] and formally justified by the present authors [27]. The SISO

algorithm is a nice generalization of the Bahl et al. (BCJR) algorithm [28], taking a priori probabilities of both information symbols and code symbols, and computing extrinsic APPs of both information symbols and code symbols. An extrinsic APP is an APP from which the a priori probability has been divided.

For the inner SISO, channel observations are used as a priori probabilities of the code symbols of the CPM system, as described in Section 3.4. The inner SISO then computes extrinsic APPs of the information bits of the CPM system; these are deinterleaved and used as a priori information on the code bits of the CC in the outer SISO. Applying the constraints of the CC, the outer SISO updates this information to extrinsic APPs that are interleaved and used as a priori information on the information bits of the CPM system in the next iteration. At the same time, the outer SISO computes APPs of the information bits of the CC; these are used by the decision device to select the bit with maximum APP in the last iteration. As can be seen from Fig. 3, those inputs/outputs not needed are left unconnected. A uniform distribution of the bits is assumed for an unconnected input. Note that at the first iteration, no a priori information on the CPM information bits is available in the inner SISO.

3.3. Description of CPM

Rimoldi [29] gave a concise description of CPM as the concatenation of two separate devices: a continuous-phase encoder (CPE) and a memoryless modulator (MM). He used the concept of a tilted-phase representation of CPM, which we repeat here for convenience.

Consider a CPM system with M-ary information symbols $u \in \{0, 1, \ldots, M-1\}$ transmitted every symbol interval T with energy E. To match it to the CC, let $\log_2 M$ be an integer number. The symbols are phase-modulated using a positive normalized frequency pulse $g(t)$ containing no impulses and being nonzero for L symbol intervals; thus the system is full response ($L = 1$) or partial response ($L > 1$). The phase response $q(t)$ is the integral of the frequency pulse, and it is normalized to $q(LT) = \frac{1}{2}$. The LREC and LRC families [2] are examples of frequency pulses.

The modulation index h is assumed to be rational and irreducible, $h = K/P$, since then the system can be described by a trellis [1]. The tilted-phase $\psi(t)$ during symbol interval n ($t = \tau + nT$) is given by [29]

$$\psi(\tau + nT) = \left[2\pi h \left[\sum_{i=0}^{n-L} u_i \mod P \right] + 4\pi h \sum_{i=0}^{L-1} u_{n-i} q(\tau + iT) + W(\tau) \right] \mod 2\pi \qquad 0 \le \tau < T \quad (1)$$

where $\{u_i\}$ are information symbols and the data-independent function $W(\tau)$ is given by

$$W(\tau) = \pi h (M-1) \frac{\tau}{T} - 2\pi h (M-1) \sum_{i=0}^{L-1} 9(\tau + iT) + \pi h (M-1)(L-1) \quad (2)$$

The transmitted signal during the same interval is

$$s(\tau + nT) = \mathrm{Re}\left\{ s_b(\tau + nT) \cdot \exp\left[j(2\pi f_1(\tau + nT) + \varphi_0) \right] \right\} \quad (3)$$

where $s_b(\tau + nT) = (2E/T)^{1/2} \cdot \exp[j\psi(\tau + nT)]$ is the complex baseband equivalent, $f_1 = f_0 - h(M-1)/2T$ is a shift of the center frequency f_0, and φ_0 is the initial phase of the carrier. Note from these relations that the transmitted signal during symbol interval n is completely specified by the current symbol u_n, the $L-1$ previous data symbols $u_{n-L+1}, \ldots, u_{n-1}$, and the accumulated value

$$v_n = \sum_{i=0}^{n-1} u_i \mod P \quad (4)$$

which can take only P values. Hence, in each symbol interval n, the CPE produces the code symbol (vector) $\boldsymbol{c}_n = [v_n, u_{n-L+1}, \ldots, u_n]$ to the MM, which transmits one of $P \cdot M^L$ signals $s_b(t, \boldsymbol{c}_n)$.

3.4. Channel Observations in CPM SISO Algorithm

As mentioned above, the SISO algorithm takes *a priori* probabilities of both information symbols, $\Pr(u_n = U)$, and code symbols, $\Pr(\boldsymbol{c}_n = \boldsymbol{C})$. For the inner CPM SISO, the latter are replaced by channel observations $p(\mathbf{r}_n | \mathbf{c}_n = \mathbf{C})$, where $\mathbf{r}_n$ is a sufficient statistic [30] obtained from $r(t)$. It can be shown [12] that the channel observations are proportional to $\exp[\mathrm{Re}\{r_k\}/N_0]$, where r_k is the sampled output of a filter matched to $s_b(t, \mathbf{c}_n)$,

$$r_k = \int_{nT}^{(n+1)T} r(t) s_b^*(t, \mathbf{c}_n)\, dt \quad (5)$$

and where the superscript * denotes complex conjugate. This sufficient statistic is exactly the same as when performing maximum-likelihood sequence detection (MLSD) in uncoded CPM [2], for example, using the Viterbi algorithm.

4. ANALYSIS OF SCCPM

In this section, an SCCPM system like the one in Fig. 3 is analyzed with the overall goal to demonstrate its ability to provide performance gains similar to Turbo codes and SCCs. These latter codes were analyzed through average transfer function bounds to the BER by Benedetto et al. in Refs. 6 and 9, respectively. As demonstrated in Ref. 12, this type of bound can be applied also to SCCPM. However, for all these codes, several arguments can be raised against the use of the transfer function bound:

1. Since this bound is based on a union bound, it diverges below a signal-to-noise ratio (SNR) threshold of approximately 3 dB, well above the operating point of many Turbolike systems.
2. The bound is a theoretical one in the sense that it is a bound on the average BER over all possible interleavers, each appearing with equal probability (the so-called uniform interleaver [6]). A specific

interleaver may yield substantially lower or higher BER. In fact, more recent interleaver design methods have shown improvements by orders of magnitude compared to the uniform interleaver [e.g., [31]].

3. The transfer function bound is based on optimal MLSD, which has a tremendous complexity for Turbolike systems because of their large and nearly random interleavers. The performance of the suboptimal iterative decoder can deviate substantially from MLSD, especially for low SNRs. Note that MLSD is equivalent to maximum APP sequence detection (MAPSD) because of the independent and equiprobable information bits.

Still, in this article we adhere to the transfer function bound because (1) performance above 3 dB is also of interest, and the transfer function bound is easy to obtain; (2) our primary goal is to demonstrate the Turbolike performance of SCCPM in general, as well as to compare different combinations of CCs and CPM systems; and (3) MLSD/MAPSD represents the ultimate performance limit of the system with any decoder, iterative or not. We do not address the convergence of the iterative decoder in this article, but refer the reader to other articles [e.g., 32–34,38]. The points presented above should anyway be kept in mind when bounds are compared with simulation results.

4.1. Transfer Function Bound for SCCPM

Similar to an SCC [9], the combination of a rate-R_{cc} CC, a bitwise length-N interleaver, and M-ary CPM is a block code with information words of length $k = N \cdot R_{cc}$. However, it does not possess the uniform error property, because CPM is not a linear code. Hence we cannot assume the all-zero information word to be transmitted; instead all pairs of transmitted and hypothesized (candidate) information words in the MLSD receiver must be considered. This is simplified by the fact that error events in CPM do not in general depend on the specific transmitted sequence, but only on the difference sequence between transmitted and hypothesized CPM symbols. Thus we need only enumerate the set of difference sequences in the inner CPM system, if we keep track of the fraction of transmitted sequences they correspond to. (Details of CPM error events can be found in Ref. 2.)

A union bound on the BER $P_b(e)$ in SCCPM with MLSD is

$$P_b(e) \le \frac{1}{2} \sum_{\mathcal{D}} B_{\mathcal{D}} \exp\left(-\frac{\mathcal{D} R_{cc} E_b}{2N_0}\right) \tag{6}$$

where E_b is the energy per information bit entering the CC and

$$B_{\mathcal{D}} = \sum_{w} \frac{w}{N \cdot R_{cc}} \overline{A}_{w,\mathcal{D}} \tag{7}$$

is the bit error multiplicity for error events with normalized squared Euclidean distance (NSED) $\mathcal{D}$ on the channel. The latter is assumed to be normalized to the energy per bit entering the CPM system; thus we use the same NSED as in uncoded CPM (2). $\overline{A}_{w,\mathcal{D}}$ is the number of concatenated error events with input weight w (Hamming distance $d_H(\cdot) = w$ between information words) and output weight (NSED) $\mathcal{D}$, averaged with respect to the number of transmitted codewords they correspond to. Thus $\overline{A}_{w,\mathcal{D}}$ is the input/output weight enumerating function (IOWEF), or input/output weight spectrum of the system. Assuming a uniform interleaver of length N bits, it can be shown that $\overline{A}_{w,\mathcal{D}}$ is obtained — the same as for SCCs — as

$$\overline{A}_{w,\mathcal{D}} = \sum_{l=0}^{N} \frac{A_{w,l}^{out} \cdot \overline{A}_{l,\mathcal{D}}^{in}}{\binom{N}{l}} \tag{8}$$

from the IOWEFs of the outer convolutional code and the inner CPM system. However, due to the linearity of the CC, its IOWEF need not be averaged. The denominator counts the number of distinct permutations that the uniform interleaver can produce from an output weight l codeword from the outer encoder, entering the inner CPM system as an input weight l codeword [9].

4.2. Error Events in CPM

Clearly, in concatenated and interleaved systems the input weight of error events is of interest. In the outer CC, the input weight counts the number of bit errors made during an error event, as given by the factor w in Eq. (7). But even more important, the input weight of error events in the inner CPM system determines not only which concatenated error events will be possible, but also their multiplicity, as can be seen from Eq. (8).

As an example, consider minimum shift keying (MSK) (binary 1REC, $h = \frac{1}{2}$). Error events in MSK start when the transmitted and hypothesized CPM symbols differ from each other for the first time in the block (see Fig. 4). When this occurs, there will be a Hamming distance of 1 between the bit labels of the transmitted and hypothesized symbols, that is, the input weight will increase by 1. At the same time, the NSED of the error event also will increase by 1. If the symbols differ again, the input weight and NSED will increase by the same amount, and the error event will end. However, if they do not differ, the Hamming distance will be 0 and the input weight will not increase, but the error event will continue. In this case, the NSED will increase by 2. Eventually, the symbols will differ once again, and a total input weight of 2 will be obtained, while the NSED will be any multiple of 2 depending on the length of the error event. Concluding, isolated error events in MSK will always have input weight 2, and the minimum NSED will also be 2.

For SCCs, the inner code should be a recursive systematic convolutional (RSC) code to avoid input weight 1 error events. This is because these events nullify

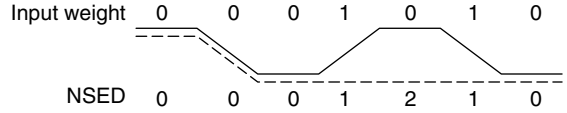


Figure 4. Input weight of an error event in MSK. Transmitted (solid) and hypothesized (dashed) symbol sequences.

the performance gain of an increased interleaver size N, the *interleaver gain* [9]. For CPM with $M > 2$, the input weight depends on the bit labeling (mapping) of the CPM symbols, including natural mapping or gray mapping. In addition, certain combinations of mapping and modulation index $h = K/P$ yield input weight 1 error events, despite the inherent recursive nature of CPM. More specifically, the following conditions apply:

- Binary ($M = 2$) CPM: integer modulation indices $h = K$. These systems are seldom used in practice.
- Quaternary ($M = 4$) CPM with natural mapping: $h = K/2$. With gray mapping: $h = K/3$.
- Octal ($M = 8$) CPM with natural mapping: $h = K/4$. With gray mapping: $h = K/3$, $h = K/5$, and $h = K/7$.
- Hexadecimal ($M = 16$) CPM with natural mapping: $h = K/8$. With gray mapping: $h = K/5$, $h = K/7$, $h = K/9$, $h = K/11$, $h = K/13$, and $h = K/15$.

Thus it is always possible to avoid input weight 1 error events, and hence to obtain interleaver gain, by a careful selection of the mapping.

4.3. Transfer Function Bound for Coded and Interleaved MSK

To find the IOWEF of the CC for all combinations of w and l, the methods described by Benedetto and Montorsi [6] or Divsalar et al. [35] can be used. The former is a recursive algorithm suitable when the output weights are integers. Regarding the inner CPM system, the IOWEF generally contains a manifold of real-valued NSEDs $\mathcal{D}$. However, for the important special case of minimum shift keying (MSK) (binary 1REC, $h = \frac{1}{2}$), all NSEDs are integers and thus the algorithm in Divsalar et al. [35] can be applied.

As an example, consider coded MSK with a uniform interleaver of length $N = 128$, 512, 2048, or 8192. The outer code is the 4-state, nonrecursive, nonsystematic rate-$\frac{1}{2}$ CC specified by the octal connection polynomial (7,5) (generating matrix $G(D) = [1 + D + D^2, 1 + D^2]$) and having free distance 5. Using the IOWEFs discussed above, Eqs. (7) and (8) give the total bit error multiplicity for each output weight $\mathcal{D}$ in the concatenated system, and Eq. (6) gives the upper bound on the average BER, shown in Fig. 5. Note the so-called divergence of the bound for E_b/N_0 (SNR) values below approximately 3 dB. This phenomenon has been observed also for Turbo codes [6] and SCCs [9], and is an inherent effect of the union bound. Above 3 dB, the bound exhibits a rather steep slope similar to those of SCCs. This promises for good practical performance, even with a suboptimal iterative decoder and a real, pseudorandom interleaver. Comparing the bounds for different interleaver sizes, it can be noted that when N increases by a factor of four, the BER bound decreases with roughly a factor 64; thus the interleaver gain appears to be approximately N^{-3}.

4.4. Detailed Analysis of Error Events

With a careful examination of the BER bound in Eqs. (6)–(8), several observations can be made regarding the behavior of SCCPM systems, when a uniform interleaver and MLSD is employed. For large SNRs, the minimum NSED $\mathcal{D}_{\min}$ dominates the bound. Thus SCCPM (as well as Turbo codes and SCCs) is no different from other coded modulation systems. However, the true strength of turbo-like systems is the spectral thinning of the distance (weight) distribution that occurs for large interleavers. This means that the lowest weights (distances) have small bit error multiplicities $B_{\mathcal{D}}$, effectively causing a displacement of the distance distribution toward larger weights. A "thin" weight spectrum gives improved BER performance for small-to-medium SNRs provided that the dominance of the exponential function in Eq. (6) is not too strong (i.e., that E_b/N_0 is not too large). As discussed

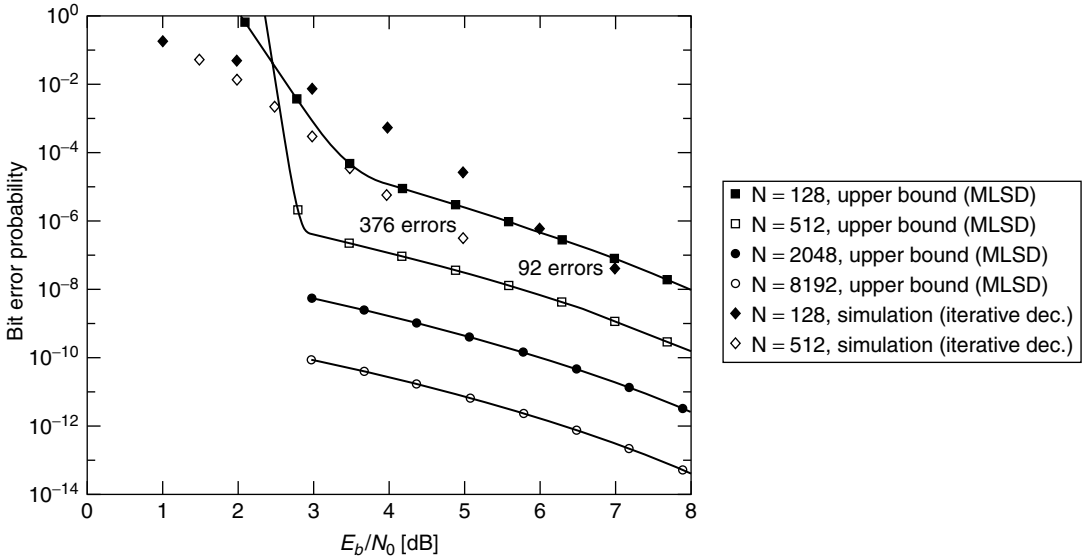


Figure 5. BER bounds and simulation results for (7,5)-coded MSK for various interleaver sizes N. The bounds for $N = 2048$ and $N = 8192$ were calculated from truncated weight spectra.

earlier, Turbolike systems have interleaver gain such that a larger interleaver decreases the bit error multiplicity [cf. Eq. (8)], and thus achieves spectral thinning.

As demonstrated for SCCs by Benedetto et al. [9], the dependency of $B_\mathcal{D}$ on the interleaver size N can be investigated with the aid of the BER bound. Specifically, $B_\mathcal{D}$ can be bounded from above, for each output weight $\mathcal{D}$, by a polynomial in N whose maximum exponent is $\alpha(\mathcal{D})$. When $\alpha(\mathcal{D}) < 0$, there is interleaver gain for the specific output weight $\mathcal{D}$. It can be shown that in general, $\alpha(\mathcal{D}_{\min}) < 0$ such that there is asymptotic (large E_b/N_0) interleaver gain.

However, the overall maximum value of the $\alpha(\mathcal{D})$, $\alpha_{\max}$, is more interesting because it determines whether there is interleaver gain over the whole SNR range. For SCCs, Benedetto et al. distinguish two separate cases [9]: (1) block and nonrecursive convolutional inner codes and (2) recursive convolutional inner codes. In case (1), input weight 1 events exist, yielding $\alpha_{\max} \geq 0$; thus, interleaver gain cannot be assured for all SNRs. In case (2), $\alpha_{\max} = -\lfloor (d_{\text{free}} + 1)/2 \rfloor < 0$ for an outer code with free distance d_{free}, because there are no input weight 1 events in the inner code. The crucial point is thus the existence of these events. We have already seen in Section 4.2 that certain CPM systems in fact do have such events — and we will also see that this affects the BER performance negatively. MSK does not have input weight 1 error events, so $\alpha_{\max} = -3$ and $B_\mathcal{D} \leq N^{-3}$ when it is combined with the (7,5) CC, for which $d_{\text{free}} = 5$. This is consistent with the bounds for different interleaver sizes in Fig. 5.

We can also study the NSED associated with $\alpha_{\max}$, $\mathcal{D}(\alpha_{\max})$, as an alternative to $\mathcal{D}_{\min}$ for small-to-medium SNRs. (For large SNRs, $\mathcal{D}_{\min}$ dominates the bound.) Let $\mathcal{D}^{\text{in}}(i)$ denote the minimum NSED for input weight i error events in the inner CPM system, and if no such event exists, let $\mathcal{D}^{\text{in}}(i) = \infty$. Thus, the classic minimum distance in CPM is the minimum of $\mathcal{D}^{\text{in}}(i)$ over all i. $\mathcal{D}^{\text{in}}(i)$ is given directly by the IOWEF of the CPM system, but it can also be obtained from a search in the difference trellis [12]. For inner CCs in SCCs, $\mathcal{D}^{\text{in}}(2)$ is called the *effective free distance* [9]. It is important since it can be shown [9] that $\mathcal{D}(\alpha_{\max})$ is given by a concatenation of input weight 2 events in the inner code, such that

$$\mathcal{D}(\alpha_{\max}) = \left\lfloor \frac{d_{\text{free}} + 1}{2} \right\rfloor \cdot \mathcal{D}^{\text{in}}(2) \tag{9}$$

(This equation is not valid when d_{free} is odd and $\mathcal{D}^{\text{in}}(3)$ is finite [12].) Clearly, performance can be improved not only by increasing the interleaver size, but also by a larger free distance in the outer CC, or a larger effective free distance in the inner CPM system. However, one should be careful with conclusions from $\mathcal{D}(\alpha_{\max})$ since it does not necessarily dominate the bound. As shown in another study [12], there are more combinations of error events that may play a role.

Example 1. Consider once again (7,5)-coded MSK. For the inner MSK system, $\mathcal{D}^{\text{in}}(2) = 2$ and $\mathcal{D}^{\text{in}}(i) = \infty$ for all other i. Hence $\mathcal{D}(\alpha_{\max}) = 6$ which coincides with $\mathcal{D}_{\min}$ for this specific system. It is interesting to note that this NSED is of the same order as that in many traditional coded CPM systems. However, the associated bit error multiplicity B_6 is overbounded by N^{-3}, which can be made as small as desired by enlarging the interleaver. This spectral thinning is not so easily obtained in traditional coded CPM.

Example 2. (7,5)-coded binary 2RC, $h = \frac{3}{4}$. Here, there are no odd input weight events in the CPM system, while $\mathcal{D}^{\text{in}}(2) = 2.66$. Therefore, $\mathcal{D}(\alpha_{\max}) = 7.97$ which is larger than $\mathcal{D}_{\min} = 5.21$. There is also a combination with $\mathcal{D} = 6.59$. The bit error multiplicities for these events are $B_{7.97} \leq N^{-3}$, $B_{6.59} \leq N^{-4}$, and $B_{5.21} \leq N^{-5}$. Thus the resulting performance is a mixture of the contributions to the bound of these events. Still, we can compare this system with Example 1 by observing that the event with lowest NSED (5.21) has a factor N^2 lower multiplicity, while the other events have larger NSEDs. Thus we can expect it to perform better than Example 1, at least above the divergence threshold around 3 dB. From the simulations in section V, this appears to be the case also for smaller SNRs.

4.5. Summary

Our observations can be summarized as follows. Note that they are generally valid only above the divergence threshold around 3 dB for systems employing a uniform interleaver and MLSD, although we will see in Section 5 that this appears to be the case also for more practical scenarios where a pseudo-random interleaver and iterative decoding is used in the low-SNR region.

- Like SCCs, SCCPM systems are capable of providing interleaver gain, if input weight 1 error events are avoided. This is always possible by a change in the mapping, such as from natural to gray, or vice versa. The order of the interleaver gain is determined by the free distance of the outer code, d_{free}.
- For a general inner CPM system, several combinations of the minimum NSEDs per input weight contribute to the BER bound. In order to compare different CPM systems, these combinations must be examined in detail for a given outer CC.
- It is always possible to avoid input weight 1 events by selecting a different mapping, specifically, natural binary instead of gray, or vice versa.

5. EXAMPLE SYSTEMS

The analysis in Section 4 is based on a union bound on the BER with MLSD, assuming a uniform interleaver. In practice, a suboptimal iterative decoder and a pseudorandom interleaver are used instead. Still, above the divergence threshold at E_b/N_0 around 3 dB, we shall see that the iterative decoder indeed performs quite close to the upper bound for MLSD. Below this value, the bound diverges and thus observations based on it may not be accurate; furthermore, the iterative decoder may not even converge toward the MLSD decision. All these points should be taken into account when performance of the iterative decoder is studied.

In this section, some examples of SCCPM systems are presented together with computer simulation results. We use (7,5)-coded MSK as our reference system, and then study the influence of other outer codes and inner CPM systems. Finally, a system with inner input weight 1 events is evaluated. In all simulations, the number of iterations[1] (8–12) as well as the delay of the sliding-window APP algorithms (10–15 symbols) are chosen such that no practical deterioration of performance could be noticed. The interleavers are chosen at random.

5.1. Reference System: (7,5)-Coded MSK

For this simple SCCPM scheme (Example 1, above), the MLSD transfer function bound is compared with simulation results for the iterative decoder in Fig. 5 for $N = 128$ and 512. For large SNRs, the actual interleaver and iterative decoding used in the simulations give slightly better performance than does the uniform interleaver in the MLSD bound, although the BER is not very accurate in the region below 10^{-6}. In fact, experience with Turbo codes tells us that most interleavers perform better than the average represented by the uniform interleaver. However, the discrepancy around the divergence threshold only be explained can by the fact that a suboptimal iterative decoder is used instead of MLSD.

Figure 6 shows more simulation results for different interleaver sizes, and also without any interleaver (i.e., a traditional coded CPM system, which was treated by Lindell [17] assuming MLSD detection). The interleaved system is evaluated using the iterative decoder (which has $4 + 2$ states) for $N = 128$, 512, 2048, 8192, and 32,768. For the noninterleaved system, both the iterative decoder and MLSD (8 states) is employed. Clearly, without interleaver, iterative decoding is inferior to MLSD by approximately 1.5 dB. At a BER of 10^{-3}, this transforms to a gain of more than 3 dB when a large interleaver is inserted.

[1] Here, and throughout the article, one decoder iteration corresponds to a pass through both the inner and the outer SISO module.

5.2. Influence of the Outer CC

According to the analysis above, the main influence of the outer code is the impact of its free distance d_{free} on the interleaver gain. Here, four different rate-$\frac{1}{2}$ outer CCs concatenated with MSK are investigated, whereby one is the reference system above. The codes are the 2-state (2,3) code, the 4-state (7,5) code, the 16-state (23,35) code, and the 64-state (133,171) code, and all codes are nonrecursive, nonsystematic convolutional codes. They have $d_{\text{free}} = 3$, 5, 7, and 10, yielding interleaver gains between N^{-2} and N^{-5}, and implying $\mathcal{D}(\alpha_{\max}) = 4$, 6, 8, and 10.

In Fig. 7, the codes are compared at an input delay of 4096 information bits ($N = 8192$). As can be seen, the required E_b/N_0 for a given BER $\geq 10^{-4}$ actually *increases* with d_{free}. This is in contrast to the analysis, where a larger d_{free} would give a larger interleaver gain. However, as pointed out earlier, this is true only above the divergence threshold, and for sufficiently large interleavers. In the

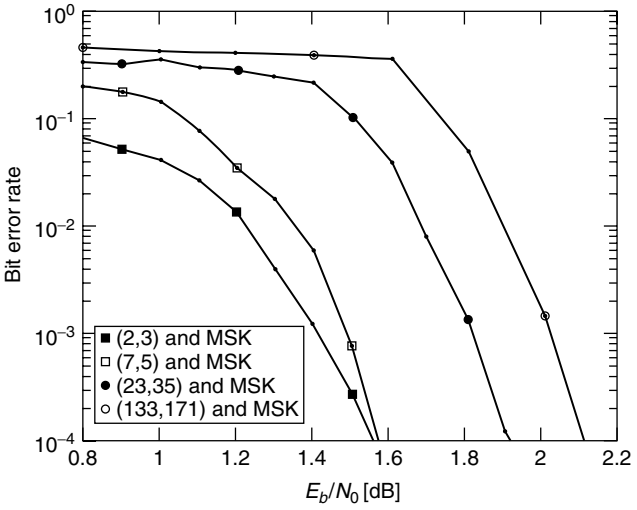


Figure 7. Simulation results for four different outer convolutional codes (CCs) with MSK, at a delay of 4096 information bits ($N = 8192$). All systems utilize the same bandwidth.

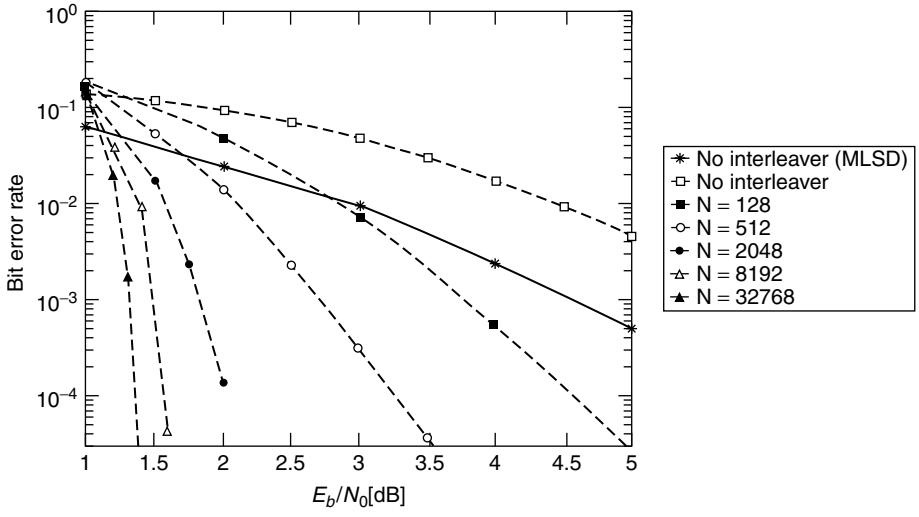


Figure 6. Simulation results for (7,5)-coded MSK ($4 + 2$ states in the iterative decoder).

low-SNR region, outer codes with more states suffer from convergence problems, as has been reported for SCCs [33]. The gain of a larger d_{free} can thus not be exploited in this SNR region. Still, it can be noticed that the 4-state code has a steeper BER slope than the 2-state code above 1.4 dB. Thus we can expect it to perform better asymptotically.

5.3. Influence of the Inner CPM System

We saw in section IV that the BER bound for SCCPM is made up of several combinations of error events in the inner CPM system, which depend on the outer CC. Therefore, in this section, we compare four different CPM systems, including MSK, concatenated through an $N = 8192$ interleaver with the same outer (7,5) CC. The systems certainly have different bandwidths, but we defer that comparison to Section 6. Their dominant error event combinations, as calculated in an earlier study [12], are shown in Table 1.

System 1 is the reference system (MSK, 2 states). System 2 is Example 2 (in Section 4) (binary 2RC, $h = \frac{3}{4}$, 8 states), which is supposed to perform better than system 1. System 3 is binary 3RC, $h = \frac{2}{3}$ (12 states), and system 4 is binary 3RC, $h = \frac{4}{5}$ (20 states). Studying the error events in Table 1 we find that system 3 should be slightly better than system 1, but worse than system 2, and that system 4 should perform about as well as system 2. These conclusions generally hold only above the divergence threshold of the BER bound, but the simulation results in Fig. 8 indicate that the relations stay the same in the low-SNR region as well, although the difference between systems 1 and 3 is larger than expected. Note that system 1 has the largest $\mathcal{D}_{\min}$, but it performs worst of these four systems.

Figure 9 shows simulation results from Ref. 36 for some SCCPM systems with $M = 4$, 8, and 16, together with system 1. The rate of the CC is adapted accordingly, such that $R_{cc} = \frac{1}{2}$, $\frac{2}{3}$, and $\frac{3}{4}$, respectively. In order to compare the systems at equal encoder delays (4096 information bits), interleaver sizes of $N = 8192$, 6144, and 5460 bits were chosen. As can be seen, increasing h and M gives improved performance, although this need not be true for arbitrary modulation indices. As demonstrated earlier [36], 3RC systems generally give worse BER performance — keeping everything else constant — than do their 2RC counterparts, but they consume less bandwidth. We will return to bandwidth in Section 6. In Fig. 9, the best BER performance is obtained by rate-$\frac{3}{4}$ coded $M = 16$ 2RC, $h = \frac{1}{2}$, which reaches BER $= 10^{-3}$ at $E_b/N_0 = 0.35$ dB. This is even better than the rate-$\frac{1}{2}$ Turbo codes with BPSK presented by Berrou et al. [4], but the bandwidth is slightly larger.

Concluding, for SCCPM systems utilizing the same outer code, performance below the divergence threshold is roughly in line with the error event analysis, and the convergence problem is not very pronounced. Still, methods of examining the convergence behavior [32–34] could give insights beyond the traditional union bound approach.

Table 1. Dominant Error Events for Some SCCPM Systems

System	$B_{\mathcal{D}} \sim N^{-5}$	$B_{\mathcal{D}} \sim N^{-4}$	$B_{\mathcal{D}} \sim N^{-3}$
1. (7,5) and binary 1REC, $h = \frac{1}{2}$ (MSK)	—	—	$B_6 \sim 480N^{-3}$
2. (7,5) and binary 2RC, $h = \frac{3}{4}$	$B_{5.21} \sim 2880N^{-5}$	$B_{6.59} \sim 1440N^{-4}$	$B_{7.97} \sim 480N^{-3}$
3. (7,5) and binary 3RC, $h = \frac{2}{3}$	—	$B_{3.55} \sim 120N^{-4}$	$B_{6.06} \sim 60N^{-3}$
4. (7,5) and binary 3RC, $h = \frac{4}{5}$	—	$B_{5.54} \sim 120N^{-4}$	—

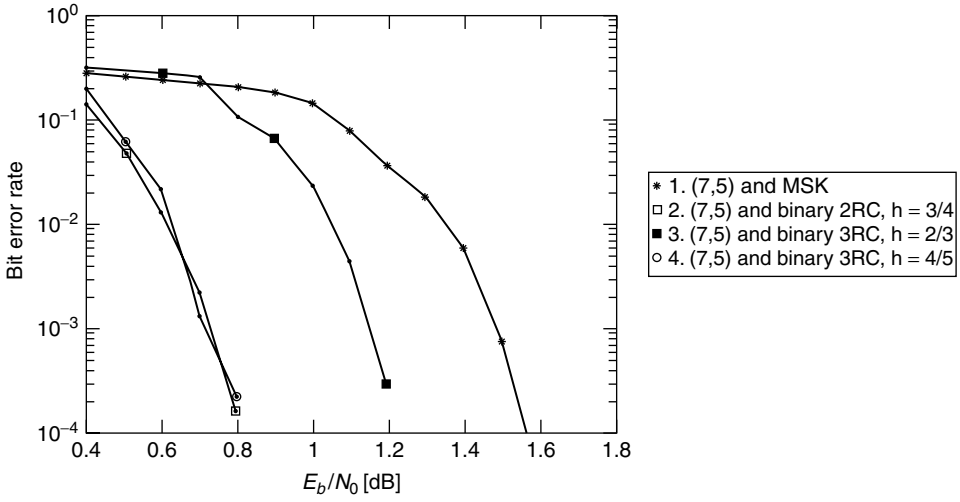


Figure 8. Simulation results for four different inner CPM systems with the (7,5) CC, at a delay of 4096 information bits ($N = 8192$). See Fig. 11 for a bandwidth comparison of these systems.

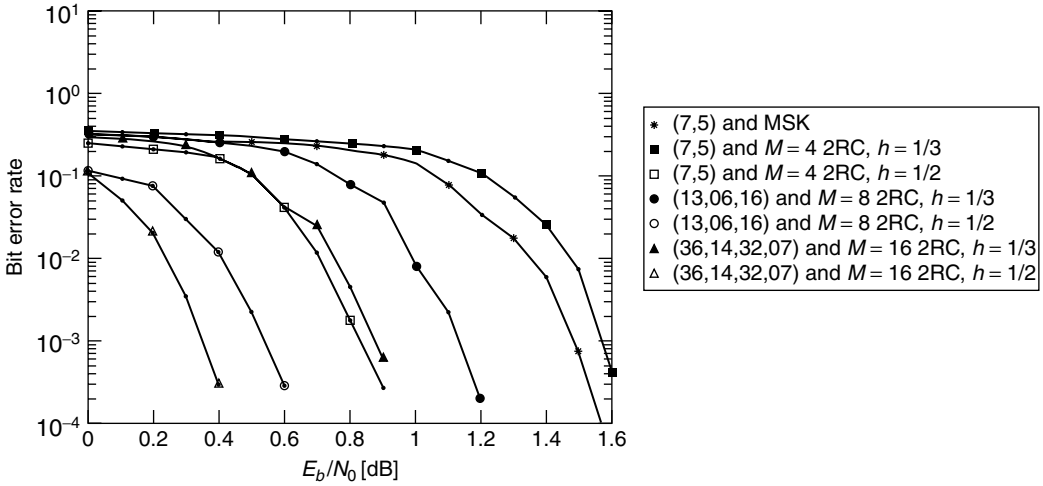


Figure 9. Simulation results for some $M = 4$, 8, and 16 SCCPM systems with rates $\frac{1}{2}$, $\frac{2}{3}$, and $\frac{3}{4}$ outer CCs, respectively, and interleaver sizes $N = 8192$, 6144, and 5460. See Fig. 11 for a bandwidth comparison of some of these systems.

5.4. Influence of Inner Input Weight 1 Error Events

It was stated in Section 4 that if the inner CPM system has input weight 1 error events, the interleaver gain would not be assured for all SNRs. We investigate this phenomenon by considering quaternary 1REC, $h = \frac{1}{3}$ with natural and gray mapping (3 states) together with the (2,3) and (7,5) outer CCs (2 and 4 states, respectively). Figure 10 shows BER simulation results with an $N = 8192$ interleaver. Clearly, the systems with Gray mapping — and thus with input weight 1 error events — do not provide any spectral thinning caused by the random interleaver; hence their BER curves have a modest slope that is due only to the distance profile in Eq. (6). Notice that for the 4-state code, gray mapping still gives better performance down to a BER of 10^{-5}.

6. POWER SPECTRAL DENSITY AND BANDWIDTH

Bandwidth is at least as important as BER performance in many communication systems. The demand for higher data rates in wireless data communications and increased capacity in digital mobile telephony networks is growing constantly. Spectrum reuse in the form of small-cell networks can meet these requirements to a certain extent, but bandwidth-efficient coding and modulation is nevertheless essential. In this section, we evaluate the bandwidth efficiency of SCCPM in general, and compare it with power efficiency (BER) for some selected examples.

For a coded CPM system without interleaving, Ho and McLane [20] have calculated the true power spectral density. However, since random bit interleaving is used in SCCPM, an accurate estimate (see also Ref. 17 and 20) of the baseband power spectral density $S(f)$ is

$$S_{\text{coded}}(f) \approx R_{cc} S_{\text{uncoded}}(R_{cc} f) \tag{10}$$

The power spectral density of the carrier-modulated SCCPM signal is given by

$$P_c(f) = \frac{P}{2}\left[S_{\text{coded}}(f - f_0) + S_{\text{coded}}(-f - f_0)\right] \tag{11}$$

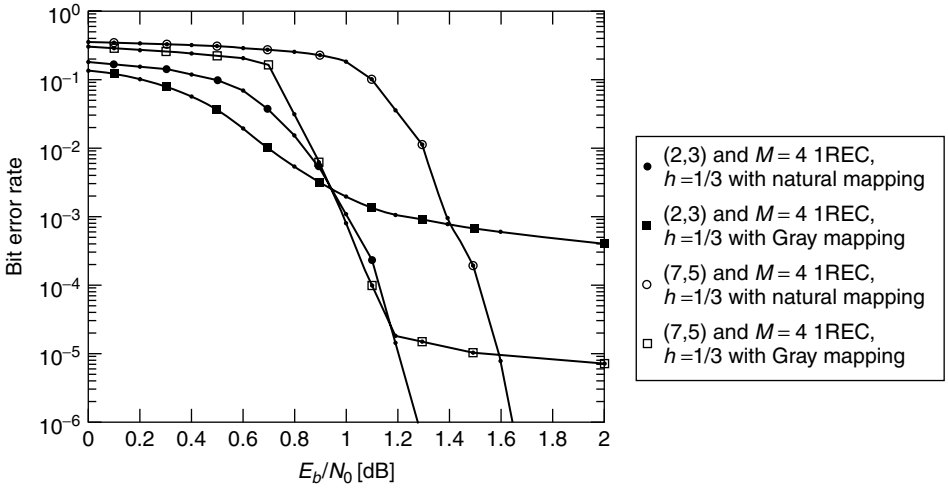


Figure 10. The effect of input weight 1 error events in the inner CPM system, $N = 8192$. Gray mapping yields such events; natural mapping does not.

where $P = E/T$ is the average transmitted signal power. Similarly, the normalized double-sided bandwidth of the coded CPM system $2B_{\text{coded}}T_b$ can be expressed in terms of the corresponding quantity of the uncoded CPM system, $2B_{\text{uncoded}}T_b$:

$$2B_{\text{coded}}T_b \approx \frac{2B_{\text{uncoded}}T_b}{R_{cc}} \tag{12}$$

Note that the rate of the CC is included in the normalized bandwidth in Eq. (12). The bandwidths are defined as 99% in-band power, equivalent with the −20 dB level in the fractional out-of-band power function [2,17]. The uncoded bandwidth of a CPM system can be calculated using the general method described in a previous publication [37].

The normalized bandwidths (bandwidth efficiencies) of some SCCPM systems are compared with their error performance in Fig. 11. Interleaver sizes corresponding to an input delay of 4096 information bits were used, namely, $N \approx 4096/R_{cc}$. Error performance is measured in terms of the required E_b/N_0 for a simulated BER of 10^{-3}, and plotted against the normalized double-sided bandwidth $2B_{\text{coded}}T_b$. Lines are used to interconnect points for similar systems but with different modulation indices. Notice that these lines do not provide any information on modulation indices in between the points. As can be seen, over the whole bandwidth range shown in Fig. 11, the lowest E_b/N_0 among these systems is obtained for the rate-$\frac{2}{3}$, (13,06,16) CC, a 6144-bit interleaver, and $M = 8$ 2RC. At $2B_{\text{coded}}T_b \approx 2.0$, it requires only 0.5 dB, and at $2B_{\text{coded}}T_b \approx 1.15$, it requires roughly 1.6 dB. The latter can be compared to uncoded MSK ($2B_{\text{coded}}T_b \approx 1.20$), which requires 7.3 dB for a BER of 10^{-3}. Thus the gain over uncoded MSK is around 5.7 dB. For smaller BERs, the gain increases because the SCCPM system has a much steeper BER slope. However, if a low-complexity system is the goal, the quaternary 1REC systems may provide a good tradeoff. The rate-$\frac{3}{4}$-coded $M = 16$ 2RC systems are not shown here because they are less power/bandwidth-efficient than are the rate-$\frac{2}{3}$-coded $M = 8$ 2RC systems, as demonstrated earlier [36].

How does SCCPM compare with a typical Turbo code with BPSK? The answer to this question depends on the bandwidth required to transmit one BPSK symbol. While theoretically, a normalized bandwidth of 1 is sufficient if signals extend over infinite time, we here assume a 99% bandwidth of 1.2 as being more practical. Then, a rate-$\frac{1}{2}$ Turbo code will have $2B_{\text{coded}}T_b = 2.4$, while typically yielding a BER of 10^{-3} at roughly 0.8 dB for an input delay of 4096 bits (Berrou et al. [4] obtained 0.6 dB with 65536 bits). This is slightly inferior to the rate-$\frac{2}{3}$-coded $M = 8$ 2RC, $h = \frac{1}{2}$ SCCPM system, which requires around 0.55 dB with $2B_{\text{coded}}T_b \approx 2$.

Finally, we compare some of the coded, noninterleaved CPM systems with MLSD, investigated by Lindell [17]. These systems have large asymptotic gains (upto dB over uncoded MSK), but here we still use simulated BER $= 10^{-3}$ as the performance measure. As can be seen from Fig. 11, SCCPM is able to provide gains of 1.5–2.0 dB at this BER level, although with increased receiver complexity. At a lower BER, the gain would increase even further because the SCCPM systems generally have a much steeper slope than do the noninterleaved systems. For example, rate-$\frac{3}{4}$ and $M = 16$ 1REC, $h = \frac{2}{13}$ with MLSD requires 5.7 dB for BER $= 10^{-5}$, while rate-$\frac{2}{3}$ and $M = 8$ 2RC, $h = \frac{1}{4}$ with a 6144-bit interleaver and iterative decoding requires 1.9 dB, yielding a gain of 3.8 dB.

7. CONCLUSIONS

SCCPM with iterative decoding is an exciting new coding and modulation technique, merging the principles of both coded CPM and Turbo codes. With the simple insertion of a random interleaver between the outer code and the CPM system, dramatic performance gains

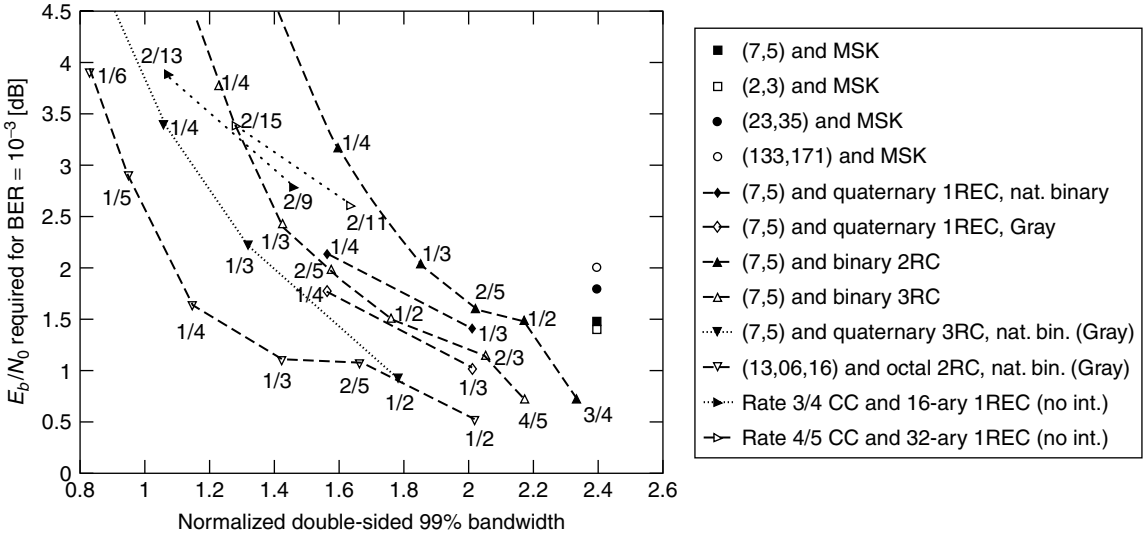


Figure 11. Bandwidth/performance comparison for an input delay of 4096 bits and BER $= 10^{-3}$. The numbers indicate modulation indices. "Nat. bin. (Gray)" refers to gray mapping only for modulation indices that yield input weight 1 error events with natural mapping.

become feasible. Since MLSD (Viterbi decoding) is practically unreasonable, this requires a rather different iterative decoder that consists of SISO algorithms for the constituent code and modulation, passing APPs between each other. This decoder is very similar to that used in SCCs with memoryless modulation.

SCCPM can be analyzed through a transfer function bound to the BER, using a number of assumptions. This includes the averaging over all possible interleavers (the uniform interleaver assumption), the use of an MLSD receiver instead of the iterative decoder, and the assumption of an SNR above 3 dB, which is the divergence threshold of the bound. Still, the transfer function bound gives theoretical insights such as the following:

1. Similar to Turbo codes and SCCs, SCCPM is capable of providing interleaver gain, the order of which is determined by the free distance of the outer CC.
2. Several error events in the inner CPM system contribute to the overall performance; hence it is difficult to give general design guidelines for the inner CPM system.
3. However, CPM systems with input weight 1 error events should be avoided since they do not give interleaver gain. This is always possible by a different choice of mapping.

Simulation results for a variety of SCCPM systems indicate that, in addition to these three conclusions, the convergence of the iterative decoder also affects performance in the low-SNR (waterfall) region. For unequal outer CCs with similar CPM systems, performance in the waterfall region actually deteriorates with increasing free distance of the outer code. This is consistent with conclusions from SCCs. On the other hand, for similar outer CCs with different CPM systems, the transfer function bound analysis appears to give satisfactory conclusions. Also, we demonstrated that CPM systems with input weight 1 error events indeed fail to provide Turbolike performance. We also demonstrated a number of higher-order SCCPM systems with $M = 4$, 8, and 16. These systems both have attractive BER performance and bandwidth efficiency. Comparing power and bandwidth efficiency, it is seen that the $M = 8$ 2RC systems are the most efficient as known today. These systems can provide gains of almost 4 dB at BER $= 10^{-5}$, at the same bandwidth, compared to the best known coded CPM systems without interleaving.

In conclusion, wherever a constant envelope of the transmitted signal is desired, SCCPM should be able to provide appealing power and bandwidth efficiencies. As of today, the decoding complexity may seem high, but we are strongly convinced that in a near future, this will not pose any major obstacles. Of course, much work remains to be done on implementation issues such as timing and carrier phase recovery, reduced sampling rates, and quantization to fix-point representation.

BIOGRAPHIES

Pär Moqvist received his M.S. and Lic.Eng. degrees in electrical engineering from Chalmers University of Technology, Göteborg, Sweden, in 1993 and 1999, respectively. In 1994, he was awarded the John Ericsson Medal for outstanding scholarship. He joined Ericsson AB in 1993 as a radio system engineer, working on the design of digital radio modems for land-mobile applications. Since 1997 he has been a Ph.D. student with the Telecommunication Theory Group at Chalmers University. His research is in the general area of signaling and detection for digital communication systems, with a current focus on iterative decoding methods.

Tor M. Aulin received his M.S. degree in electrical engineering from the University of Lund, Lund, Sweden, in 1974 and the Dr. Techn. (Ph.D.) from the Institute of Telecommunication Theory, University of Lund, Göteborg, Sweden in November 1979. He became a docent there in 1981 and he was also a visiting scientist at the ECSE Department at Rensselaer Polytechnic Institute, Troy, New York. One year was spent at the European Space Agency (ESA, ESTEC) as an ESA research fellow. In 1983, he became a research professor (docent) in information theory at Chalmers University of Technology, Göteborg, Sweden. In 1991, he formed the Telecommunication Theory group there and became a docent in computer engineering in 1995. During 1995 he was a visiting fellow at the telecommunications engineering department, Australian National University, Canberra, ACT, Australia. Some of Dr. Aulin's research interests are communication theory, combined modulation/coding strategies (such as CPM and TCM), analysis of general sequence detection strategies and digital radio channel characterization. Dr. Aulin has authored the book *Digital Phase Modulation*, Plenum Press 1986. He is a fellow of the IEEE and an editor for *IEEE Transactions on Communications* in the area of communication theory and coding. In 1997, he was awarded the prestigious Senior Individual Grant, handed over by the Prime Minister of Sweden.

BIBLIOGRAPHY

1. T. Aulin, *CPM—a Power and Bandwidth Efficient Digital Constant Envelope Modulation Scheme*, Ph.D. dissertation, Univ. Lund, Lund, Sweden, 1979.
2. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
3. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 623–656 (1948).
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo codes, *Proc. Int. Conf. Communications* (*ICC'93*), 1993, pp. 1064–1070.
5. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (1996).
6. S. Benedetto and G. Montorsi, Unveiling turbo codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**: 409–428 (1996).
7. S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* **44**: 591–600 (1996).
8. S. Benedetto and G. Montorsi, Serial concatenation of block and convolutional codes, *Electron. Lett.* **32**: 887–888 (1996).

9. S. Benedetto et al., Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (1998).
10. *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Multiplexing and Channel Coding* (FDD) (Release 1999), (3GPP TS 25.212 V3.5.0), 2000.
11. *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, TIA/EIA/IS-2000.2-A, 2000.
12. P. Moqvist, *Serially Concatenated Systems: An Iterative Decoding Approach with Application to Continuous Phase Modulation*, thesis, Chalmers Univ. Technology, Göteborg, Sweden, 1999; *http://www.ce.chalmers.se/TCT/*.
13. C. Brutel and J. Boutros, Serial concatenation of interleaved convolutional codes and M-ary continuous phase modulations, *Ann. Telecommun.* **54**: 235–242 (1999).
14. K. R. Narayanan and G. L. Stüber, Performance of trellis coded CPM with iterative demodulation and decoding, *Proc. Global Telecommunications Conf.* (*GLOBECOM'99*), 1999, pp. 2346–2351.
15. E. Biglieri et al., *Introduction to Trellis-Coded Modulation with Applications*, Macmillan, New York, 1991.
16. S. Benedetto et al., Serial concatenated trellis coded modulation with iterative decoding, *Proc. Int. Symp. Information Theory* (*ISIT'97*), 1997, p. 8.
17. G. Lindell, *On Coded Continuous Phase Modulation*, Ph.D. dissertation, Univ. Lund, Lund, Sweden, 1985.
18. S. V. Pizzi and S. G. Wilson, Convolutional coding combined with continuous phase modulation, *IEEE Trans. Commun.* **33**: 20–29 (1985).
19. F. Morales-Moreno and S. Pasupathy, Structure, optimization and realization of FFSK trellis codes, *IEEE Trans. Inform. Theory* **34**: 730–741 (1988).
20. P. Ho and P. J. McLane, The power spectral density of digital continuous phase modulation with correlated data symbols, *IEE Proc.* **133**(Pt. F): 95–114 (1986).
21. F. Morales-Moreno, W. Holubowicz, and S. Pasupathy, Optimization of trellis coded TFM via matched codes, *IEEE Trans. Commun.* **42**: 1586–1594 (1994).
22. R. H.-H. Yang and D. P. Taylor, Trellis-coded continuous-phase frequency-shift keying with ring convolutional codes, *IEEE Trans. Inform. Theory* **40**: 1057–1067 (1994).
23. B. Rimoldi and Q. Li, Coded continuous phase modulation using ring convolutional codes, *IEEE Trans. Commun.* **43**: 2714–2720 (1995).
24. G. Karam, I. Fernandez, and V. Paxal, New coded 8-ary CPFSK schemes, *Proc. Int. Conf. Communications,* (*ICC'93*), 1993, pp. 1059–1063.
25. P. Moqvist and T. Aulin, Trellis termination in CPM, *Electron. Lett.* **36**: 1940–1941 (2000).
26. S. Benedetto et al., A soft-input soft-output APP module for iterative decoding of concatenated codes, *IEEE Commun. Lett.* **1**: 22–24 (1997).
27. P. Moqvist and T. Aulin, Certain aspects on MAP algorithms for turbo codes, *Proc. 17th Swedish Conf. Radio Science and Communication* (*RVK'99*), 1999, pp. 623–625.
28. L. R. Bahl et al., Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (1974).
29. B. Rimoldi, A decomposition approach to CPM, *IEEE Trans. Inform. Theory* **34**: 260–270 (1988).
30. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
31. M. Breiling, S. Peeters, and J. Huber, Interleaver design using backtracking and spreading methods, *Proc. Int. Symp. Information Theory* (*ISIT'2000*), 2000, pp. 451.
32. H. El Gamal, *On the Theory and Application of Space-Time and Graph Based Codes*, Univ. Maryland, College Park, MD, 1999.
33. S. ten Brink, Convergence of iterative decoding, *Electron. Lett.* **35**: 806–808 (1999).
34. D. Divsalar, S. Dolinar, and F. Pollara, Iterative turbo decoder analysis based on Gaussian density evolution, *Proc. IEEE Military Communications Conf.* (*MILCOM 2000*), 2000, pp. 202–208.
35. D. Divsalar et al., Transfer Function Bounds on the Performance of Turbo Codes, TDA Progress Report, Jet Propulsion Lab., Pasadena, CA, 42-122, 1995, pp. 44–55.
36. P. Moqvist and T. Aulin, Power and bandwidth efficient serially concatenated CPM with iterative decoding, *Proc. IEEE Global Telecomm. Conf.* (*GLOBECOM'00*), 2000, pp. 790–794.
37. T. Aulin and C.-E. Sundberg, An easy way to calculate power spectra for digital FM, *IEE Proc.* **130**(Pt. F): 519–526 (1983).
38. P. Moqvist and T. Aulin, Convergence Analysis of SCCPM with iterative Decoding, *Proc. IEEE Global Telecomm. Conf.* (*GLOBECOM'01*), 2001, pp. 1048–1052.

SERVICES VIA MOBILITY PORTALS

DANIEL RALPH
CHRIS SHEPHARD
BTexact Technologies
Ipswich, Suffolk, United Kingdom

1. INTRODUCTION

Mobility portals look set to become the window through which the user will access a whole range of innovative services [4]. Wherever and whenever, the mobility portal will warn you, inform you or just entertain you.

There is a distinction between the mobility portal and the mobile portal, the concept of mobility extends to include terminal independence through user profiles, additional value add services such as "find me, follow me" call routing and integration with existing systems in the fixed network. The mobile portal however is being expressed as a cut-down version of existing Web-based applications such as news, weather, and email.

The explosion in use of Short Message Service (SMS) for sending text messages demonstrates a natural evolution path to WAP services on smartphone and personal digital assistant (PDA) devices. Support from many vendors has created an environment where the applications delivered through the mobile portal will increasingly substitute access through the fixed network.

However, although the Internet and mobile phones have independently been successful (see Figs. 1 and 2), this

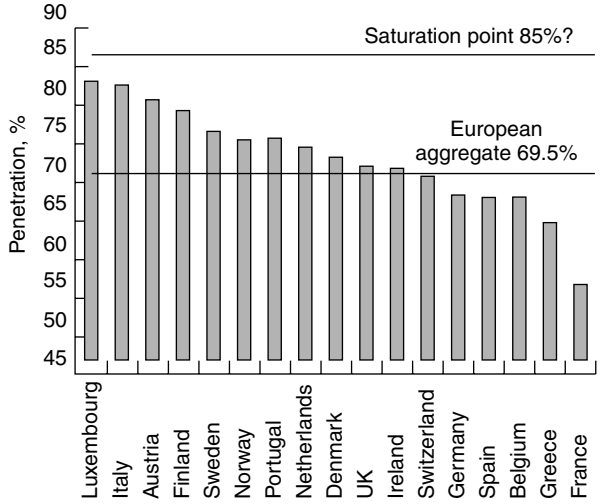


Figure 1. European mobile penetration by country, June 2001 (%) (*Source*: Morgan Stanley Research).

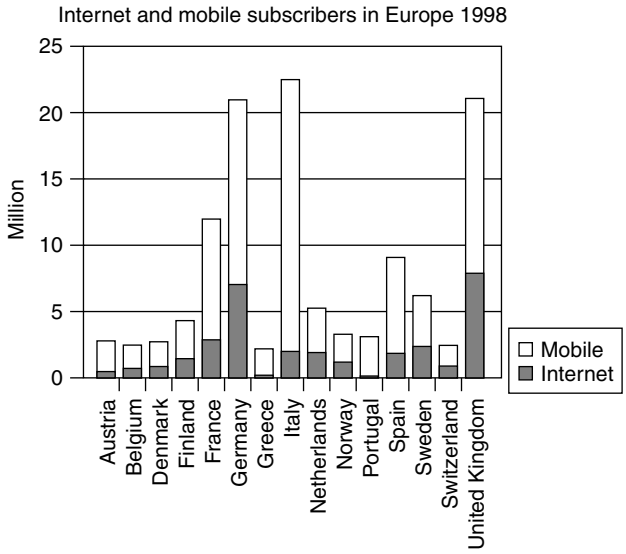


Figure 2. Internet and mobile subscribers in Europe, 1998. (*Source*: Dataquest, Mobile Communications International, Computer Industry Almanac [10]).

does not always mean that combining these technologies will also be successful. Evidence of this is clear in the combination of TV and phone technologies, both successful in their own right but the uptake of personal videophones has not happened. Mobile videophone is considered a killer application. However, it will become clear from this article that to achieve a high takeup is unlikely in the foreseeable future due to bandwidth limitations and device battery life. A more likely scenario is that a short videoclip downloaded in non–real time could be played back to the user to provide visual information.

In addition, users may subscribe to specialized information feeds for stock trading, weather, ski and snow conditions, and horoscopes. Finally, the services will offer specialized alerting through partners such as auction, travel sites, banking, and messaging providers.

2. THE RISE OF THE MOBILITY PORTAL

Without doubt the explosive growth rate of Internet users and the significance of accessing content through the use of Web-based portal services will continue. While this highlights the demand for content from the fixed network, it cannot be assumed that the same type or level of demand will be present in the mobile environment. The differences in network and device capability will require a different solution in providing content through the mobile portal.

In assessing why the mobility portal is important, it can be seen from looking across the value chain that a number of key elements are available to be exploited in delivering new mobility services:

- The existing infrastructure will support these services, in terms of both core IP networks and content availability. Whether this is from the customers Internet service provider or other content providers, it will enable the rise of the mobility portal.
- The prospect of "always on" packet-based services delivered over 2.5G and 3G networks such as GPRS and UMTS will enhance the user experience in the way content is delivered over the next-generation infrastructure provided by the mobile operator.
- The first-generation consumer equipment is readily available and supports basic web access and WAP services. This is in the form of smartphone and PDA devices.
- In the short term, existing content can be redeveloped for delivery to the smartphone or PDA device, this will enable the rapid deployment of mobile services. Although this will only be through provision of existing generic Web portal services on the mobile device, this will include search engines, personalization, snap-shot text information and basic messaging.

It is the very issue of content and what services the user will require that will determine the successful uptake of service access through the mobility portal. Figure 3 illustrates some of the anticipated benefits from a wireless portal.

Questions have to be answered to determine the customer's needs when mobile: What do they want? When? Why? Providers of mobility portals will need to assess customer requirements and actual usage to develop context awareness so as to present the most relevant content. If the network knows the weather conditions, user location, time, and whom the user is accompanied by, then this gives the network an opportunity to provide information ahead of the user requirement for this information, further developing the mobile device as a lifestyle tool [12].

Future applications need to develop the unique attributes of the mobile device. The opportunity for the network to provide location-based services, and the personal nature of the mobile device lend it to acting as a payment device.

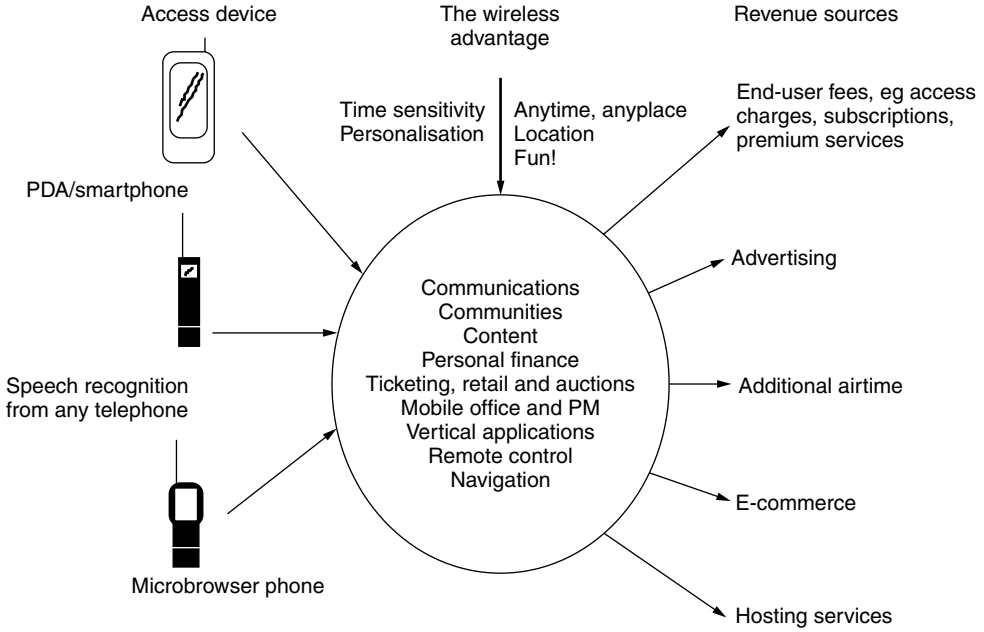


Figure 3. Benefits of a wireless portal.

3. INSIDE THE MACHINE

The current UK digital mobile network has evolved from the first-generation analog network and is based on the Global System for Mobile communications (GSM).

Network technologies are evolving toward the 3G mobile vision, this will require an evolution through extensions to the existing 2G mobile network technologies. In the near future this will involve

- GSM (2G) component technologies:
- SMS and cell broadcast
- GPRS (general packet Radio Service)
- EDGE (enhanced data rates for GSM Evolution)
- UMTS [3G] (Universal Mobile Telecommunications System)
- Bluetooth
- GPS and other location-based techniques (Mobile Positioning Protocol)
- Mobile agents and intelligent networks

The next-generation mobile system will rely heavily on emerging application technologies such as

- Wireless Application Protocol (WAP)
- XML/XSL (eXtensible Markup Language)
- JAVA Technology (Java 2 Micro Edition)
- SIM Application Toolkit
- Lightweight Efficient Application Protocol (LEAP) [1]
- Compact HTML
- XHTML

It is clear that the Internet will play a pivotal role, requiring increasing quality of service (QoS) and allowing better integration of applications across different terminals and bearers, probably using APIs to provide access to functionality.

The importance of voice must not be underestimated as a mechanism for controlling services by natural language commands, such as adding an appointment to your calendar or more simply as an alternative to typing a response to an email. Voice browsing of a mobility portal will be standardized through the use of VoiceXML and provide a development environment to deliver powerful niche applications [2].

Before service providers can offer content or applications using mobility portals, a number of issues require further consideration.

3.1. Wireless Application Protocol

Wireless Application Protocol (WAP) is a specification for a set of communication protocols to standardize the way that wireless devices, such as cellular telephones and radio transceivers, can be used for Internet access, including email, the World Wide Web, newsgroups, and Internet Relay Chat (IRC) [3].

While Internet access has been possible in the past, different manufacturers have used different technologies. In the future, devices and services that use WAP will be able to interoperate (see Fig. 4).

The WAP layers are

- Wireless application environment (WAE)
- Wireless session layer (WSP)
- Wireless transport layer (WTP)
- Wireless transport layer security (WTLS)
- Wireless datagram protocol (WDP)

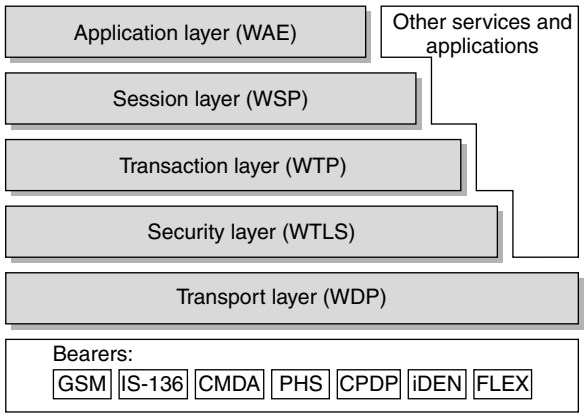


Figure 4. The WAP protocol stack.

Following independent development, four companies proposed the WAP standard: Ericsson, Motorola, Nokia, and Openwave (previously known as *phone.com*).

The use of transcoding engines (described as the HTML filter in Fig. 5) is being proposed as a "quick win" in providing services from existing Web content to the first generation of wireless terminals. This has met with limited success due to the complexity present in much of the multimedia-rich Web content. The use of JavaScript, Frames, and imagemaps causes in many cases a failure to transcode the html into a suitable WML page that can be displayed on a smartphone.

3.2. i-mode

The deployment of WAP in Europe must be contrasted with the developments in Japan, where the i-mode service has significantly better functionality. The service uses a Compact Hypertext Markup Language (CHTML) and is provided over the Personal HandyPhone System (PHS), which is a packet-based "always on" service, operating at 9.6 kbps (kilobits per second). The use of CHTML permits color graphics to be displayed or 10 lines of text, it still has limitations similar to the Wireless Markup Language as used in WAP [11].

There are many reasons for the success of i-mode. It relies on open technology, allowing any Internet site to join in; content is written in a simplified version of HTML (CHTML); and no fees are charged for placement on the i-mode portal. The services are positioned as a unique mobile service and not as "the Internet on your phone," thereby avoiding unfavorable comparisons with service from the fixed Internet. The low penetration of PCs in Japanese homes probably also helps control user expectations. Finally, NTT DoCoMo has concentrated on growing core revenues from airtime usage and has chosen not to support the service through advertising or transaction revenue.

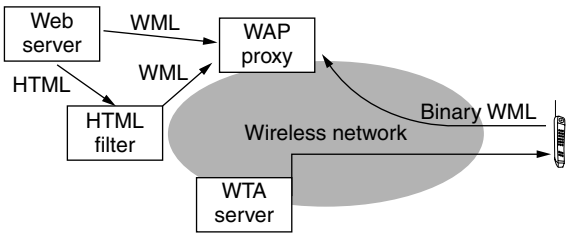


Figure 5. Example WAP network.

Mobile Internet services to compete with i-mode have now been introduced by rival operators. However, they have thus far failed to attract significant numbers of subscribers or content providers, despite offering greater bandwidth.

By being first to market, i-mode may now have an important head start, and competitors offering WAP based services cannot tap into the wealth of available i-mode content because of format incompatibilities.

i-mode is a brand name for NTT DoCoMo's Internet service. The technology and service offerings behind i-mode will continue to evolve to take advantage of improving network capabilities, including the introduction of 3rd generation mobile networks. In addition, DoCoMo has plans to roll out i-mode service in other countries, particularly in mobile networks where NTT owns a share. One example is the cooperation with Hutchinson in Hong Kong. At the same time, DoCoMo is active in the *WAP Forum*, *World Wide Web Consortium* (W3C), and Internet Engineering Task Force (IETF) whose standards, in combination, will map the way forward for a globally compatible mobile Internet.

The deployment of global mobility portals are essential to support the global traveler, they need to support the roaming capabilities of GPRS and GSM; and remove the added complexity for the user to change profiles to connect to the same services in different countries. The geographic distribution of network elements and the seamless access to services across different countries will strengthen the mobile operator's position as a service provider.

3.3. Terminals

The look and feel of the mobile device will have an impact on the user's perception of services, the usability of the device will have an impact on the usability of any service delivered over that device. Regardless of the content the portal provides, the terminal usability will prove significant in the promotion of services through mobility portals.

The types of mobile device range from the smartphone complete with NaviRoller to a pen based PDA device (see Figs. 6 and 7) and upto a subnotebook with keyboard and 1024×480 pixel display. Each of these devices has different attributes related to

- Screen size (resolution)
- Input method, such as through keyboard, touch screen, handwriting recognition
- Presence or absence of a speaker and microphone

The multiple methods for input and display are compounded by the variable and often low bit rate of mobile networks. In order to provide suitable content, device characteristics will need to be available to the service provider to allow appropriate formatting and display to

Figure 6. D503i mobile phone.

Figure 7. Next-generation PDA device with pen and speech based input, integral GPS and GPRS connectivity [9].

the range of devices outlined. The W3C have a working group on device capability and profile called Composite Capabilities/Preference Profiles (CC/PP) that is driving standards efforts to implement device profiles to enable content providers to determine device capability.

Further proposals from the Salutation consortium [6], Jini and Universal PnP (UpnP) [5] are intended to deliver service discovery and device recognition capabilities to enable the network to seek the services that the user requires and that are suitable for the terminal capability. For instance, the user may require a print service, but have available only an IrDa port with which to communicate with the printer.

The limitations of device capabilities in terms of storage and access to information will increase the need for additional support from the network. This may be provided through network based backup and synchronization services or server processing for voice recognition to authenticate to a service.

3.4. Addressing

It has been a concern for some time that the lack of IPv4 address space would prevent connections to the Internet. This problem will only become worsen with an explosion of mobile devices that have an "always on" connection to the Internet. Ultimately the proposed migration to IPv6 with its 128-bit address scheme will remove this problem with available address space for devices connected to fixed or mobile networks. The size of the IPv6 address range will allow 10^{20} addresses (or thousands of IPv6 address for every square inch of the earth.

Each mobile device that a user carries or interacts with will be uniquely identified by an IPv6 address and may well communicate to the user using Bluetooth devices connected within one's "personal bubble" or to an external network using UMTS. It is estimated that 1 billion Bluetooth addresses will be in use by 2005 [7].

The home environment is where a significant portion of this addressing will be required; in particular, the ability for devices around the home to communicate, to share information on your whereabouts, and to anticipate the family's needs. This could be extended to the fridge ordering fresh milk from the supermarket on your arrival home from holiday.

3.5. Billing

Getting information to the handset of a user is the simple part. Devising a mechanism that meets the commercial aspirations of all the players involved in making this happen is much harder. Many network operators will partner with a publisher to manage the day-to-day operations of their portal. Service providers will want recompense for the information provided.

Even in the simple case of portal provision, a number of revenue models have been developed to describe the different ways in which the value can be expressed. For example, a simple revenue share might be appropriate for a traffic-congestion report, the operator charging the user "per click" for information and passing on a share to the information provider.

Initially, WAP services are expected to be expensive to use since the tendency is to be online for a long circuit-switched data (CSD) call since features such as interactivity and selection of more information are used by the end user. With the introduction of GPRS handset the billing model is based on the number of packets sent or received by the user.

In contrast, i-mode uses an efficient micro billing system via the mobile phone bill, which makes it easy for subscribers to pay for value added services and premium sites and is also attractive for site owners, to sell information to users.

3.6. Security

There are several areas of WAP security that will need additional development to ensure the end-to-end security requirements for transactions involving payments or

instructions requiring nonrepudiation. These include what is being called "premature encryption endpoint." The WAP gateway acts as a proxy and decrypts the incoming WTLS session and encrypts as SSL3.0 for the outgoing connection to the Origin server.

The WAP specification 1.1 supports the use of certificates in a way similar to X.509 and encryption is supported at 56 and 128 bits. In the future the WAP specification will be extended to include the wireless identity module (WIM), which may be a replacement for your SIM card containing a personal digital signature to identify and authenticate the User. The addition of a crypto library will enable the developer of WAP services access to a range of functions that allow the User to digitally sign and encrypt a message before it is sent from the mobile device.

3.7. Interoperability

The provision of mobile data services in the past has always been using proprietary technologies [8]. The standardization of WAP allows developers and services to be less restricted in their range of mobile networks or mobile devices. The first release of gateway and device products have the inevitable problems associated with new technology and significantly more testing will be required between different vendors products. This is an area where the service provider can add value in providing an integrated service to an agreed level of quality and compatibility with other services.

3.8. Quality of Service

The issue of quality of service (QoS) has plagued IP networks since it became commercialized. Several IETF standards have been proposed to satisfy the user requirements, these include Resource Reservation Protocol (RSVP) with IETF Integrated Services (IntServ), Differentiated Services (Diffserv), and multiprotocol label switching (MPLS).

The mobile network suffers from QoS issues such as dropped connections, lack of resources, and reduced bandwidth due to too many users sharing the available capacity. These factors will significantly restrict the availability of mobile services that compete directly with fixed network services, such as videoconferencing or high-speed file transfer.

As users are demanding higher levels of QoS from IP networks for fixed services delivered on these networks, so, too, will the mobile user. The demand for real-time services may be sufficiently low to cause less difficulty, videostreaming can use buffering to delay the transmission. Videoconferencing cannot accept high network latency, but it may be that this is not the "killer application."

Improved quality of service will be available with the advent of 3G networks. The rollout plans for many 3G operators are progressing at a rapid pace, with statements from several outlining available data rates between 64 and 384 kbps on initial launch. With these achieving commercial service in the near future, NTT DoCoMo began trials over their installed 3G network in July 2001.

4. CONCLUSION

Future developments of the mobility portal must embrace the entire range of existing web based content if it is to succeed as the default method of access to the network. This will include all applications from gaming to mobile commerce.

Services should be designed so the relevant elements of a service are available over to the particular client device, the requirement for the network to recognize the device capability and network connection characteristics.

There will be no single "killer application" for the mobile device, although its unique attributes, including location positioning and personal nature, provide an opportunity for the mobile device to become the portal through which the user interacts whilst "on the move."

It is highly likely that WAP will be superseded by evolution of existing Internet standards, the requirement for legacy support will remain for WAP enabled mobile phone devices.

The reasons for the limited lifetime of WAP are due to

Low bandwidth
High latency
Low-resolution monochrome displays
Dropped calls and other quality of service issues
Low device processing power

All of these factors are not limitations of WAP; however, once they have been improved or overcome, the tendency will be to use existing Internet standards end to end, as opposed to the WAP architecture of creating a wireless version of these standards. Even if successful, once all devices and networks support WAP, it will cease to differentiate and therefore will not reduce subscriber churn. However, it will provide the user with convenient access to services whenever and wherever required.

With so many heavyweight carriers, equipment manufacturers and software and applications developers backing the Wireless Application Forum, a momentum is being generated that will drive WAP-based equipment and services forward into the future.

Despite the drawbacks discussed in this article, WAP certainly seems to be shaping up to play a major role in facilitating the brave new world of the personalized mobility portal. WAP is a powerful tool. It enables "anytime, anywhere" connectivity to a wealth of information, whether for leisure or business applications.

It is impossible to predict who will win between WAP and i-mode; clearly, DoCoMo has a significantly larger market share at the time of writing. However, if XML becomes the dominant content standard, there may be situations where both standards are absorbed into existing Internet protocols.

This article has demonstrated through the discussion of mobility portals some of the problems involved in providing content, applications, and filtering information to a new generation of wireless devices.

Acknowledgments
The author would like to thank the Institute of Electrical Engineers (IEE) for reproduction of this article.

BIOGRAPHIES

Daniel Ralph joined BTexact Technologies in 1996 following a period working in the biotechnology industry. Since joining BT he has worked on projects as diverse as the deployment of the trial global VoIP network for Concert and the design and implementation of the call processing engine "powering" the network intelligence platform. He is now responsible for a number of technical developments within the arena of mobile Internet technologies. He is a graduate of the Open University, and has recently completed the BT MSc in telecommunications. He is a corporate member of the BCS.

Chris Shephard read theoretical physics at UEA and then University of Birmingham, United Kingdom. He joined BTexact Technologie in 1980 and helped to develop signaling, call control, and maintenance software for System X PSTN and ISDN switches. After two years with Siemens in Florida, where he was part of a large project designed to adapt their range of digital public switching systems (EWSD) to the USA regulatory and market requirements. He returned to BTexact Technologies where he worked initially on the development and delivery of local network switches (LA-30) and subsequently on the design and management of two European broadband ISDN projects. He then switched to the study of network centric computing and the use of thin client devices. He now works in the terminals and applications unit within Multimedia Applications where he leads a team developing XML applications.

BIBLIOGRAPHY

1. The Lightweight & Efficient Application Protocol (LEAP) Manifesto, *http://www.freeprotocols.org/leap*.
2. Voice portals: Ready for Prime Time? (12/7/00), *http://www.zdnet.com/anchordesk/stories/story/0,10738,2601898,00.html*.
3. WAP Forum, WAP Architecture Specification (WAPARCH), April 30, 1998; URL: *http://www.wapforum.org*.
4. European Wireless Portal Use to Boom (7/7/00), *http://www.allnetdevices.com/industry/market/2000/07/07/europeanwireless.html*.
5. UPnP, Jini, Salutation, *http://www.cswl.com/whiteppr/tech/upnp.html*.
6. Service discovery and management, *http://www.salutation.org*.
7. Bluetooth SIG Adds Protocol (10/7/00), *http://www.allnetdevices.com/developer/news/2000/07/10/bluetoothsig.html*.
8. B. Johnston, C. Fenton, and D. Gilliand, Mobile data services, *BT Technol. J.* **14**(3): 92–108 (July 1996) (*http://www.bt.com/bttj*).
9. Future phones, *http://www.futurefonezone.com*.
10. UK demographics from CIA publications, *http://www.odci.gov/cia/publications/factbook/geos/uk.html*.
11. i-mode, *http://imodelinks.com/desktop/faq.html*.
12. W. N. Schilit, N. I. Adams, and R. Want, Context-aware computing applications, *Proc. Workshop on Mobile Computing Systems and Applications*, IEEE Computer Society Press, Santa Cruz, CA, 1994, pp. 85–90.

FURTHER READING

Official Wireless Application Protocol 2.0, *The Complete Standard with Searchable CD-ROM*, Wireless Application Protocol Forum, Ltd.

Holma H. and A. Toskala, eds., *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, rev. ed., New York, Wiley, 2000.

SESSION INITIATION PROTOCOL (SIP)

HENNING SCHULZRINNE
Columbia University
New York, New York

1. OVERVIEW

The Session Initiation Protocol (SIP) is a *signaling* protocol for setting up, tearing down, and modifying multimedia sessions. These sessions can consist of any number of media, such as audio, video, and shared applications, and can be either unicast (point-to-point) or multicast. SIP does not transport media itself, but rather allows endpoints (Internet hosts) to coordinate the exchange of media.

This coordination function becomes necessary since endpoints, called *user agents* by SIP, need to discover one another even if they change their Internet address. Endpoints also need to agree on the characteristics of the session, such as the number and type of media streams, the codecs to be used for audio and video, the properties of shared applications, or the availability of quality of service (QoS) mechanisms. SIP is based on the notion that an endpoint identifier, a SIP URI, is associated with a person or telephone number. A person keeps the same SIP URI even if they change their IP address or use different devices, such as a office PC, wireless 3G phone, or traditional telephone. The ability to reach a person with different devices under the same name is sometimes referred to as *personal mobility* [1].

SIP supports five aspects of managing multimedia sessions:

User location — given a SIP URI, SIP protocol entities find all relevant end systems that are registered for that URI.

User availability — based on information in the SIP request, the called end systems indicate their willingness to participate in the session or may indicate alternate locations.

User capabilities — through protocol exchanges, both end systems learn about the media capabilities of the other side.

Session setup — if there is an intersecting set of capabilities, the called party alerts the user ("rings") and, if a human controls the endpoint, the session may get established.

Session management — SIP requests can transfer sessions to others, terminate sessions, and modify session parameters.

SIP itself does not describe the media content; rather, it relies on external protocols for this task. Currently, SIP applications almost exclusively use the Session Description Protocol (SDP) [2] for this purpose, although other mechanisms can be negotiated between session participants. A newer version of the Session Description Protocol, called SDPng, is under development [3], but it is unclear how fast it can replace SDP.

SIP messages typically, but not necessarily, travel through a sequence of intermediaries, called *SIP proxies*. These intermediaries serve many functions, from user location and providing other services to policing reachability to opening firewalls for media streams. Often, the initial message traverses all such proxies, while subsequent messages within a session are exchanged directly between the two user agents.

SIP can use a variety of network-layer and transport-layer protocols. SIP works equally well over IPv4 [4] and IPv6 [5]; the latter is particularly important for its application to next-generation wireless networks. Currently, operation has been specified for UDP [6] and TCP [7], as well as for SCTP [8,9]. For reliable transport protocols such as TCP and SCTP, SIP can also benefit from transport-layer hop-by-hop security between end systems and proxies as well as between proxies. A single SIP request may traverse a path consisting of a number of different protocols.

Unlike traditional signaling protocol approaches, SIP does not try to model services directly, but rather provide a set of interoperable building blocks that can be used to build common telephony services [10] as well as new services that are beyond the capabilities of the existing circuit-switched network.

While useful for conference establishment, the core SIP specification does not address conference control features such as floor control, specifically, the coordination of access to a shared resource. However, SIP events (Section 5.6) and possibly common remote procedure call mechanisms may be used for this purpose.

SIP is maintained by two working groups within the Internet Engineering Task Force (IETF) and is currently an Internet Proposed Standard [11]. It was published in March 1999 and reissued with corrections and enhancements in March 2002.

2. SIP USAGE

SIP was originally designed to be used with multicast multimedia conferences in the Internet, allowing participants to invite others to join the multicast group. This usage is still supported, but has been overtaken in practical importance by its use in voice over IP (VoIP) applications. Voice over IP applications using SIP can be roughly divided into four areas: enterprise, cable, landline carrier, and third-generation (3G) wireless networks. In enterprise networks, traditional analog or digital handsets are replaced by a mixture of Ethernet-connected IP phones, such as the ones pictured in Fig. 1, and software running on desktop PCs [12]. These devices then use SIP to set up internal calls or to reach a gateway operated by the enterprise, a PBX with an IP interface, or a third party.[1] This arrangement makes it easy to add devices to a local dialing plan regardless of their physical location.

Traditional landline carriers can use SIP as a rough replacement for their current backbone signaling protocols, such as ISUP. In such an arrangement, local or gateway switches offer the same analog or digital circuit-switched service to the carrier's customers, but instead of a dedicated signaling network, the carrier uses SIP to interconnect these switches. SIP has been extended to transport ISUP messages across SIP networks, so that midcall information that is relevant only to ISUP networks is not lost in the protocol translation across an ISUP-SIP-ISUP signaling path [13]. The set of specifications for facilitating the interworking of SIP and ISUP is called SIP-T [14]. This is not a different protocol, just a set of guidelines for interworking.

Cable carriers can also use SIP as the signaling protocol in the Distributed Call Signaling (DCS) version of the PacketCable specification [15]. A number of

[1] The latter arrangement is often called *IP centrex*.

Figure 1. SIP phones.

extensions [16–18] have been proposed to make DCS mimic traditional telephone behavior more closely.

Third-generation wireless networks, using both GSM-evolved and CDM2000-based wireless networks, are being standardized by the 3GGP and 3GPP2 consortia. Both have chosen SIP as the signaling protocol for multimedia sessions. For 3GPP, SIP is slated to appear in release 5 of their UMTS framework.

In addition to VoIP and multimedia session setup, SIP has also been proposed for signaling events, that is, asynchronous notifications of state changes. One of the first applications is instant messaging and presence (Section 5.6). SIP events are also useful to unify many of the features found in traditional telephone systems, such as voicemail notification, users joining and leaving conference calls, or the transmission of DTMF digits [19]. Even the use of SIP events to control home appliances have also been proposed [20].

3. RELATED PROTOCOLS

SIP is related to other signaling protocols such as H.323 [21,22] and ISUP [23]. Like these, it operates out-of-band, that is, using a two separate associations for the signaling information and the actual media data, such as voice. Unlike ISUP, SIP is designed for IP networks and for multimedia sessions. A full comparison with H.323 is beyond the scope of this article, but a number of papers offer perspectives [24–27]. Interworking between the two protocols is also possible [28].

As noted above, SIP uses the Session Description Protocol (SDP) to describe the characteristics of the multimedia streams. Typically, these multimedia sessions use RTP [29,30] to carry audio and video information across IP networks.

RTSP [31] is a related control protocol that sets up streaming media sessions. It shares some characteristics with SIP, such as the protocol format and the use of SDP, but supports the control of stored media, through requests such as pause and play, rather than setting up interactive sessions. It does not emphasize finding locations, as it is assumed that the location of multimedia objects is much more stable than those of people. RTSP can be used in a SIP-based telecommunication system for playing announcements, recording voicemail [32] and other so-called media server functionalities.

SIP can interact directly with Internet telephony gateways that translate IP voice packets into circuit-switched calls. Alternatively, these gateways, called *media gateways*, can be controlled by a media gateway controller (MGC), using a master–slave protocol such as MGCP [33] or MEGACO/H.248 [34]. The media gateway controller translates SIP requests into MGCP or H.248 requests, and vice versa. MGCs may also terminate ISUP or other PSTN signaling. MGCs are sometimes called *soft switches*, although the definition is not very precise. Unlike these media control protocols, SIP is a peer-to-peer protocol, where both endpoints are equal.

As indicated above, SIP can use telephone numbers [35] to reach destinations. Proxies can translate these numbers into SIP URIs, possibly via multiple translations, using the ENUM mechanism [36]. ENUM uses NAPTR DNS records [37] to translate a telephone number such as +1-917-555-1234 to a DNS name, here *4.3.2.1.5.5.5.7.1.9.1.enum.arpa*. This record then contains a SIP URI, for example, allowing subscribers to keep their existing telephone number as they migrate to a SIP-based telephone system.

4. PROTOCOL DESCRIPTION

4.1. Protocol Layers

SIP is an application-layer protocol, with several logical sublayers. The lowest layer describes how SIP requests and responses are encoded as messages (Section 4.4). The second layer is the transport layer, defining how clients send requests and receive responses (Section 4.3). Above this transport layer, the transaction layer deals with the notion of group of requests and responses that form a single SIP *transaction*, namely, a request, its retransmissions, and all provisional and final responses (Section 4.4). While all SIP elements (Section 4.2) share similar encoding, transport, and transaction layers, they are distinguished by their role-specific *core*.

4.2. SIP Elements

There are three principal SIP elements: user agents (UAs), proxies, and registrars. User agents can act as clients, initiating requests and receiving responses, and/or servers, receiving requests, while proxies always combine a client and a server functionality. Registrars receive only REGISTER requests and update bindings between stable names and shorter-term contact addresses (see Section 4.3). These roles are logical only, so that a single software implementation can act as a user agent and a proxy server for different requests.

Proxies can be *stateful* or *stateless*. A stateless proxy simply forwards requests and responses one by one, but does not have to concern itself with transactions or associating responses with requests. Stateful proxies are transaction-aware. Stateful proxies are needed if a single inbound request can generate multiple outbound requests going to different destinations ("forking"). Unlike a normal telephone switch, both types of proxies do not have to keep state for the duration of the call. Some proxies do keep call state, for example, for accounting purposes or to control a firewall, and are referred to as *call-stateful*. Naturally, user agents are always call-stateful. Reducing state in SIP network elements helps with scaling and improves robustness, as any host in a server farm, for example, can easily take over should one host fail (Section 4.3).

While not a different protocol element as such, it has become common to refer to back-to-back user agents (B2BUA) in creating services. B2BUAs can be thought of as two user agents where one receives a request and the other issues a related request. This arrangement can be useful for certain types of media-handling firewalls or to create services where two participants have SIP dialog with the call controller, but media are exchanged between them directly. This arrangement is called *third-party call control* [38].

4.3. Locating Users and Servers

SIP endpoints are identified by SIP URIs that are similar to email addresses. (Users may often be able to use their email address as a SIP URI.) SIP URIs can vary in specificity; they may identify a particular user at a host located at an IP address, or, more commonly, identify a user generically by name and domain. An example of the latter is *sip:alice@example.com*. Such a generic URI is then translated via one or more steps to zero or more host locations, possibly identified by SIP URIs. For example, a call to *sip:alice@example.com* may reach her at *sip:alice@128.59.16.1* and *sip:asmith592@aol.com*. SIP URIs with the "sips" scheme indicate that the destination insists on being contacted via transport-layer security (TLS) [39].

The externally visible address of a person or other endpoint is known as the *address-of-record* (AOR). A single person or SIP telephone may have several such AORs. The AOR is bound to any number of contact addresses using the SIP REGISTER method. These contact addresses are typically SIP URIs, but can be any URI, commonly including mail to URIs and http URIs. Each of the user's hosts sends periodic REGISTER request to the registrar identified by the AOR, refreshing the address bindings. For example, for the AOR *sip:alice@example.com*, all of Alice's SIP-enabled communications devices would send updates to the registrar at *example.com*. SIP registrars are one example of a SIP location service that is then used by proxies for the domain to route calls for the AOR. While less common, implementors can choose any other mechanism to populate the location service. For example, a Webpage might allow manual updates of telephone numbers or other non-SIP URIs.

SIP requests may also carry telephone URLs [40] identifying telephone numbers. An example of such a URL describing a telephone terminal in New Jersey is tel:+1-201-555-1234. However, since this URL does not identify an Internet host, a SIP entity needs to translate it to a SIP URI, either a user name or a telephone number at a particular destination domain. In the example, a SIP proxy may translate the number to *12015551234@bigcarrier.com* if it wants the call to be routed to a gateway operated by *bigcarrier.com*. The actual routing may be determined by routing protocols such as TRIP [41,42] that distribute information about the reachability of telephone numbers and their associated gateways.

When receiving an initial SIP request, a SIP proxy looks at the request URI contained in the first line of the SIP request. (In Fig. 2, this is *bob@biloxi.com*) If the domain name is managed by the proxy, the proxy maps the user name to one or more contact addresses and sends replicas of the request to those addresses, creating *branches*. The requests may be sent sequentially, after a previous branch has failed, or in parallel. This procedure is referred to as *forking* and greatly simplifies reaching one individual that may be using multiple devices to communicate. It roughly mimics, on a global scale, the operation of multiple phones in a household, all ringing at once. Each branch may also return a redirection response, indicating one or more alternate addresses to be tried. As soon as somebody picks up at one of the addresses, that SIP user agent returns a success response and the proxy sends a CANCEL request to all other active branches, terminating ringing there.

If the request URI identifies a domain not handled by the proxy, the proxy is acting as an *outbound proxy*. Outbound proxies are typically used by the initiator of the request to handle all outbound requests, regardless of destination. For example, Alice might use an outbound proxy in her home domain, *atlanta.com*. Outbound proxies may be necessary if firewall policies restrict inbound requests or might be useful to offer additional outbound services, such as custom abbreviated dialing.

Unlike HTTP, SIP URIs attempt to avoid identifying a single physical server, but rather make it possible to direct a request to one of the servers handling SIP requests for a particular domain. This approach was first taken for email, using DNS MX [43] records to list a set of mail transport agents for a domain. SIP employs a two-step process [44], where first DNS NAPTR records [37] indicate the set of transport protocols (UDP, TCP, TCP with TLS, etc.) available for the domain. For each suitable protocol, DNS SRV [45] resource records list a set of candidate servers, qualified by preference and weight. The client looking for a server picks a random server among the highest-preference group of servers, performing a simple form of client-based load balancing, albeit without load feedback.

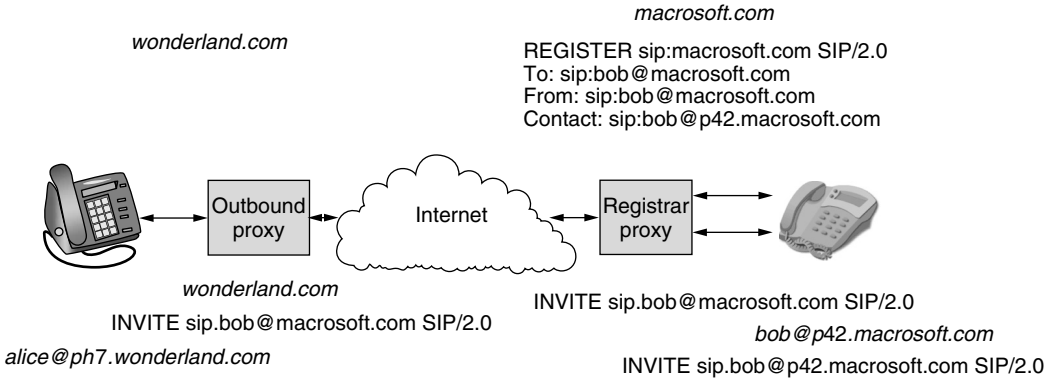


Figure 2. SIP architecture.

4.4. SIP Requests and Responses

SIP is a request-response protocol similar to HTTP. Requests and responses are expressed as plain text, consisting of a header or status line, a set of header fields and a message body. The message body can contain any information, including binary data. Just like HTTP and email, SIP uses MIME [46] to carry multiple distinct message bodies in a single request.

Each request is identified by a *method* that describes its basic functionality. The basic protocol defines six methods — INVITE, ACK, BYE for call setup and termination, OPTIONS for discovering capabilities, CANCEL for terminating pending calls, and REGISTER for establishing address bindings, but other specifications add additional methods. For example, the INFO method [47] can carry midcall ISUP PSTN call signaling information [48]. A request may trigger zero or more provisional responses that indicate call progress and one or more final responses. Each SIP method is qualified by a set of parameters expressed as header fields, similar to email or HTTP headers. Indeed, SIP inherits a large number of HTTP and email header fields.

The details of the header field syntax can be found in the protocol specification [49]. Header fields can be roughly divided into header fields that describe the message body, identify the request or response, help in routing requests, assist in authentication, negotiate protocol capabilities, and provide auxiliary information to end systems and proxies. The content-describing header fields identify the media type, language, handling, and encoding and measure the length of the body. The request is identified by components of the From, To, Call-ID and CSeq header fields. As described in Section 4.5, the Via, Record-Route, and Route header fields trace request paths and allow requests to revisit the same set of servers as a previous request. A set of header fields, mostly borrowed from HTTP [50], provide a challenge-response authentication framework, as described in Section 4.7. Headers also negotiate protocol capabilities, such as the support for media types in message bodies, human languages understood for error messages, and which protocol features are supported or required. A large set of optional header fields identify the caller or call, for example, the purpose of the call, the organization the caller or callee are affiliated with, or additional hints for the call, such as customized alerting or a Webpage related to the caller.

Responses are classified by their three-digit response code. The first digit is sufficient to identify the type of response; the first digit of "1" indicates a provisional response such as "trying" or "ringing", a first digit of "2," success; "3," a redirection; "4," a client error; "5," a server error; and "6," a global failure. (A global failure indicates that further searches for an instance of the callee are fruitless, for example, because the callee refuses to talk to the caller.)

A sample INVITE request and the corresponding final response is shown in Fig. 3. In the example, Alice, using her SIP device located at *pc33.atlanta.com*, calls Bob, at *biloxi.com*. Alice tells Bob that she can receive PCMU (that is, μ-law) audio over RTP using the AVP profile [30] at port 49172 and IP address 100.101.102.103; Bob responds to Alice that he is willing to talk and waiting to receive audio at port 33280 on address 110.111.112.113. The Call-ID field is a randomly chosen identifier for the call; the randomly chosen tags in the From and To fields identify the participants, allowing multiple terminals of Bob and Alice to be distinguished. The CSeq identifier allows Alice to unambiguously associate Bob's response with her request. The example does not show the third and final request in the call setup, with method ACK, that Alice sends to Bob to confirm that she has received Bob's "200 OK" response. The final ACK request establishes a SIP *dialog* between Alice and Bob. Only INVITE requests use an ACK request for confirmation.

As noted, SIP can operate directly on top of unreliable transport protocols such as UDP. To ensure that the call is set up, SIP implements its own retransmission mechanism. Due to forking at proxies, requests are retransmitted and acknowledged on each branch, not end to end. Requests are retransmitted at exponentially increasing intervals starting with the estimate for the round-trip time, if known, and 0.5 s if the round-trip time is unknown. The user agent client retransmits until it receives a provisional or final response or after seven retransmissions.

INVITE transactions differ somewhat in their behavior from all other requests. Non-INVITE requests are assumed to take only a small amount of time to process, so that the client simply retransmits until it gets a final response. It slows down retransmission to once every few seconds after it has received a provisional response, to deal with lost final responses. For INVITE transactions, the final answer can take tens of seconds or minutes since a human may need to answer the ringing phone. Here, the server retransmits the final response until the client confirms its receipt with an ACK request.

During the call, either Alice or Bob can send additional INVITE requests to update the media session, for example, to change the set of media or to adjust the media destination IP address. If either Alice or Bob hang up, their SIP terminal sends a BYE request to the other side. Typically, both midcall INVITE requests and the BYE request bypass any intermediate proxies since Alice and Bob now know each other's IP address.

4.5. Routing Requests and Responses

The routing of requests depends on the request URI and, for all but the first request in a dialog, on the Route header fields that may be present (see below). The response to the first request will contain a Contact header field that describes the location to send subsequent requests to. Thus, subsequent requests will contain that address as the request URI.

Route header fields lists proxies that indicated a desire to stay in the path of subsequent requests within a dialog, for example, any reINVITE requests to change session parameters or the final BYE request terminating the dialog. Requests in both directions, from caller to callee and vice versa, traverse this set of proxies. Route headers offer a functionality somewhat similar to IP source routing. A proxy that wants to see subsequent requests for a dialog

Request:

```
INVITE sip:bob@biloxi.com SIP/2.0
Via: SIP/2.0/UDP pc33.atlanta.com;branch=z9hG4bK776asdhds
To: Bob <sip:bob@biloxi.com>
From: Alice <sip:alice@atlanta.com>;tag=1928301774
Call-ID: a84b4c76e66710@pc33.atlanta.com
CSeq: 314159 INVITE
Contact: <sip:alice@pc33.atlanta.com>
Content-Type: application/sdp
Content-Length: 145

v=0
o=alice 2890844526 2890844526 IN IP4 atlanta.com
s=Session SDP
c=IN IP4 100.101.102.103
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

Response:

```
SIP/2.0 200 OK
To: Bob <sip:bob@biloxi.com>;tag=8321234356
From: Alice <sip:alice@atlanta.com>;tag=1928301774
Via: SIP/2.0/UDP pc33.atlanta.com;branch=z9hG4bK776asdhds
Call-ID: a84b4c76e66710@pc33.atlanta.com
CSeq: 314159 INVITE
Contact: <sip:alice@pc33.atlanta.com>
Content-Type: application/sdp
Content-Length: 140

v=0
o=bob 2890844526 2890844526 IN IP4 biloxi.com
s=Session SDP
c=IN IP4 110.111.112.113
t=0 0
m=audio 33280 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

Figure 3. Sample INVITE request and response.

adds itself to the *route set* for a dialog by adding Record-Route header fields to the initial request. The callee user agent then copies those Record-Route header fields into the response. The callee also keeps the list, in the same order, as its route set. The caller reverses the list and makes that its route set. The route set is then inserted as a Route header into subsequent requests by both caller and callee in the same session. Each proxy inspects the top-most route header and uses it to find the next hop toward the destination. If the Route header names the proxy itself, the proxy removes the header field.

Responses always traverse the same set of proxy servers as the corresponding request, guided by the Via header fields inserted into the request by the proxy servers.

4.6. Extending SIP

Since VoIP and multimedia conferencing are still in their infancy, it is likely that SIP will evolve. Also, some areas where SIP is being applied have specific requirements whose fulfillment should not burden other applications. SIP was designed to be extensible while maintaining a maximum of backward compatibility. Allowing for extensions requires defining suitable behavior when an implementation discovers an unknown protocol element and the ability to ascertain the capabilities of SIP elements.

Since SIP proxies forward requests independent of their method, additional methods can be added without upgrading all proxies. SIP user agents can indicate the methods that they support in an Allow header.

User agents and proxies can add new header field types without coordination, as proxies simply copy them into outgoing requests and receiving user agents ignore unknown header fields. If a user agent client wants to ensure that a particular feature is supported by a proxy or user agent server, it adds a Require header field indicating the name of the feature. In turn, the client or server can summarize its own capabilities with the Supported header field.

4.7. SIP Security

SIP poses a number of security challenges. It carries potentially sensitive information, such as subject information about calls, media encryption keys, and reachability information. The existence of communication relationships themselves may be considered private. Registrations need to be protected against malicious modifications, as such modifications would allow the attacker to redirect requests to any location. Also, attackers must be prevented from injecting requests into existing calls, as those fake requests could be used to terminate or redirect the call. The security challenges are increased by the use of proxies that may be operated by either the caller's network provider, the callee's network provider, or possibly the operator of the network being visited by either caller or callee.

Since it must be possible to make calls to parties with whom the caller does not have a preexisting security relationship, standard shared secret security is of only limited use. A public key infrastructure for large user populations does not currently exist, but organizations can easily obtain public key certificates, for example, for Web servers.

Given those constraints, SIP uses three security mechanisms, namely transport-layer security (TLS) [39,51], digest authentication [52], and secure/multipurpose Internet mail extensions (S/MIME) [53]. Naturally, IPsec can be deployed as well, but it is effectively invisible to SIP. TLS protects exchanges "hop by hop," that is, between user agent and proxy or registrar, and between proxies. With standard server certificates, the user agent can ascertain the identity of the registrar or proxy. Digest authentication is a challenge–response authentication protocol based on a shared secret. It is commonly used to protect registrations, although it offers header integrity only in combination with TLS. Finally, S/MIME is used in a fashion similar to email, but encapsulating the SIP message parts that are to be encrypted or signed in an outer SIP wrapper that remains visible to proxies. For encryption, S/MIME requires that the sender knows the recipients public key or has access to a shared public key infrastructure.

5. SIP EXTENSIONS

The IETF is extending SIP in a variety of ways to accommodate special requirements. Many of the extensions are motivated by the need to interoperate with the PSTN, but a major extension, SIP events (Section 5.6), offers new services and extends SIP to provide event notification. SIP extensions are designed to be backward-compatible, so that peers can establish sessions, with possibly reduced functionality, even if one or both of them do not support a particular extension. Many extensions are specific to particular deployment scenarios or applications and thus are likely to be supported by only a subset of implementations.

5.1. Session Keep-Alive

Once a SIP dialog has been established, the two peers have no direct way to discover if the other one has crashed or has been disconnected from the network. Usually, the lack of media data may provide a clue, but, unlike the situation in traditional phone calls, SIP sessions can persist even if no media are being exchanged. In circumstances where endpoints want to remove state if the other side is no longer reachable, the SIP session timer mechanism [54] has been proposed. The initial INVITE request indicates how often the caller would like the session to be refreshed. The callee can claim the role of refresher or leave this to the caller. The party responsible for refreshes periodically issues another INVITE request. If it receives no response, it terminates the call. Similarly, the party expecting periodic reINVITE requests will terminate the call if they are not forthcoming. The session interval can be lowered by proxies en route, but proxies can reject session intervals that are too short.

As with TCP liveness detection, SIP session timers must be used with caution, as they may unnecessarily terminate a long-lived session during brief network outages.

5.2. Conferencing

SIP supports a variety of multiparty conferencing architectures, including Internet multicast, end-system mixing, dialin conference servers, ad hoc centralized conferences, dialout conferences, or centralized signaling with peer-to-peer or multicast media [55,56]. For multicast and end-system mixing, there are no special servers. However, multicast is not yet widely available in the Internet; end-system mixing is likely to scale only to modest-size conferences, since one of the participants needs to send a mixed audio- or videostream to all other participants. The most common architectures are likely to be dialin and dialout conferences with a centralized conference server, or possibly a hierarchy of such servers.

Conferences are treated just like normal users; they are addressed using SIP URLs such as *sip:staff-meeting@example.com*. For ad hoc conferences, the initiator sends an INVITE request to the conference bridge, with a randomly chosen conference identifier.

The manner of adding users depends on the conference model. For multicast and end-system mixed conferences, regular INVITE requests suffice. For dialin conferences, conference members can ask others to join the conference by issuing a REFER request to the candidate member.

5.3. Multiparty Calls

A number of common telephony features, such as call waiting, blind and attended transfer, conference calls, call parking, call pickup, music on hold, call monitoring, barge-in, whispered call waiting, and single-line extensions, or Internet-oriented features such as presence-enabled conferencing all generally involve multiple participants, both human and machines (media servers). Often, media of these participants are mixed.

Many of these features can be implemented using third-party call control [38], where a central controller maintains and terminates dialogs, using INVITE, reINVITE, and BYE, with the participants and mixes media as appropriate. This approach has the advantage that it requires only basic SIP capabilities in the end systems, but it means that features have to implemented in a single point for each call.

Alternatively, a peer-to-peer approach can be used [57], where all functionality is implemented "at the edges," that is, by terminals. This approach requires a set of primitives to replace an existing dialog, join an existing dialog with another one, perform media replication by the media origin, and allow a user agent to ask another user agent to send a request on its behalf. The latter functionality is accomplished with the REFER request [58] that asks the recipient to construct another SIP request. The SIP request and its parameters and contained in the Refer-To header as a SIP URI.

5.4. Caller Preferences

In SIP, terminal addresses often identify a generic destination, such as a person, and not a particular terminal or phonejack. Thus, it is helpful to provide the caller with the ability to indicate which of the devices reachable by a particular SIP URI the caller would prefer to reach. This preference is expressed through properties that the destination should have, not its address [59]. For example, the caller can indicate a desire to reach voicemail instead of a human being, might prefer not to talk to a mobile phone or might like to be connected to a Chinese-speaking operator. SIP proxies then use this information in making call routing decisions.

5.5. Quality of Service

Since SIP requests do not necessarily traverse the same route as the data packets for the session they establish, SIP cannot directly control quality of service on the data path. Instead, this is left to mechanisms such as the Resource Reservation Protocol (RSVP) [60] or differentiated services. However, SIP can help negotiate the use of such mechanisms and determine when the resource reservation has succeeded. This is needed to prevent scenarios where the called phone rings, but then the resource reservation fails because of lack of network bandwidth, leading to a call without media or with only best-effort media streams. An additional SIP method, COMET (condition met), signals the successful completion of the resource reservation [61].

5.6. SIP for Events, Presence, and Instant Messaging

Events, that is, changes in state, are a useful abstraction in many telecommunication services. Examples include automatic callback, message waiting, and multi party conference management. Also, they underlie the notion of presence (usually combined with instance messaging) that has become a popular Internet service. Event notifications are often approximated by email messages, but since email is picked up only by the recipient, there can be a large and unpredictable latency between the event notification and its receipt. SIP can readily be extended to signal events. Here, SIP follows the common subscribe–publish model. A SIP entity sends a SIP SUBSCRIBE request to the source of the events. The subscriber is then notified, using NOTIFY, each time the event occurs. The subscription has a limited lifetime and needs to be refreshed before it expires.

SIP is suitable since many of the same properties needed for setting up calls apply. For example, the destination of the event subscription and event notification should be identified by a long-term address, while the actual destination may change network addresses.

Beyond signaling events related to SIP sessions, SIP events have also been defined for describing the presence state of a user [62], that is, whether the user is available to communicate by phone, SIP call, text chat, or other means. Since this information needs to be available even if the user terminal is not connected to the network, a presence agent acts as a standin for the user. It also merges presence information from multiple devices. The presence agent can use information from SIP binding updates (REGISTER) to detect changes in presence state.

5.7. Programming SIP Services

While not part of the protocol specification itself, there have been a number of proposals on how to program SIP-based proxies and end systems. Among these are JAIN SIP, SIP servlets [63], sip-cgi [64], and the Call Processing Language (CPL) [65,66]. JAIN SIP is a Java API that allows to construct SIP messages and extract information about SIP headers and responses. SIP servlets are similar to Java servlets for Web servers, offering a model where the servlet handles a full transaction. Sip-cgi attempts to transfer the cgi programming model found in Web servers to SIP proxies. SIP requests are passed as environment variables to scripts or programs, with each request causing a new process to be invoked. The scripts can be written in any language, including such common Web services scripting languages such as Perl, Python, or Tcl. Unlike HTTP requests, which are completely stateless, sip-cgi offers some support for associating multiple requests and responses within the same transaction. The script is responsible for parsing SIP headers and generating responses, but can simplify its task by invoking default behaviors.

CPL is an XML-based tree for handling a SIP transaction in proxies. It is useful primarily for call routing, admission, rejection, and logging. Each transaction is logically handled by a single CPL script, acting as a form of decision tree. The traversal of the tree can depend on the status responses returned by branches. Extensions of CPL to presence services and to end-system services are in progress. For voice-oriented services, voice menu systems such as VoiceXML are available.

5.8. SIP Performance

There currently is no generally accepted, standard way to measure the performance of SIP clients and servers. However, SIPstone [67] has been proposed as a simple metric for common operations, such as calls and registrations.

5.9. Emergency Services

One of the principal functions of the existing PSTN is to summon emergency help. In many countries, a system has emerged where callers can call a single, nationwide emergency number, such as 911 in the United States or 112 in many parts of Europe. The call is then directed to the nearest emergency call center, where the dispatcher

can see the caller's geographic location on her console. A similar architecture has been proposed for SIP [68], with a global emergency identifier, "sos," and a special proxy that maps user location to the nearest emergency call center. Unlike in the PSTN, the user location cannot be directly determined from the device address, as IP subnets can span large geographic regions, particularly with virtual private networks and dialup connections.

5.10. Configuring SIP Networks

SIP attempts to make it possible to set up a SIP network with minimal manual configuration. It is sufficient for a SIP device to know the address-of-record of its owner. The domain part of that address identifies the registrar for the owner. If desired, an outbound proxy can be discovered using DHCP [69]. Configuring other parameters, such as dial plans or SIP protocol timing parameters, is currently under discussion [70].

6. SUMMARY AND CONCLUSION

SIP is a signaling protocol that allows peers to establish multimedia sessions across the Internet. Since some of the requirements, such as mapping from long-term stable user identities to addresses, are similar, SIP can also be used to convey events to users, including changes of presence information. SIP is currently used primarily by voice over IP applications. A number of extensions have been standardized, with many more under discussion.

BIOGRAPHY

Henning Schulzrinne received his undergraduate degree in economics and electrical engineering from the Darmstadt University of Technology, Germany, his M.S.E.E. degree as a Fulbright scholar from the University of Cincinnati, Ohio, and his Ph.D. degree from the University of Massachusetts in Amherst, Massachusetts. He was a member of the technical staff at AT&T Bell Laboratories, Murray Hill and an associate department head at GMD-Fokus (Berlin), before joining the Computer Science and Electrical Engineering Departments at Columbia University, New York. His research interests encompass real-time, multimedia network services in the Internet and modeling and performance evaluation.

He is a division editor of the *Journal of Communications and Networks* and an editor of the *IEEE/ACM Transactions on Networking* and former editor of the *IEEE Internet Computing Magazine* and *IEEE Transactions on Image Processing*. He is member of the Board of Governors of the IEEE Communications Society and the ACM SIGCOMM Executive Committee, former chair of the IEEE Communications Society Technical Committees on Computer Communications and the Internet, and has been technical program chair of Global Internet, Infocom, NOSSDAV, and IPtel. He also was a member of the Internet Architecture Board (IAB).

Protocols codeveloped by him are now Internet standards, used by almost all Internet telephony and multimedia applications. His research interests include Internet multimedia systems, quality of service, and performance evaluation.

BIBLIOGRAPHY

1. H. Schulzrinne, Personal mobility for multimedia services in the Internet, in *European Workshop on Interactive Distributed Multimedia Systems and Services* (*IDMS*), Berlin, Germany, March 1996.
2. M. Handley and V. Jacobson, *SDP: Session Description Protocol*, RFC 2327, Internet Engineering Task Force, April 1998.
3. D. Kutscher, J. Ott, and C. Bormann, *Session Description and Capability Negotiation*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).
4. J. Postel, *Internet Protocol*, RFC 791, Internet Engineering Task Force, Sept. 1981.
5. S. Deering and R. Hinden, *Internet Protocol, Version 6* (*ipv6*) *Specification*, RFC 2460, Internet Engineering Task Force, Dec. 1998.
6. J. Postel, *User Datagram Protocol*, RFC 768, Internet Engineering Task Force, Aug. 1980.
7. J. Postel, *Transmission Control Protocol*, RFC 793, Internet Engineering Task Force, Sept. 1981.
8. R. Stewart et al., *Stream Control Transmission Protocol*, RFC 2960, Internet Engineering Task Force, Oct. 2000.
9. J. Rosenberg, H. Schulzrinne, and G. Camarillo, *SCTP as a Transport for SIP*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).
10. J. Lennox, H. Schulzrinne, and T. F. L. Porta, *Implementing Intelligent Network Services with the Session Initiation Protocol*, Technical Report CUCS-002-99, Columbia Univ., New York, NY, Jan. 1999.
11. S. Bradner, *The Internet Standards Process—Revision 3*, RFC 2026, Internet Engineering Task Force, Oct. 1996.
12. W. Jiang, J. Lennox, H. Schulzrinne, and K. Singh, Towards junking the PBX: deploying IP telephony, *Proc. Int. Workshop on Network and Operating System Support for Digital Audio and Video* (*NOSSDAV*), Port Jefferson, NY, June 2001.
13. International Telecommunication Union, *Introduction to CCITT Signalling System No. 7*, Recommendation Q.700, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, March 1993.
14. A. Vemuri and J. Peterson, *SIP for telephones* (*SIP-t*): *Context and Architectures*, Internet Draft, Internet Engineering Task Force, April 2002 (work in progress).
15. E. Miller, F. Andreasen, and G. Russell, The PacketCable architecture, *IEEE Commun. Mag.* **39**: (June 2001).
16. W. Marshall et al., *SIP Extensions for Supporting Distributed Call State*, Internet Draft, Internet Engineering Task Force, Aug. 2001 (work in progress).
17. W. Marshall, F. Andreasen, and D. Evans, *SIP Extensions for Media Authorization*, Internet Draft, Internet Engineering Task Force, May 2002 (work in progress).
18. W. Marshall et al., *SIP Extensions for Network-Asserted Caller Identity and Privacy within Trusted Networks*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).

19. B. Culpepper, R. Fairlie-Cuninghame, and J. Mule, *SIP Event Package for Keys*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).
20. A. Roychowdhury and S. Moyer, Instant messaging and presence for network appliances using SIP, *Internet Telephony Workshop 2001*, New York, April 2001.
21. J. Toga and J. Ott, ITU-T standardization activities for interactive multimedia communications on packet-based networks: H.323 and related recommendations, *Comput. Networks ISDN Syst.* **31**: 205–223 (Feb. 1999).
22. International Telecommunication Union, *Visual Telephone Systems and Equipment for Local Area Networks Which Provide a Non-Guaranteed Quality of Service*, Recommendation H.323, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1996.
23. International Telecommunication Union, *Functional Description of the ISDN User Part of Signalling System No. 7*, Recommendation Q.761, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 1994.
24. H. Schulzrinne and J. Rosenberg, A comparison of SIP and H.323 for Internet telephony, *Proc. Int. Workshop on Network and Operating System Support for Digital Audio and Video* (*NOSSDAV*), Cambridge, UK, July 1998, pp. 83–86.
25. I. Dalgic and H. Fang, Comparison of H.323 and SIP for IP telephony signaling, *Proc. Photonics East*, Boston, MA, SPIE, Sept. 1999.
26. Nortel Networks, *A Comparison of H.323v4 and SIP*. 3GPP contribution, Jan. 2000.
27. T. Eyers and H. Schulzrinne, Predicting internet telephony call setup delay, *Proc. 1st IP-Telephony Workshop* (*IPtel 2000*), Berlin, Germany, April 2000.
28. K. Singh and H. Schulzrinne, Interworking between SIP/SDP and H.323, *Proc. 1st IP-Telephony Workshop* (*IPtel 2000*), Berlin, Germany, April 2000.
29. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Internet Engineering Task Force, Jan. 1996.
30. H. Schulzrinne, *RTP Profile for Audio and Video Conferences with Minimal Control*, RFC 1890, Internet Engineering Task Force, Jan. 1996.
31. H. Schulzrinne, A. Rao, and R. Lanphier, *Real Time Streaming Protocol* (*RTSP*), RFC 2326, Internet Engineering Task Force, April 1998.
32. K. Singh and H. Schulzrinne, Unified messaging using SIP and RTSP, *Proc. IP Telecom Services Workshop*, Atlanta, GA, Sept. 2000 pp. 31–37.
33. M. Arango et al., *Media Gateway Control Protocol* (*MGCP*) *Version 1.0*, RFC 2705, Internet Engineering Task Force, Oct. 1999.
34. F. Cuervo et al., *Megaco Protocol Version 1.0*, RFC 3015, Internet Engineering Task Force, Nov. 2000.
35. International Telecommunication Union, *The International Public Telecommunication Numbering Plan*, Recommendation E.164, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1997.
36. P. Faltstrom, *E.164 Number and DNS*, RFC 2916, Internet Engineering Task Force, Sept. 2000.
37. R. Daniel and M. Mealling, *Resolution of Uniform Resource Identifiers Using the Domain Name System*, RFC 2168, Internet Engineering Task Force, June 1997.
38. J. Rosenberg, J. Peterson, H. Schulzrinne, and G. Camarillo, *Third Party Call Control in SIP*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).
39. T. Dierks and C. Allen, *The TLS Protocol Version 1.0*, RFC 2246, Internet Engineering Task Force, Jan. 1999.
40. A. Vaha-Sipila, *URLs for Telephone Calls*, RFC 2806, Internet Engineering Task Force, April 2000.
41. J. Rosenberg and H. Schulzrinne, *A Framework for Telephony Routing over IP*, RFC 2871, Internet Engineering Task Force, June 2000.
42. J. Rosenberg, H. Salama, and M. Squire, *Telephony Routing over IP* (*TRIP*), RFC 3219, Internet Engineering Task Force, Jan. 2002.
43. C. Partridge, *Mail Routing and the Domain System*, RFC 974, Internet Engineering Task Force, Jan. 1986.
44. J. Rosenberg and H. Schulzrinne, *SIP: Locating SIP Servers*, RFC 3263, Internet Engineering Task Force, May 2002.
45. A. Gulbrandsen, P. Vixie, and L. Esibov, *A DNS RR for Specifying the Location of Services* (*DNS SRV*), RFC 2782, Internet Engineering Task Force, Feb. 2000.
46. N. Borenstein and N. Freed, *MIME* (*multipurpose Internet Mail Extensions*) *Part One*: *Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, RFC 1521, Internet Engineering Task Force, Sept. 1993.
47. S. Donovan, *The SIP INFO Method*, RFC 2976, Internet Engineering Task Force, Oct. 2000.
48. E. Zimmerer et al., *MIME Media Types for ISUP and QSIG Objects*, RFC 3204, Internet Engineering Task Force, Dec. 2001.
49. J. Rosenberg et al., *SIP: Session Initiation Protocol*, RFC 3261, Internet Engineering Task Force, May 2002.
50. R. Fielding et al., *Hypertext Transfer Protocol — HTTP/1.1*, RFC 2616, Internet Engineering Task Force, June 1999.
51. E. Rescorla, *HTTP over TLS*, RFC 2818, Internet Engineering Task Force, May 2000.
52. J. Franks et al., *HTTP Authentication: Basic and Digest Access Authentication*, RFC 2617, Internet Engineering Task Force, June 1999.
53. B. Ramsdell, *S/MIME Version 3 Message Specification*, RFC 2633, Internet Engineering Task Force, June 1999.
54. S. Donovan and J. Rosenberg, *SIP Session Timer*, Internet Draft, Internet Engineering Task Force, Oct. 2001 (work in progress).
55. J. Rosenberg and H. Schulzrinne, *Models for Multi Party Conferencing in SIP*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).
56. K. Singh, G. Nair, and H. Schulzrinne, Centralized conferencing using SIP, *Proc. Internet Telephony Workshop 2001*, New York, April 2001.
57. R. Mahy et al., *A Multi-Party Application Framework for SIP*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).
58. R. Sparks, *The Refer Method*, Internet Draft, Internet Engineering Task Force, Oct. 2001 (work in progress).
59. H. Schulzrinne and J. Rosenberg, *SIP Caller Preferences and Callee Capabilities*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).

60. R. Braden et al., *Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, Internet Engineering Task Force, Sept. 1997.

61. W. Marshall, G. Camarillo, and J. Rosenberg, *Integration of Resource Management and SIP*, Internet Draft, Internet Engineering Task Force, April 2002 (work in progress).

62. J. Rosenberg et al., *Session Initiation Protocol (SIP) Extensions for Presence*, Internet Draft, Internet Engineering Task Force, April 2002 (work in progress).

63. A. Kristensen and A. Byttner, *The SIP Servlet API*, Internet Draft, Internet Engineering Task Force, Sept. 1999 (work in progress).

64. J. Lennox, H. Schulzrinne, and J. Rosenberg, *Common Gateway Interface for SIP*, RFC 3050, Internet Engineering Task Force, Jan. 2001.

65. J. Lennox and H. Schulzrinne, *CPL: A Language for User Control of Internet Telephony Services*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).

66. J. Lennox and H. Schulzrinne, The call processing language: User control of internet telephony services, *Proc. Lucent Technologies XML Day*, Murray Hill, NJ, Feb. 2000.

67. H. Schulzrinne, S. Narayanan, J. Lennox, and M. Doyle, *SIPstone—Benchmarking SIP Server Performance*, Technical Report CUCS-005-02, Dept. Computer Science, Columbia Univ., New York, March 2002.

68. H. Schulzrinne and K. Arabshian, Providing emergency services in internet telephony, *IEEE Internet Comput.* **6**: 39–47 (May 2002).

69. H. Schulzrinne, *DHCP Option for SIP Servers*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).

70. C. Stredicke and I. Butcher, *SIP End Point Configuration Data Format*, Internet Draft, Internet Engineering Task Force, Feb. 2002 (work in progress).

SHALLOW-WATER ACOUSTIC NETWORKS*

JOHN G. PROAKIS
JOSEPH A. RICE
ETHEM M. SOZER
MILICA STOJANOVIC
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Since the early 1980s, the underwater acoustic (UWA) communications technology has progressed significantly. Communication systems with increased bit rate and reliability now enable real-time point-to-point links between underwater nodes such as ocean-bottom sensors and autonomous underwater vehicles (AUVs). Current research is focused on combining various point-to-point links within a network structure to meet the emerging demand for applications such as environmental data collection, offshore exploration, pollution monitoring, and military surveillance [1].

The traditional approach for ocean-bottom or ocean-column monitoring is to deploy oceanographic sensors, record the data, and recover the instruments. This approach has several disadvantages:

- The recorded data cannot be recovered until the end of the mission, which can be several months.
- There is no interactive communication between the underwater instruments and the onshore user. Therefore, it is not possible to reconfigure the system as interesting events occur.
- If a failure occurs before recovery, data acquisition may stop, or all the data may be lost.

The long delay between data acquisition and recovery can be reduced by using expendable or reusable communication probes as in the EMMA [2] and GEOSTAR [3] systems. Both systems have ocean-bottom sensors and a number of probes that have radiocommunication equipment. The data collected by sensors are carried to the surface with probes at preprogrammed intervals or as soon as some interesting events occur. After surfacing, the probe sends the data to an onshore user via satellite. The release of the probes can also be forced by sending acoustic signals from a nearby ship. These systems provide quasi-real-time data collection. However, lack of bidirectional communication links and the high cost of probes limit their usage.

The ideal solution for real-time monitoring of selected ocean areas for long periods of time is to connect various instruments through wireless links within a network structure. Basic underwater acoustic networks are formed by establishing bidirectional acoustic communication between nodes such as autonomous underwater vehicles (AUVs) and fixed sensors. An RF link connects the network to a surface station that can be further connected to terrestrial networks, such as the Internet, through an RF link. Onshore users can extract real-time data from multiple distant underwater instruments. After evaluating the obtained data, they can send control messages to individual instruments. Since data is not stored in the underwater instruments, data loss is prevented as long as isolated node failures can be circumvented by reconfiguring the network.

A major constraint of UWA networks is the limited energy supply. Whereas the batteries of a wireless modem can be easily replaced on land-based systems, the replacement of an underwater modem battery involves ship time and retrieval of the modem from the ocean bottom, which is costly and time-consuming. Therefore, transmission energy is precious in underwater applications. Network protocols should conserve energy by reducing the number of retransmissions, powering down between transactions, and minimizing the energy required per transmission.

* This work was supported by the Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564, by Small-Business Innovative Research (SBIR) Program, and by ONR 321.

Some underwater applications require that the network be deployed quickly without substantial planning, such as in rescue and recovery missions. Therefore, the network should be able to determine the node locations and configure itself automatically to provide an efficient data communication environment. Also, if the channel conditions change or some of the nodes fail during the mission, the network should be capable of reconfiguring itself dynamically to continue its operation.

2. UNDERWATER ACOUSTIC COMMUNICATIONS

Unlike digital communications through radio channels where data are transmitted by means of electromagnetic waves, acoustic waves are used primarily in underwater channels. The propagation speed of acoustic waves in UWA channels is five orders of magnitude less than that of radiowaves. This low propagation speed increases the latency of a packet network.

The available bandwidth of an UWA channel depends critically on transmission loss, which increases with both range and frequency, and severely limits the available bandwidth [4,5]. For example, long-range systems that operate over several tens of kilometers may have a bandwidth of only a few kilohertz, while a short-range system operating over several tens of meters may have more than a 100 kHz bandwidth [6]. Within this limited bandwidth, the acoustic signals are subject to time-varying multipath [4], which may result in severe intersymbol interference (ISI) and large Doppler shifts and spreads, relative to radio channels, especially in shallow-water channels. Multipath propagation and Doppler effects degrade of acoustic signals and limit the data throughput. Special processing techniques are needed to combat these channel impairments.

Until the year 1990, as a result of the challenging characteristics of UWA channels, modem development was focused on employing noncoherent frequency shift keying (FSK) signals for achieving reliable communication. Since FSK demodulation is based on energy detection, it does not require phase tracking, which is a very difficult task in high Doppler spread environments. The multipath effects are eliminated by inserting guard periods between successive pulses to ensure that all the reverberation vanishes before each subsequent pulse is to be received. In addition, to avoid Doppler effects, some guard bands are employed between frequency tones. By varying the values of the guard bands, the communication signals can be matched to the channel characteristics, providing an adaptive modem structure. Table 1 presents some data on the noncoherent FSK modems described in the literature.

Although noncoherent FSK systems are effective in UWA channels, their low bandwidth efficiency makes them inappropriate for high-data-rate applications such as multiuser networks. The need for high-throughput, long-range systems has resulted in a focus toward coherent modulation techniques.

Today, with the availability of powerful digital signal processing devices, we are able to employ fully coherent PSK modulation in underwater communications. Equalizers are used to undo the effects of ISI, instead of trying to avoid or suppress it. When combined with explicit phase tracking loops, such as phase-locked loops (PLLs), decision feedback equalizers can provide high data throughput [7]. Other similar structures that use transversal filters and various adaptation algorithms are also reported in the literature. Coherent systems are summarized in Table 1.

Table 1. Summary of Performance Metrics for Some UWA Modems Presented in the Literature [5]

Type	Year	Data Rate (bps)	Bandwidth (kHz)	Range (km)[a]
FSK	1984	1200	5	3.0_S
FSK	1991	1250	10	2.0_D
FSK	1997	2400–600	5	10.0_D–5.0_S
Coherent	1989	500,000	125	0.06_D
Coherent	1993	600–300	0.3–1	89_S–203_D
Coherent	1994	20	20	0.9_S
Coherent	1998	1670–6700	2–10	4.0_S–2.0_S

[a] Subscript S indicates a shallow-water result; D indicates a deep-water result, generally a vertical channel.

Current research is focused on DSP algorithms with decreased complexity and multiuser modems that can operate in a network environment.

3. UNDERWATER ACOUSTIC NETWORKS

Information networks are designed in the form of a layered architecture [8]. The first three layers of this hierarchical structure are the physical layer, the data-link control layer, and the network layer.

The function of physical layer is to create a virtual link for transmitting a sequence of logical information (bits 0 and 1) between pairs of nodes. The information bits are converted into acoustic signals (in case of UWA networks), which are transmitted through the acoustic channel. At the receiving node, the physical layer converts the channel corrupted signals back into logical bits. The modem structures that can be used in the physical layer of an acoustic network were discussed in the previous section.

The second layer in the hierarchical structure is the *data-link control* (DLC) layer, which is responsible for converting the unreliable bit pipe of the physical layer into a higher-level error-free link. For this purpose DLC employs two mechanisms: framing and error correction control. Framing is accomplished by adding header information, which consists of a synchronization preamble, with source and destination addresses at the beginning of the information sequence, and cyclic redundancy check (CRC) bits at the end. The CRC bits are formed from the bits in the packet and are used for error correction control.

At the receiver side, the DLC performs a check using the CRC field to detect errors in a packet. If the CRC fails, it may ask for a retransmission depending on the automatic repeat request (also known as the Automatic Repeat reQuest) (ARQ) protocol. Some widely used ARQ schemes are stop and wait, go-back-N, and selective repeat. These protocols control the logical sequence of transmitting packets between two nodes

and acknowledging the correctly received packets. ARQ procedures form the logical link control (LLC) sublayer of the DLC. If the network is based on multiaccess links, rather than point-to-point links, additional measures must be taken to orchestrate the access of multiple sources to the same medium. These measures are called *media access control* (MAC). Commonly used MAC protocols are the ALOHA protocol, the carrier sense media access (CSMA) protocols, and token protocols. These protocols form the MAC sublayer of DLC.

The layer above the DLC layer is the network layer. The main function of the network layer is to transfer information packets to their final destination, which is called *routing*. Routing involves finding a path through the network and forwarding the packets from the source to the destination along this path. If a route is established at the beginning of a transaction and all the packets follow this path, the network is called a *virtual circuit-switching network*. If a new path is determined for each packet of the same transaction, the network employs datagram switching. In case of datagram switching, packets may arrive to the destination out of order. Therefore, the network layer should reorder the packets before passing them to a higher level.

The network layer selects optimal routes by minimizing the end-to-end path distance. The distance metric can be the delay, the number of hops, required energy, or some other "distance" measure. Some well-known static routing algorithms are the Dijkstra algorithm and the Bellman–Ford algorithm. In dynamic environments, the routes provided by these static algorithms are modified as the "distance" metrics change.

In the following paragraphs, we review the methods and protocols used in the DLC layer and network layer, together with their applicability to UWA networks. We also discuss possible network topologies, which is an important constraint in designing a network protocol.

3.1. Network Topologies

Three basic topologies can be used to interconnect network nodes: centralized, distributed, and multihop [9]. In a *centralized* network, the communication between nodes takes place through a central station, which is sometimes called the "hub" of the network. The network is connected to a backbone at this central station. This configuration is suitable for deep-water networks, where a surface buoy with both an acoustic and an RF modem acts as the hub and controls the communication to and from ocean-bottom instruments. A major disadvantage of this configuration is the presence of a single failure point [9]. If the hub fails, the entire network shuts down. Also, because of the limited range of a single modem, the network cannot cover large areas.

The next two topologies support peer-to-peer links. A fully connected peer-to-peer topology provides point-to-point links between every node of the network. Such a topology eliminates the need for routing. However, the output power needed for communicating with widely separated nodes is excessive. Also, a node that is trying to send packets to a far-end node can overpower and interfere with communication between neighboring nodes, which is called the *near–far problem* [9].

Multihop peer-to-peer networks are formed by establishing communication links only between neighboring nodes. Messages are transferred from source to destination by hopping packets along a multinode route. Routing of the messages is handled by intelligent algorithms that can adapt to changing conditions. Multihop networks can cover relatively larger areas since the range of the network is determined by the number of nodes rather than the modem range.

One of the UWA network design goals is to minimize the energy consumption while providing reliable connectivity between the nodes in the network and the backbone. Network topology is an important parameter that determines the energy consumption [10]. A simplified scenario in which a number of nodes and a master node arranged linearly along a line is considered. The nodes are uniformly distributed, and each node tries to send its packets to the master node. Two extreme communication strategies are possible in this scenario. In the first strategy, each node has direct access to the master node (fully connected topology). In the second strategy, each node transmits only to its nearest neighbor, who then relays the information toward the master node (multihop peer-to-peer topology). The energy consumption curves for both of these cases is plotted as a function of the total distance spanned by the network shown in Fig. 1. Dashed curves represent the case of direct access, which obviously requires more energy. For direct access, inclusion of each additional node results in an increase in total energy. For relaying, the situation is reversed; inclusion of each additional node decreases the total energy consumption because the additional node serves as an additional relay along the same distance.

Hence, the strategy that minimizes energy consumption is multihop peer-to-peer topology. The price paid for the decrease in energy consumption is the need for a sophisticated communication protocol and an increase in packet delay. Therefore, special attention should be given to applications that are sensitive to delays.

3.2. Multiple-Access Methods

In many information networks, communication is bursty, and the amount of time a user spends transmitting over the channel is usually smaller than the amount of time it remains idle. Thus, network users should share the available frequency and time in an efficient manner by means of a multiple access method. Frequency-division multiple access (FDMA) divides the available frequency band into subbands and assigns each subband to an individual user. Because of the severe bandwidth limitations and vulnerability of narrowband systems to fading, FDMA systems do not provide an efficient solution for UWA applications. Instead of dividing the frequency band, time-division multiple access (TDMA) divides a time interval, called a "frame," into time slots. Collision of packets from adjacent time slots is prevented by including guard times that are proportional to the propagation delay present in the channel. TDMA systems require very precise synchronization for proper utilization of the time

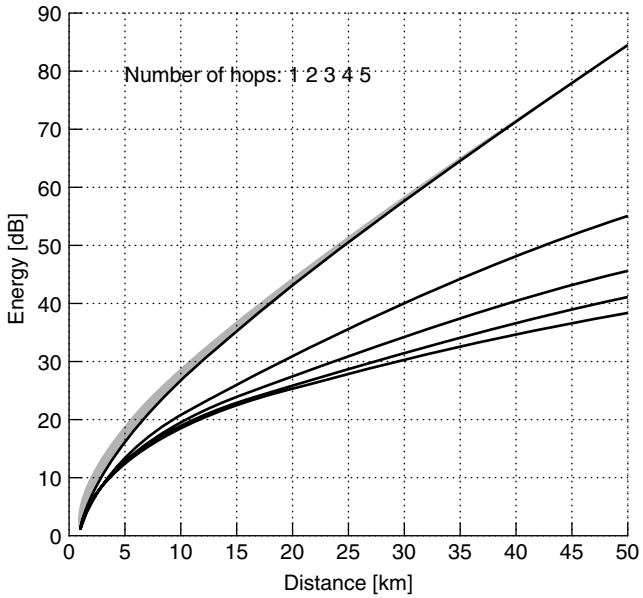


Figure 1. Total (normalized) energy needed to transmit a packet from each network node to the master node. Solid curves represent relaying; dotted curves represent direct access. The parameter on the curves is the number of hops. For relaying, the number of hops increases fro the top curve downward; although more packets are sent when there are more hops, total energy consumption is lower. For direct access, there is little difference between the curves, and the situation is reversed: the number of hops increases from 1 for the lowest energy consumption to 5 for the highest.

slots. High latency present in UWA channels requires long guard times that limits the efficiency of TDMA. Also, establishing a common timing reference is a difficult task. Code-division multiple access (CDMA) allows multiple users to transmit simultaneously over the entire frequency band. Signals from different users are distinguished by means of pseudonoise (PN)codes that are used for spreading the user messages. The large bandwidth of CDMA channels provides resistance to frequency-selective fading and exploits the time diversity present in the UWA channel by employing RAKE filters at the receiver in the case of DS-CDMA (direct-sequence CDMA) [11]. Spread-spectrum signals can be used for resolving collisions at the receiver by using multiuser detectors [12]. In this way, the number of retransmissions and energy requirements of the system can be reduced. This property both reduces battery consumption and increases the throughput of the network. Also, the power requirement of CDMA systems may be less than one-tenth that of TDMA [13]. In conclusion, CDMA appears to be a promising multiple-access technique for shallow-water acoustic networks.

3.3. Media Access Control (MAC) Protocols

Since the number of channels (time frames, frequency bands, or spreading codes) offered by a multiple access method can be much less than the total number of users in a network environment, the same channel is assigned to more than one user. If these users access the channel at the same time, their signals overlap and may be lost (packet collision). Likewise, most underwater acoustic modems are half-duplex in nature, and signals arriving during a transmission are lost, and must also be treated as packet collisions. Media access control (MAC) protocols are used to avoid information loss as a result of packet collisions.

A group of MAC protocols, such as the ALOHA protocol, do not try to prevent collisions but detect collisions and retransmit lost packets. The original ALOHA protocol is based on random access of users to the medium [14]. Whenever a user has information to send, it transmits it immediately. An acknowledgment (ACK) is sent back by the receiver if the packet is received without errors. Because of the arbitrary transmission times, collisions occur and packets are lost. Slotted ALOHA is an enhanced version of the ALOHA protocol, where the time frame is divided into time slots. When a node wants to send a packet, it waits until the next time slot and then begins transmission. Restricting packet transmission to predetermined time slots decreases the probability of collisions [9]. As in the case of TDMA, the ALOHA protocol is inefficient for UWA environment due to slow propagation. Also, the need for retransmissions increases the power consumption of the network nodes and reduces the lifetime of the network.

The number of retransmissions can be reduced if the MAC protocol uses a priori information about the channel state. The media access methods based on this idea are called *carrier sense multiple access* (CSMA) [9]. Details and various forms of this method can be found in papers by Kleinrock and Tobagi [15–18]. The CSMA method tries to avoid the collisions by listening for a carrier in the vicinity of the transmitter. However, this approach does not avoid collisions at the receiver [19]. Let us consider a network formed by three users as shown in Fig. 2. The circles around each node show the communication range of that node. Assume that node A is sending a packet to node B. At the same time, node C listens to the channel and because it is out of the range of A and does not detect the carrier of A, it begins transmission. This creates a collision at B, which is the receiver node. Node A was hidden from node C. This situation is called the "hidden node" scenario [19]. To enable B to hear both messages, node C should defer its transmission. However, if the destination of the packet of C is not B, there is no reason to defer the transmission, provided that node B has the capability to deal with the interference generated by the signal from C [19]. In the case of B sending a packet to A,

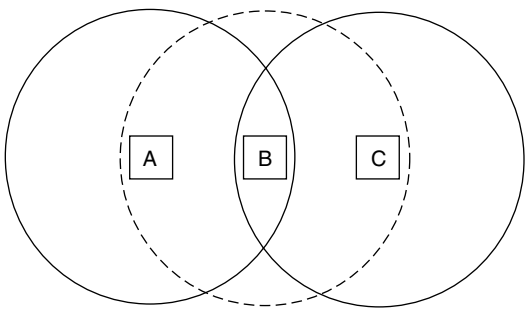


Figure 2. Node A can communicated with node B, but not with node C. Node B can communicated with both A and C.

node C detects a carrier. This creates an "exposed node" situation, where C is exposed to node B.

CSMA cannot solve these problems without adding a guard time between transmissions that is proportional to the maximum propagation time present in the network. The extensive propagation delays in underwater channels can cause this method to become very inefficient. If we consider an underwater acoustic network with a maximum range of 10 km, a data rate of 1 kbps, and a packet size of 1000 bits, the transmission delay and the maximum propagation delay become 1 and 6.7 s, respectively. In this situation, most of the time the channel will be idle, which results in low throughput.

The multiple access with collision avoidance (MACA) protocol was proposed by Karn [20] to detect collisions at the receiver as an alternative to CSMA. This protocol uses two signaling packets: "request to send" (RTS) and "clear to send" (CTS). When node A wants to send a message to B, it first issues an RTS command that contains the length of the message that is to be sent. If B receives the RTS, it sends back a CTS command that also contains the length of the message. As soon as A receives CTS, it begins transmission of the data packet. A node that overhears an RTS (C in this case) defers long enough to let node A receive the corresponding CTS. Also, any node that overhears a CTS defers its transmission for the length of the data packet to avoid collision. If a node overhears an RTS but not a CTS, it decides that it is out of range of the receiver, and transmits its own packet. Therefore, this protocol can solve both the hidden- and the exposed-node problems. The nodes can probe the channel during the RTS–CTS exchange [20]. The channel state information can be used to set the physical-layer parameters, such as output power and modulation type, These properties of the MACA protocol are essential for efficient UWA network design. MACA provides information for reliable communication with minimum energy consumption and can prevent collisions before they occur. The RTS–CTS exchange adds overhead but the reduction of retransmissions can compensate for this increase.

The MACAW protocol was proposed by Bharghavan [19] to improve performance and reliability of the MACA protocol. Instead of creating error-free, reliable point-to-point links with the DLC layer by use of acknowledgments, the MACA protocol ensures the reliability of the end-to-end link with the network layer. If some packets of a message are lost because of errors, the final destination node will ask the originating source to retransmit the lost packets. On highly reliable links, this approach increases throughput, since it eliminates the need to send individual acknowledgments for each hop. In case of poor-quality communication channels, a message will most probably contain erroneous packets. Recovering the errors in the data packet at the network layer will require excessive delay. Generally, error correction is better performed at the data-link layer for channels of low reliability, such as radio or shallow-water acoustic channels. For this purpose, an ACK packet is transmitted after each successful transaction. Including an extra packet in the transaction increases the overhead, which decreases the throughput. However, it has been shown [19] that, for radio channels, the gain in throughput exceeds the increase in overhead. This result may also apply to UWA channels. The MACAW protocol ignores power control and asymmetries that can occur. Its performance under power control needs to be investigated. Also, the effect of adding more overhead to the protocol in an environment where propagation delays are excessive needs to be assessed.

Deng and Haas [21] show that the performance of protocols based on RTS–CTS exchange can degrade as a result of the collision of control packets (RTS–CTS), especially in the case of high propagation and transmission delay. The dual-busy-tone multiple access (DBTMA) protocol is proposed to reduce the packet collision probability. In this protocol, the single shared channel is divided into two subchannels: a data channel and a control channel. Control packets are transmitted on the control channel, while data are carried on the data channel. In addition, two out-of-band tones are introduced to indicate that a node is transmitting on the data channel (BT_t) and a node is receiving on the data channel (BT_r). Researchers face two important limitations for employment of this protocol in an UWA network where energy and bandwidth are scarce: (1) the use of additional tones increases the energy consumption of network nodes, and (2) the divided bandwidth decreases the data throughput of nodes. However, because of the reduced number of collisions, the total energy consumption may be reduced, while network throughput can be increased.

3.4. Automatic Repeat Request (ARQ) Methods

Automatic repeat request (ARQ) is used in the data-link control layer to request the retransmission of erroneous packets. The simplest ARQ scheme that can be directly employed in a half-duplex UWA channel is the stop-and-wait ARQ, where the source of the packet waits for an ACK from the destination node for the confirmation of error-free packet transmission. Since the channel is not utilized during the round-trip propagation time, this ARQ scheme has low throughput. In go-back-N and selective-repeat ARQ schemes, nodes transmit packets and receive ACKs at the same time, and therefore require full-duplex links. Dividing the limited bandwidth of the UWA channels into two channels for full-duplex operation can significantly reduce the data rate of the physical layer. However, the effect on the overall network throughput needs to be investigated.

The selective-repeat ARQ scheme can be modified to work on half-duplex UWA channels. Instead of acknowledging each packet individually at reception time, the receiver will wait for N packet durations and send an ACK packet with the identification number of packets received without errors. Accordingly, the source of the packets will send N packets and wait for the ACK. Then, the source will send another group of N packets that contains the unacknowledged packets and new packets.

Acknowledgments can be handled in two possible ways. In the first approach, which is called *positive acknowledgment*, on reception of an error-free packet, the destination node will send an ACK packet to the source node. If the source does not receive an ACK packet before

a preset timeout duration, it will retransmit the data packet. In the case of a *negative acknowledgment*, the destination sends a packet if it receives a corrupted packet or does not receive a scheduled data packet. A negative acknowledgment may help conserve energy by eliminating the need to send explicit ACK packets and retransmission of data packets in case of a lost ACK packet. When combined with a MACA-type MAC protocol, the negative acknowledgment scheme may provide highly reliable point-to-point links due to the information obtained during RTS–CTS exchange as discussed in Section 3.3.

3.5. Routing Algorithms

As previously indicated, there are two basic methods used for routing packets through an information network: *virtual circuit* routing, where all the packets of a transaction follow the same path through the network, and *datagram* routing, where packets are allowed to pass through different paths. Networks using virtual circuits decide on the path of the communication at the beginning of the transaction. In datagram switching, each node that is involved in the transaction makes a routing decision, which is to determine the next hop of the packet.

Many of the routing methods are based on the *shortest-path* algorithm. In this method, each link in the network is assigned a cost that is a function of the physical distance and the level of congestion. The routing algorithm tries to find the shortest path, that is, the path with lowest cost, from a source node to a destination node. In a distributed implementation each node determines the cost of sending a data packet to its neighbors and shares this information with the other nodes of the network. In this way, every node maintains a database that reflects the cost of possible routes.

For routing, let us consider the most general problem where network nodes are allowed to move. This situation can be viewed as an underwater network with both fixed ocean-bottom sensors and AUVs. The instruments temporarily form a network without the aid of any preexisting infrastructure. These types of networks are called *ad hoc networks* [22].

In ad hoc networks the main problem is to obtain the most recent state of each individual link in the network, so as to decide on the best route for a packet. However, if the communication medium is highly variable as in the shallow-water acoustic channel, the number of routing updates can be very high. Current research on routing focuses on reducing the overhead added by routing messages while at the same time finding the best path, which are two conflicting requirements. Broch et al. [23] compared four ad hoc network routing protocols presented in the literature:

- Destination sequence distance vector (DSDV) [24]
- Temporally ordered routing algorithm (TORA) [25]
- Dynamic source routing (DSR) [26]
- Ad hoc on-demand distance vector (AODV) [27]

DSDV maintains a list of *next hops* for each destination node that belongs to the shortest-distance route. The protocol requires each node to periodically broadcast routing updates to maintain routing tables.

TORA is a distributed routing algorithm. The routes are discovered on demand. This protocol can provide multiple routes to a destination very quickly. The route optimality is considered as a second priority, and the routing overhead is reduced.

DSR employs source routing; that is, the route of each packet is included in its header. Each intermediate node that receives the packet checks the header for the next hop and forwards the packet. This eliminates the need for intermediate nodes to maintain best routing information to route the packets.

AODV uses the on-demand route discovery and maintenance characteristic of DSR and employs them in a hop-by-hop routing scheme instead of source routing. Also, periodic updates are used in this protocol.

In a mobile radio environment DSR provides the best performance in terms of reliability, routing overhead, and path optimality [23]. The effect of long propagation delays and channel asymmetries caused by power control are issues that need to be addressed when considering application of these network routing protocols to UWA channels.

4. EVOLUTION OF UWA NETWORKS

Two types of applications have guided the evolution of underwater networks so far: gathering of environmental data and surveillance of an underwater area. In the first case, the network consists of several types of sensors, some mounted on fixed moorings and others mounted on moving vehicles. This type of network is called an *autonomous ocean sampling network* (AOSN), where the word "sampling" implies collecting the samples of oceanographic parameters such as temperature, salinity, and underwater currents. For surveillance applications, the network consists of a larger number of sensors, typically bottom-mounted or on slowly crawling robots, that can be quickly deployed, and whose task is to map a shallow-water area. In particular, mapping may focus on detection of warfare objects. An example of such a network, called *Seaweb*, will be described in more detail in Section 5.

An AOSN is formed by a number of autonomous underwater vehicles (AUVs), moorings, and surface buoys. The AUVs traverse an ocean area spanned by the network nodes (moorings and surface buoys) collecting scientific data. The coordination of the AUVs is handled from a central location, which can be either at one of the network nodes and/or on shore. AUVs relay key observations and their status to the central location. After evaluating the incoming data, the central location sends control signals to the AUVs through the network nodes. The acoustic communication between the AUVs and the network nodes is designed so that it does not require high data throughput. More complete data sets are transferred to the onshore control center through a radio channel when AUVs dock to a mooring. The control center is connected to a backbone such as the Internet, so that a scientist can reach the sampling network in real time. An important

limitation observed during the tests of the AOSN was the impossibility for the AUVs to instantly respond to commands due to the high-latency environment [28]. As a result of the highly variable acoustic channel, network connections to AUVs were occasionally lost. Therefore, some level of automation is needed in the AUVs to avoid disastrous events, which may occur if the last command sent to an AUV directs it toward an obstacle and the connection is lost.

A deep-water acoustic local-area network (ALAN) was deployed in Monterey Canyon, California, for long-term data acquisition and ocean monitoring from multiple ocean-bottom sources [1]. A centralized network topology was employed with a hub on the surface. The MAC protocol was based on TDMA, where time slots were determined adaptively on the basis of estimated latency. Because this protocol relies on correct estimation of the round-trip propagation times, any error in the estimation process decreases the throughput of the system by causing retransmissions.

The evolution of research on underwater acoustic networks has followed the usual layered architecture. Most work to date has been performed on the physical layer and multiple-access techniques. The data-link-layer protocols have been addressed to a lesser extent, and the work has only begun on the network layer and routing algorithms [5]. In all of these areas, the focus of research has been on adapting the well-known theoretical concepts to the requirements and constraints of the underwater acoustic channels.

Typically, packet transmission in a store-and-forward network is considered in most of the underwater acoustic networks. The design of automatic repeat request (ARQ) protocols is influenced by the long propagation times in underwater channels. Talavage et al. [29] proposed a shallow-water acoustic local-area network (S-ALAN) protocol, which is a modified version of ARPA-supported packet radio network (PRN) protocol. The S-ALAN differs from PRN in the routing algorithm and the data transmission medium. In contrast to PRN, which uses datagram switching over a single channel, S-ALAN employs virtual circuit switching using three separate channels (frequency bands) and selective-repeat ARQ. When a network node gathers enough information to send to the control center, it issues a request to set up a virtual circuit. When the setup request reaches the destination node, the destination node assigns transmit data, receive data, and acknowledgment channels for all the nodes in the virtual circuit and reports the final configuration back to the source node. The use of three separate channels enables the network to fully utilize the ARQ scheme.

A peer-to-peer communication protocol has been developed to control AUVs [30] based on carrier sense multiple access with collision avoidance (CSMA/CA). Since the CSMA/CA protocol relies on acknowledgments, the channel remains idle for an amount proportional to the round-trip propagation time. Because of the long propagation delays in UWA channels, this protocol has a low throughput. On the other hand, it is highly reliable.

A media access protocol proposed for shallow-water acoustic networks is presented in Ref. 10. The protocol is based on the MACA protocol and employs a stop-and-wait ARQ scheme. The RTS–CTS exchange is used to determine the channel conditions, and this information is used to set the acoustic modem parameters such as output power level. The details of this network are given in the following section.

The routing optimization problem for a shallow-water acoustic network is addressed in Ref. 31. The genetic algorithm based routing protocol tries to maximize the lifetime of the battery-powered network by minimizing the total energy consumption of the network. The minimum energy required to establish reliable communication between two nodes is used as the link distance metric. A master node collects the link cost information from the network nodes, determines optimum routes, and sends the routing information back to the nodes. The authors showed that the optimization algorithm favors multihop links at the expense of increased delay.

5. SEAWEB

Seaweb is an acoustic network for communications and navigation of deployable autonomous undersea systems [32]. The U.S. Navy incorporated Seaweb networking in the June 2001 Fleet Battle Experiment India (FBE-I). The Seaweb installation charted in Fig. 3 was part of the overall FBE-I joint forces architecture for command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR) providing wide-area connectivity, enhanced bandwidth, and reachback capability. Seaweb reliably supported asynchronous networking for an improved 688-class (688I) fast-attack nuclear submarine mobile node and two deployable autonomous distributed system (DADS) nodes. Two moored radio-acoustic communications (racom) buoy gateway nodes provided line-of-sight radio links to a shore station having terrestrial Internet connection to an antisubmarine warfare (ASW) command center (ASWCC) located ashore. In addition, 10 undersea repeater nodes highlighted the flexible architecture, indicating expandability and potential area coverage. The network performed reliably with no hardware failures and no lost transmissions. Seaweb supported Internet Protocol (IP) delivery of automated ASW contact reports from the DADS sensors to shore, and command and control (C2) on the IP backlink. Naval messages to and from the submarine via Seaweb protocols permitted assured access at tactical depth. The 15-node FBE-I Seaweb system extended Naval network-centric operations into the undersea battlespace. Analysts executed numerous communication and navigation tests, and proved that the FBE-I network design was overly conservative and could have supported even greater area coverage and traffic load.

5.1. Concept of Operations

Telesonar wireless acoustic links interconnect distributed undersea instruments, potentially integrating them as a unified resource. Seaweb is the realization of an undersea telesonar network [10] of fixed and mobile nodes, with various interfaces to manned command centers.

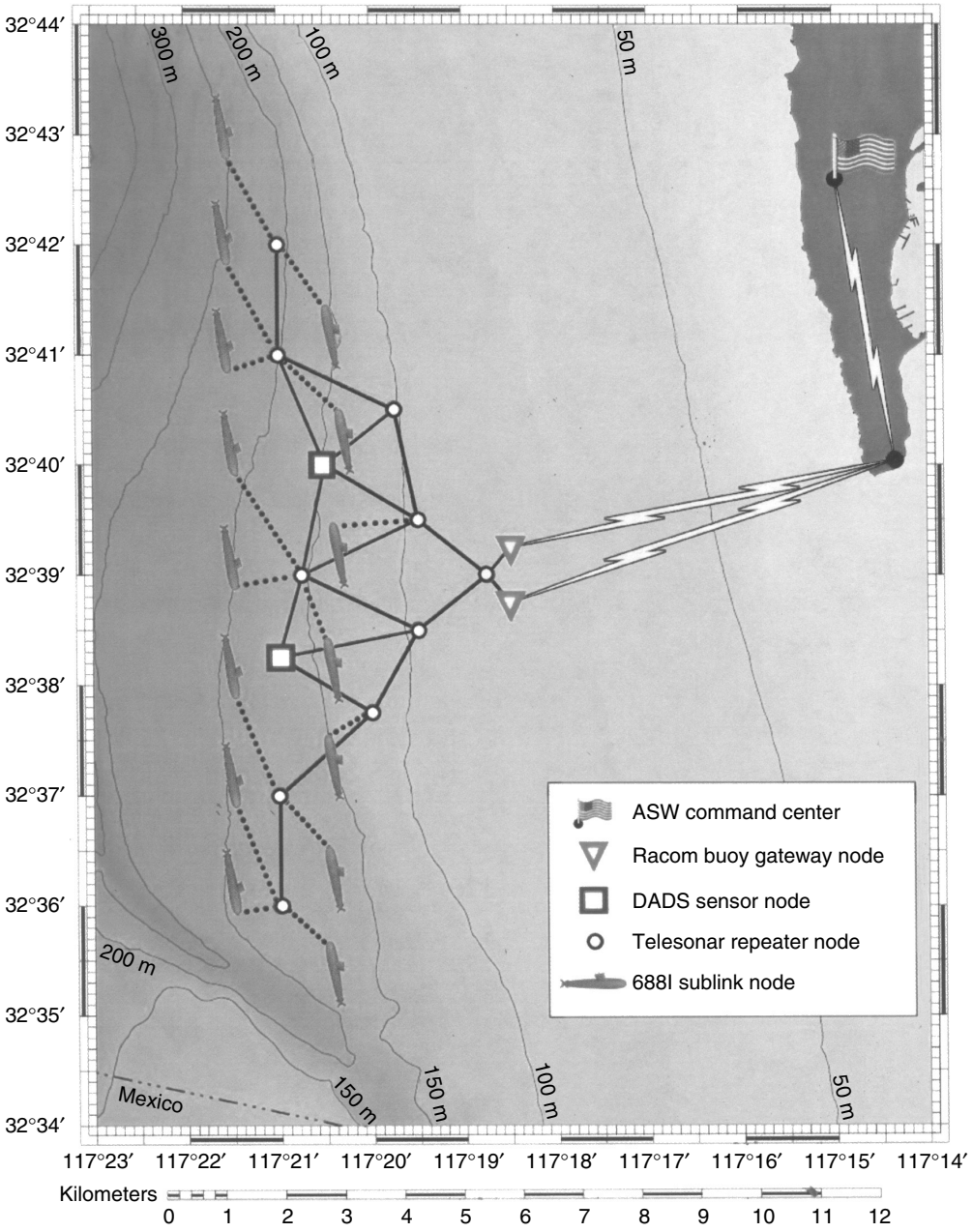


Figure 3. The FBE-I Seaweb network was a 14-node undersea grid. Two nodes were prototype deployable autonomous distributed system (DADS) sensors for littoral ASW, and two nodes were moored radio/acoustic communications (racom) gateway nodes. A mobile submarine node with sublink capacity had full interoperability with the Seaweb network. An ASW command center served as the ashore site. U.S. Navy personnel exercised the complete Seaweb installation for 4 days with high reliability and no component failures.

The architectural flexibility afforded by Seaweb wireless connections permits the network designer to allocate an arbitrary mix of node types with a node density and area coverage appropriate for the given telesonar propagation conditions and for the mission at hand. The concept of operations assumes that the majority of network nodes are inexpensive, autonomous, battery-limited devices deployed from submarine, ship, and aircraft, or from unmanned undersea vehicles (UUVs) and unmanned aerial vehicles (UAVs).

Seaweb networks support asynchronous data communications from autonomous nodes to command centers. On the backlink, Seaweb allows remote command and control of instruments associated with the autonomous nodes. Additionally, network activity supports acoustic navigation and geolocalization of undersea nodes as a natural byproduct of telesonar ranging signals. More generally, Seaweb networking permits wireless transmissions between member nodes in the network using established routes or via an intervening cellular node.

Seaweb enables future Naval capabilities in littoral ASW and undersea autonomous operations. A significant dual use of Seaweb is communication and navigation for oceanographic surveys and environmental assessment. Certainly, a major potential benefit of the technology is cross-system, cross-platform, cross-mission interoperability, providing enormous added value to otherwise solitary systems. For example, a UUV mobile node operating within a grid of fixed sensor nodes benefits from the established network topology for situational awareness, navigation, and communications via gateway nodes to distant command centers. Conversely, UUVs add value to the fixed grid for sensor deployment, search, survey, water-column sampling, popup racom gateway communications, and other functions.

The initial motivation for Seaweb is a requirement for wide-area surveillance in littoral waters. These sensors operate in 50- to 300-m waters with node spacing of 2–5 km. Sensor nodes generate concise ASW contact reports that Seaweb routes to a master node for field-level data fusion [33]. Primary network packets are contact reports with about 1000 information bits [34]. Sensor nodes asynchronously produce these packets at a variable rate dependent on the receiver operating characteristics (ROCs) for a particular sensor suite and mission. The master node communicates with manned command centers via encrypted gateway nodes such as a racom sea-surface buoy linked with space satellite networks. Following ad hoc deployments, the Seaweb network self-organizes, including node identification, clock synchronization on the order of 0.1–1.0 s, node geolocalization on the order of 100 m, assimilation of new nodes, and self-healing following node failures.

As a fixed grid of inexpensive interoperable sensor nodes and repeater nodes, this is the most fundamental Seaweb operating mode based on a stable topology that periodically adjusts itself to optimize overall network endurance and quality of service (QoS). The fixed Seaweb topology provides an underlying cellular network suited for supporting an AOSN [35], including communication and navigation for UUV mobile nodes. The cellular architecture likewise provides seamless connectivity for submarine operations at speed and depth in a manner not unlike those for terrestrial cellular telephone service for automobiles.

5.2. Developmental Approach

The concept of operations emphasizes simplicity, efficiency, reliability, and security, and these attributes therefore govern the design philosophy for Seaweb development. Research is advancing telesonar modem technology for reliable underwater signaling by addressing the issues of (1) adverse transmission channels, (2) asynchronous networking, (3) battery energy efficiency, (4) transmission security, and (5) cost.

Despite a concept of operations emphasizing simplicity, Seaweb is a multifaceted system, and its development is a grand challenge. The high cost of sea testing and the need for many prototype nodes motivate extensive engineering system analysis. Simulations using an optimized network engineering tool (OPNET) with simplified ocean acoustic propagation assumptions permit laboratory refinement of networking protocols [34] and initialization methods [36]. Meanwhile, controlled experimentation in actual ocean conditions incrementally advances telesonar signaling technology [37].

Seaweb development balances the desire for rapid increases in capability and the need for stable operation in support of applications that are themselves developmental. This balance is achieved by an annual cycle culminating with the late-summer Seaweb experiment (Seaweb '98, Seaweb '99, Seaweb 2000, etc.).

The objective of the annual Seaweb experiments is to exercise telesonar modems in networked configurations where various modulation and networking algorithms can be assessed. In the long term, the goal is to provide for a self-configuring network of distributed assets, with network links automatically adapting to the prevailing environment through selection of the optimum transmit parameters. A full year of hardware improvements and in-air network testing helps ensure that the incremental developments tested at sea will provide tractable progress and mitigate overall developmental risk. In preparation for Seaweb experimentation, multiple contributing projects conduct relevant research during the first three-quarters of the fiscal year. The fourth-quarter Seaweb experiment then implements and tests the results from these research activities with a concentration of resources in prolonged ocean experiments. The products of the annual Seaweb experiment are major capability upgrades for integrated Seaweb server software, telesonar modem Seaweb firmware, and telesonar modem hardware. The annual Seaweb experiment also transitions these upgrades into participating application systems. After the annual Seaweb experiment yields a stable level of functionality, the firmware product can be further exercised and refinements instituted during system testing and by spinoff applications throughout the year. For example, in year 2001, Seaweb 2000 technology enabled the March–June FRONT-3 ocean observatory on the continental shelf east of Long Island, New York [38]. These applications afford valuable long-term performance data that are not obtainable during Seaweb experiments when algorithms are in flux and deployed modems are receiving frequent firmware upgrades. At the conclusion of the Seaweb experiment, the upgraded Seaweb capability reaches stability suitable for use with the continuing development of the various application systems during the following year. Meanwhile, the annual cycle repeats, beginning with research and preparations for the next Seaweb experiment. And so, Seaweb capability increases in an incremental manner.

The Seaweb architecture of interest includes the physical layer, the media access control (MAC) layer, and the network layer. These most fundamental layers of communications functionality support higher layers, collectively identified here as the "client" layer. The client layer tends to be application-specific and is not the direct responsibility of telesonar modems or the Seaweb network.

5.3. Telesonar Modems and the Physical Layer

The U.S. Navy has been developing *telesonar* modems designed to function at low bit rates with high reliability and modest processing [39,40]. The basis for this approach is the need for low-cost, energy-efficient workhorse modems suitable for the development of networking technologies [41]. From an interoperability perspective, the low-bit-rate modem offers the lowest common denominator for cross-system networks that may include low-cost, expendable nodes [42,43]. As a pair of modems establishes a low-bit-rate link, they may adaptively negotiate higher-bit-rate modulations if warranted by favorable propagation and available processing resources.

The present telesonar modem [44] normally uses 5 kHz of acoustic bandwidth encompassing 120 discrete MFSK bins and 8 tracking bins [45]. A basic 1-of-4 MFSK modulation carries 2400 bps but lacks data protection and error correction coding (ECC). A constraint-length-9, rate-$\frac{1}{2}$ convolutional code very effectively corrects bit errors by representing binary information across multiple symbols; the rate-$\frac{1}{2}$ reduces throughput by a factor of 2. For protection against multipath-induced inter-symbol interference (ISI), the MFSK chip duration may be lengthened; the modem allows for a doubling from 25 to 50 ms, resulting in another factor of 2 reduction in bit rate. A "Doppler-tolerant" mode skips alternate MFSK bins to allow greater latitude for tracking Doppler shifts caused by node-to-node range rate. The Doppler-tolerant mode also increases robustness by doubling the acoustic energy per chip, but it also causes another halving in bit rate. Finally, a Hadamard MFSK modulation carries 6 Hadamard codewords of 20 tones each. Interleaving the codewords across the band increases immunity to frequency-selective fading, and Hadamard coding yields a frequency diversity factor of 5 for adverse channels having low or modest spectral coherence. Hadamard signaling is effective in channels having frequency-selective fading or narrowband noise. Any combination of the above mentioned options is possible. For FBE-I, a conservative modulation choice combined Hadamard MFSK, rate-$\frac{1}{2}$ convolutional coding, and 50-ms chip lengths to yield a net 300 bps information rate.

For all operational modes, receiving modems process the data noncoherently using a fixed-point TMS320C5410 DSP. Directional transducers can further enhance the performance of these devices [46,47]. The present telesonar modem includes provision for a watchdog function hosted aboard a microchip independent of the DSP. The watchdog resets the DSP on detection of supply voltage drops or on cessation of DSP activity pulses. The watchdog provides a high level of fault tolerance and permits experimental modems to continue functioning in spite of system errors. A watchdog reset triggers the logging of additional diagnostics for thorough troubleshooting after modem recovery.

Low-bandwidth, half-duplex, high-latency telesonar links limit Seaweb quality of service (QoS). Occasional outages from poor propagation or elevated noise levels can disrupt telesonar links [48]. Ultimately, the available energy supply dictates service life, and battery-limited nodes must be energy-conserving [49]. Moreover, Seaweb must ensure transmission security by operating with low bit-energy per noise-spectral-density (E_b/N_0) and by otherwise limiting interception by unauthorized receivers.

Spread-spectrum modulation is consistent with the desire for asynchronous multiple access to the physical channel using CDMA networking [50]. Nevertheless, the Seaweb concept does not exclude TDMA or FDMA methods and is in fact pursuing hybrid schemes suited to the physical-layer constraints. In a data transfer, for example, a concise asynchronous CDMA dialog could queue data packets for transmission during a time slot or within a frequency band such that multiaccess interference (MAI) collisions are avoided altogether.

At the physical layer, an understanding of the transmission channel is obtained through at sea measurements [51] and numerical propagation models [52]. Knowledge of the fundamental constraints on telesonar signaling translates into increasingly sophisticated modems [53]. DSP-based modulators and demodulators permit the application of modern digital communications techniques to exploit the unique aspects of the underwater channel. To aid understanding of telesonar performance, modems automatically log physical-layer diagnostics, including signal-to-noise ratio (SNR), automatic gain control (AGC), bit error rate (BER), and the number of corrected and uncorrected errors.

5.4. Handshake Protocols and the MAC Layer

Developmental Seaweb modem firmware implements the core features of a compact, structured protocol for secure, low-power, point-to-point, connectivity. The protocol efficiently maps MAC-layer functionality onto a physical layer based on channel-tolerant, 64-bit utility packets and channel-adaptive, arbitrary-length data packets. Seaweb firmware implements utility packet types using the basic Hadamard MFSK physical layer. These utility packet formats permit data transfers and node-to-node ranging. A richer set of available utility packets is being investigated with OPNET simulations prior to modem implementation, but seven core utility packets provided substantial networking capability for FBE-I.

The telesonar handshake protocol is suited to wireless half-duplex networking with slow propagation. Handshaking [20] asynchronously establishes adaptive telesonar links [54]. The initial handshake consists of the transmitter sending a request-to-send (RTS) packet and the receiver replying with a clear-to-send (CTS) packet. A busy signal (BSY) packet may be issued in response to an RTS when the receiver node decides to defer data reception in favor of other traffic. Following a successful RTS–CTS handshake, the data packet(s) are sent. This RTS–CTS round trip establishes the communications link and probes the channel to gauge optimal transmit power. Future firmware enhancements will support power control and the adaptive choice of data modulation method, with selection based on channel estimates derived from the RTS role as a probe signal. Telesonar links eventually will be environmentally adaptive [55], with provision for bidirectional asymmetry. Handshaking permits addressing, ranging, channel estimation, adaptive modulation, and power control.

The Seaweb 2000 core protocol implemented stop-and-wait ARQ scheme by providing either positive or negative acknowledgment of a data message. The choice of acknowledgment type depends on the traffic patterns associated with a particular network mission. Handshaking provided the means for resolving packet collisions automatically using retries from the transmitter or automatic repeat request (ARQ) packets from the receiver. For FBE-I, a purely negative acknowledgment was supported by the modem, implemented as an ARQ utility packet. At the client layer, the DADS client system supports positive acknowledgment through its IP implementation. Figure 4 illustrates the MAC layer protocol.

If two nodes send an RTS to each other, unnecessary retries may occur because both nodes will ignore the received RTS command. Each node will then wait for the other node to send a CTS for a timeout duration, and retransmit their RTS packet. This problem is solved by assigning priority to the packets that are directed towards the master node, as explained below (Fig. 5).

Assume that node A is a lower level node than node B; that is, Node A is the parent of B. Node A and node B both send RTS to each other. As a result of transmission delays, packets arrive to their destinations while both nodes are waiting for a CTS packet. When node B receives the RTS, it notices that the packet is from its own destination, node A. Node B checks whether node A is its parent or child. Since node A is its parent, node B has the priority and sends a CTS packet immediately. By that time, node A receives the RTS of node B, does the same check and decides that it should wait for node B to complete its data transmission, since node B is its child. Therefore, node A puts its own data packet into a queue and waits for the CTS packet of node B.

Future implementations of Seaweb firmware will retain the purely negative acknowledgment approach, as analysis has shown this to be the appropriate MAC layer implementation for long-latency, half-duplex links. For communications requiring positive acknowledgments, the Seaweb 2001 firmware includes provision for efficient delivery of a receipt (RCPT) utility packet from the destination node to the source node.

The RTS–CTS approach anticipates eventual implementation of adaptive modulation and secure addressing. The initiating node transmits a RTS waveform with a frequency-hopped, spread-spectrum (FHSS) [44] pattern or direct-sequence spread-spectrum (DSSS) [11] pseudo-random carrier uniquely addressing the intended receiver.

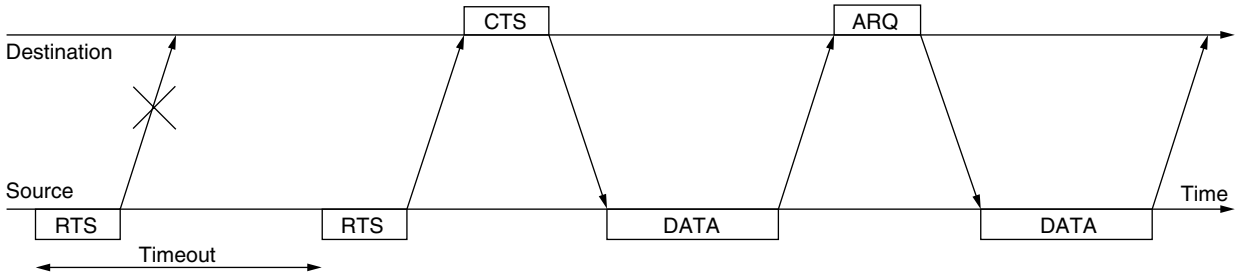


Figure 4. The source node starts the MAC layer handshake protocol by sending an RTS packet to the destination node. If the RTS packet is lost in the channel, the source node retransmits the RTS packet after a timeout duration equal to the round-trip time of an header-only packet (e.g., RTS, CTS, or ARQ), and calculated using the range information in the neighboring tables. When the destination node receives the RTS, it replies with a CTS packet. On receipt of the CTS packet by the source, the DATA packet, which contains a header and the information, is sent to the destination. The destination node issues an ARQ packet if it does not receive the DATA packet before a timeout occurs.

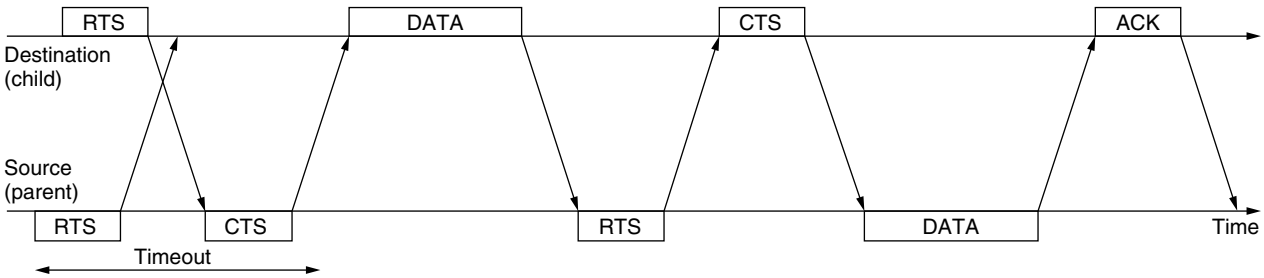


Figure 5. If two nodes send RTS packets to each other with some delay, both nodes receive the packets. When they recognize that the RTS is received from their destination, they check for the priority of the nodes. Higer-level nodes, or children, have higher priority. Therefore, the parent node sends CTS telling the child to send its data. When the parent receives the data, it replies with an RTS, which also act as an acknowledgments. Then the parent sends its data to the child. This example uses positive acknowledgments (ACK).

(Alternatively, the initiating node may transmit a universal code for broadcasting or when establishing links with unknown nodes.) The addressed node detects the request and awakens from an energy-conserving sleep state to demodulate. Further processing of the RTS signal can provide an estimate of the channel scattering function and signal excess. An adaptive power-control technique determines the source level that will deliver sufficient but not excessive SNR. The addressed node then acknowledges receipt with a FHSS or DSSS acoustic response. This CTS reply specifies appropriate modulation parameters for the ensuing message packets based upon the measured channel conditions. Following this RTS–CTS handshake, the initiating node transmits the data packet(s) with nearly optimal bit rate, modulation, coding, and source level.

5.5. Seaweb Server and the Network Layer

The Seaweb backbone is a set of autonomous, stationary nodes (e.g., deployable surveillance sensors, repeaters). Seaweb peripherals are mobile nodes (e.g., UUVs, including swimmers, gliders, and crawlers) and specialized nodes (e.g., bistatic sonar projectors). Seaweb gateways provide connections to command centers submerged, afloat, aloft, and ashore. Telesonar-equipped gateway nodes interface Seaweb to terrestrial, airborne, and space-based networks. For example, a telesonobuoy serves as a racom interface permitting satellites and maritime patrol aircraft to access submerged, autonomous systems. Similarly, submarines can access Seaweb with telesonar signaling through the underwater telephone band or other high-frequency sonar [56]. Seaweb provides the submarine commander digital connectivity at speed and depth with bidirectional access to all Seaweb-linked resources and distant gateways.

At the physical and MAC layers, adaptive modulation and power control are keys to maximizing both channel capacity (bps) and channel efficiency (bit-kilometer/joule). At the network layer, careful selection of routing is required to minimize transmit energy, latency, and net energy consumption, and to maximize reliability and security. Seaweb experimentation underscores the differences between acoustic networks and conventional networks. Limited power, small bandwidth, and propagation latencies dictate that the Seaweb network layer be simple and efficient. For compatibility with Seaweb networks, the higher client layers must utilize lookup tables, data compression, forward error correction, and data filtering to minimize packet sizes and retransmissions, and to avoid congestion at the network layer.

A very significant development was the introduction of the Seaweb server [57]. A Seaweb server resides at manned command centers and is the graphical user interface to the undersea network. It interprets, formats, and routes downlink traffic destined for undersea nodes. On the uplink, it archives incoming data packets produced by the network, retrieves the information for an operator, and provides Web-based read-only database access for client users. The server manages Seaweb gateways and member nodes. It monitors, displays, and logs the network status. The server manages the network routing tables and neighbor tables and ensures network interoperability. Seaweb '99 modem firmware permitted the server to remotely reconfigure routing topologies, a foreshadowing of future self-configuration and dynamic network control. The Seaweb server is a suite of software programs implemented under Linux on a laptop PC with a LabVIEW graphical user interface. A single designated "super" server controls and reconfigures the network.

Network supervisory algorithms can execute either at an autonomous master node or at the Seaweb server. Seaweb provides for graceful failure of network nodes, addition of new nodes, and assimilation of mobile nodes. Essential byproducts of the telesonar link are range measurement, range rate measurement, and clock synchronization. Collectively, these features will support network initialization, node localization, route configuration, resource optimization, and maintenance.

Node-to-node ranging employed a new implementation of a round-trip travel time measurement algorithm with 0.1-ms resolution linked to the DSP clock rate.

As a network analysis aid, all modems now include a data-logging feature. All output generated by the telesonar modem and normally available via direct serial connection is logged to an internal buffer. Thus, the behavior of autonomous nodes can be studied in great detail after recovery from sea. Seaweb 2000 firmware logs diagnostics related to channel estimation (SNR, multipath duration, range rate, etc.), demodulation statistics (BER, AGC, intermediate decoding results, power level, etc.), and networking (data packet source, data packet sink, routing path, etc.). For Seaweb applications, the data-logging feature can also support the archiving of data until such time that an adjacent node is able to download the data. For example, a designated sink node operating without access to a gateway node can collect all packets forwarded from the network and telemeter them to a command center when interrogated by a gateway (such as a ship arriving on station for just such a data download).

Since the network in consideration is an ad hoc network, an initialization algorithm is needed to establish preliminary connections autonomously. This algorithm is based on polling, and as such it guarantees connectivity to all the nodes that are acoustically reachable by at least one of their nearest neighbors. During initialization, the nodes create *neighbor tables*. These tables contain a list of each node's neighbors and a quality measure of their link, which can be the received SNR from the corresponding neighbor. The neighbor tables are then collected by the master node and a routing tree is formed. Table 2 is an example node table for node 3 of the network given in Fig. 6.

The master node decides on the primary (and secondary) routes to each destination, with routing optimization based on the genetic algorithm. Initialization ends when the master node sends primary routes to the nodes. The initialization algorithm provides either a single set of connections, or multiple connections between the nodes. Multiple connections are desirable to provide greater robustness to failures. A possible routing tree with backup routes for the network of Fig. 6 is given in Fig. 7.

Optimum routes are determined with the help of a genetic algorithm-based routing protocol [31]. The routing protocol tries to maximize the lifetime of the

Table 2. Node Table[a] of Node 3 for Network Given in Fig. 6

Neighbor ID	Range (km)	Output Power (dB)
1	6	−9 dB
2	7	−9 dB
4	9	−3 dB

[a] This table contains the ID of the nodes with which node 3 can communicate with a direct link. For each neighbor, the range of the node and the minimum output power required to communicate with that node are entered. The output power is in decibels with respect to the maximum output power of the modem. Because of the channel characteristics, nodes at different ranges may require the same amount of output power.

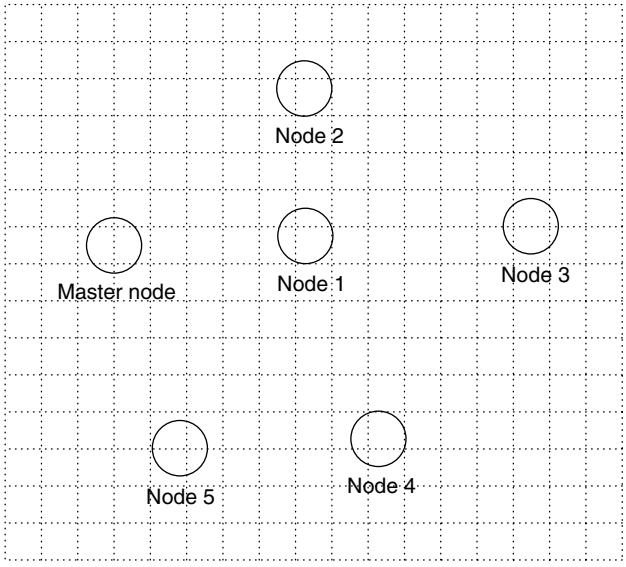


Figure 6. The network consists of a master node and five sensor nodes. The sensor nodes send information packets to the master node, which is the connection point of the network to a backbone. The master node can also send control packets to the sensor nodes.

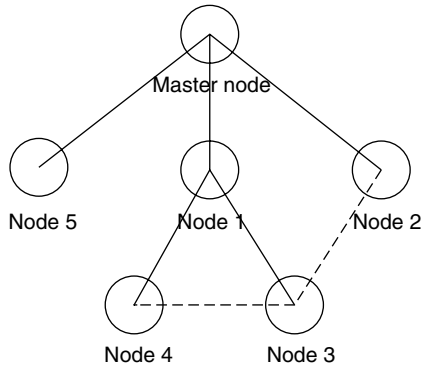


Figure 7. The routing tree is created by the master node. The master node is at the top of the tree. Sensor nodes are the leafs. Nodes 1, 2, and 5 are the children of the master node and form the first layer of the network. Node 1 is the parent of nodes 3 and 4, which form the second layer of the network. Dashed lines represent backup links that can be used in case of a failed link.

battery-powered network by minimizing the total energy consumption of the network. The minimum energy required to establish reliable communication between two nodes is used as the link distance metric. A master node collects the link cost information from the network nodes, determines optimum routes, and sends the routing information back to the nodes. The optimization algorithm favors multihop links at the expense of increased delay.

The performance of acoustic links between nodes can degrade, and even a link can be permanently lost as a result of node failure. In such cases, the network should be able to adapt itself to the changing conditions without interrupting the packet transfer. This robustness can be obtained by updating the routes periodically.

In the current implementation, the network tables are created manually at the Seaweb server with the help of the ranging packets. Also, the initial routes and updates are determined and reported to the network nodes manually through the Seaweb server.

6. CONCLUDING REMARKS

In this article, we presented an overview of basic principles and constraints in the design of reliable shallow-water acoustic networks that may be used for transmitting data from a variety of undersea sensors to onshore facilities. Major impediments in the design of such networks were considered, including (1) severe power limitations imposed by battery power; (2) severe bandwidth limitations; and (3) channel characteristics, including long propagation times, multipath, and signal fading. Multiple-access methods, network protocols, and routing algorithms were also considered.

Of the multiple-access methods considered, it appears that CDMA, achieved either by frequency hopping or by direct sequence, provides the most robust method for the underwater network environment. Currently under development are modems that utilize these types of spread-spectrum signals to provide the multiple-access capability to the various nodes in the network. Simultaneously with current modem development, there are several investigations on the design of routing algorithms and network protocols.

The design example of the shallow-water network employed in Seaweb embodies the power and bandwidth constraints that are so important in digital communication through underwater acoustic channels. As an information system compatible with low bandwidth, high latency, and variable quality of service, Seaweb offers a blueprint for the development of future shallow-water acoustic networks. Experimental data that will be collected in the near future will be used to assess the performance of the network and possibly validate a number of assumptions and tradeoffs included in the design. Over the next decade (at the time of writing), significant improvements are anticipated in the design and implementation of shallow water acoustic networks as more experience is gained through at-sea experiments and network simulations.

BIOGRAPHIES

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

Joseph A. Rice joined the Naval Postgraduate School's Physics Department in 2001 as the SPAWAR Systems Center San Diego Engineering Acoustics Chair. He is Principal Investigator for numerous ONR projects and tasks collectively performed as the Seaweb Initiative. He is the Technical POC for three SBIR contracts producing undersea acoustic modems, networks, and directional transducers. He has been a U.S. Navy research engineer since 1981, developing digital signal processing and numerical modeling concepts for solving undersea acoustics problems. He received the U.S. Navy Meritorious Civilian Service Award for experimental demonstration of undersea acoustic matched-field processing (MFP) during the Swellex acoustic propagation studies. He is interested in acoustic localization and navigation, having developed the prototype array element localization (PAEL) sonobuoy and associated processing for the Sonobuoy Thinned Random Array Program (STRAP). He holds a B.A. in Mathematics and a B.S. in Engineering Science, both from the University of Cincinnati. He holds an M.S.E.E. in Applied Ocean Science from UCSD. He is a member of IEEE and ASA.

Ethem M. Sozer received his B.S. and M.S. degrees in Electrical Engineering from the Middle East Technical University, Ankara, Turkey, in 1994 and 1997, respectively. He is working toward his Ph.D. in Electrical Engineering at Northeastern University, Boston. His research interests include iterative equalization techniques, underwater acoustic communications, and underwater acoustic networks. He is currently working for Delphi Communication Systems, Inc, Maynard, Maryland.

Milica Stojanovic graduated from the University of Belgrade, Belgrade, Yugoslavia, in 1988, and received the M.S. and Ph.D. degrees in Electrical Engineering from Northeastern University, Boston, in 1991 and 1993, respectively. She is a Principal Research Scientist at the Massachusetts Institute of Technology, and a Guest Investigator at the Woods Hole Oceanographic Institution. Her research interests include digital communications theory and statistical signal processing and their applications to wireless communication systems.

BIBLIOGRAPHY

1. J. Catipovic, D. Brady, and S. Etchemendy, Development of underwater acoustic modems and networks, *Oceanography* **6**: 112–119 (March 1993).
2. R. Conogan and J. P. Guinard, Observing operationally in situ ocean water parameters: The EMMA system, *Proc. OCEANS'98*, Nice, France, Sept. 1998, pp. 37–41.
3. J. Marvaldi et al., GEOSTAR — development and test of a communication system for deep-sea benthic stations, *Proc. OCEANS'98*, Nice, France, Sept. 1998, pp. 1102–1107.
4. M. Stojanovic, Recent advances in high-speed underwater acoustic communications, *IEEE J. Ocean. Eng.* **21**: 125–136 (April 1996).
5. D. B. Kilfoyle and A. B. Baggeroer, The state of the art in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **25**: 4–27 (2000).
6. J. Catipovic, Performance limitations in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **15**: 205–216 (July 1990).
7. M. Stojanovic, J. Catipovic, and J. Proakis, Phase-coherent digital communications for underwater acoustic channels, *IEEE J. Ocean. Eng.* **19**: 100–111 (1994).
8. A. Tannenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.

9. K. Pahlavan and A. H. Levesque, *Wireless Information Networks*, Wiley, New York, 1995.
10. E. M. Sozer, M. Stojanovic, and J. G. Proakis, Underwater acoustic networks, *IEEE J. Ocean. Eng.* **25**: 72–83 (Jan. 2000).
11. E. M. Sozer et al., Direct sequence spread spectrum based modem for under water acoustic communication and channel measurements, *Proc. OCEANS'99*, Nov. 1999.
12. Z. Zvonar, D. Brady, and J. A. Catipovic, Adaptive decentralized linear multiuser receiver for deep water acoustic telemetry, *J. Acoust. Soc. Am.* 2384–2387 (April 1997).
13. A. C. Chen, Overview of code division multiple access technology for wireless communications, *Proc. IECON'98*, 1998, Issue 24, pp. T15–T24.
14. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
15. L. Kleinrock and F. A. Tobagi, Carrier sense multiple access for packet switched radio channels, *Proc. ICC'74*, June 1974, pp. 21B-1–21B-7.
16. L. Kleinrock and F. A. Tobagi, Packet switching in radio channels, Part I: Carrier sense multiple access modes and their throughput-delay characteristics, *IEEE Trans. Commun.* 1400–1416 (1975).
17. F. A. Tobagi and L. Kleinrock, Packet switching in radio channels, Part II: The hidden terminal problem in carrier sense multiple access and busy tone solution, *IEEE Trans. Commun.* 1417–1433 (1975).
18. H. Takagi and L. Kleinrock, Correction to "Throughput analysis for persistent CSMA systems," *IEEE Trans. Commun.* 243–245 (1987).
19. V. Bharghavan, A. Deers, S. Shenker, and L. Zhang, MACAW: A media access protocol for wireless LAN's, *Proc. ACMSIGCOMM*, Aug. 1994, pp. 212–225.
20. P. Karn, MACA — a new channel access method for packet radio, *Proc. ARRL/CRRL Amateur Radio 9th Computer Network Conf.*, Sept. 1990.
21. J. Deng and Z. J. Haas, Dual busy tone multiple access (DBTMA): A new medium access control for packet radio networks, *Proc. IEEE 49th Vehicular Technology Conf.*, Houston, TX, May 1998, pp. 973–977.
22. D. B. Johnson, Routing in ad hoc networks of mobile hosts, *Proc. Workshop on Mobile Computing and Applications*, Dec. 1994, pp. 159–163.
23. J. Broch et al., A performance comparison of multi-hop wireless ad hoc network routing protocols, *Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking*, Oct. 1998.
24. C. E. Perkins and P. Bhagwat, Highly dynamic destination sequence distance vector routing (DSDV) for mobile computers, *Proc. SIGCOMM'94*, Aug. 1994, pp. 234–244.
25. V. D. Park and M. S. Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, *Proc. INFOCOM'97*, April 1997, pp. 1405–1413.
26. D. B. Johnson and D. A. Maltz, Protocols for adaptive wireless and mobile networking, *IEEE Pers. Commun.* (Feb. 1996).
27. C. E. Perkins, *Ad Hoc on Demand Distance Vector* (*AODV*) *Routing*, Internet-Draft, *draft-ietf-manet-aodv-00.txt*, Nov. 1997.
28. J. H. Kim et al., Experiments in remote monitoring and control of autonomous underwater vehicles, *Proc. OCEANS'96*, Fort Lauderdale, FL, Sept. 1996, pp. 411–416.
29. J. L. Talavage, T. E. Thiel, and D. Brady, An efficient store-and-forward protocol for a shallow-water acoustic local area network, *Proc. OCEANS'94*, Brest, France, Sept. 1994, pp. I883–I888.
30. S. M. Smith and J. C. Park, A peer-to-peer communication protocol for underwater acoustic communication, *Proc. OCEANS'97*, Oct. 97, pp. 268–272.
31. E. M. Sozer, M. Stojanovic, and J. G. Proakis, Initialization and routing optimization for ad hoc underwater acoustic networks, *Proc. Opnetwork'00*, Washington, DC, Aug. 2000.
32. J. A. Rice et al., Seaweb underwater acoustic nets, *SSC San Diego Biennial Review*, Aug. 2001.
33. E. Jahn, M. Hatch, and J. Kaina, Fusion of multi-sensor information from an autonomous undersea distributed field of sensors, *Proc. FUSION'99 Conf.*, Sunnyvale, CA, July 1999.
34. S. McGirr, K. Raysin, C. Ivancic, and C. Alspaugh, Simulation of underwater sensor networks, *Proc. IEEE OCEANS'99 Conf.*, Seattle, WA, Sept. 1999.
35. T. B. Curtin, J. G. Bellingham, J. Catipovic, and D. Webb, Autonomous oceanographic sampling networks, *Oceanography* **6**: 86–94 (1993).
36. J. G. Proakis, M. Stojanovic, and J. A. Rice, Design of a communication network for shallow-water acoustic modems, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1150–1159.
37. V. K. McDonald, J. A. Rice, M. B. Porter, and P. A. Baxley, Performance measurements of a diverse collection of undersea acoustic communication signals, *Proc. IEEE OCEANS'99 Conf.*, Seattle, WA, Sept. 1999.
38. D. L. Codiga, J. A. Rice, and P. S. Bogden, Real-time delivery of subsurface coastal circulation measurements from distributed instruments using networked acoustic modems, *Proc. IEEE OCEANS 2000 Conf.*, Providence, RI, Sept. 2000.
39. S. Merriam and D. Porta, DSP-based acoustic telemetry modems, *Sea Technol.* (May 1993).
40. D. Porta, DSP-based acoustic data telemetry, *Sea Technol.* (Feb. 1996).
41. J. A. Rice and K. E. Rogers, Directions in littoral undersea wireless telemetry, *Proc. TTCP Sympo. Shallow-Water Undersea Warfare*, Halifax, NS, Canada, Oct. 1996, Vol. 1, pp. 161–172.
42. M. D. Green, J. A. Rice, and S. Merriam, Underwater acoustic modem configured for use in a local area network, *Proc. IEEE OCEANS'98 Conf.*, Nice, France, Sept. 1998, Vol. 2, pp. 634–638.
43. M. D. Green, J. A. Rice, and S. Merriam, Implementing an undersea wireless network using COTS acoustic modems, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1027–1031.
44. M. D. Green, New innovations in underwater acoustic communications, *Proc. Oceanology International*, Brighton, UK, March 2000.
45. K. F. Scussel, J. A. Rice, and S. Merriam, A new MFSK acoustic modem for operation in adverse underwater channels, *paper presented at the OCEANS'97*, Halifax, NS, Canada, 1997.
46. N. Fruehauf and J. A. Rice, System design aspects of a steerable directional acoustic communications transducer for autonomous undersea systems, *Proc. OCEANS 2000 Conf.*, Providence, RI, Sept. 2000.

47. A. L. Butler, J. L. Butler, W. L. Dalton, and J. A. Rice, Multimode directional telesonar transducer, *Proc. IEEE OCEANS 2000 Conf.*, Providence, RI, Sept. 2000.

48. J. A. Rice, Acoustic signal dispersion and distortion by shallow undersea transmission channels, *Proc. NATO SACLANT Undersea Research Centre Conf. High-Frequency Acoustics in Shallow Water*, Lerici, Italy, July 1997, pp. 435–442.

49. J. A. Rice and R. C. Shockley, Battery-energy estimates for telesonar modems in a notional undersea network, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1007–1015.

50. M. Stojanovic, J. G. Proakis, J. A. Rice, and M. D. Green, Spread-spectrum methods for underwater acoustic communications, *Proc. IEEE OCEANS'98 Conf.*, Nice, France, Sept. 1998, Vol. 2, pp. 650–654.

51. V. K. McDonald and J. A. Rice, Telesonar testbed advances in undersea wireless communications, *Sea Technol.* **40**(2): 17–23 (Feb. 1999).

52. P. A. Baxley, H. P. Bucker, and J. A. Rice, Shallow-water acoustic communications channel modeling using three-dimensional Gaussian beams, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1022–1026.

53. M. B. Porter, V. K. McDonald, J. A. Rice, and P. A. Baxley, Relating the channel to acoustic modem performance, *Proc. European Conf. Underwater Acoustics*, Lyons, France, July 2000.

54. J. A. Rice and M. D. Green, Adaptive modulation for undersea acoustic modems, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 850–855.

55. J. A. Rice, V. K. McDonald, M. D. Green, and D. Porta, Adaptive modulation for undersea acoustic telemetry, *Sea Technol.* **40**(5): 29–36 (May 1999).

56. J. A. Rice, Telesonar signaling and seaweb underwater wireless networks, *Proc. NATO Sympo. New Information Processing Techniques for Military Systems*, Istanbul, Turkey, Oct. 9–11 2000.

57. C. L. Fletcher, J. A. Rice, R. K. Creber, and D. L. Codiga, Undersea acoustic network operations through a database-oriented server client interface, *Proc. IEEE OCEANS 2001 Conf.*, Waikiki, HI, Nov. 2001.

SHELL MAPPING

HENRY K. KWOK
Cisco Systems, Inc.
San Jose, California

DOUGLAS L. JONES
University of Illinois at Urbana — Champaign
Urbana, Illinois

1. CONSTELLATION SHAPING

Constellation shaping [1,2] is a technique that improves the efficiency of high-rate digital communications by reducing the average power without compromising the data rate or bit error rate. It can provide moderate gain (called *shaping gain*) of up to 1.53 dB on top of any coding gain (such as by using convolutional codes) with modest complexity. For instance, it is used in the V.34 high-rate analog voiceband modem standard to reduce the average power of the QAM constellation.

The concept and purpose of shaping is easily illustrated by the following simple example. Consider two 256-QAM constellations (see Figs. 1 and 2). Both constellations have the same number of distinct points, allowing 8 bits to be transmitted with each symbol. The error rate is primarily a function of the distance between constellation points, so it is virtually identical for both constellations. However, the average energy is reduced from 170 to 162.75 by choosing

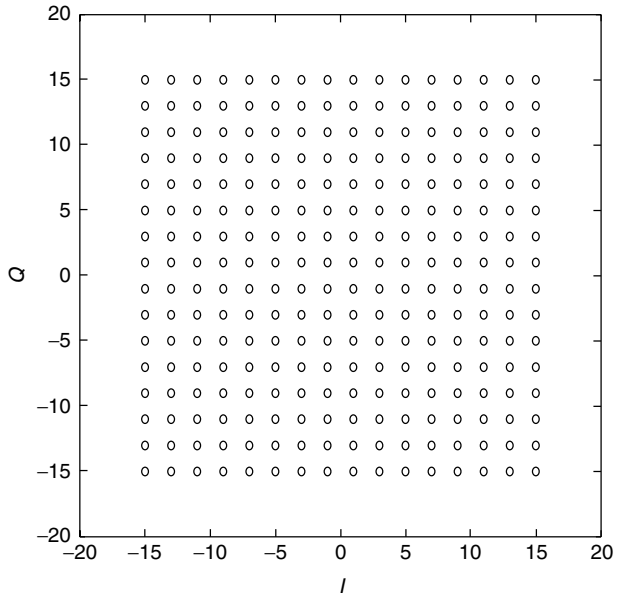


Figure 1. The original (unshaped) 256-QAM constellation has an average energy of 170.

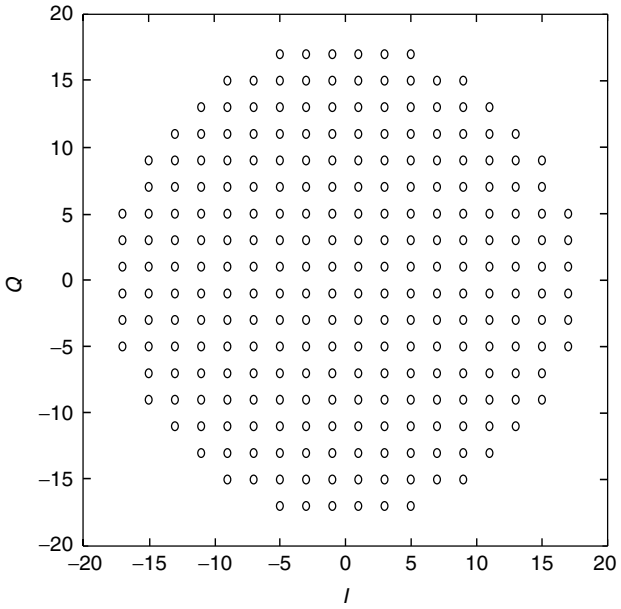


Figure 2. The shaped 256-QAM constellation has an average energy of 162.75 — a reduction of 0.379 dB.

a circular boundary instead of a square one. The reduction in average power is 0.379 dB without compromising the data rate or bit error rate.

Higher-dimensional constellations can provide even more average power reduction (up to 1.53 dB) [1–3]. For example, V.34 modems use a 16-dimensional constellation that provides about 1 dB of average power reduction. The idea is essentially the same as for two dimensions; conceptually, a block of N successive complex-valued symbols are grouped into a single $2N$-dimensional "hypersymbol," and a "hyperconstellation" in this space encodes a large number of bits simultaneously. Use of an approximately hyperspherical constellation with the same distance between constellation points reduces the average energy while preserving the same noise margin.

The primary practical challenge of shaping is to efficiently index the very large number of points in a high-dimensional, shaped constellation. Beyond two dimensions, simple tabulation becomes infeasible. Shell mapping is an efficient algebraic technique for performing this mapping that allows shaping to be used in practical communication systems. Shell mapping groups constellation points of roughly equal energy from a quadrature (QAM) constellation into concentric rings, or shells. A block of successive QAM symbols jointly code a large number of bits, which are divided into two groups: those mapped to the particular constellation points within the chosen shell for each QAM symbol, and another set of bits that indicates the combination of shells used for that block. For example, a system encoding an average rate of 6 bits per QAM symbol might use a block of eight symbols with 16 constellation points per shell; in this case, $8 \times 4 = 32$ bits would be encoded in the specific constellation point used within the shells selected for the eight successive QAM symbols, and the remaining 16 bits map to the 2^{16}th lowest-energy combinations of shells for the block. The total rate per eight symbols is thus $32 + 16 = 48$, for an average rate of 6 bits per QAM symbol. Shell mapping is a mathematical algorithm that efficiently identifies and indexes the lowest-energy sets of shells in a block of several successive QAM symbols using a technique based on combinatorics, generating functions, and finite-field algebra.

2. SHELL MAPPING

Shell mapping provides an efficient way to index lattice points inside a regular solid (such as a hypersphere or a hyperdiamond). It was first used in reducing the average power of QAM-based modems, in which it divides a QAM constellation into concentric rings with equal areas. A certain number of input bits are used to select the rings from which the constellation points will be chosen. A method due to Laroia [3] is used in the V.34 modems. We present a general version of this algorithm from a combinatorial viewpoint.

We use superscript $^{(N)}$ to indicate the dimension of a particular quantity. (The superscript may be suppressed if the dimension is 1.) Bold face represents a vector or a collection of lower-dimensional quantities. A subscript is used to differentiate different quantities of the same dimensionality. For example, the ith ring index is denoted as s_i [or $s_i^{(1)}$], and the N-element combination is denoted as $\mathbf{s}^{(N)} = (s_1, s_2, \ldots, s_N)$.

2.1. Generating Function

In combinatorics [4,5], a generating function is defined as a formal series

$$g(x) = a_0 + a_1 x + \cdots + a_N x^N \tag{1}$$

Generating functions are often used to solve problems involving a collection of objects with costs. We illustrate the method of generating functions with an example.

Example 1. How many ways are there to add three integers from the set of $\{0, 1, 2, 3\}$ such that the sum equals C? To solve this via the generating function, we first define

$$g^{(1)}(x) = 1 + x + x^2 + x^3 \tag{2}$$

The power of each coefficient represents the value of the first number. Then, $g^{(3)}(x) = [g^{(1)}(x)]^3 = 1 + 3x + 6x^2 + 10x^3 + 12x^4 + 12x^5 + 10x^6 + 6x^7 + 3x^8 + x^9$ represents the generating function of the three-integer sum. There is one three-integer combination that adds up to 0 (or 9), three combinations that add up to 1 (or 8), and so on. Table 1 lists the combinations that result in those costs and the corresponding terms in the generating function.

The exponent of the terms indicates the *cost* of selecting a certain object, and the coefficients of the terms denote the *number* of combinations that achieve a particular cost. To compute the N-element generating function, one simply multiplies the generating polynomial N times and locates the coefficient for the term representing the desired cost.

Table 1. Each Term in the Generating Function and Its Associated Inputs and Outputs

Terms in $g^{(3)}(x)$	Combinations
1	(0, 0, 0)
$3x$	(1, 0, 0), (0, 1, 0), (0, 0, 1)
$6x^2$	(1, 1, 0), (0, 1, 1), (1, 0, 1), (2, 0, 0), (0, 2, 0), (0, 0, 2)
$10x^3$	(1, 1, 1), (2, 1, 0), (2, 0, 1), (0, 2, 1), (1, 2, 0), (0, 1, 2), (1, 0, 2), (3, 0, 0), (0, 3, 0), (0, 0, 3)
$12x^4$	(0, 1, 3), (0, 3, 1), (3, 0, 1), (1, 1, 2), (1, 2, 1), (2, 1, 1), (1, 0, 3), (1, 3, 0), (3, 1, 0), (2, 2, 0), (2, 0, 2), (0, 2, 2)
$12x^5$	(3, 2, 0), (3, 0, 2), (0, 3, 2), (2, 2, 1), (2, 1, 2), (1, 2, 2), (2, 3, 0), (2, 0, 3), (0, 2, 3), (1, 1, 3), (1, 3, 1), (3, 1, 1)
$10x^6$	(2, 2, 2), (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1), (3, 3, 0), (3, 0, 3), (0, 3, 3)
$6x^7$	(3, 3, 1), (3, 1, 3), (1, 3, 3), (3, 2, 2), (2, 3, 2), (2, 2, 3)
$3x^8$	(3, 3, 2), (3, 2, 3), (2, 3, 3)
x^9	(3, 3, 3)

2.2. Indexing Algorithm for Sets with Integral Summable Cost

The method of generating functions has been used to solve numerous counting problems. However, for this article, we focus on a specific class of problems.

Let S be a set of objects. Let each element $s \in S$ have an integral cost of $|s|$. Denote the cost of an N-element combination, $\mathbf{s}^{(N)} = (s_1 \cdots s_N) \in S^N$, as

$$|\mathbf{s}^{(N)}| = \sum_{i=1}^{N} |s_i| \tag{3}$$

It is important that the cost of an element take on integral values and that the cost of an N-element combination be the sum of the cost of individual elements. Equation (3) has the same form as the l_1-norm and often leads to geometrical interpretation.

Example 2. In the previous example, $S = \{0, 1, 2, 3\}$ and $|s| = s$. Therefore

$$|\mathbf{s}^{(N)}| = \sum_{i=1}^{N} |s_i| = \|\mathbf{s}^{(N)}\|_1 \tag{4}$$

Graphically, $\|\mathbf{s}^{(N)}\|_1 = c$ depicts a hyperplane in the all-positive quadrant.[1] We denote

$$S_c^{(N)} = \{\mathbf{s}^{(N)} \in S^N : |\mathbf{s}^{(N)}| = c\} \tag{5}$$

We illustrate the first four hyperplanes, $S_0^{(3)}$, $S_1^{(3)}$, $S_2^{(3)}$, and $S_3^{(3)}$, with the lattice points that belong to each hyperplane shown in Fig. 3. The lattice points on $S_c^{(3)}$ represent three-integer combinations that sum up to c. Each layer corresponds to a term in $g^{(3)}(x)$.

We assume that it is desirable to select N-element combinations within a certain range of costs. From this point on, we assume (without loss of generality) that the cost is nonnegative. This assumption implies that the generating functions will consist only of terms x^n where $n \in \{0\} \cup \mathbb{Z}^+$.

[1] Strictly speaking, the set defined is not a hyperplane, but rather a set of lattice points that lie on a common hyperplane. Throughout this article; however, we loosely use the term "hyperplane" to denote a set of lattice points on a common hyperplane.

2.3. Counting the Sets

First, we count the number of combinations that satisfy the cost constraint. To that end, we define the one-element generating function

$$g^{(1)}(x) = \sum_{i=\min_{s\in S}|s|}^{\max_{s\in S}|s|} a_i x^i \tag{6}$$

where a_i represents the number of elements in S that have the cost of i. Next, we compute the N-element generating function

$$g^{(N)}(x) = [g^{(1)}(x)]^N = \sum_i a_i^{(N)} x^i \tag{7}$$

where $a_i^{(N)}$ represents the number of N-element combinations in S^N that have the cost of i. To seek the number of combinations within a cost range, we can sum the coefficients over the range $C_l \le |\mathbf{s}^{(N)}| \le C_h$ as

$$K = \sum_{i=C_l}^{C_h} a_i^{(N)} \tag{8}$$

2.4. Mapping Algorithm

A mapping procedure from an input value to an N-element combination is needed in an actual communication system. Suppose that there are K N-element combinations that satisfy the cost constraint. Let R be an input integer from the set $I^{(N)} = \{i \in \mathbb{Z} : 0 \le i < K\}$, and let the set of the K N-element combinations be $S^{(N)} = \{\mathbf{s} \in S^N : C_l \le |\mathbf{s}| \le C_h\}$. The goal is to find a bijective mapping, $\mathbf{f} : I^{(N)} \to S^{(N)}$. The basic approach is to decompose the N-dimensional problem into multiple lower-dimensional problems. Specifically, Laroia [3] provides a way to decompose the N-dimensional problem into two $\frac{N}{2}$-dimensional problems and another way to decompose the problem into a 1D (one-dimensional) problem and an $(N-1)$-dimensional problem. Arbitrary decompositions

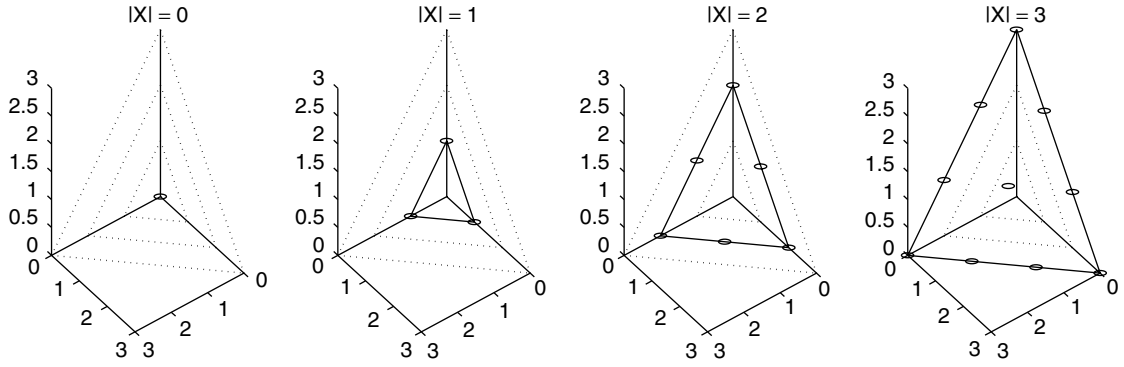


Figure 3. The layering of the generating function.

offering different trade-offs between computational cost and memory requirements can be found in Ref. 6.

2.5. Identifying the Cost of the Combination

First, we partition $S^{(N)}$ into a set of hyperplanes, $\{S_c^{(N)}\}_{c=C_l}^{C_h}$. Since these hyperplanes are parallel, they are also disjoint.[2] Thus, the $S^{(N)}$ set is partitioned into

$$S^{(N)} = S_{C_l}^{(N)} \cup S_{C_l+1}^{(N)} \cup \cdots \cup S_{C_h}^{(N)} \tag{9}$$

Next, we associate ranges of integers from $I^{(N)}$ to these N-dimensional hyperplanes. This is done by successively assigning the set of $a_c^{(N)}$ lowest available input integers in $I^{(N)}$ to the hyperplane of cost c, $S_c^{(N)}$. At the end of this process, we associate each input $R \in I^{(N)}$ to one of the $S_c^{(N)}$. If the input R is associated with $S_c^{(N)}$, we define the N-D *cost* of the input R as

$$C^{(N)} \triangleq |\mathbf{f}(R)| = c \tag{10}$$

This is called the N-dimensional cost because all N-element combinations $\mathbf{s}^{(N)}$ from this hyperplane have the cost $|\mathbf{s}^{(N)}|$ of $C^{(N)}$. Next, we partition the input set into

$$I_c^{(N)} = \{r \in I^{(N)} : \mathbf{f}(r) \in S_i^{(N)}\} \tag{11}$$

In other words, $I_c^{(N)}$ is the subset $I^{(N)}$ that is associated with $S_c^{(N)}$. Since each input is associated to only one hyperplane, the collection of $I_c^{(N)}$'s is disjoint, and we can write

$$I^{(N)} = \cup_c I_c^{(N)} \tag{12}$$

From the example above, it is easy to see that each $I_c^{(N)}$ consists of a range of consecutive integers. We define the N-dimensional *residue* as

$$R^{(N)} = R - \min I_{C^{(N)}}^{(N)} \tag{13}$$

Physically, $R^{(N)}$ uniquely indices a combination in $S_{C^{(N)}}^{(N)}$. (The uniqueness can be shown from the fact that $0 \le R^{(N)} < a_{C^{(N)}}^{(N)}$.) So far, we have created a mapping $(g_c, g_s) : I^{(N)} \to \mathbb{Z} \times \mathbb{Z}$ that maps an input to an N-dimensional cost and an N-dimensional residue. The cost selects the hyperplane $S_{C^{(N)}}^{(N)}$, and the residue points to an N-element combination in $S_{C^{(N)}}^{(N)}$.

Example 3. From Example 2, we have $N = 3$, $S = \{0, 1, 2, 3\}$, $|s| = s$, $C_l = 0$, and $C_h = 3$. Also, let the input $R = 14$. We have

$$K = \sum_{i=C_l}^{C_h} a_i^{(N)} = \sum_{i=0}^{3} a_i^{(3)} \tag{14}$$

$$= \begin{array}{ccccccc} 1 & + & 3 & + & 6 & + & 10 \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ x^0 & & x^1 & & x^2 & & x^3 \end{array} \tag{15}$$

$$= 20 \tag{16}$$

[2] Again, what we mean here is that since the two sets lie on two parallel hyperplanes, there is no element that belongs to both sets.

We associate the range of input to the subset of outputs as

$$\begin{aligned} I_0^{(3)} &= \{0\} \to S_0^{(3)} \\ I_1^{(3)} &= \{1, 2, 3\} \to S_1^{(3)} \\ I_2^{(3)} &= \{4, 5, \ldots, 9\} \to S_2^{(3)} \\ I_3^{(3)} &= \{10, 11, \ldots, 19\} \to S_3^{(3)} \end{aligned} \tag{17}$$

From (17), we find that $R = 14$ belongs to the range of 10–19. This means $C^{(3)} = 3$ and $R^{(3)} = 14 - 10 = 4$ [or we are now considering the fifth combination in the hyperplane of $|\mathbf{s}^{(3)}| = C^{(3)} = 3$; see Fig. 3].

2.6. Decomposition of the Problem

We decompose the N-dimensional problem into a P-dimensional problem and a Q-dimensional problem. The idea is to split $(C^{(N)}, R^{(N)})$ into $(C^{(P)}, R^{(P)})$, and $(C^{(Q)}, R^{(Q)})$. The obvious constraint is that $P + Q = N$.

Now, we perform the actual decomposition of the problem. We break up the N-dimensional hyperplane into the union of a collection of Cartesian product sets of P-dimensional and Q-dimensional hyperplanes, where $Q = N - P$.[3] For $\forall c \in \mathbb{Z}$, we can write

$$\begin{aligned} S_c^{(N)} &= (S_0^{(P)} \times S_c^{(Q)}) \cup (S_1^{(P)} \times S_{c-1}^{(Q)}) \cup \cdots \cup (S_c^{(P)} \times S_0^{(Q)}) \\ &= \bigcup_{i=0}^{c} (S_i^{(P)} \times S_{c-i}^{(Q)}) \end{aligned} \tag{18}$$

Since all product sets on the right side are disjoint, we can relate the size of these sets as follows:

$$\begin{aligned} |S_c^{(N)}| &= |S_0^{(P)}| \cdot |S_c^{(Q)}| + |S_1^{(P)}| \cdot |S_{c-1}^{(Q)}| + \cdots + |S_c^{(P)}| \cdot |S_0^{(Q)}| \\ &= \sum_{i=0}^{c} |S_i^{(P)}| \cdot |S_{c-i}^{(Q)}| \end{aligned} \tag{19}$$

We may also reach the following conclusion by noting that $a_i^{(n)} = |S_i^{(n)}|$ and

$$a_c^{(N)} = \sum_{i=0}^{c} a_i^{(P)} a_{c-i}^{(Q)} \tag{20}$$

In other words, (19) simply states the rule for evaluating the coefficients of $g^{(N)}(x)$ from the coefficients of $g^{(P)}(x)$ and $g^{(Q)}(x)$. Now, we partition the hyperplane $S_c^{(N)}$ into a collection of Cartesian product sets of lower-dimensional hyperplanes. Meanwhile, we partition the corresponding input range, $I_c^{(N)}$, similar to the way described above. The $|S_0^{(P)} \times S_c^{(Q)}| = |S_i^{(P)}| \cdot |S_{c-i}^{(Q)}|$ lowest available integers in $I_c^{(N)}$ are assigned to the Cartesian-product set, $S_i^{(P)} \times S_{c-i}^{(Q)}$. This range of input is denoted as $I_{i,c-i}^{(N)}$. Once again, we have

$$I_c^{(N)} = \bigcup_{i=0}^{c} I_{i,c-i}^{(N)} \tag{21}$$

[3] As mentioned earlier, usually $P = N - P = N/2$. However, for the sake of visual illustration, we choose N to be 3, which is not a power of 2. The shell mapper still operates properly; however, it is less efficient.

Once we decide that an input R is within $I^{(N)}_{i,C^{(N)}-i}$ and is associated with $S^{(P)}_i \times S^{(Q)}_{C^{(N)}-i}$, we have split the cost into

$$C^{(P)} = |\mathbf{s}^{(P)}| = i \tag{22}$$

$$C^{(Q)} = |s^{(Q)}| = C^{(N)} - i \tag{23}$$

In addition, we need to adjust the residue so that it now indexes a combination in $S^{(P)}_i \times S^{(Q)}_{C^{(N)}-i}$ instead of $S^{(N)}_{C^{(N)}}$. This is done by subtracting from $R^{(N)}$ the smallest value in $Z^{(N)}_{i,C^{(N)}-i}$. Now a residue of 0 indicates the first combination in $S^{(P)}_i \times S^{(Q)}_{c-i}$. We denote the new residue as

$$R^{(N)}_{i,c-i} = R^{(N)} - (\min I^{(N)}_{i,c-i} - \min I^{(N)}_c) = R - \min I^{(N)}_{i,c-i} \tag{24}$$

where $c = C^{(N)}$ in our specific case.

Example 4. Continuing with our previous example of $S = \{0, 1, 2, 3\}$, $N = 3$, $C_l = 0$, $C_h = 3$, $K = 20$, and $R = 14$, recall that $(C^{(3)}, R^{(3)}) = (3, 4)$. We now search for the Cartesian product set with which the input is associated. Let $P = 2$, and $Q = N - P = 1$. We associate the input range $I^{(3)}$ to the Cartesian product sets as

$$\begin{aligned} I^{(3)}_{0,3} &= \{10\} \rightarrow S^{(2)}_0 \times S^{(1)}_3 \\ I^{(3)}_{1,2} &= \{11, 12\} \rightarrow S^{(2)}_1 \times S^{(1)}_2 \\ I^{(3)}_{2,1} &= \{13, 14, 15\} \rightarrow S^{(2)}_2 \times S^{(1)}_1 \\ I^{(3)}_{3,0} &= \{16, 17, 18, 19\} \rightarrow S^{(2)}_3 \times S^{(1)}_0 \end{aligned} \tag{25}$$

We search in the following order: $S^{(2)}_0 \times S^{(1)}_3$, $S^{(2)}_1 \times S^{(1)}_2$, and then $S^{(2)}_2 \times S^{(1)}_1$. The search sequence is graphically depicted in Fig. 4. We find that $R = 14$ belongs to $I^{(3)}_{2,1}$. Thus, the two costs are $C^{(2)} = 2$, $C^{(1)}_3 = 1$.

In order to complete the decomposition into two lower-dimensional problems, we need two residues, $R^{(P)}$ and $R^{(Q)}$, for the two costs, $C^{(P)}$ and $C^{(Q)}$. Observe that there are several constraints. First, $R^{(P)}$ (or $R^{(Q)}$) is an index to a P-element (or Q-element) combination in $S^{(P)}_{C^{(P)}}$ (or $S^{(Q)}_{C^{(Q)}}$). Thus, $R^{(P)}$ and $R^{(Q)}$ must satisfy the following inequalities:

$$0 \le R^{(P)} < |S^{(P)}_{C^{(P)}}| \tag{26}$$

$$0 \le R^{(Q)} < |S^{(Q)}_{C^{(Q)}}| \tag{27}$$

In addition, the mapping of $|S^{(P)}_{C^{(P)}}| \cdot |S^{(Q)}_{C^{(Q)}}|$ values of $R^{(N)}_{C^{(P)},C^{(Q)}}$ to $(R^{(P)}, R^{(Q)})$ must be bijective. One such mapping is

$$\mathbf{g}_{|S^{(P)}_{C^{(P)}}|}(R^{(N)}) = \left(R^{(N)}_{C^{(P)},C^{(Q)}} \bmod |S^{(P)}_{C^{(P)}}|, \left\lfloor \frac{R^{(N)}_{C^{(P)},C^{(Q)}}}{|S^{(P)}_{C^{(P)}}|} \right\rfloor \right) \tag{28}$$

This is the same as arranging the $|S^{(P)}_{C^{(P)}}| \cdot |S^{(Q)}_{C^{(Q)}}|$ values of $R^{(N)}$ into an $|S^{(P)}_{C^{(P)}}|$-column $\times$ $|S^{(Q)}_{C^{(Q)}}|$-row array. Then, the column and row index of each array element are taken as $R^{(P)}_1$ and $R^{(Q)}$, respectively. Now, we have transformed $(C^{(N)}, R^{(N)})$ into $(C^{(P)}, R^{(P)})$ and $(C^{(Q)}, C^{(Q)})$. We may recursively apply the procedure to each cost–residue pair. Eventually, the dimension of the cost (and residue) will become 1. If $a^{(1)}_{C^{(1)}} = 1$, we are guaranteed that $R^{(1)} = 0$, signifying the first (and only) element in S with the cost $C^{(1)}$. If $a^{(1)}_{C^{(1)}} > 1$, it is possible that $R^{(1)} > 0$. In that case, we must resort to a lookup table to find the specific element in S.

Example 5. To conclude this section, we finish the shell mapping of the input $R = 14$. Note that $R = 14$ is the second combination ($R^{(3)}_{C^{(2)},C^{(1)}} = 1$) in $S^{(2)}_{C^{(2)}} \times S^{(1)}_{C^{(1)}}$, so the lower-dimensional residues should be

$$R^{(2)} = 1 \bmod 3 = 1 \tag{29}$$

$$R^{(1)} = \lfloor \tfrac{1}{3} \rfloor = 0 \tag{30}$$

We perform the same decomposition procedure on the cost-residue pair of $(C^{(2)}, R^{(2)}) = (2, 2)$. We know that $S^{(2)}_{C^{(2)}} = S^{(2)}_2 = (S^{(1)}_0 \times S^{(1)}_2) \cup (S^{(1)}_1 \times S^{(1)}_1) \cup (S^{(1)}_2 \times S^{(1)}_0)$. Since $S^{(1)}_i$ is a simple point in one dimension, $S^{(1)}_i \times S^{(1)}_j$ is a 2D point. We can also derive this fact from $|S^{(1)}_i \times S^{(1)}_j| = |S^{(1)}_i| \cdot |S^{(1)}_j| = 1 \cdot 1 = 1$. In addition, we know from the generating function $g^{(2)}(x)$ that $|S^{(2)}_{C^{(2)}}| = 3$. On the other hand, from Eq. (19), we have

$$\begin{aligned} |S^{(2)}_{C^{(2)}}| &= |S^{(1)}_0 \times S^{(1)}_2| + |S^{(1)}_1 \times S^{(1)}_1| + |S^{(1)}_2 \times S^{(1)}_0| \\ &= |S^{(1)}_0| \cdot |S^{(1)}_2| + |S^{(1)}_1| \cdot |S^{(1)}_1| + |S^{(1)}_2| \cdot |S^{(1)}_0| \\ &= 1 + 1 + 1 = 3 \end{aligned} \tag{31}$$

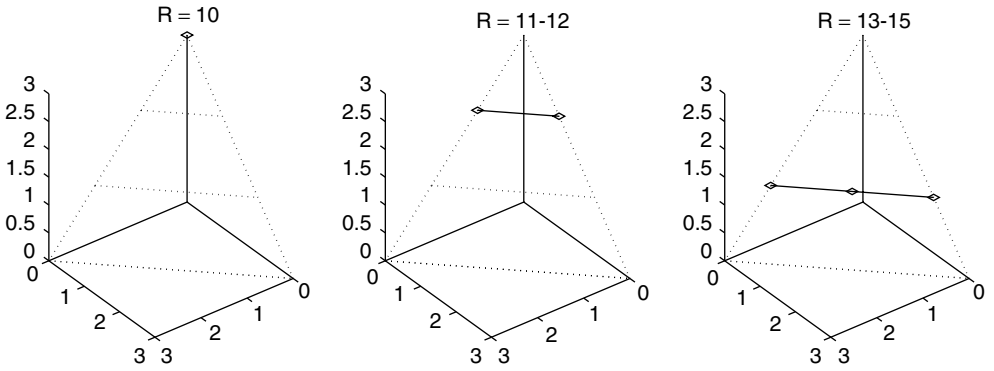


Figure 4. Decomposition of $S^{(3)}_3$ into a collection of Cartesian-product sets.

The corresponding input partitioning is

$$\begin{aligned} I_{0,2}^{(2)} &= \{13\} \rightarrow S_0^{(1)} \times S_2^{(1)} \\ I_{1,1}^{(2)} &= \{14\} \rightarrow S_1^{(1)} \times S_1^{(1)} \\ I_{2,0}^{(2)} &= \{15\} \rightarrow S_2^{(1)} \times S_0^{(1)} \end{aligned} \tag{32}$$

Thus, we have three Cartesian product sets, each with only one element. Since we are seeking the second combination ($R^{(2)} = 1$), we split the 2D cost $C^{(2)}$ into

$$C_1^{(1)} = 1 \tag{33}$$

$$C_2^{(1)} = 2 - 1 = 1 \tag{34}$$

The residue $R^{(2)}$ is reduced to

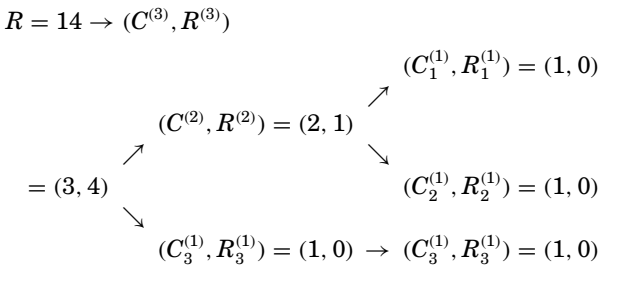
$$\begin{aligned} R_{1,1}^{(2)} &= R^{(2)} - (\min I_{1,1}^{(2)} - \min I_2^{(2)}) \\ &= 1 - (14 - 13) = 0 \end{aligned} \tag{35}$$

$$R_1^{(1)} = R_{1,1}^{(2)} \bmod |S_1^{(1)}| = 0 \bmod 1 = 0 \tag{36}$$

$$R_2^{(1)} = \left\lfloor \frac{R_{1,1}^{(2)}}{|S_1^{(1)}|} \right\rfloor = \left\lfloor \frac{0}{1} \right\rfloor = 0 \tag{37}$$

The final cost-residue pairs are $(C_1^{(1)}, R_1^{(1)}) = (1, 0)$, $(C_2^{(1)}, R_2^{(1)}) = (1, 0)$, and $(C_3^{(1)}, R_3^{(3)}) = (1, 0)$. We may now find the elements that correspond to each cost-residue pair. For example, we need to find the first combination with the cost of 1 for $(C_1^{(1)}, R_1^{(1)})$. Coincidentally, the elements have the same value as the costs. Now, the three 1D costs form the final combination $\mathbf{s}^{(3)} = (C_1^{(1)}, C_2^{(1)}, C_3^{(1)}) = (1, 1, 1)$. The answer is verified graphically in Fig. 5.

The decomposition results in a tree diagram as



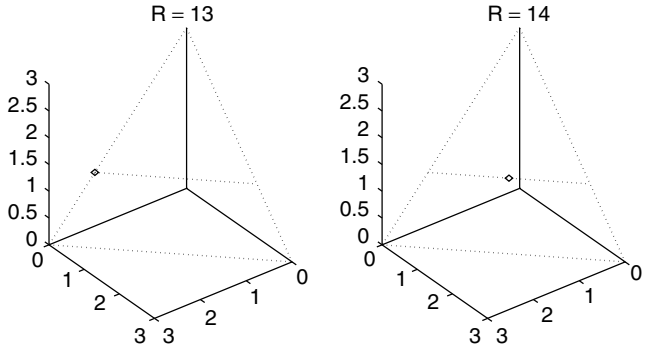


Figure 5. Decomposition of $S_2^{(2)}$ into individual lattice points. The final answer is $s^{(3)} = (1, 1, 1)$.

Note that at each column all costs sum up to 3 ($3 = 2 + 1 = 1 + 1 + 1$). Algebraically, this is equivalent to the decomposition of $g^{(3)}(x) \rightarrow g^{(2)}(x) \cdot g^{(1)}(x) \rightarrow g^{(1)}(x) \cdot g^{(1)}(x) \cdot g^{(1)}(x)$.

2.7. Demapping Algorithm

The demapping operation is the inverse operation of the mapping process. We solve for the inverse of each step and reverse the order of the steps. It is simpler than the mapping process in the sense that there is no need to search for the proper-cost decomposition.

First, given a vector $\mathbf{s}^{(N)} = (s_1 \cdots s_N)$, we find (usually via lookup tables) $\{(C_i^{(1)}, R_i^{(1)})\}_{i=1}^{N}$ such that s_i is the $(R_i^{(1)} + 1)$-th element in S with a cost of $C_i^{(1)}$. Next, for each pair of cost–residue pairs, $(C^{(P)}, R^{(P)})$ and $(C^{(Q)}, R^{(Q)})$, we combine the costs by

$$C^{(P+Q)} = C^{(P)} + C^{(Q)} \tag{38}$$

and the residues by

$$R_{C^{(P)}, C^{(Q)}}^{(P+Q)} = |S_{C^{(P)}}^{(P)}| R^{(Q)} + R^{(P)} \tag{39}$$

$$R^{(P+Q)} = R_{C^{(P)}, C^{(Q)}}^{(P+Q)} + (\min I_{C^{(P)}, C^{(Q)}}^{(P+Q)} - \min I_{C^{(P+Q)}}^{(P+Q)}) \tag{40}$$

We repeatedly apply this procedure until we combine all cost–residue pairs to one pair: $(C^{(N)}, R^{(N)})$. Then, the original index can be recovered by

$$R = \sum_{i=0}^{C^{(N)}} g^{(N)}(i) + R^{(N)} \tag{41}$$

3. APPLICATION AND PRACTICAL CONSIDERATION

Shell mapping can be applied in situations in which the cost (i.e., energy in the case of constellation shaping) per ring is proportional to the ring index and the total cost is the sum of the individual ring costs. This is nearly exact for large-QAM constellations subdivided into many rings. The shaping gain from shell mapping tends to improve as the number of rings increases, thereby reducing the number of symbols in each ring. However, the complexity and cost of shell mapping also increases with the number of rings, so practical systems make a judicious trade-off between the accuracy of the approximation and the number of rings. For example, the V.34 modem standard uses 8 rings of QAM symbols to create a 16-dimensional-shaped constellation. Each QAM symbol is divided into as many as 16 rings from a constellation as large as 1664-QAM.

The maximal shaping gain is obtained asymptotically as the number of dimensions grows arbitrarily large. However, the complexity of the shell mapping algorithm also grows with the dimension, so practical systems select a moderate block size. For example, the V.34 modem uses 16-dimensional blocks; larger block sizes provide incremental performance benefits at the cost of rapidly increasing complexity.

Shell mapping reduces the average energy of a block of QAM symbols by using only the 2^b combinations of

rings with the lowest total energies over the block. For instance, it would not use the high-energy constellation points for every QAM symbol in a single block. To provide the freedom to reject high-energy combinations of symbols, at least a modest overdeterminacy in the individual QAM constellations is required. For example, if a data rate of 8 bits per QAM symbol is desired, a shell-mapped QAM constellation of more than 256 points is required to support shaping. The larger constellation increases both complexity and the peak-to-average power ratio, however, so some compromise between expansion and reduced shaping gain is generally made. Factors of 1.2–1.5 times the base size are usually sufficient to realize most of the available shaping gain.

The ITU V.34 modem standard includes shell mapping. It uses eight successive QAM symbols to create a 16-dimensional-shaped constellation. Each QAM symbol is divided into as many as 16 rings from a base constellation as large as 1664-QAM. The incoming bit stream is segmented into blocks of bits. Within each block, a portion are used for the trellis coder. The output selects one of four QAM constellations. Some bits are sent to the shell mapper, which yields eight ring indices. The remaining bits are used to select the actual constellation points within each ring.

BIOGRAPHY

Douglas L. Jones received th B.S.E.E., M.S.E.E., and Ph.D. degrees from Rice University in 1983, 1985, and 1987, respectively. During the 1987–1988 academic year, he was at the University of Erlangen — Nuremberg in Germany on a Fulbright postdoctoral fellowship. Since 1988, he has been with the University of Illinois at Urbana — Champaign, where he is currently a Professor in Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute. In the Spring semester of 1999 he served as the Texas Instruments Visiting Professor at Rice University. He is an author of the laboratory textbook *A Digital Signal Processing Laboratory Using the TMS32010*, over 150 conference and journal publications, and several patents. His research interests are in digital signal processing and communications, including nonstationary signal analysis, adaptive processing, multisensor data processing, OFDM, and various applications such as advanced hearing aids.

BIBLIOGRAPHY

1. G. D. Forney and L. Wei, Multidimensional constellations — Part I: Introduction, figures of merit, and generalized cross constellation, *IEEE J. Select. Areas Commun.* **7**: 877–892 (1989).
2. G. D. Forney and L. Wei, Multidimensional constellations — Part II: Voronoi constellations, *IEEE J. Select. Areas Commun.* **7**: 941–958 (1989).
3. R. Laroia, On optimal shaping of multidimensional constellations, *IEEE Trans. Inform. Theory* **40**: 1044–1056 (1994).
4. P. J. Cameron, *Combinatorics: Topics, Techniques, Algorithms*, Cambridge Univ. Press, Cambridge, UK, 1994.
5. R. A. Brualdi, *Introductory Combinatorics*, Prentice-Hall, Englewood Cliffs, NJ, 1994.
6. H. K. Kwok, *Shape Up: Peak-Power Reduction via Constellation Shaping*. Ph.D. thesis, Univ. Illinois at Urbana — Champaign, 2000.

SIGMA-DELTA CONVERTERS IN COMMUNICATION SYSTEMS

FRED HARRIS
San Diego State University
San Diego, California

1. THE WHY OF SIGMA-DELTA CONVERTERS

Correlation between a sequence of sample values can be used to quantize, to a specified fidelity, a signal with significantly fewer bits per sample than that required by an instantaneous quantizer. The sigma-delta (Σ-Δ) converter uses feedback to shape the quantizing noise spectrum of an oversampled low-resolution quantizer to obtain low levels of in-band noise in exchange for higher levels of out-of-band-noise. In effect, the quantizer arranges for the quantizing noise spectrum and input signal spectrum to occupy nearly distinct spectral regions. Filtering that rejects the out-of-band-shaped quantizing noise converts the signal correlation to additional bits by the ratio of coherent gain to incoherent gain of the filtering process.

A major application of the sigma-delta process is in the area of analog-to-digital conversion, particularly in the conversion of audio signals, of instrumentation signals, and modem input signals. High-performance converters operating with 1-bit quantizers at an input rate 64 times the desired output rate of 100 kHz can supply 24-bit samples with 120 dB SNR (signal-to-noise ratio). The second major application of the Σ-Δ process has been in the area of digital-to-analog conversion for audio signals, control signals, and modem output signals. It is standard practice to use multirate digital filters to raise the sample rate of a 16-bit sampled data signal by a factor of 64, say, from 48 kHz to 3.072 MHz, then convert the 16-bit data samples to 1-bit data samples with a digital Σ-Δ converter, which is then presented to a 1-bit DAC for conversion to an analog signal.

The enhanced analog-to-digital conversion application uses analog sampled data components in the Σ-Δ modulator. These are implemented with switched capacitor or continuous-time integrators, a pair of low-resolution A-to-D (A/D) and D-to-A (D/A) converters in the feedforward–feedback path, and digital filters to reduce noise bandwidth and signal sample rate. By way of comparison, the enhanced D/A conversion application uses standard digital signal processing blocks in its Σ-Δ modulator. It also uses standard DSP blocks to raise the sample rate, a single low-resolution D/A converter to form an analog signal, and analog filters to reduce the noise bandwidth.

The third major application of the Σ-Δ process is in the area of digital-to-digital (D/D) conversion. In this

application, the input signal, collected and quantized with a conventional A/D, is pre-processed with a digital Σ-Δ to reduce the number of bits required to represent the signal in the bandwidth to be extracted by subsequent processing. The Σ-Δ modulator shapes the quantizing noise spectrum to preserve low noise levels over the frequency span of the passband while permitting significant increase of noise level in the frequency span of the stopband of the subsequent digital filter. Processing resources, such as multiplier widths required to implement digital filters after the digital Σ-Δ modulator, are reduced considerably. Reduction in data bit width from 16 to 4 bits can convert filter multipliers to lookup tables for FPGA implementations, while reductions from 16 bits to 1 bit permits arbitrary FIR filtering with no multipliers at all.

Processing flow diagrams matching the three applications we have described are shown in Fig. 1. The material covered in the remainder of this presentation will be

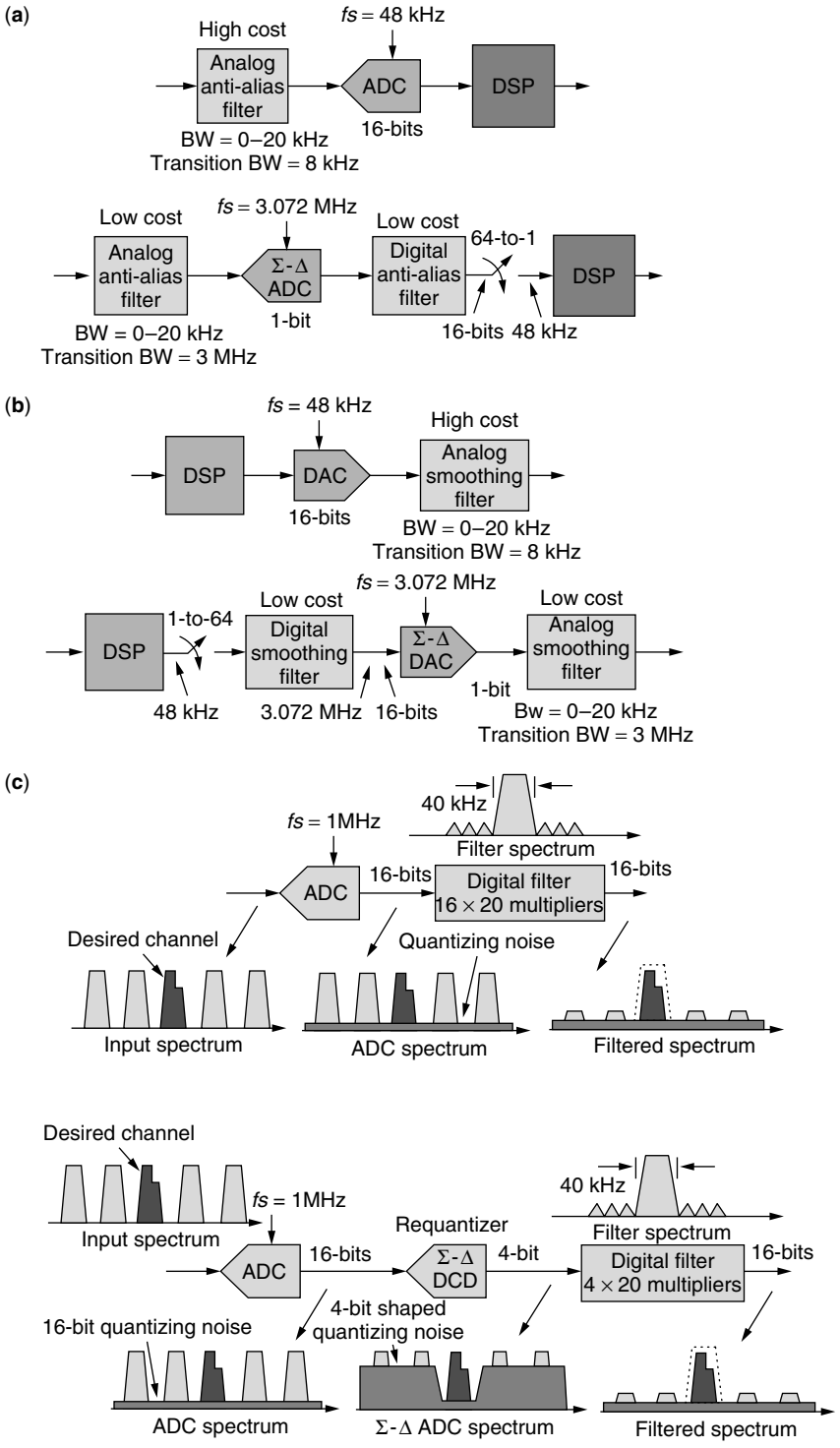


Figure 1. (**a**) A/D conversion with memoryless converter compared to oversampled sigma-delta A/D modulator with low-cost digital filter; (**b**) D/A conversion with memoryless converter compared to low-cost digital filter with oversampled sigma-delta D/A modulator; (**c**) digital signal processing of sampled data with full-bandwidth A/D conversion compared to processing with sigma-delta preprocessed bandwidth-limited data.

generic descriptions of the Σ-Δ process without regard to implementation. Applications presented near the end of this article will emphasize the third application, with use of the all-DSP digital-to-digital process in communication systems.

2. BACKGROUND

We start by reviewing the model of the ideal amplitude quantizer that performs instantaneous mapping from input amplitude x to output amplitude x_q in accord with a specified input–output profile $x_q = Q(x)$. As shown in Eq. (1) and in Fig. 2, the difference between the input and output of the quantizer represents an error, and to control the size of this error, the profile should be close to the errorless profile $x_q = x$, the unit slope line through the origin.

$$x_q = Q(x)$$
$$e_q = Q(x) - x = x_q - x \tag{1}$$

The quantization profile, reminiscent of a staircase, is defined by the locations of its treads and risers. Figure 2 shows the nonlinear mapping profile and the error profile of a uniform quantizer, one exhibiting equally spaced treads and risers. Such a quantizer is described as being a uniform or, paradoxically, a linear quantizer.

Note from the error profile that the quantization error is a deterministic and nonlinear function of the input signal level with a known error for a known input. A linear model of the quantizer is that of a zero-mean, independent, uniformly distributed, white-noise source added to each input sample. Both models are shown in Fig. 3. The linear model is often substituted for the nonlinear model for ease of analysis. In our subsequent discussions, we make this substitution with the awareness that this may a poor representation of the quantizing noise when the number of output levels is less than 16, or when the input samples are highly correlated. These are precisely the conditions we encounter in the Σ-Δ process. It is for this reason that simple linear analysis of a Σ-Δ system does not predict nonlinear modes of behavior, which we discuss in a later section.

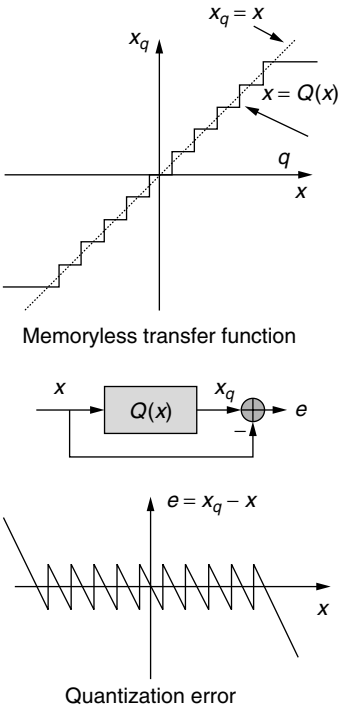


Figure 2. Quantization profile and error profile for uniform quantizer.

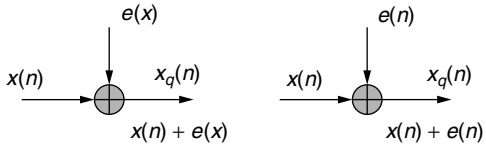


Figure 3. Nonlinear model and linear additive model of quantizer.

The noise in the linear additive noise model of the quantizer is uniformly distributed over a quantile interval of width q, from which we can compute the mean-square quantizing noise as shown in Eq. (2) to be $q^2/12$. We can similarly identify the mean-square level of the signal component at the quantizer output for any input amplitude distribution. The simplest such distribution is uniform over the range of $-Mq$ to $+Mq$, where M is the maximum number of positive and negative quantile increments. The mean-square signal level for this distribution is shown in Eq. (3) to be $(2Mq)^2/12$:

$$\sigma_q^2 = \int_{-q/2}^{+q/2} e^2 f_q(e)\, de = \frac{1}{q}\int_{-q/2}^{+q/2} e^2\, de = \frac{q^2}{12} \tag{2}$$

$$\sigma_s^2 = \int_{-Mq}^{+Mq} s^2 f_s(s)\, ds = \frac{1}{2Mq}\int_{-Mq}^{+Mq} s^2\, ds = \frac{(2M)^2 q^2}{12} \tag{3}$$

The signal : quantizing noise ratio of the quantizer is shown in Eq. (4) to be $(2M)^2$. Equation (4) also shows that when we replace $2M$, the number of levels in a quantizer with a power of 2 of the form 2^b, the SNR is 2^{2b}. Finally we see that the SNR in decibels is seen to be $6b$ dB, from which we obtain the standard quantizer rule of 6 dB per bit. The SNR of a quantizer, for any input amplitude distribution, is of the form shown in Eq. (5), where the offset factor $K_{\text{density}}(\sigma^2)$, is a parameter that varies with amplitude density and signal variance. K_{density} is negative for most densities and becomes more so as we decrease the input signal variance relative to quantizer dynamic range.

The primary message presented by Eq. (5) is that quantizer fidelity or SNR can be purchased with a linear quantizer by increasing the number of bits involved in the quantization process. The first rule of quantizing is: "If you want a higher fidelity representation of the signal, get more bits." In the next section we derive a corollary to this rule: "If you can't get more bits, get correlated samples and convert them to more bits."

$$\text{SNR} = \frac{\sigma_s^2}{\sigma_q^2} = \frac{(2M)^2 q^2/12}{q^2/12} = (2M)^2 = (2^b)^2 = 2^{2b}$$

$$\begin{aligned}\mathrm{SNR}_{\mathrm{dB}} &= 10\log_{10}(\mathrm{SNR}) = 10\log_{10}(2^{2b}) \qquad (4)\\ &= 20b\log_{10}(2) = 6b \text{ dB}\end{aligned}$$

$$\mathrm{SNR}_{\mathrm{dB}} = 6b + K_{\mathrm{density}}(\sigma^2) \qquad (5)$$

3. SIGMA-DELTA MODEL

The model of a Σ-Δ converter can be derived from a number of equivalent perspectives; the two most common are that of an error feedback modulator and that of a standard two-input one-output feedback loop. We first examine the error feedback model and then convert this model to other feedback models. Figure 4 presents the structure of a noise feedback modulator or coder. The coder consists of a prediction filter that computes, from previous quantization errors, an estimate $\hat{q}(n)$ of the next quantization error $q(n)$. This estimate is subtracted from the input and presented to the internal quantizer that adds the actual quantization error, $q(n)$, an error modeled as an additive noise source. The size of this added error is computed as the difference between the output and input of the quantizer and is delivered to the predictor filter for use in the next prediction cycle. From Fig. 4, we can determine the input and output of the quantizer that is shown in Eq. (6).

$$\begin{aligned}&\text{Quantizer input:} && x(n) - \hat{q}(n)\\ &\text{Quantizer output:} && y(n) = [x(n) - \hat{q}(n)] + q(n) \qquad (6)\\ & && = x(n) + [q(n) - \hat{q}(n)]\end{aligned}$$

The transfer function of the noise feedback coder is presented in Eq. (7). Here we see that the output of the system contains the signal input to the system, $X(Z)$, plus a filtered version of the noise input $Q(Z)[1 - P(Z)]$. The term $[1 - P(Z)]$ is denoted the noise transfer function (NTF):

$$\begin{aligned}Y(Z) &= X(Z) + [Q(Z) - \hat{Q}(Z)]\\ &= X(Z) + Q(Z)[1 - P(Z)] \qquad (7)\end{aligned}$$

Rather than continuing with this general model with arbitrary $P(Z)$, we choose to examine a specific example, in particular, the simplest prediction filter from which we will derive insight and guidance to other filter structures. We reason as follows. We want the output $\hat{q}(n)$ of the prediction filter to be a good approximation of $q(n)$, and further we want the computational burden required to perform the prediction to be small. We can realize these conditions if the successive samples of the error are highly correlated and, of course, we obtain high correlation when the input signal is significantly oversampled. This is the justification for the requirement to deliver over sampled data to the Σ-Δ converter. It is common, for instance, for the input data to be sampled at 64 times the signal's Nyquist rate. With significant oversampling, we can easily argue that the correlation between successive input samples is high and consequently the correlation between successive quantization errors is also high. For this condition, a good approximation of the next quantization error $\hat{q}(n)$ is the previous error $q(n-1)$, and the filter that supplies the delayed noise sample is $P(Z) = Z^{-1}$. Figure 5 shows the noise feedback coder with the prediction filter replaced with a single delay element Z^{-1}.

We note that there are two feedback loops in Fig. 5. The first loop starts at the input summing junction, negotiates the lower summing junction, and passes through the delay line back to the input junction. The second loop starts at the input summing junction, passes through the quantizer summing junction, the sign reversal of the lower summing junction, through the delay line and back to the input summing junction. Figure 6 presents the noise feedback quantizer redrawn to explicitly show the two loops just traversed, and then recognizing that the minor loop is a digital integrator, replacing it with the transfer function $Z/(Z-1)$. This figure is the conventional model of a Σ-Δ converter comprising an integrator and quantizer in a feedback loop.

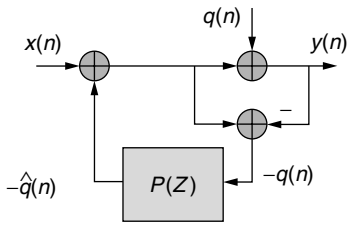


Figure 4. Simple model of noise feedback quantizer.

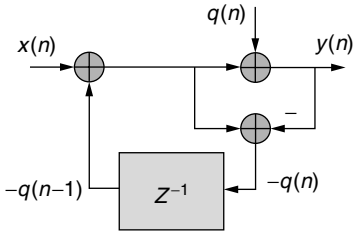


Figure 5. Noise feedback quantizer with delay-line predictor.

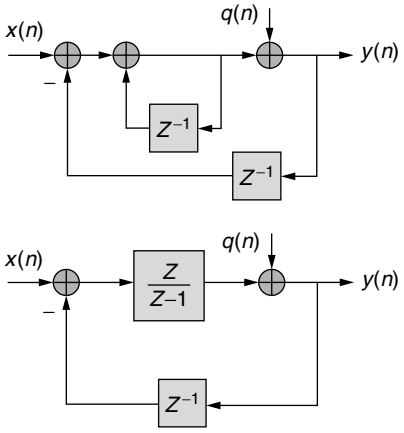


Figure 6. Noise feedback quantizer drawn with inner and outer feedback loops and with inner loop, a digital integrator represented by its transfer function.

Replacing $P(Z)$ with Z^{-1} in Eq. (7) leads to Eq. (8), which describes the input–output relationship of the Σ-Δ converter. Note that the NTF for the quantizing noise input is that of a simple differentiator $[1 - Z^{-1}]$ and that this filter has a transmission zero at $Z = 1$ or at DC. The pole-zero diagram of this NTF and its power spectral response along with the spectra of a typical input signal is shown in Fig. 7. The power spectral response of the NTF is derived in Eq. (9), where sampled data frequency is denoted by the parameter θ, where $\theta = \omega T = 2\pi(f/fs)$ and has units of radians per sample.

$$\begin{aligned} Y(Z) &= X(Z) + Q(Z)[1 - Z^{-1}] \\ &= X(Z) + Q(Z)\left[\frac{Z-1}{Z}\right] \end{aligned} \tag{8}$$

$$\begin{aligned} |\mathrm{NTF}(\theta)|^2 &= \left[\frac{e^{j\theta}-1}{e^{j\theta}}\right]\left[\frac{e^{-j\theta}-1}{e^{-j\theta}}\right] \\ &= 2[1-\cos(\theta)] = 4\sin^2(\theta/2) \end{aligned} \tag{9}$$

In Fig. 7, note that the zero of the NTF is located at DC, which suppressed the quantization noise in the neighborhood of DC. The signal spectral is restricted by the significant oversampling to reside in a small neighborhood of DC with two-sided width on the order of 1.5% of the sample rate. The combination of the over sampling and noise spectral shaping has arranged, to first order, for the signal and the noise spectra to occupy distinct spectral intervals. The signal spectra can be retrieved from the composite signal by a lowpass filter that passes the signal but rejects the noise residing beyond the signal bandwidth. Increasing the number of NTF zeros in the signal bandwidth can further improve the in-band noise suppression. We will examine Σ-Δ converters with multiple NTF zeros. The zeros of the NTF significantly suppress the quantizer noise levels of a quantizer embedded in Σ-Δ, consequently, low levels of output quantizing noise can be obtained from quantizers with relatively large levels of input quantizing noise. Most Σ-Δ converters are designed to operate with 1-bit quantizers with a number of high-performance converters operating with 4-bit quantizers.

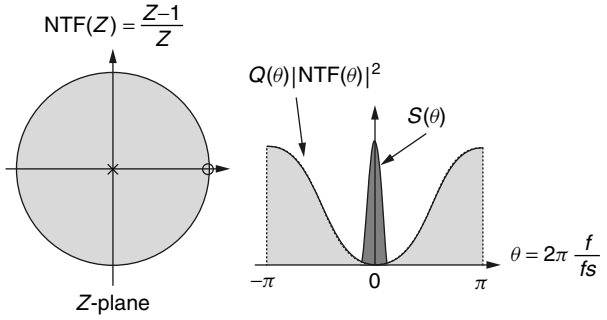


Figure 7. Pole-zero diagram of noise transfer function and power spectrum of input signal and of quantizing noise shaped by NTF.

4. NOISE PERFORMANCE OF SIGMA-DELTA CONVERTERS

We now review how filtering an oversampled and quantized data signal can improve that signal's quantizing SNR. The spectrum of nonshaped quantizing noise is uniformly distributed over a spectral width equal to the sample rate. If the signal is oversampled, say, by a factor of 2, half the quantizing noise power is in-band and half the noise power is out-of-band. We can pass the oversampled signal through a half-band filter without affecting the signal but effectively rejecting half the noise power. Rejecting half the noise bandwidth improves the signal to quantizing noise ratio by 3 dB. Quantizing noise is measured at 6 dB per bit, so reducing the noise bandwidth of uniformly distributed noise by a factor 2 improves the SNR by half a bit. To realize a 1-bit improvement in SNR, a signal would have to be oversampled by a factor of 4 and have its quantizing noise bandwidth reduced by the same factor of 4.

We now address the relationship between the SNR of a Σ-Δ converter with a shaped noise spectrum and its oversample ratio. In many modulators, the NTF of the Σ-Δ contains one or more zeros located at DC. We expand the power spectral response of a single-zero NTF in a Taylor series about DC and truncate the series after the first nonzero term to obtain an approximation to the filter response valid in the neighborhood of the signal spectrum:

$$\begin{aligned} |\mathrm{NTF}(\theta)|^2 &= 4\sin^2\frac{\theta}{2} = 2[1-\cos(\theta)] \\ &= 2\left\{1-\left[1-\frac{\theta^2}{2!}+\cdots+\right]\right\} = \theta^2 \end{aligned} \tag{10}$$

The fraction of the shaped noise spectrum that contributes to the final output of the Σ-Δ converter is that part that survives the filtering operation of the lowpass filter following the initial conversion process. The noise contained in the filter bandwidth is

$$\sigma_q^2 = \frac{N_0}{2}\int_{-\theta_{\mathrm{BW}}}^{\theta_{\mathrm{BW}}}\theta^2 d\theta = \frac{N_0}{2}\frac{1}{3}\theta^3\Big|_{\theta=-\theta_{\mathrm{BW}}}^{\theta=\theta_{\mathrm{BW}}} = \frac{N_0}{3}(\theta_{\mathrm{BW}})^3 \tag{11}$$

where we see that the noise contained in the filtered Σ-Δ output is proportional to the cube of the ratio of bandwidth-to-sample rate. If we reduce this ratio by a factor of 2, the noise contribution is reduced by a factor of 8, or by 9 dB, or equivalently by an improvement of 1.5 bits. Figure 8 illustrates how reduction in bandwidth (relative to sample rate) effects the reduction in output noise. Table 1 lists the rate at which the output SNR increases as a function of oversample ratio for a quantizer with 0, 1, 2, and 3 zeros in the NTF. The next section describes architectures for Σ-Δ converters that have NTFs with multiple zeros.

Figure 9 presents the power spectrum of a Σ-Δ's two-zero noise transfer function along with the spectrum of its 1-bit time series formed when processing an input signal containing two in-band sinusoids. Figure 10 show a segment of the input time series overlaid with the corresponding one-bit output series as well as the spectrum obtained by filtering the one-bit output series of the Σ-Δ

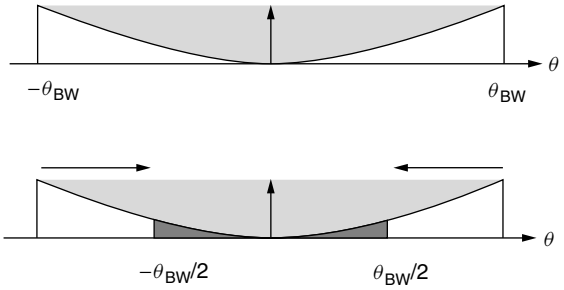


Figure 8. Area under spectrally shaped noise for two different ratios of output bandwidth-to-sample rate ratio.

converter. Note that the normalized bandwidth of the modulator and filter is $\frac{1}{100}$th of the input sample rate, a ratio equal to 6.64 octaves. From Table 1 we see that a 2-zero NTF exhibits a noise processing gain of 15 dB per oversampling octave, from which we expect a 6.64 × 15-dB or 99.6-dB improvement in SNR. With the spectrum normalized to the peak response of the windowed FFT (fast Fourier transform) we observe that the noise level of the filtered output is on the order of 98–105 dB with 103.6 dB as the actual SNR determined by integrating over the filter bandwidth.

5. SIGMA-DELTA ARCHITECTURES

A small number of common structures describe the architecture of the majority of Σ-Δ converters. We now examine the design philosophy common to many of these architectures. Most Σ-Δ architectures are formed about a feedback loop containing a loop filter and low-resolution quantizer that forms an oversampled, spectrally shaped data sequence that is processed by an external band-limiting filter to reject out-of-band noise. The feedback loop of the Σ-Δ quantizer is called the "Σ-Δ modulator," which, when combined with the filter, forms the sigma-delta converter.

Figure 11 is an extension of the redrawn noise feedback structure of Fig. 6 that cast the Σ-Δ as a two-input one-output feedback system. That first system employed an integrator and quantizer in a unity feedback control loop. What we have done here is replace the feedback integrator with an arbitrary filter $H(Z)$ and have also placed filter $G(Z)$ in the input path to enable private zeros for the input signal. We now consider appropriate constraints for the two filters to obtain the desired Σ-Δ performance.

As shown in Eq. (12), the transfer functions from the two inputs, $X(Z)$ and $Q(Z)$, to the common output $Y(Z)$ is computed as their distinct forward gains divide by one minus the loop gain:

$$Y(Z) = \frac{G(Z)}{1+H(Z)}X(Z) + \frac{1}{1+H(Z)}Q(Z) \tag{12}$$

The transfer functions of $H(Z)$ and $G(Z)$ are ratios of numerators to a common denominator polynomial defined by $N_1(Z)/D_1(Z)$ and $N_2(Z)/D_1(Z)$, respectively. Substituting these ratios in Eq. (12) leads to Eq. (13), and clearing the resulting denominators results in Eq. (14):

$$Y(Z) = \frac{N_2(Z)/D_1(Z)}{1+N_1(Z)/D_1(Z)}X(Z) + \frac{1}{1+N_1(Z)/D_1(Z)}Q(Z) \tag{13}$$

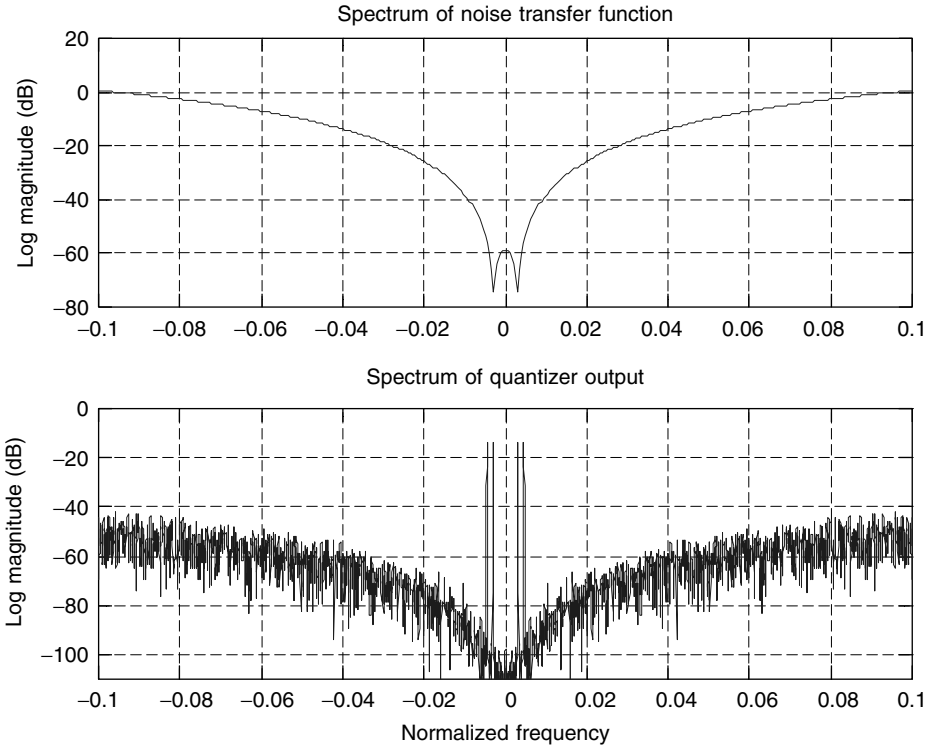


Figure 9. Noise transfer function of a 2-zero sigma-delta modulator and power spectrum of the output series obtained from modulator.

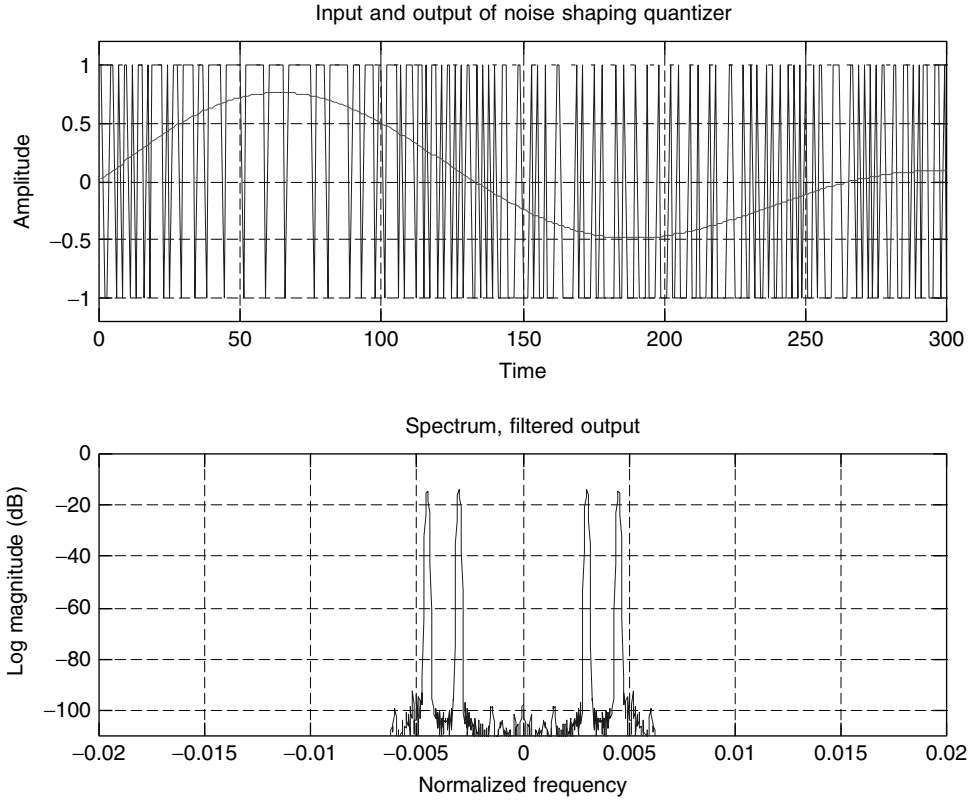


Figure 10. Segment of input and 1-bit output time series and power spectrum of filtered 1-bit output series from modulator.

$$Y(Z) = \frac{N_2(Z)}{D_1(Z) + N_1(Z)} X(Z) + \frac{D_1(Z)}{D_1(Z) + N_1(Z)} Q(Z) \tag{14}$$

5.1. MULTIPLE FEEDBACK LOOPS

We identify the two transfer functions shown in Eq. (14) as the signal transfer function (STF) and as the noise transfer function (NTF), respectively. As expected, the poles, $D(Z)$, of the loop filter transfer function become the zeros of the NTF. To assure good stopband performance of the NTF, $D(Z)$ must have its roots on the unit circle in the signal passband. Poles located at $Z = 1$, obtained by local feedback, form integrators that lead to the desired NTF zeros. Local feedback between the integrators can redistribute the zeros along the unit circle to realize an equal-ripple stopband to exchange excess attenuation for wider stopband bandwidth. Typical NTFs that can be realized with the structure of Eq. (14) are illustrated in Eq. (15); by way of example, these NTFs are third-order NTFs implementing Chebyshev (also transliterated as Tchebyschev), Butterworth, and derivative stopbands. These NTFs have distributed zeros and active poles, repeated zeros and active poles, and repeated zeros with inactive poles; "active" poles are finite poles that affect the spectral magnitude response, and hence are finite poles not at the origin. The frequency responses of these sample NTFs are shown in Fig. 12, where we see that the active poles have the desirable effect of reducing the

Table 1. Improvement in Quantizer SNR in dB and Effective Number of Bits for Each Doubling of Sample Rate Relative to Signal Bandwidth for Sigma-Delta Converters with 0, 1, 2, and 3 Noise Transfer Function Zeros

Number of NTF Zeros	SNR Improvement	
0	3 dB/double	0.5 bit/double
1	9 dB/double	1.5 bit/double
2	15 dB/double	2.5 bit/double
4	21 dB/double	3.5 bit/double

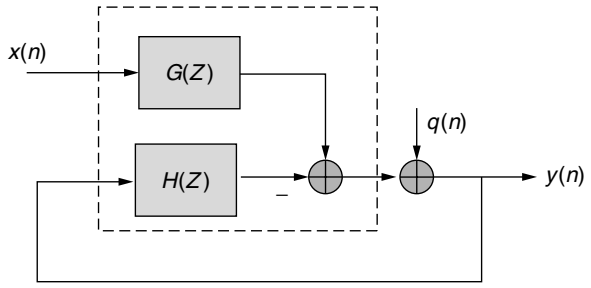


Figure 11. Two-input one-output feedback model of sigma-delta modulator.

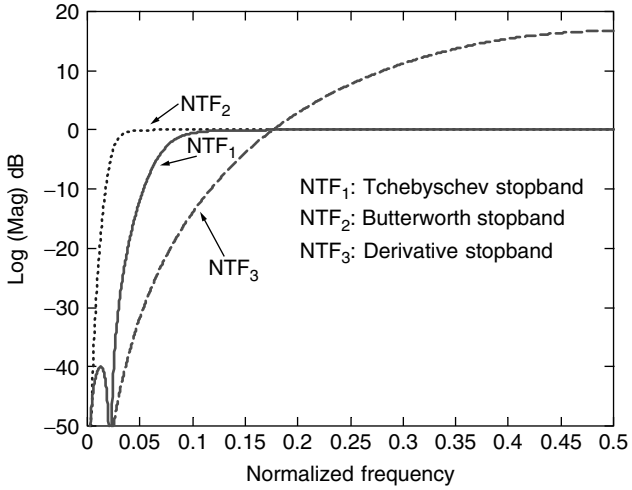


Figure 12. Spectral responses of typical noise transfer functions.

high-frequency gain of the NTF and that the distributed zeros widens the stopband bandwidth of the NTF.

$$\begin{aligned} \mathrm{NTF}_1 &= \frac{(Z-1)(Z^2+c_1Z+1)}{(Z-1)^3+(Z^2+b_1Z+b_2)} \\ \mathrm{NTF}_2 &= \frac{(Z-1)^3}{(Z-1)^3+(Z^2+b_1Z+b_2)} \\ \mathrm{NTF}_2 &= \frac{(Z-1)^3}{Z^3} \end{aligned} \tag{15}$$

The zeros $N_1(Z)$ of the feedback filter $G(Z)$ are selected so that the system poles match a selected prototype transfer function such as an elliptic, Chebyshev, or Butterworth stopband filter.

Figure 13a shows a three-stage version of the general multiple feedback–feedforward filter structure with local feedback forming the discrete integrators. The integrator poles become, via the major feedback loops, the desired NTF zeros. In this configuration, the input data $x(n)$ enter the filter through the feedforward path while the output data $y(n)$ leave the filter at the feedback path. Figure 13b presents the dual three-stage version of the pole structure presented in Fig. 13a. Here the input data $x(n)$ enter the filter at the feedback path while the output data $y(n)$ also leave the filter at the feedback path in which the quantizer must reside. This form of the filter does not offer a feedforward path to form private zeros for the signal transfer function.

5.2. Cascade Converters

The second major architecture for Σ-Δ modulators is that of cascade low-order Σ-Δ modulators. The cascade form is called a MASH converter, for *multiple sample and holds*, a description of the analog implementation. The low-order modulators can be formed by any structure but is usually implemented with one or two integrator loops and a 1-bit quantizer as originally shown in Fig. 6. We derived the expression for the output of the first stage of a single-loop modulator in Eq. (8). The output contains the loop's quantizer noise differentiated by the loop NTF. We obtain an improved NTF by the use of a second Σ-Δ modulator to measure and cancel the noise of the first modulator. This structure is shown in Fig. 14.

The Z transform of the first loop's output is shown in Eq. (16), and that of the second loop's output is shown in Eq. (17). Note that the first output contains $Q_1(Z)[1-Z^{-1}]$, the first loop's noise filtered by the loop NTF, a derivative, while the second output contains $Q_1(Z)$, the first loop's noise without the derivative. Applying a derivative as a $(1-Z^{-1})$ operator to the output of the second loop leads to the terms shown in Eq. (18), and forming the sum of $y_1(n)$ and $[y_2(n)-y_2(n-1)]$ leads to the terms shown in Eq. (19):

$$Y_1(Z) = X(Z) + [1-Z^{-1}]Q_1(Z) \tag{16}$$

$$Y_2(Z) = -Q_1(Z) + [1-Z^{-1}]Q_2(Z) \tag{17}$$

$$[1-Z^{-1}]Y_2(Z) = -[1-Z^{-1}]Q_1(Z) + [1-Z^{-1}]^2Q_2(Z) \tag{18}$$

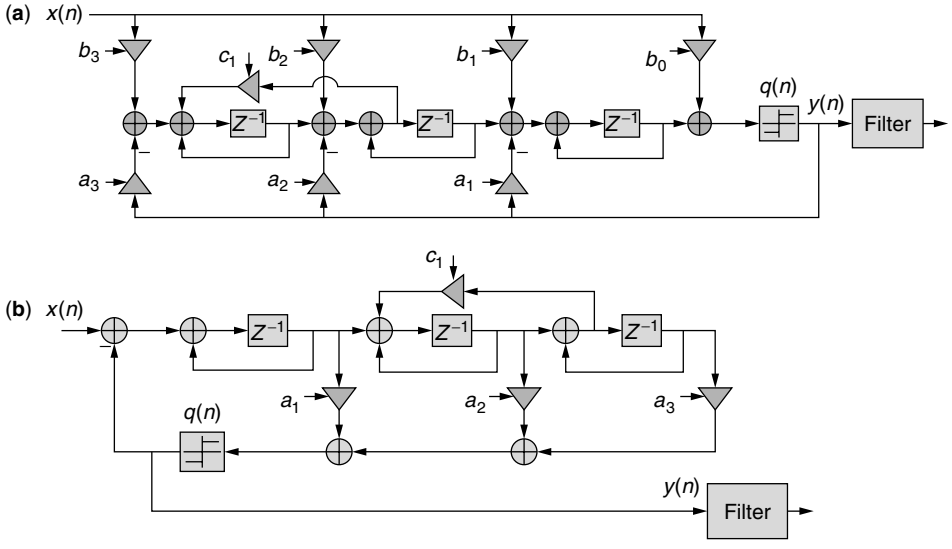


Figure 13. (**a**) Three-stage example of feedback–feedforward filter using cascade discrete integrators: zeros formed at input, poles formed at output; (**b**) dual three-stage example of feedback using cascade discrete integrators: poles formed at input.

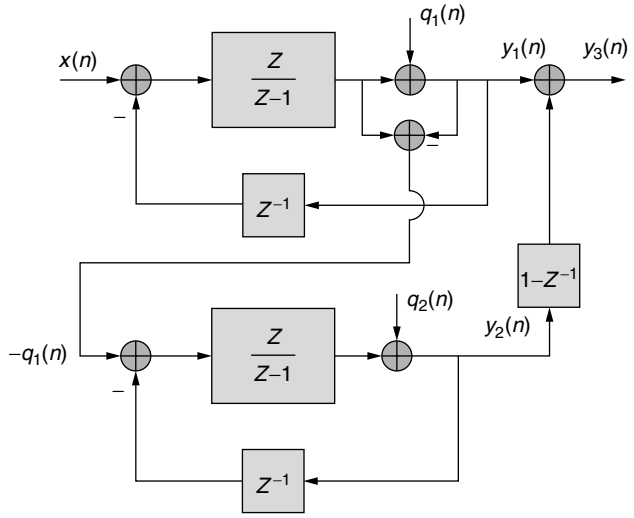


Figure 14. Cascade single-loop, single-bit sigma-delta modulators.

$$\begin{aligned} Y_1(Z) + [1 - Z^{-1}]Y_2(Z) &= X(Z) + [1 - Z^{-1}]Q_1(Z) \\ &\quad + \{-[1 - Z^{-1}]Q_1(Z) \\ &\quad + [1 - Z^{-1}]^2 Q_2(Z)\} \\ &= X(Z) + [1 - Z^{-1}]^2 Q_2(Z) \end{aligned} \tag{19}$$

The output of the cascade modulators contains the input $X(Z)$ and the doubly differentiated noise $Q_2(Z)[1 - Z^{-1}]^2$ of the second modulator loop. The double zero in the NTF has the effect of further suppressing the in-band quantization noise, which improves the output SNR for a given over sample rate. The process of canceling the first loop's noise with differentiated output of the second loop results in bit growth due to the two summations. If each of loop uses a 1-bit quantizer, the modulator outputs are ± 1. The output of the discrete derivative contains the three levels, 0 and ± 2, and the output of the sum of the two paths contains the four levels, ± 1 and ± 3. The result of combining the output of the cascade 1-bit modulators results in an equivalent 2-bit modulator. The cascade can be extended to include three stages by using a third Σ-Δ modulator to measure and cancel the noise of the second Σ-Δ modulator. The output of the third stage is double differentiated and added to the output of the two-stage cascade, which replaces the doubly differentiated second stage noise with the triply differentiated third-stage noise. Here again, the double derivative of a 1-bit sequence results in an increased output bit width. The output levels of the three cascade modulators are ± 1, ± 3, ± 5, and ± 7, the equivalent of a 3-bit modulator. Figure 15a–c presents the time series and the spectrum obtained from successive stages of a cascade of three single-loop modulators. Note the bit growth at each successive output and the enhanced depth of spectral noise suppression with the increased number of NTF zeros of the cascade modulator. For comparison, Fig. 15d presents the time series and spectrum from a single 3-loop, 3-bit sigma delta modulator. The spectral responses of the two 3-loop systems are essentially the same even though the time series from the two systems are very different.

5.3. Noise Prediction Loops

The third and the last major architecture for a Σ-Δ modulators is an enhanced version of the prediction filter introduced as our first model of the Σ-Δ modulator and presented in Fig. 4. We initially replaced the prediction filter with a delay element Z^{-1}, and then elected to operate the loop at rates far in excess of the signal Nyquist rate to assure high correlation between successive errors. We now return to this structure and describe one technique of designing efficient prediction filters, which permit significant reduction in the system over sample rate.

The design of a prediction filter that forms successive estimates of the next input sample is a standard task in signal estimation. The optimum prediction filter processes a sequence of N successive input samples with weights that minimize the mean-squared error between the prediction of the next sample and the ensuing measurement of that sample. This structure is expressed in Eq. (20) and shown in Fig. 16:

$$\begin{aligned} \hat{q}(n) &= \sum_{k=1}^{N} b(k) q(n-k) \\ e(n) &= q(n) - \hat{q}(n) \end{aligned} \tag{20}$$

The process of minimizing the mean-squared error leads to the standard normal equations shown in Eq. (21a) along with the augmented power of the prediction error shown in Eq. (21b). The set of N normal equations can be represented in matrix form as shown in Eq. (22), where the column vector $\overline{b}$ of dimension $(N+1)$ is the augmented coefficient vector $\{1 - b_1 - b_2 \cdots - b_N\}^T$, and the column vector $\overline{r}_e$ is the error correlation vector:

$$r_{qq}(m) = \sum_{k=1}^{N} b(k) r_{qq}(m-k) \quad m = 1, 2, 3, \ldots, N \tag{21a}$$

$$r_{ee}(0) = r_{qq}(0) - \sum_{k=1}^{N} b(k) r_{qq}(k) \tag{21b}$$

$$\begin{bmatrix} r_{qq}(0) & r_{qq}(1) & \cdots & \cdots & r_{qq}(N-1) \\ r_{qq}(1) & r_{qq}(0) & \cdots & \cdots & r_{qq}(N-21) \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ r_{qq}(N-1) & r_{qq}(N-2) & \cdots & \cdots & r_{qq}(0) \end{bmatrix} \times \begin{bmatrix} 1 \\ -b(1) \\ \vdots \\ \vdots \\ -b(N) \end{bmatrix} = \begin{bmatrix} r_{ee}(1) \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \tag{22}$$

Equation (22) can be written concisely in vector–matrix form as shown in Eq. (23) and then solved for the optimum weights as shown in Eq. (24):

$$\mathbf{R}_{qq}\overline{\boldsymbol{b}} = \overline{\boldsymbol{r}}_e \tag{23}$$

$$\overline{\boldsymbol{b}} = \mathbf{R}_{qq}^{-1}\overline{\boldsymbol{r}}_e \tag{24}$$

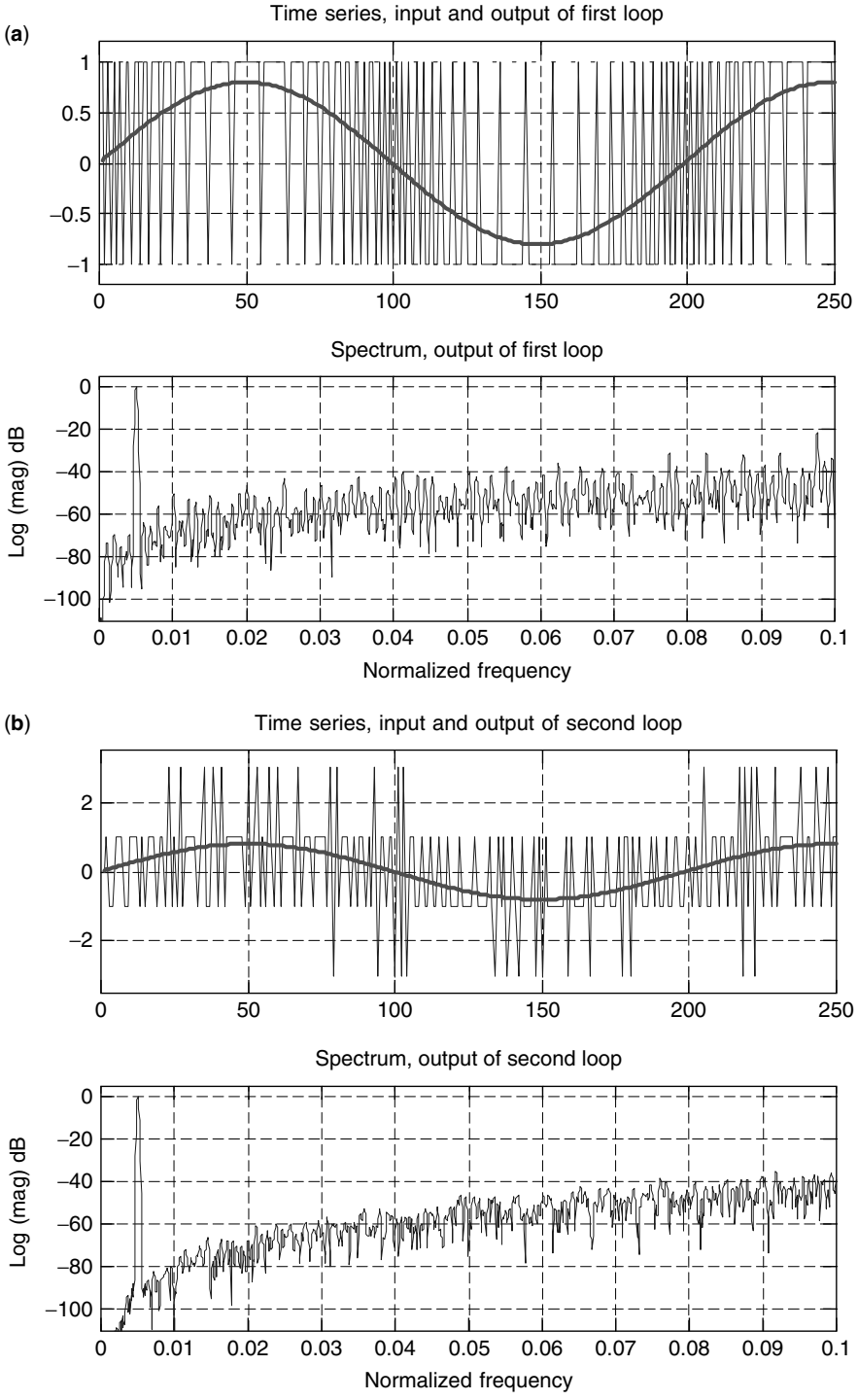


Figure 15. (**a**) Time series and spectral response of (**a**) first 1-bit modulator; (**b**) two-cascade 1-bit modulators; (**c**) three-cascade 1-bit modulators; and (**d**) 3-zero 3-bit modulator.

We have no problem with the existence of, and the process for, computing the weight set that minimizes the mean squared error for a given signal from which we can extract the second-order statistics. The problem is that we do not know the signal a priori hence do not have the required signal statistics. We fall back to a minimax solution, that of finding the solution for the signal with the worst statistics and apply that solution to the arbitrary signal. This signal is band limited white noise, and of course it is the band limiting that permits the successful prediction. The power spectrum of the bandlimited noise is shown in Fig. 17 and expressed as follows:

$$P_{qq}(\theta) = \begin{cases} 1 & -\theta_0 \le \theta \le \theta_0 \\ 0 & \text{elsewhere} \end{cases} \tag{25}$$

The correlation function of the modeled band-limited white noise, found as the inverse DTFT of Eq. (25), is

$$r_{qq}(n) = \frac{\theta_0}{\pi} \frac{\sin(n\theta_0)}{(n\theta_0)} \tag{26}$$

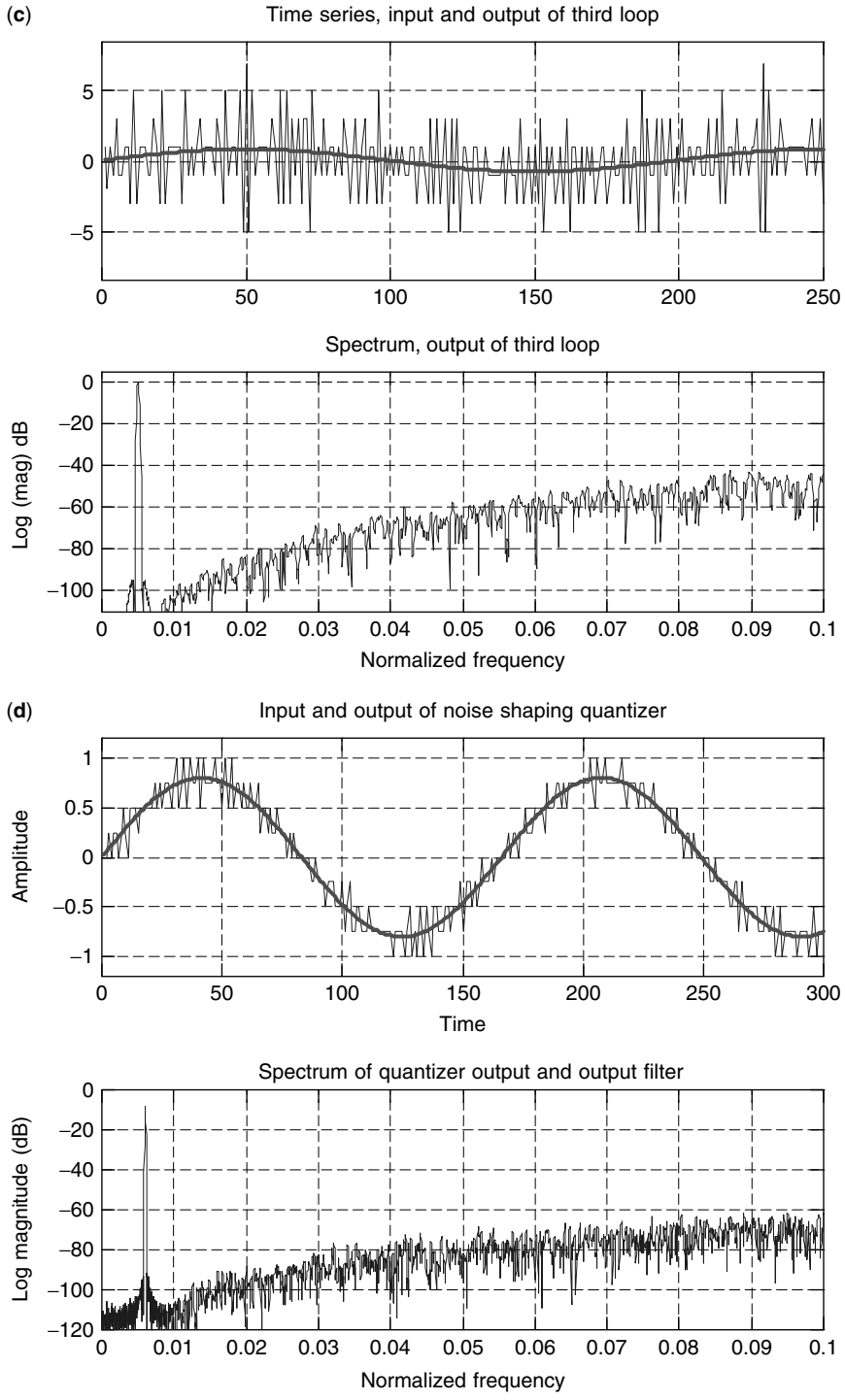


Figure 15. (*Continued*)

Using sample values of the correlation function presented in Eq. (26) to solve for the optimum weights of Eq. (24) leads to an ill-conditioned correlation matrix and to a set of weights with a large spread of coefficient values. This large spread is undesirable, and is due to the fact the optimum predictor is also the whitening filter. This filter tries to whiten the spectrum, and since there is no energy in the out-of-band region of Eq. (25), the filter can and does set arbitrarily large gains in this region. To control the out-of-band spectral gain of the predictor, we overlay the entire frequency band with a low-level white-noise spectrum of amplitude ε, modifying Eq. (25) to take the following form:

$$P_{qq}(\theta) = \begin{cases} 1+\varepsilon & -\theta_0 \le \theta \le \theta_0 \\ \varepsilon & \text{elsewhere} \end{cases} \tag{27}$$

The modified power spectrum has the following correlation function:

$$r_{qq}(n) = \frac{\theta_0}{\pi}\frac{\sin(n\theta_0)}{(n\theta_0)} + \varepsilon\delta(n) \tag{28}$$

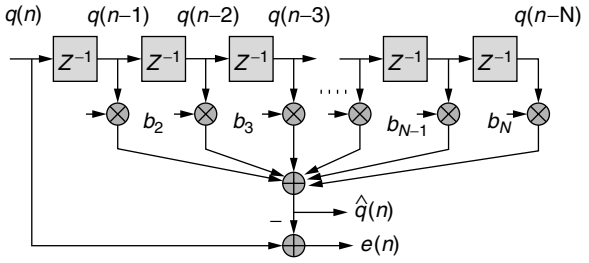


Figure 16. Tapped delay-line prediction filter.

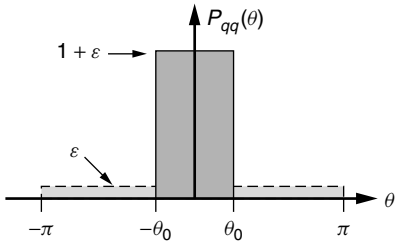


Figure 17. Power spectrum of band-limited white noise with ε white-noise overlay.

The effect of the added low-level white noise is to improve the condition number of the correlation matrix by adding the small value ε to the diagonal terms of the matrix. Adjustment of this parameter is an effective way to control out-of-band gain to improve the modulator stability margin.

Figure 18a presents the spectrum of a 20% bandwidth 10-tap prediction filter designed with the conditioning parameter $\varepsilon = 10^{-5}$. This is an unusually wide bandwidth Σ-Δ modulator with sample rate only 5 times the signal's Nyquist rate. Also shown is the spectrum formed from the Σ-Δ modulator 4-bit output sequence obtained with this prediction filter. Figure 18b shows a segment of the input series and the 4-bit output series from this modulator as well as a zoom to the passband of the output signal spectrum.

5.4. Bandpass Sigma-Delta Modulators

The Σ-Δ modulators we have examined thus far have been baseband since they were originally designed around the spectral characteristics of the pole of a digital integrator. The standard method for quantizing a passband signal involves a downconversion of the center frequency to baseband with a pair of quadrature mixers, a pair of filters to remove the sum frequencies formed by the mixing operation, and finally quantization with a pair of matched converters. It is logical to perform this translation when the Σ-Δ modulator is the conversion device, since the NTF zeros reside at base band. As expected, two converters are required to service the complex base band process. Another option is to move the NTF zeros from baseband to the center frequency of the narrowband input and perform the desired conversions with the signal residing at a low intermediate frequency.

When we move the sigma-delta integrator poles from the neighborhood of zero frequency, the resultant poles are called *resonators*. A resonator formed by structures with real coefficients exhibits both positive- and negative-frequency images, and hence requires twice the number of integrators to build the positive- and the negative-frequency NTF zeros. Doubling the number of integrators is equivalent to building and using two filters in a pair of modulators, so the increase in implementation complexity is not a concern. Feedback around resonator pole pairs results in NTF spectral zeros at these locations; hence a Σ-Δ modulator can be designed to operate at any frequency within the spectral span defined by the sample rate. Rather than pursue the straightforward synthesis problem, we consider techniques that can be applied directly to the structure of the prototype baseband modulator. A number of such techniques can be used to translate the poles of a baseband prototype modulator to an arbitrary center frequency.

We can effect a frequency transformation of an existing filter structure by replacing all-pass networks in the structure with another all-pass network. The simplest such transformation, shown in Eq. (29), replaces a delay line, represented by Z^{-1}, with a phase rotated delay line, represented by $Z^{-1}e^{j\theta}$. When applied to the delays in a filter, the spectral response of the filter is translated to the center frequency θ radians per sample. This relationship, shown in Eq. (30), is known as the *modulation property* of the Z transform:

$$Z^{-1} \Rightarrow Z^{-1}e^{j\theta} \tag{29}$$

$$\text{If} \quad h(n) \Leftrightarrow H(Z)$$

$$\text{Then} \quad h(n)e^{jn\theta} \Leftrightarrow H(Ze^{-j\theta}) \tag{30}$$

Figure 19 is a third-order multiple feedback Σ-Δ modulator with the tuning substitution described in Eq. (29) converting the prototype integrators into complex resonators. This substitution results in complex data due to the complex scalars, requiring the use of complex registers in place of the delay lines of the prototype, and similarly, the need for two real quantizers in place of the original quantizer. Figure 20 shows the frequency response of time series obtained from a baseband prototype, from a complex resonated version, and from a real resonated version of the 3-loop, 1-bit modulator presented in Fig. 19. The real resonator version is described next.

One might object to the computational burden due to the use of a complex phase rotator in each complex resonator, but remember that the scalar phase rotators replace the pair of input mixers and the DDS required for the downconversion. Judicious choice of center frequency can also lead to trivial operators that replace the complex phase rotators. In particular, when the center frequency is the quarter-sample rate, the phase rotator defaults to $\exp(j\,\pi/2)$ or j, a simple data transfer and possible sign reversal between registers.

A related all-pass transformation that can be applied directly to the baseband prototype modulator is the lowpass to bandpass transformation shown in Eq. (31). This transformation appends to each delay element in the filter a sign change, and an all-pass tuning filter with a parameter $c = \cos(\theta c)$, where θc is the center frequency.

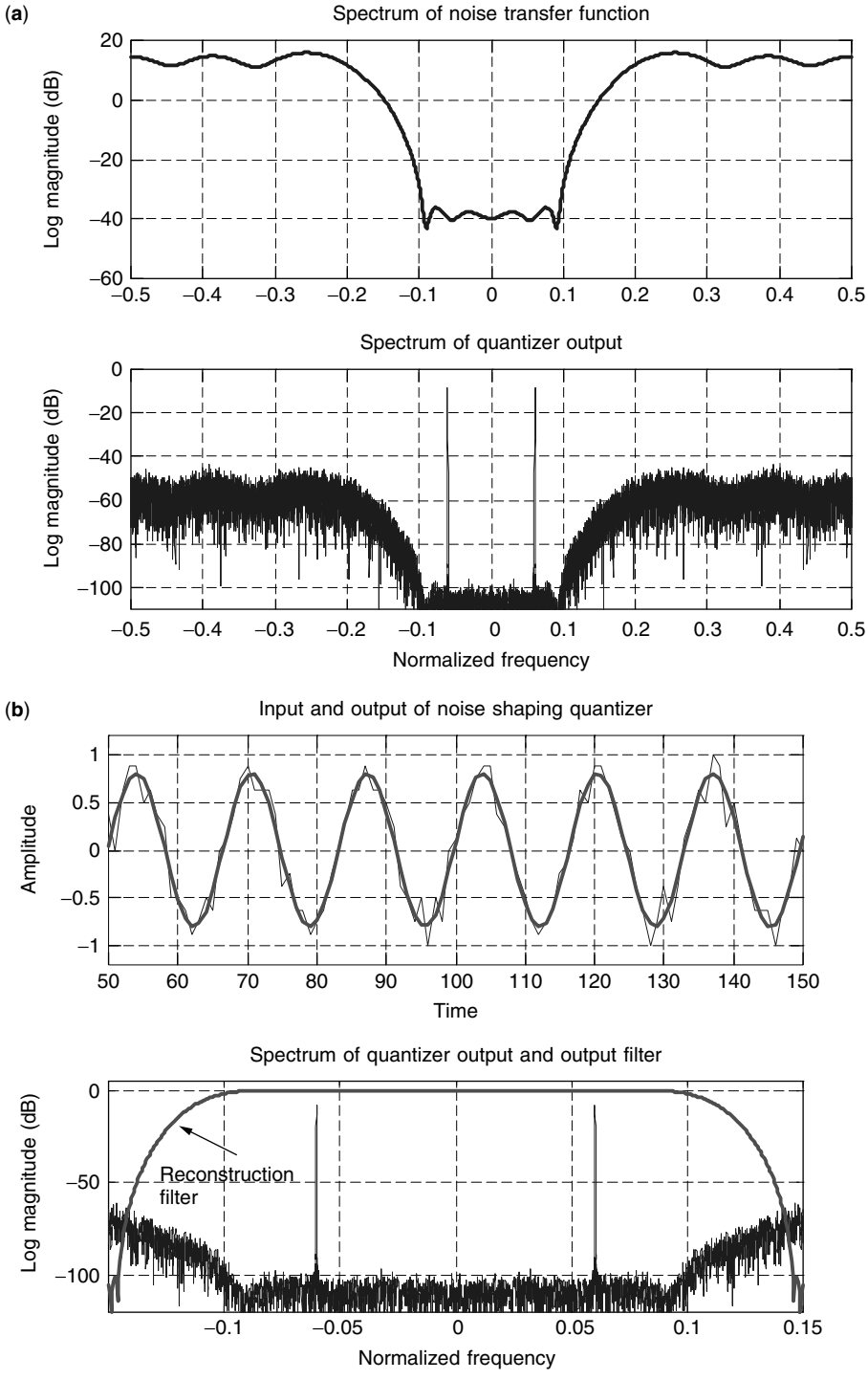


Figure 18. (**a**) Spectrum of noise transfer function obtained with 10-tap prediction filter and spectrum of sigma-delta output 4-bit series; (**b**) segment of input and output time series obtained with 10-tap prediction filter with 4-bit sigma-delta modulator and detail of output series spectrum.

The most common form of this frequency transformation is the default case $\theta c = \pi/2$, which sets c to zero, and consequently replaces Z^{-1} with $-Z^{-2}$. This substitution results in two delays with negative feedback in place of the baseband integrators:

$$\frac{1}{Z} \Rightarrow -\frac{1}{Z}\frac{1-cZ}{Z-c}, \qquad \text{where} \quad c = \cos(\theta_c) \tag{31}$$

Figure 21 illustrates the progression from the base band integrator, through the quarter-sample rate resonator, to the arbitrary all-pass tuned resonator. The lowpass to bandpass generalized delay in the tuning resonator requires two delays, two additions, and a single multiplier to implement both numerator and denominator as well as the cascade delay shown in Eq. (31). The third subplot of Fig. 19 presents the spectrum of a time series formed by the all-pass tuned variant of the modulator shown in Fig. 18.

As a final note on resonated Σ-Δ modulators, we comment that the predictive noise filters of the previous

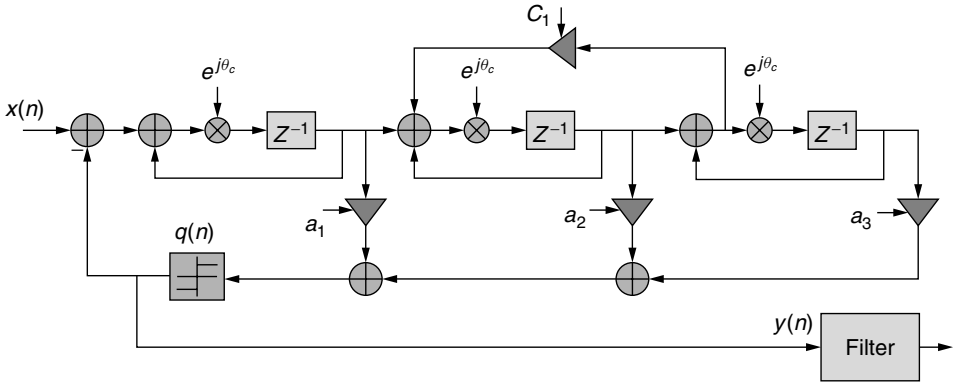


Figure 19. Bandpass sigma-delta converter with complex, phase-rotated resonators.

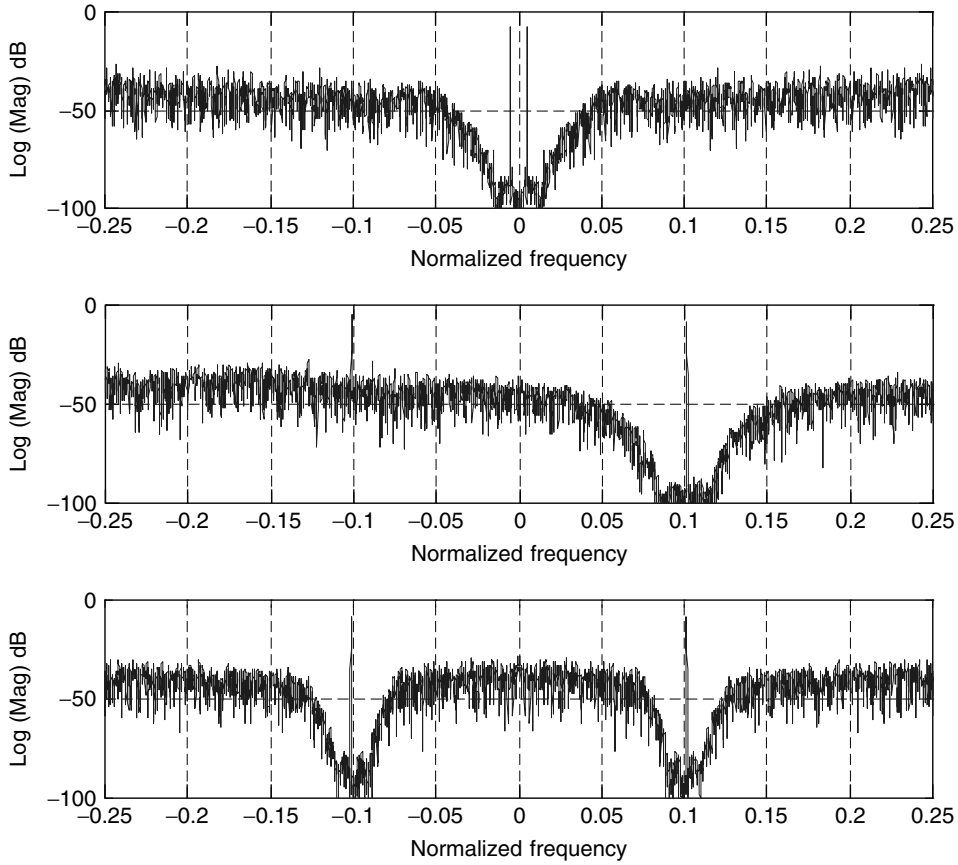


Figure 20. Spectrum of time series from 3-loop, 1-bit baseband prototype, from complex resonator, and from real resonator versions of prototype modulator.

section can also be trivially tuned to become bandpass Σ-Δ modulators. We manage this by spectrally shifting, as a symmetric or asymmetric translation, the spectra of the band-limited noise power spectrum in Eq. (27) used to define the correlation sequence of Eq. (28). As a consequence of this translation, the filter designed by the normal equation will be have a bandstop NTF rather than a baseband NTF.

6. STABILITY CONSIDERATIONS

We now address models of the quantizer in the Σ-Δ modulator. Our first order model of the quantizer is that

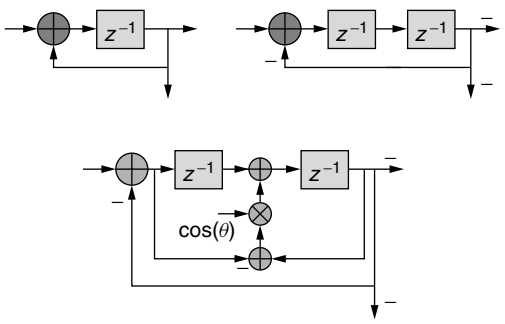


Figure 21. Prototype integrator, quarter-sample rate resonator, and all-pass tuned arbitrary frequency resonator.

of an independent additive noise source, the standard approximation used in a memoryless quantizer. When we have a small number of quantization levels, this model is poor because of the high correlation between the quantized signal and the quantization error. The quality of the model is improved by adding a random dither signal to the data samples prior to the quantization process. When embedded in a feedback loop, the model must also include a gain that may be amplitude-dependent. One linear model of the nonlinear quantizer $Q[u(n)]$ is shown in Fig. 22, where the input signal $u(n)$ is partitioned into two components, $u_{DC}(n)$ and $u_{AC}(n)$, subjected to their separate gains of K_0 and K_1 and then added to the additive noise source. Reasonable questions to ask are (1) whether this is a valid model and (2) over what range of input signal amplitudes is this a valid model.

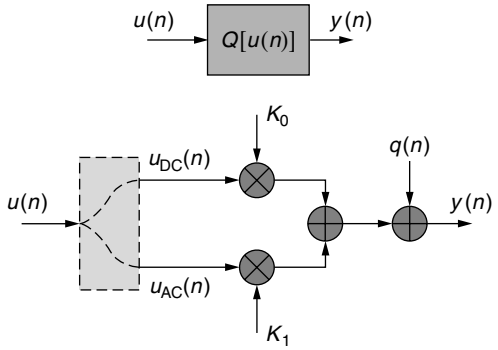


Figure 22. Linear model of quantizer in sigma-delta modulator feedback loop.

We answer the first question, concerning model validity by operating two 3-loop feedback models, one containing a standard 1-bit quantizer, and one an additive noise source in place of the quantizer. Figure 23a presents the

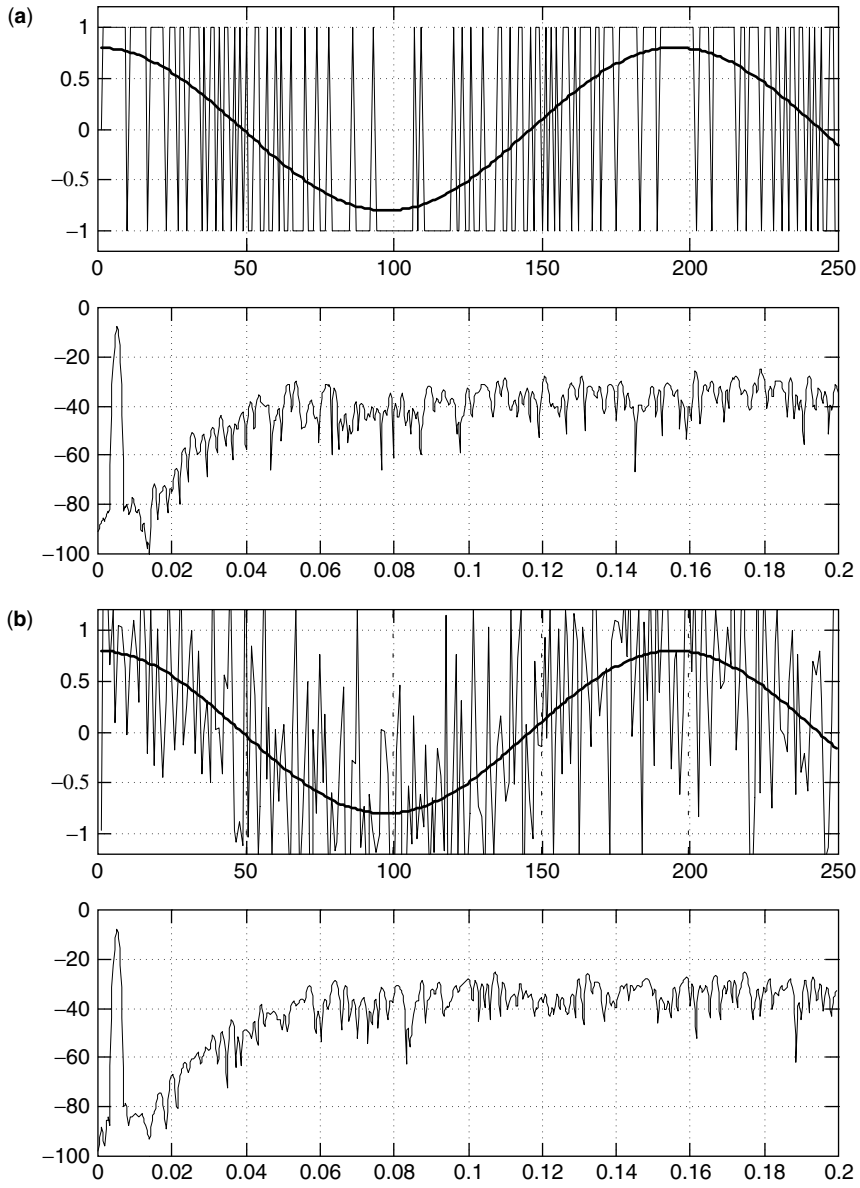


Figure 23. Input and output time series, and output spectra of a (**a**) 3-loop, 1-bit sigma-delta modulator and (**b**) 3-loop, linear additive noise model of sigma-delta modulator.

input and output time sequences with associated output spectrum of the 3-loop, 1-bit modulator. Figure 23b shows the corresponding figures for the same system containing the additive noise source with comparable variance. As projected, the spectra of the two loops is identical; the shaped inserted noise has the same spectrum as the shaped quantization noise. It is interesting to see how well the feedback loop suppresses the nonlinear quantizer behavior, enabling the linearized model to approximate the performance of the nonlinear system.

The next question we address is the range of input amplitudes for which the linearized model is a valid description of the nonlinear system. The standard approach to this inquiry is to collect statistics on the maximum level of the modulator's internal registers as a function of the input signal level. Figure 24a presents curves showing the maximum register levels observed in a 3-loop, 1-bit Σ-Δ modulator for test runs of length 16,384 samples with fixed (DC) levels over the range of 0–1, where 1 is the full-scale 1-bit quantizer output level. Also shown are the maximum register levels as a function of fixed input levels for the linear model of the loop with the inserted noise rather than quantizing noise. We note that the quantized loop and the linear model exhibit similar responses for fixed DC input levels spanning the range 0 to 0.6, and that the quantized loop exhibits poor stability for input signal levels beyond 0.6 and in fact becomes unstable at approximately 0.74.

Figure 24b presents a set of similar curves for the maximum register levels observed in the same 3-loop,

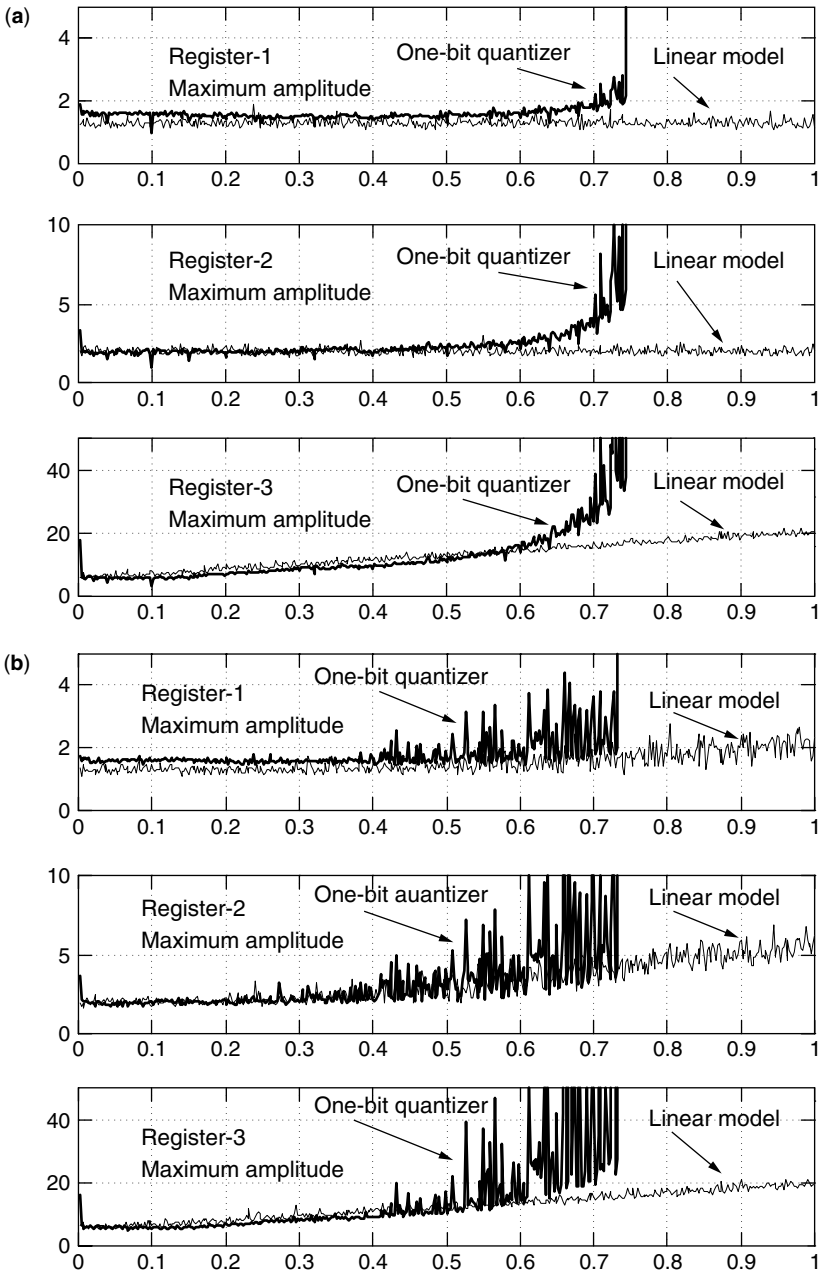


Figure 24. Amplitudes of internal states of 3-loop sigma-delta modulator and linearized noise model as function of (**a**) DC input level and (**b**) in-band AC input level.

1-bit Σ-Δ modulator for the same measurement conditions except that the input signal is an in-band sinusoid with fixed amplitude levels varied over the range of 0–1. As in the companion figure, also shown are the maximum register levels as a function of input levels for the linear model of the loop. We note that the quantized loop and the linear model exhibit similar responses for fixed AC input levels spanning the reduced range 0–0.4, with the quantized loop exhibiting increasingly poorer stability for input signal levels in the range 0.4–0.74 and in fact becoming unstable at 0.74. Note that the range of input AC levels for which the internal states avoid instability is smaller than the range of input DC levels. It is standard practice to restrict the range of input levels to half the input range to avoid the operating at the edge of unstable behavior.

The relationship that couples the instability of the sigma-delta modulator to the amplitude of the input resides in the fact that the linear model gain terms, K_0 and K_1, introduced in Fig. 22, decrease as the input amplitude increases. We can include the gain K_1 in the closed-loop expression for the noise transfer function, as shown in Eq. (32) to illustrate how the denominator of the transfer function changes with K_1, and hence varies with the input amplitude. From conventional feedback analysis we see that the closed-loop zeros are the open-loop poles, and that as K_1 varies from 0 to 1, the closed root poles migrate from the closed root zeros to the unity-gain closed root poles. The locus of the root migration as a function of K_1 is shown in Fig. 25. Note that for values of K_1 less than 0.276, the system poles are outside of the unit circle and the system is unstable:

$$\mathrm{NTF}(Z) = \frac{1}{1 + K_1 N(Z)/D(Z)} = \frac{D(Z)}{D(Z) + K_1 N(Z)} \qquad (32)$$

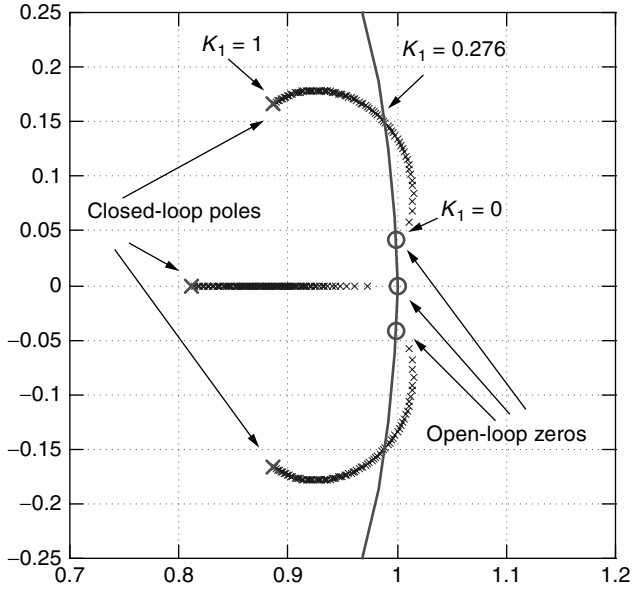


Figure 25. Root locus of closed-loop poles for third-order sigma-delta modulator demonstrating cause of instability as quantizer gain K_1 is reduced.

7. DIGITAL SIGNAL PROCESSING APPLICATIONS OF SIGMA-DELTA CONVERTERS

7.1. Σ-Δ Modulator in Data Preprocessor

The Σ-Δ converter can be used to improve the fidelity of any quantization process in a DSP system, many of which occur naturally and many by virtue of enlightened design. In the spirit of the latter option, the Σ-Δ modulator can be used as a preprocessor in any filtering task for which the filter bandwidth is a small fraction of its sample rate. Under this condition, the bandwidth of interest is already oversampled and can be requantized to a smaller number of bits to reduce the arithmetic resources required for the ensuing filtering process. One example we cite is that of a digital FIR filter to extract a downconverted 0.6-MHz-wide color subcarrier of a composite 6-MHz-wide NTSC signal sampled at 12.0 MHz. We note that the signal bandwidth processed by the filter is already oversampled by a factor of 20 occupying only 5% of the sample rate. The Σ-Δ can requantize the data from 10 bits to 1 bit with a shaped noise spectrum. The shaped noise spectrum preserves the quantizing noise level of the input signal in the signal bandwidth while permitting increased quantizing noise levels in the band to be rejected by the filtering process. Block diagrams of the two approaches to this processing task, conventional and sigma-delta preprocessing, are shown in Fig. 26. Here, a 4-tap preprocessor converts the 10-bit data to 1-bit so that the following 80-tap filter can be implemented without multiplies. Figure 27 shows the input and output time series as well as the output spectrum of the preprocessor. As expected, the 1-bit process successfully preserves the signal fidelity in the important band, the band processed by the filter.

7.2. Σ-Δ Modulator in DC Canceler

A second application of the Σ-Δ modulator as a DSP preprocessor is its insertion in the common signal task of canceling DC. DC components are generated and inserted in the signal a number of mechanisms, these include analog insertion due to untrimmed offsets in A/D converters, and digital insertion due to arithmetic truncation of two's-complement products and summations in various signal processing operations. The bias, or DC offset, on the order of a fraction of a bit per sample, does not appears to be a problem at first glance, but in

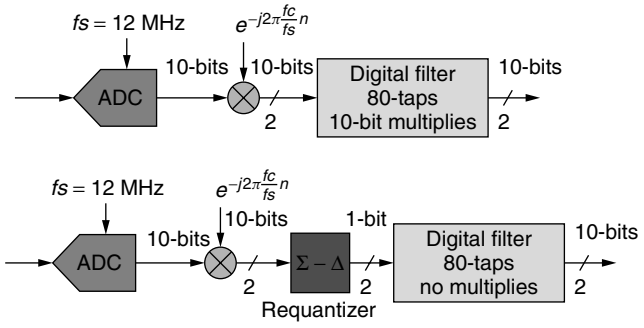


Figure 26. Narrowband processing with conventional and unconventional sigma-delta preprocess filtering.

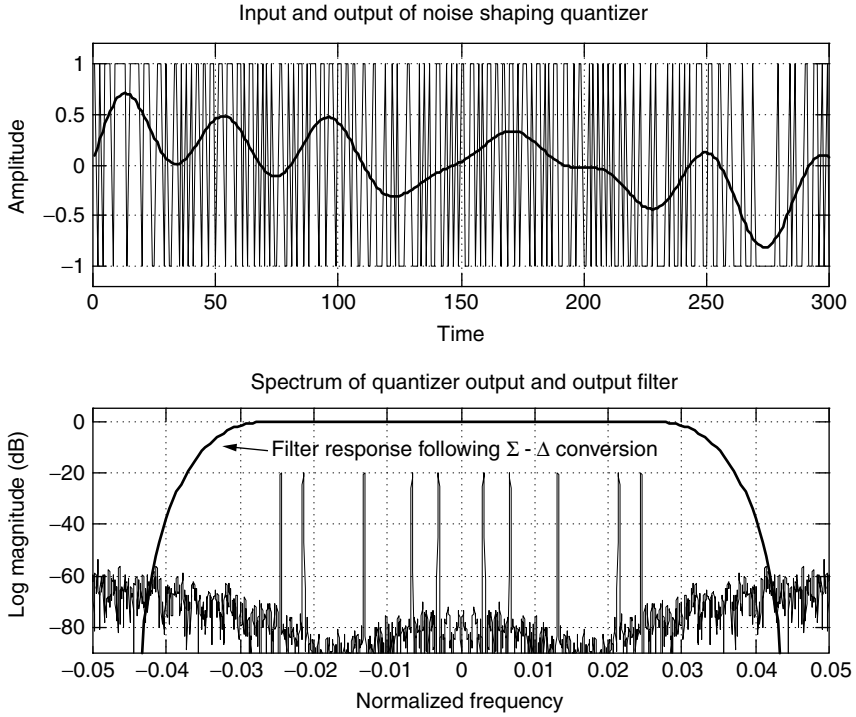


Figure 27. Time series and spectrum of signal preprocess quantized to 1 bit for reduced workload in subsequent filtering process.

fact becomes a problem when the samples are coherently processed. The fractional bit offset can grow to a many-bit offset as the signal experiences the coherent gain of subsequent filters as they reduce signal bandwidth. In some applications, the DC is of no consequence and can be ignored, but in others, the DC offset must be removed early in the signal processing path. We do this to preserve the dynamic range of a fixed-point data set and in digital receivers to avoid decision biases in the detection process following matched-filter processing.

Removal of the DC is performed by a DC notch filter usually implemented as a DC canceler of the form shown in Fig. 28. The integrator in the feedback loop of the filter becomes the transmission zero of this filter that also exhibits a nearby pole at $Z = (1 - \mu)$. The notch filter has the transfer function shown in Eq. (33), and we note that the distance, μ, between the zero and pole defines the bandwidth of the notch. The parameter μ is a small binary number, on the order of 2^{-10}, implemented by a right data shift. The integrator estimates the DC in the series, and the filter subtracts the DC from the input sequence. In order to cancel a DC term whose value is a fraction of a bit, the output of the canceler filter must grow additional bits to the right of the input data's binary point. On leaving the canceler, the lower-order data bits are discarded by the output quantizer that returns the number of output bits to the number of input bits for the benefit of subsequent arithmetic processing. This quantization discards the fractional part of the correction inserted by the canceler, so that the combined canceler and quantizer is capable of rejecting only the integer number of bits of the DC offset:

$$H(Z) = \frac{Z - 1}{Z - (1 - \mu)} \tag{33}$$

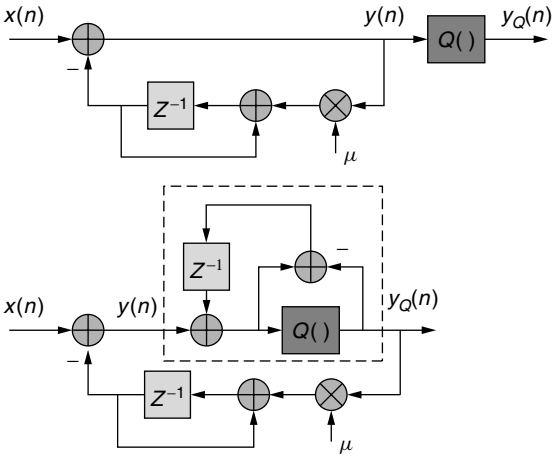


Figure 28. DC canceler with integrator in feedback path and external quantizer and same DC canceler with internal quantizer with noise feedback loop.

To preserve the fractional part of the corrected DC cancellation, we move the quantizer into the feedback loop and wrap a noise feedback loop around the quantizer. This modified structure is shown in the lower section of Fig. 28. The noise feedback quantizer can be placed in either the feedforward or in the feedback portion of the noise canceler. Figure 29 shows the spectrum of the input and output of the DC canceler before and after the external quantizer. We see the DC term present in the input spectrum and its absence in the output spectrum. A DC term is reinserted at the output of the external quantizer as the data samples are returned to 8-bit datawords. In the last figure we note that the internal quantizer with its

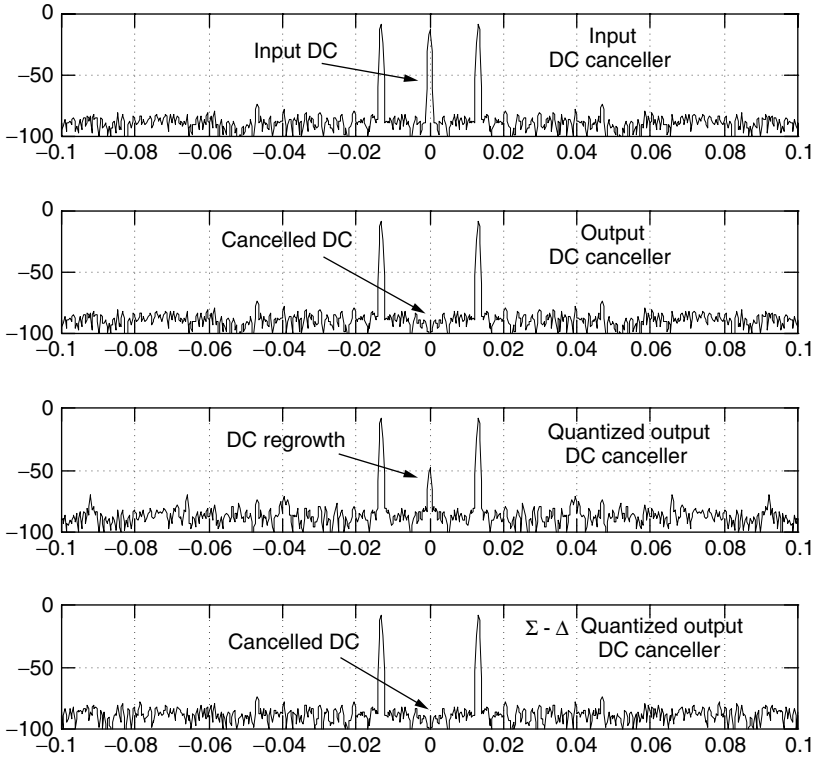


Figure 29. Spectra of input and output of DC canceler, and then output of canceler with 8-bit external quantizer and 8-bit internal noise feedback quantizer.

noise feedback loop does not reinsert the DC as the data are truncated to 8 bits.

7.3. Σ-Δ Regulator in DDS

The last example we cite for the use of a Σ-Δ requantizer is in a direct digital synthesizer (DDS). The DDS uses a fine-resolution overflowing phase accumulator to synthesize a specified phase–time profile for an output complex sinusoid. In a typical system, the output phase word, drawn from a 32-bit accumulator, is quantized to an 8-bit word used as an address to access the sine–cosine values stored in its lookup table. This structure is shown as the first of the three block diagrams in Fig. 30. The quantization forms a correlated error sequence, in fact a sawtooth-shaped periodic phase error of peak amplitude equal to the least significant output bit. The phase error sequence, the difference between the input and output of the quantizer, is shown as the top segment of Fig. 31. This error sequence phase-modulates the output sinusoid, generating an undesired set of line spectra centered about the output center frequency. The amplitude of the maximum phase modulation spurious line is 6 dB per address bit below the desired spectral line. The first curve of Fig. 32 presents the spectrum formed from an 8-bit lookup table containing 16-bit values of sine and cosine of the table address. We can clearly see the line structure and the −48-dB phase modulation spurious tone.

The spectral line structure related to the periodic phase error is undesirable, and the standard remedy to suppress this line structure is the use of additive dither to break up the regularity of the error sequence. The second block diagram of Fig. 30 illustrates the location for the dither

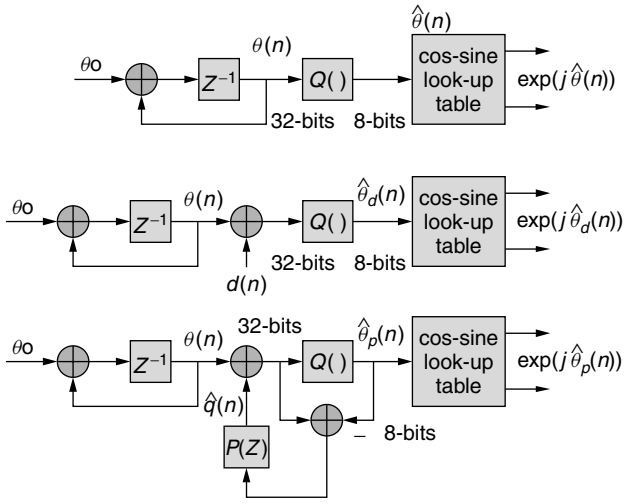


Figure 30. DDS quantization options for table addresses: option 1 — truncation; option 2 — dithered truncation; option 3 — noise feedback truncation.

insertion, the corresponding curve in Fig. 31 shows the dithered phase error structure, and the related curve in Fig. 32 shows the spectrum of the time series generated by the dithered address process. A proper dither suppresses the line structure and pulls the average phase noise level down by 12 dB or 2 bits, in this example from −48 to −60 dB.

It is an easy transition to replace the addition of random dither prior to the quantization process by the addition of correlated dither formed in a noise feedback

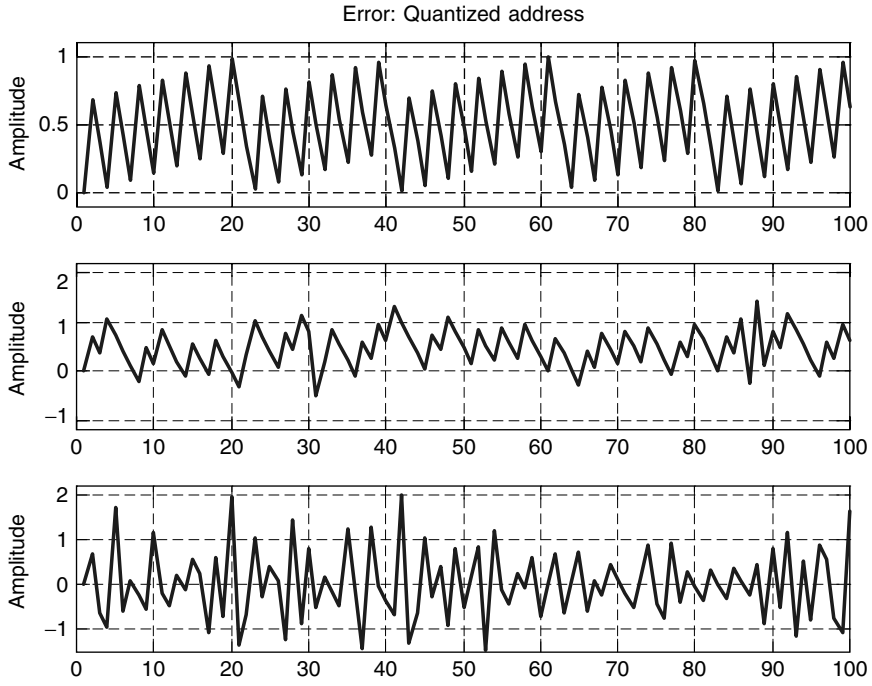


Figure 31. DDS phase error sequences obtained by different quantization options: option 1—truncation; option 2—dithered truncation; option 3—noise feedback truncation.

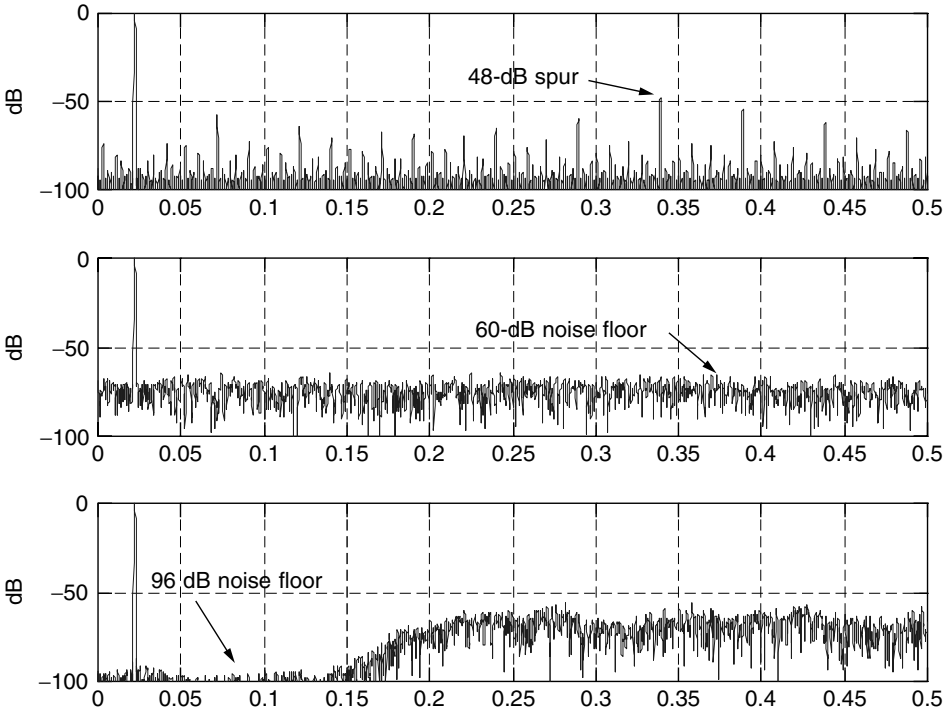


Figure 32. Spectrum of sinusoids obtained from 8-bit table with different quantization options: option 1—truncation; option 2—dithered truncation; option 3—noise feedback truncation.

loop. The third block diagram of Fig. 30 illustrates this option. The corresponding segment of Fig. 31 presents the dithered phase error generated by a 10-tap prediction filter designed using the normal equations presented in Eqs. (22)–(28) to obtain a noise suppression bandwidth equal to 25% of the sample rate. The shaped noise spectrum shown in the third segment of Fig. 32 illustrates a dramatic improvement in the level of phase noise modulation. The in-band level has been pulled down from −60 to −96 dB with rather small increases in out-of-band spectral level. This option of using the sigma-delta loop to manipulate addresses rather than signal amplitude leads to SNR improvements comparable to that available from a dithered 14-bit table with data samples extracted from an 8-bit table. Shorter tables, and predictors with smaller fractional bandwidth can form sinusoids at arbitrary center frequencies exhibiting remarkable fidelity over restricted close-in bandwidths.

8. CLOSING COMMENTS

We have reviewed the theory of operation of noise feedback, hence noise shaping, quantizers commonly called *sigma-delta modulators*. A number of architecture variants were described and their performance illustrated by examining spectra of their output series. Sigma-delta modulator architectures were described that reflect three primary design perspectives: cascade integrators, predictive feedback, and combinations of multiple simple modulators. The cascade integrator model uses minor-loop feedback to place and distribute open-loop poles on the unit circle and then major-loop feedback through a low-resolution quantizer to convert these poles to closed-loop noise transfer function zeros. The noise feedback model measures the quantizing error and uses band-limited prediction filters to predict and precancel the next value of quantizing noise contained within a selected segment of input bandwidth. The cascade models use a succession of low-order modulators to synthesize a high-order modulator. This is accomplished by measuring and canceling the shaped quantizing noise from previous stages with enhanced-shaped noise from later stages.

The tradeoff between oversample ratios, quantization SNR improvement, and order of feedback filter was described for multiple zero NTF. We examined narrow-band resonator-based models of the Σ-Δ process. In particular, we limited our discussion to transformations of baseband prototype systems that support tunable variants of existing systems. Feedback stability considerations were examined to explain the divergence of behavior between the linear and nonlinear models of the Σ-Δ modulator. Finally, a number of Σ-Δ applications were presented and illustrated in which the requantization process was enhanced or invoked to improve the performance of standard DSP-based signal processing tasks.

Material not addressed in this presentation included design and implementation of digital resampling filters that normally accompany the modulator when it is used in A/D and D/A applications. Here the emphasis was the embedding of the Σ-Δ modulator in a DSP system environment. Consistent with this emphasis, analog implementations of the noise feedback process were also intentionally not addressed in this review article.

BIOGRAPHY

Fred J. Harris received the B.S. degree in Electrical Engineering in 1961 from the Polytechnic Institute of Brooklyn, the M.S. degree in Electrical Engineering in 1967 from San Diego State University, and pursued Ph.D. work in electrical engineering from 1967 to 1971 at the University of California at San Diego. Since 1967 he has taught at San Diego State University, were he occupies the CUBIC Corp. Signal Processing Chair. Teaching and research areas include digital signal processing, multirate signal processing, communication systems, source coding, and modem design. He holds a number of patents involving multirate signal processing for satellite and cable modems as well as for sigma-delta implementations. He has contributed to a number of texts and handbooks on various aspects of signal processing. He is the traditional absentminded professor and drives secretaries and editors to distraction by requesting lowercase letters when spelling his name. He roams the world collecting old toys and sliderules and riding old railways.

FURTHER READING

Aziz P. M., H. V. Sorenson, and J. Van Der Spiegel, An overview of sigma delta converters: How a 1-bit ADC achieves more than 16-bit resolution, *IEEE Signal Process. Mag.* **13**(1): 61–84 (Jan. 1996).

Candy J. C. and G. C. Temes, *Oversampling Delta-Sigma Data Converters: Theory, Design, and Simulation*, IEEE Press, 1992.

Dick C. and F. Harris, Narrow-band FIR filtering with FPGAs using sigma-delta modulation encoding, *J. VLSI Signal Process. Signal Image Video Technol.* **14**(3): 265–282 (Dec. 1996).

Dick C. and F. Harris, FPGA signal processing using sigma-delta modulation, *IEEE Signal Process. Mag.* **17**(1): 20–35 (Jan. 2000).

Harris F. and B. McKnight, Error feedback loop linearizes direct digital synthesizers, *28th Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA. Oct. 30–Nov. 1, 1995.

Jayant N. S. and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984, Chaps. 7 and 8.

Norsworthy S. R., R. Schreier, and G. C. Temes, *Delta-Sigma Data Converters: Theory, Design, and Simulation*, IEEE Press, 1997.

Uchimura K., T. Hayashi, T. Kimura, and A. Iwata, Oversampling A-to-D and D-to-A converters with multistage noise shaping modulators, *IEEE Trans. Acoust. Speech Signal Process.* **AASP-36**: 1899–1905 (Dec. 1988).

SIGNAL PROCESSING FOR MAGNETIC-RECORDING CHANNELS

EVANGELOS ELEFTHERIOU
IBM Research
Zurich Research Laboratory
Rüschlikon, Switzerland

1. INTRODUCTION

The main driving force of progress in magnetic-recording technology has been the need for vast and reliable storage. In the past four decades, the areal density of disk drives has increased ten million fold, leading to dramatic reductions in storage cost. In particular, the storage densities of high-end disk drives have been growing at a compound growth rate of 60% annually, starting in 1991. This is due to the introduction of the magnetoresistive (MR) recording head, and the advances in high-performance data and servo channels, as well as in VLSI technology [1]. Today's commercial disk drives use longitudinal recording and can store information at a density of approximately 50 Gbits per square inch. This unprecedented areal density is achieved while maintaining the stringent on-track error rate requirements of 10^{-8} to 10^{-9} before outer error-correction coding.

It is expected that this phenomenal growth in areal density of longitudinal recording will slow down because of the superparamagnetic effect [1]. This has increased research and development efforts in perpendicular recording, which since its inception [2,3] has promised to achieve much higher areal densities than longitudinal recording can. Although indications exist that perpendicular recording may be able to achieve ultra-high areal densities, the most recent laboratory demonstrations indicate that longitudinal recording is still marginally ahead of its perpendicular counterpart. Ultimately, perpendicular recording promises areal densities that are about four to five times higher than those of longitudinal recording [4,5]. However, there are considerable engineering challenges associated with the realization of this promise [6]. A transition from longitudinal to perpendicular recording would involve changes in various disk-drive subsystems, including the head, disk, head/disk interface, and servo. It is expected, however, that from a signal-processing and coding architecture point of view, the read electronics will not undergo substantial changes.

Although advances in head and media technologies have historically been the driving force behind areal density growth, digital signal processing and coding are increasingly recognized as a cost-efficient approach in achieving substantial areal density gains while preserving the high reliability of disk drives. The general similarity of the write/read process in a hard-disk drive to transmission and reception in communication systems has led to the adoption of adaptive equalization and coding techniques to the magnetic-recording channel. The classical communication channel perspective can be applied not only to study equalization, detection, and both inner and outer coding strategies but also to estimate the ultimate information-theoretic limits of longitudinal and perpendicular recording [7–9].

In the past decade, several digital signal-processing and coding techniques have been introduced into hard-disk drives to improve the error-rate performance at ever increasing normalized linear densities as well as to reduce manufacturing and servicing costs. In the early 1990s, partial-response class-4 (PR4) shaping in conjunction with maximum-likelihood sequence detection [10] replaced the peak detection systems employing run-length-limited (RLL) (d, k)-constrained codes, and paved the way for future applications of advanced coding and signal-processing techniques. For example, at moderate storage densities, the introduction of partial-response maximum-likelihood detection (PRML) [10] requires a new class of inner constrained codes. This class of codes, collectively known as PRML(G, I) codes, facilitates timing recovery and gain control, and limits the path memory length of the sequence detector. At higher normalized linear recording densities, generalized partial-response (PR) polynomials with real coefficients reduce noise enhancement at the equalizer output. In particular, shaping polynomials of the form $(1 - D)(1 + p(D))$ and $(1 - D^2)(1 + p(D))$, where D represents the unit delay and $p(D) = \sum_{\ell=1}^{L} p_\ell D^\ell$ is a finite impulse response predictor filter with real coefficients $\{p_\ell\}$, are significant in practice. Generalized PR channels in conjunction with sequence detection give rise to noise-predictive maximum-likelihood (NPML) systems [11–15]. The extension of the NPML detection scheme to handle data-dependent medium noise was proposed in [16–20].

Parallel to the developments on NPML detection for magnetic recording, maximum transition run (MTR) (j, k) codes were introduced as a means to provide coding gains for extended partial response (E^2PR) channels [21]. The theoretical underpinning of this new class of codes and practical constructions of codes that deal with the problem of quasi-catastrophic error propagation are presented in Ref. [22]. The unifying theory presented in Ref. [22] led to an exhaustive characterization of quasi-catastrophic error-propagation-free MTR codes, called MTR(j, k, t) codes, and revealed a connection between the conventional PRML(G, I)-constrained codes used in disk drives and the recently discovered MTR codes.

More recently, constrained codes, such as the PRML(G, I) and MTR(j, k, t) codes, have been combined with multiparity block codes to improve the bit error rate performance of the inner channel even further — at the expense, sometimes, of a slight decrease in code rate. These multiparity linear inner codes deliver substantial gains in performance when decoded by a so-called *soft* post-processor that follows the NPML detector and utilizes some form of reliability information [23–28]. Currently, a 16-state NPML detector for a generalized PR channel with a first-order null at dc followed by a post-processor for soft decoding of a combined multiparity/constrained code represents the de facto industry standard.

In this article the state of the art in signal processing and constrained coding for hard-disk drives is reviewed, with emphasis on the techniques that have been used in commercial systems. In Section 2 the magnetic recording system as a communications channel is introduced, and the various sources of noise are presented. In Section 3 the three families of modulation codes, namely, RLL(d, k), PRML(G, I), and MTR(j, k, t), that have predominantly been used in storage systems are discussed, and the latest developments on combined modulation/parity codes are presented. In Section 4 equalization and detection techniques are discussed, with emphasis on the PRML and NPML detection, and performance results are given. The soft-decoding methods for the combined modulation/parity codes via post-processors are presented in Section 5, and the NPML detector for data-dependent noise is reviewed in Section 6. Finally, Section 7 contains a brief discussion of future trends in signal processing and coding for magnetic storage.

2. MAGNETIC-RECORDING SYSTEM

2.1. Saturation Recording

In hard-disk drives, data is stored by longitudinal magnetization of a layer of magnetic media that has been deposited on a rigid disk. The data is recorded in concentric circles, called *tracks*, by applying an external field using a write head flying over the spinning disk. In general, the recording process is nonlinear, primarily because of the nonlinear nature of the magnetic medium. However, if the applied field exceeds a critical value the magnetization is

completely polarized in one direction. Therefore, at any point along a track, the magnetization can, in principle, be uniformly polarized in one of two possible directions, reflecting a recorded "0" or "1." This approach to storing information is referred to as saturation recording.

Writing is most commonly performed by inductive heads. When a current is applied, a magnetic field is generated through the head and across its gap. Depending on the amplitude of the applied current, the fringe field that escapes the gap can be strong enough to saturate the magnetic medium. By reversing the polarity of the current flowing through the coil, which is wound around the head, the direction of the magnetic field and consequently the direction of the magnetization of the medium are reversed. The size of the magnetization pattern depends on the rate by which the polarity of the current through the write head changes (data rate) and on the spinning velocity of the disk.

Magnetoresistive (MR) heads have become the prevailing magnetic-recording sensor for reading back the stored information because of the higher sensitivity and lower noise than the inductive heads. The term *magnetoresistive* refers to the physical phenomenon in which the resistivity of a metal changes in the presence of a magnetic field. Therefore, as the MR head flies over a spinning disk, the changes in the magnetic field emanating from the disk translate into changes of the resistivity of the head and consequently into changes of the voltage across the head. Because the MR head directly measures the flux from the medium, the head signal is independent of the velocity of the disk. As a consequence, the MR-head readback signal is as strong for high-RPM (revolutions per minute) server-class drives as it is for slowly rotating, mobile hard-disk drives.

2.2. A Communications Channel

One may consider a magnetic-recording system as a communications channel in which information, rather than being transmitted from one point in space to another, is stored at one point in time and retrieved at another. The objective in magnetic recording is to maximize the rate at which data is stored and retrieved as well as to maximize the density of the information stored per square inch of disk space. Figure 1 shows the general architecture of the recording system in commercial magnetic hard-disk drives.

User data are organized in sectors, with each sector typically containing 512 8-bit bytes. Upon a write request, the data are organized into sectors and fed into a Reed–Solomon (RS) encoder. The RS codewords are first byte-interleaved and then encoded by a constrained or modulation code. In storage systems, the latter code imposes constraints on the data stream being stored in the medium. In magnetic storage, modulation or constrained codes are used, for example, to facilitate the operation of the conventional peak detector, to provide timing information and eliminate quasi-catastrophic error propagation in sequence detection, or to eliminate certain predominant error events and increase the Euclidean distance between output sequences as seen by certain types of sequence detectors. Modulation codes usually employ a precoder, either of the form $1/(1 \oplus D)$ (in the RLL and MTR cases) or $1/(1 \oplus D^2)$ (in the (G, I) case), where $\oplus$ denotes modulo-2 addition. The main function of the precoder is to facilitate modulation code design. In today's recording systems the modulation encoder is followed by a parity encoder that usually appends one, two, three, or four parity bits to the modulation codeword. The output of

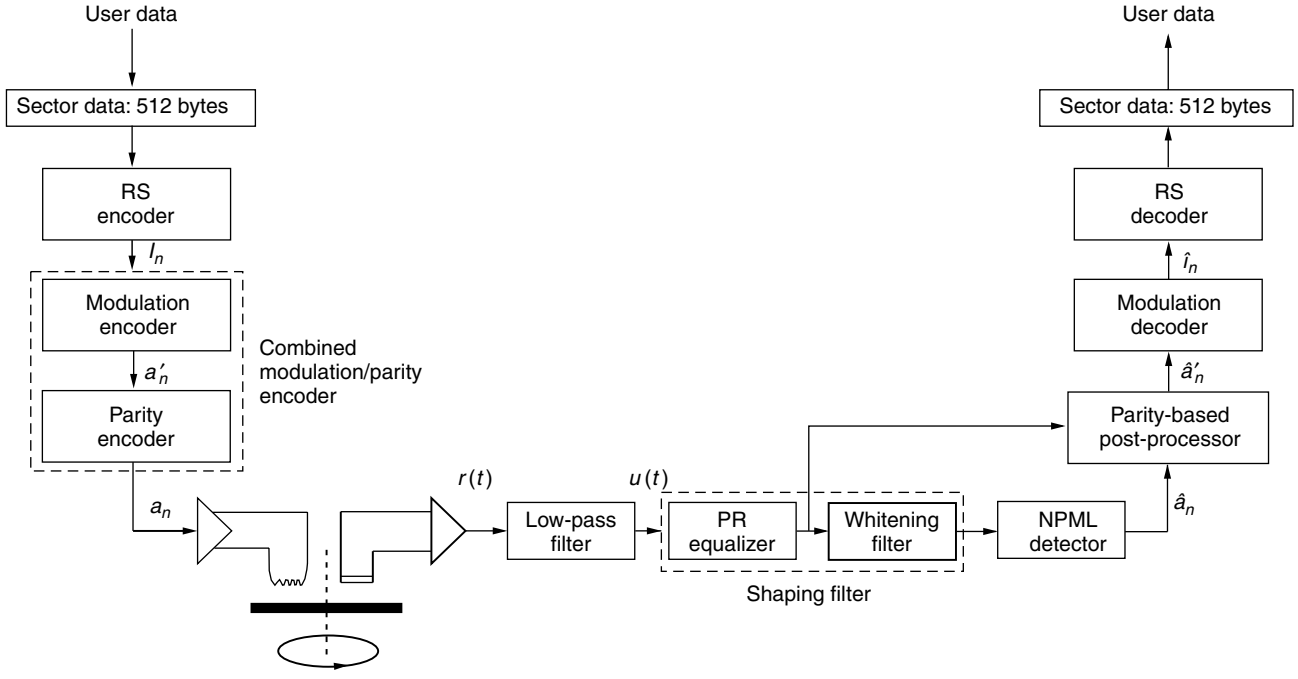


Figure 1. Recording system architecture.

the parity encoder is finally mapped into two-level channel input symbols, $a_i \in \{+1, -1\}$. These are the symbols that are actually being stored in the magnetic medium. The symbol sequence is first converted into current levels by the write head driver, and then the head stores the resulting sequence of rectangular pulses on disk tracks in the form of a series of transitions or magnetization waveforms.

For completeness, it is worthwhile to mention that there are two different formats for mapping the binary data sequence at the output of the modulation encoder to write current waveforms as shown in Fig. 2. In non-return-to-zero inverted (NRZI) recording, a binary 1 corresponds to a change in polarity of the write current, whereas a binary 0 corresponds to no change in polarity of the write current. In non-return-to-zero (NRZ) recording, the binary information sequence is mapped directly to the amplitude level of the write current waveform. From the signaling point of view, these two recording formats are related via the simple $1/(1 \oplus D)$ precoding function as shown in Fig. 3. As can be seen, the current waveform corresponding to NRZI recording of a particular data-input sequence is identical to the NRZ current waveform of the $1/(1 \oplus D)$ precoded data-input sequence. The NRZI recording format played an important role for peak-detection systems, whereas in today's sequence-detection systems, NRZ is the preferred recording format.

To read the data back from the hard disk, a read head is used that senses the changes of the magnetic flux that reflect the changes in polarity of the data sequence stored.

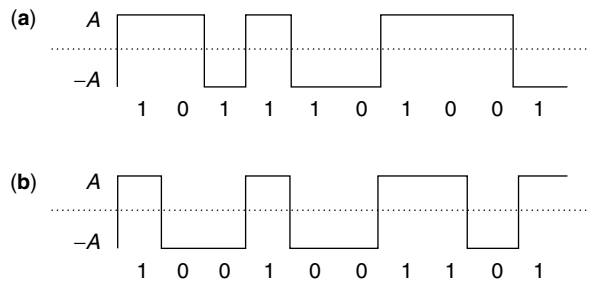


Figure 2. (**a**) NRZI and (**b**) NRZ signals.

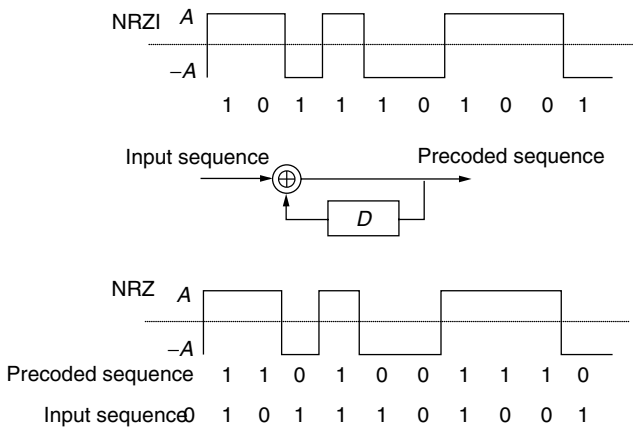


Figure 3. Relationship of NRZI and NRZ.

For an isolated positive transition, the response of the read head is a pulse $s(t)$, which is referred to as the isolated transition response. Analytically this isolated transition or step response is very well approximated by the Lorentzian pulse

$$s(t) = \frac{A}{1 + (2t/\mathrm{PW50})^2} \tag{1}$$

where A is the peak amplitude and PW50 denotes the width of the pulse at its 50% amplitude level. The value of PW50 depends on the physical width of the written transition, the characteristics of the magnetic medium, and the flying height of the head. Figure 4 shows the Lorenztian pulse response as well as the response to an isolated transition known as the Potter pulse [29] for the same value of PW50 and $A = 1$. The Lorentzian pulse is a single-parameter model, whereas the Potter pulse depends on the geometry of the head, the head-to-medium spacing, and the transition parameter [30, Chapter 6]. The Potter analytical pulse appears to be more appropriate for modeling the response of a MR head [30].

The data signal is read back via a low-pass filter (LPF) and a variable gain amplifier (VGA) as an analog signal, $r(t)$. The signal $r(t)$ is sampled periodically at times $t = iT$ to obtain a sequence of samples. The functions of the sampling device and VGA unit are controlled by the timing recovery and gain control loops, respectively. The sequence of samples is first shaped into a suitable PR signal format by the equalizer. The whitening filter then whitens the total distortion at the output of the equalizer, and the NPML sequence detector provides an initial estimate, $\hat{a}_i$, of the channel input symbols. Note that the functions of the PR equalizer and the whitening filter can be combined into a single filter. The initial estimate of the encoded data $\hat{a}_i$ coupled with the equalizer output, that is, hard and soft information, is fed to a noise-predictive parity-based post-processor. The post-processor is a suboptimum soft decision decoder for the parity code that corrects a specified number of the most likely error events at the output of the NPML detector by exploiting the parity information in the incoming sequence. The post-processor produces

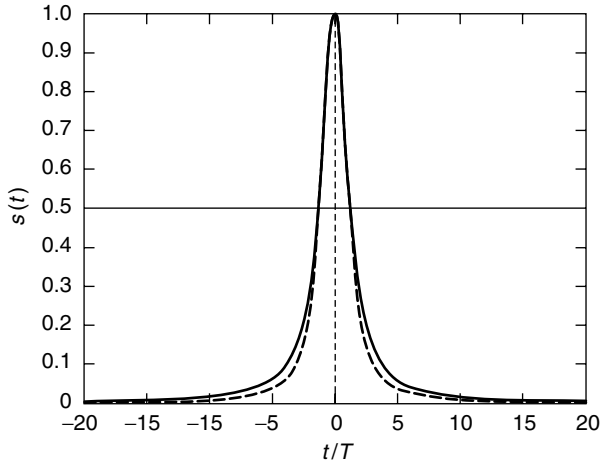


Figure 4. Isolated transition response in a magnetic-recording system. Solid line: Lorentzian pulse, dashed line: Potter pulse.

the final estimate $\hat{a}_i'$ of the modulation-encoded symbols. This sequence is fed to the modulation decoder, which delivers the final decisions. The overall performance of the recording system is very sensitive to error propagation at the modulation decoder. Thus, in practice, the design of modulation decoders with limited error propagation is crucial. Finally, after byte-deinterleaving, the RS decoder corrects any residual errors to obtain a highly reliable estimate of the original user data stored.

2.3. Linear Intersymbol-Interference Channel Model

As mentioned above, the two-level channel-input sequence, $\{a_i\}$, modulates a current source to produce a rectangular current waveform with amplitude -1 or $+1$. This rectangular current waveform can be viewed as the input to the write-head/medium/read-head assembly that produces an output voltage waveform that in essence is the differentiated and low-pass-filtered version of the input waveform corrupted by noise [31,32]. Experimental data have demonstrated that with write precompensation of magnetization transition shifts, the readback waveform is very well approximated by the sum of the appropriate series of isolated transition responses. Adopting therefore a linear model for the write/read process on a magnetic disk, the readback waveform $r(t)$ can be expressed as [31,32]

$$r(t) = \frac{d}{dt}\left[\sum_i a_i \Pi(t - iT)\right] \circledast s(t) + \eta_e(t) \tag{2}$$

where $\Pi(t)$ is a unit-amplitude rectangular pulse of duration T, $s(t)$ is the isolated transition or step response, $\eta_e(t)$ represents the additive white Gaussian electronics noise arising from read head and preamplifier, and $\circledast$ denotes convolution. The effect of medium noise in the case of well-dispersed particulate media can be taken into account by considering a second additive white Gaussian source, $\eta_m(t)$, at the channel input [33]. In this case the model in Eq. (2) becomes

$$r(t) = \frac{d}{dt}\left[\sum_i a_i \Pi(t - iT) + \eta_m(t)\right] \circledast s(t) + \eta_e(t) \tag{3}$$

It can readily be seen that the readback signal can equivalently be expressed as

$$r(t) = \sum_i a_i [s(t - iT) - s(t - T - iT)] + \eta(t) \tag{4}$$

where

$$\eta(t) = \frac{d\eta_m(t)}{dt} \circledast s(t) + \eta_e(t) \tag{5}$$

and $\dfrac{d\eta_m(t)}{dt} \circledast s(t)$ represents the medium-noise contribution to the total noise. Equation (4) describes a pulse-amplitude-modulated waveform of a binary data sequence $\{a_i\}$ transmitted at a rate of $1/T$ through a dispersive linear channel with effective impulse response $h(t) = s(t) - s(t - T)$ and received in the presence of additive noise. In a magnetic-recording system, the impulse response $h(t)$, which represents the effective impulse response of the overall magnetic-recording channel, is called pulse response because it corresponds to the response of the head/medium to a rectangular pulse. Furthermore, the quantity $D_c = \mathrm{PW50}/T$ is called normalized linear density. For a given PW50, the smaller the symbol interval T, that is, the closer the transitions, the higher the value of D_c, which implies the larger the linear density of the recording system. Conversely, the smaller the symbol interval T, the higher the interaction and overlap between the pulses, which gives rise to intersymbol interference (ISI) in a sequence of data symbols. Today's high-performance digital magnetic-recording systems operate at normalized linear densities in the range of $2.5 \le D_c \le 3.5$, where severe ISI is present in the readback signal. In such a case, the recovery of the data sequence from the readback signal requires advanced equalization and detection techniques that compensate for the presence of ISI.

Alternatively, the readback signal may be expressed in the equivalent form:

$$r(t) = \sum_i b_i s(t - iT) + \eta(t) \tag{6}$$

where $b_i = a_i - a_{i-1}$. In this case the symbol sequence $\{b_i\}$ takes values from the ternary alphabet $\{+2, 0, -2\}$. Figure 5 shows a linear ISI model of the magnetic-recording channel, where $s'(t)$ denotes the first derivative of the isolated transition response $s(t)$. The encoded NRZ data symbols a_n are passed through a $1 - D$ filter. A nonzero output of this filter corresponds to a positive or negative transition of the write current and, consequently, a transition in the magnetization pattern on the disk. The output of the $1 - D$ filter is then fed to a linear filter with impulse response $s(t)$, representing a Lorentzian, Potter, or any other analytic or experimental read-head response to an isolated transition. Finally, the readback signal is generated by adding white and colored Gaussian noise.

In the frequency domain, the overall response of the linear model may be expressed as $H(f) = S(f)[1 - e^{-j2\pi fT}]$, where $S(f)$ indicates the Fourier transform of the isolated transition response. The frequency response characteristics $H(f)$, plotted as a function of the normalized frequency fT and with D_c as a parameter, are illustrated in Fig. 6. As expected, the frequency response exhibits a spectral null at $f = 0$ because of the factor $[1 - e^{-j2\pi fT}]$. Furthermore, there is substantial high-frequency attenuation, which increases as the linear normalized density D_c is increased.

So far a linear channel model has been assumed, that is, the noise-free readback signal can be derived by linear

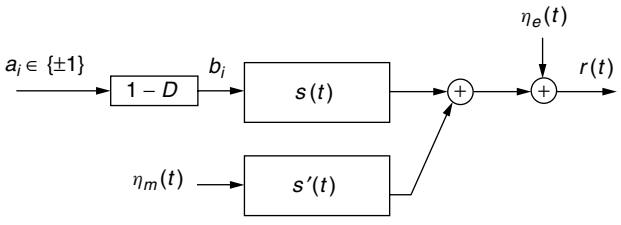


Figure 5. Linear ISI model for a magnetic-recording system.

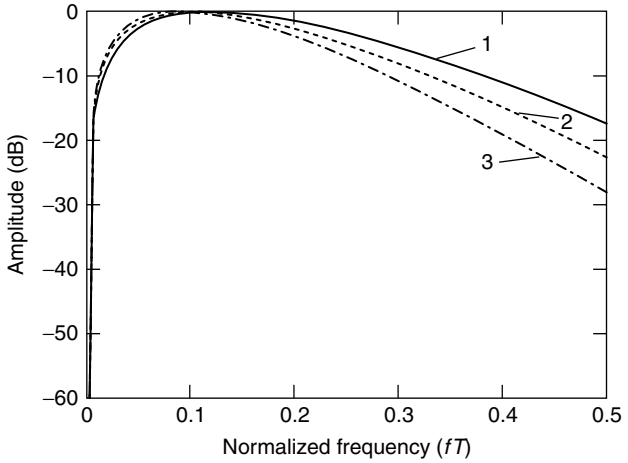


Figure 6. Frequency response to a square pulse of duration T. Curve 1: PW50/$T = 2.5$; curve 2: PW50/$T = 3.0$, and curve 3: PW50/$T = 3.5$.

superposition of isolated step responses, and the noise sources are stationary and additive. At high recording densities, the noise tends to be colored and enhanced by the linear equalizer, the nonstationary data-dependent noise becomes more dominant, and the signal becomes inherently nonlinear. Finally, today's recording systems employ thin-film media, which are dominated, as we will see below, by nonadditive noise sources.

2.4. The Microtrack Model

Recording systems using thin-film media may also exhibit a nonadditive noise known as transition noise. In general, the transition noise is due to fluctuations concentrated close to the recorded transition centers and is attributed to the random microstructural properties of the grains in thin-film recording media [34]. The effect of transition noise on equalization and data detection is more difficult to analyze than that of Gaussian noise owing to its nonstationary and data-dependent characteristics. A simple model for transition noise can be obtained by modeling the width and the position of the isolated transition response as random variables. Taking a Taylor series expansion with respect to small random deviations from the nominal value in the position and width, we can arrive at a channel model with multiplicative position jitter and pulse-widening noise sources [35].

The microtrack model developed in [18] is more general and allows a rather accurate modeling of the noise that occurs when a transition is written in thin-film magnetic-recording media. This model imitates the random zigzag effects when a transition is written. The random zigzag form of a transition is captured by dividing the recording track into N equally-sized microtracks. Figure 7 shows a track modeled by $N = 4$ microtracks, where the vertical dashed lines indicate the ideal positions of two consecutive transitions. The arrows show the direction of magnetization, whereas the short vertical lines on each microtrack indicate the corresponding positions of an instantaneous magnetization reversal. The positions of these instantaneous reversals or flips follow a specified probability density function and are chosen independently for each microtrack. Therefore, if $s(t)$ is the response to an ideal isolated transition across the entire track, then the response of the ℓ-th microtrack to a magnetization reversal at position τ_ℓ is $s(t - \tau_\ell)/N$ [18]. The noiseless output of the magnetic-recording channel to a single transition is then given by

$$\hat{s}(t) = \frac{1}{N} \sum_{\ell=1}^{N} s(t - \tau_\ell) \tag{7}$$

where τ_ℓ is the random shift associated with the ℓ-th microtrack. This random shift or jitter, τ_ℓ, is modeled as an independent and identically distributed (i.i.d.) process according to the derivative of the average cross-track magnetization profile. If the average cross-track magnetization profile has a *tanh*-shape, the jitter probability density function (pdf) is given by [34]

$$p_\tau(\tau) = \frac{1}{\pi a} \operatorname{sech}^2\left(\frac{2\tau}{\pi a}\right) = \frac{1}{\pi a} / \cosh^2\left(\frac{2\tau}{\pi a}\right) \tag{8}$$

In the case of an *erf*-shaped cross-track magnetization profile, the pdf is given by [18]

$$p_\tau(\tau) = \frac{1}{\pi a} \exp\left\{-\left(\frac{\tau}{a\sqrt{\pi}}\right)^2\right\} \tag{9}$$

In both cases a is known as the transition-width parameter, a quantity usually determined experimentally.

The microtrack model is specified by two parameters: the number of microtracks N and the transition-width parameter a. In its more general form, a third parameter, L_e, can be introduced that characterizes partial erasure and its effects. This parameter specifies the threshold below which two adjacent transitions on the same microtrack erase each other [18], see Fig. 7. The microtrack model allows a separate analysis of the write process and the transition noise. The measure of goodness of the write process is the steepness of the cross-magnetization profile. This steepness depends on the transition-width parameter a. An ideal transition

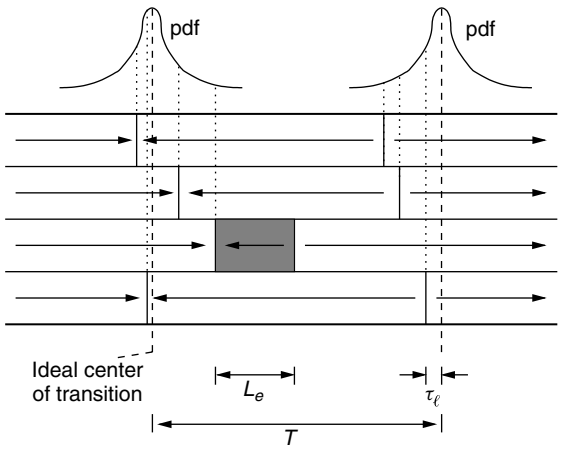


Figure 7. Microtrack model with $N = 4$ microtracks and erasure on the second microtrack.

has width zero, that is, all magnetic particles change their polarization at the same location. However, real transitions exhibit a *tanh*-shape, an *erf*-shape, or another experimentally determined cross-polarization profile. The transition- or data-dependent medium noise produced by the microtrack model is the difference between the actual output and the expected or average output with respect to the jitter distribution, that is, $\hat{s}(t) - E[\hat{s}(t)]$. As the number of microtracks increases, the transition noise decreases. By allowing $N \to \infty$ then $\hat{s}(t) \to E[\hat{s}(t)] = s(t) * p_\tau(t)$, and the transition noise is zero. In this case only electronics noise and the nonideal write process affect the performance of the magnetic-recording system via the transition-width parameter a. Another way to eliminate the effects of transition noise is to set $a = 0$. In this case $E[\hat{s}(t)] = s(t)$, and the magnetic-recording system behaves according to the linear ISI model described above.

Using Taylor's series expansion and keeping only the first two predominant terms, the transition noise for a single transition can be approximated by [18]

$$\begin{aligned}\hat{s}(t) - E[\hat{s}(t)] &= -s'(t)\left[\frac{1}{N}\sum_{\ell=1}^{N}\tau_\ell\right] \\ &\quad + s''(t)\left[\frac{1}{2N}\sum_{\ell=1}^{N}\tau_\ell^2 - \frac{1}{2}E[\tau_\ell^2]\right] \\ &\quad + \cdots \simeq s'(t)n_1 + s''(t)n_2\end{aligned} \tag{10}$$

where $n_1 = \dfrac{1}{N}\sum_{\ell=1}^{N}\tau_\ell$ and $n_2 = \dfrac{1}{2N}\sum_{\ell=1}^{N}\tau_\ell^2 - \dfrac{1}{2}E[\tau_\ell^2]$ are zero-mean random variables [18]. The random variable n_1 controls the amount of position jitter noise and is referred to as the position jitter variable, whereas the random variable n_2 controls the amount of pulse-width variation noise and is referred to as the width variable.

In general, the readback signal according to the microtrack model is expressed as

$$r(t) = \frac{1}{N}\sum_i b_i \sum_{\ell=1}^{N} s(t - iT - \tau_{\ell,i}) + \eta_e(t) \tag{11}$$

where $\tau_{\ell,i}$ is the random shift or jitter associated with the ℓ-th microtrack at the i-th symbol interval, and $\eta_e(t)$ represents additive white Gaussian noise (AWGN) characterized by its one-sided power spectral density N_0. Thus, the behavior of the magnetic-recording system can be described by the five-parameter (PW50/T, N, L_e, a, and N_0) model as shown in Fig. 8. Although for the magnetic-recording channel this model is quite versatile and encompasses the effects of both medium noise in thin-film media and electronics noise, the linear ISI model with additive Gaussian noise is still commonly used to estimate the performance of detection and coding schemes, even at ultra-high densities, and yields very accurate results.

3. MODULATION CODES

The basic principles of modulation coding, also known as coding for input-constrained channels, was established in the classic study of discrete noiseless channels [36]. With modulation coding a desired constraint is imposed on the data-input sequence so that the encoded data stream satisfies certain properties in the time or frequency domain. These codes are very important in digital-recording devices and have become ubiquitous in all data-storage applications [37].

A constrained system is represented by a labeled, directed graph, or a finite-state transition diagram (FSTD). An FSTD consists of states (or vertices) and labeled, directed transitions between states such that the allowable constrained sequences are precisely the sequences obtained by traversing paths of the diagram. The FSTD is called deterministic if at each state all outgoing transitions have distinct labels. The capacity of a constrained set of sequences represents the maximum achievable code rate of an encoder generating sequences satisfying the underlying constraint. It is given by $\log_2 \lambda_{\max}(A)$, where $\lambda_{\max}(A)$ is the largest real eigenvalue of the adjacency matrix associated with a deterministic FSTD that represents the constrained set of sequences. Among the various methods for constructing modulation codes, the state-splitting algorithm in [38] provides a systematic approach to designing finite-state encoders and sliding-block decoders for finite-type constrained systems. In practice, however, the right choices must be made during the code-construction procedure, irrespective of the approach used for code design. For example, quite often look-ahead coding techniques yield more efficient designs than the systematic state-splitting approach does. A more detailed treatment of finite-state modulation codes and their application to digital data recording can be found in [37,39].

In this section the most important classes of modulation codes will be reviewed, with emphasis on the application to magnetic recording, in particular, to hard-disk drives. Finally, the combination of the PRML(G, I) and MTR(j, k, t) codes with parity block codes in order to improve the bit error rate performance will also be discussed.

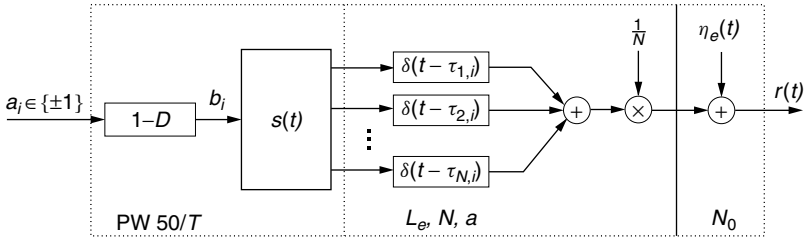


Figure 8. Data-dependent noise channel model for a magnetic-recording system. From [9], ©; IEEE 2002.

3.1. RLL(d, k) Codes

Modulation codes that impose a restriction on the number of consecutive 1s and 0s in the encoded data sequence are generally called RLL(d, k) codes. The code parameters d and k are nonnegative integers with k always larger than d, where d indicates the minimum number of 0s between two 1s and k indicates the maximum number of zeros between two 1s. At low-linear recording densities, peak-detection systems employing RLL(d, k)-constrained codes have been predominant in digital magnetic storage. RLL(d, k) codes reduce the effect of pulse interference and prevent the loss of clock synchronization. When used with the NRZ recording format, the d-constraint has the effect of spreading the transitions by at least $d + 1$ symbols further apart, thereby minimizing intersymbol interference and nonlinear distortion. The k-constraint sets an upper limit on the run of identical symbols to $k + 1$ so that useful timing information can always be extracted from the readback signal.

The set of sequences that satisfy the (d, k) constraints can be generated by the FSTD shown in Fig. 9. For (d, k)-constrained sequences the capacity can be computed as the base-2 logarithm of the largest real solution of one of the following equations [40]:

$$x^{k+2} - x^{k+1} - x^{k+1-d} = 1, \qquad k < \infty \tag{12}$$
$$x^{d+1} - x^{d} = 1, \qquad k = \infty,$$

Table 1 lists the capacity Cap(RLL(d, k)) for various values of d and k. The use of (d, k)-constrained sequences, with $d \geq 1$, allows the information density along a track to be increased while keeping the separation between adjacent recorded transitions fixed. The quantity $(d + 1)$Cap(RLL(d, k)), called the density ratio or packing density [37,41], is a direct measure of the increase in linear recording density as a function of d and of the capacity of the (d, k)-constrained sequences. Clearly, the packing density can be made arbitrarily large by increasing d. Conversely, large values of d lead to codes with very low rate, which implies high-recording symbol rates, and thus renders these codes impractical for storage systems that are limited by the clock speed. Moreover, large values of d lead to a more sensitive timing and detection window. In practical recording systems, codes with $d \leq 2$ have been used. In particular, the rate-2/3 RLL$(1, 7)$ (packing density of 4/3) and rate-1/2 RLL $(2, 7)$ (packing density of 3/2) codes have been widely used in the digital-recording industry at low-normalized densities around $D_c \simeq 1$ [37].

There are several approaches for constructing a rate-2/3 RLL$(1, 7)$ code. The simple construction in Ref. [42] is based on the code assignment of Table 2. Freely concatenating any of the four codewords from this list would violate the $d = 1$ constraint at codeword boundaries. The substitutions given in Table 3 avoid the violations at code boundaries and give rise to a *look-ahead encoder* and *sliding-block decoder*. The encoding table of the other popular rate-1/2 RLL$(2, 7)$ code is illustrated in Table 4.

3.2. PRML(G, I) Codes

The PRML scheme introduced in the early 1990s to advanced hard-disk drives operating at moderate linear recording densities, that is, approximately in the range $1.5 \leq D_c \leq 2.3$, requires a different type of constraint called (G, I) constraint [10,41]. The need for this new constraint is twofold. First, long runs of zero samples (noise-free samples) at the PR4 equalizer output can degrade the tracking performance of the timing recovery

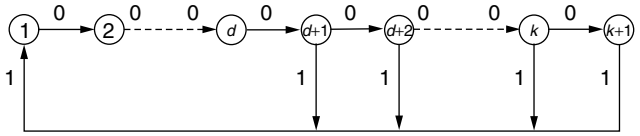


Figure 9. FSDT for RLL(d, k)-constrained sequences.

Table 1. Capacity of RLL(d, k) Constraints

k	$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d = 4$
1	0.6942				
2	0.8791	0.4057			
3	0.9468	0.5515	0.2878		
4	0.9752	0.6174	0.4057	0.2232	
5	0.9881	0.6509	0.4650	0.3218	0.1823
6	0.9942	0.6690	0.4979	0.3746	0.2669
7	0.9971	0.6793	0.5174	0.4057	0.3142
8	0.9986	0.6853	0.5293	0.4251	0.3432
9	0.9993	0.6888	0.5369	0.4376	0.3620
∞	1.00	0.6942	0.5515	0.4650	0.4057

Table 2. Encoder for the Rate-2/3 RLL(1, 7) Code

Information bits	Encoded bits
00	101
01	100
10	001
11	010

Table 3. Substitution Table for the Rate-2/3 RLL(1, 7) Code

Prohibited Pattern	Substitute Pattern
101,101	101,000
101,100	100,000
001,101	001,000
001,100	010,000

Table 4. Encoder/Decoder for the Rate-1/2 RLL(2, 7) Code

Information Bits	Encoded Bits
10	0100
11	1000
000	000100
010	100100
011	001000
0010	00100100
0011	00001000

and gain control loops. Thus, as in peak detection systems and the k-constraint described above, it is necessary to introduce a global constraint G to limit the number of zero samples at the equalizer output. Second, another constraint is necessary to limit the path memory requirements and hence force a finite decision delay in PRML detection, without incurring any significant performance degradation in the sequence of estimates produced by the detector. This constraint, referred to as the I-constraint, is related to the so-called quasi-catastrophic error propagation of PR systems [43], and imposes a limitation on the length of runs of zero samples in each of the odd and even interleaved subsequences at the PR4 equalizer output. In general, a trellis is called quasi-catastrophic if there are distinct states for which some of the output sequences starting from these states are identical and the total probability of all such indistinguishable output sequences is zero [43].

The issue of undesired sequences and the application of PRML(G, I) codes to eliminate them are more general and not restricted to PR4 shaping only. The above discussion has indicated that long strings of zeros at the PR4 channel output as well as in each of the subsequences of even and odd bit positions at the PR4 channel output constitute a set of undesired sequences that need to be eliminated. It can readily be shown that the same holds for all types of generalized PR shaping polynomials of the form $(1-D)^m(1+D)^n(1+p(D)), n, m \geq 1$, where $p(D)$ has no roots on the unit circle [22].

To facilitate the timing and gain control algorithms, in general only those channel input sequences need to be eliminated that have spectral energy at the frequencies where the generalized PR polynomial used for shaping has spectral nulls. For example, for shaping polynomials of the form $(1-D)^m(1+p(D)), m \geq 1$, exhibiting and m-th order spectral null at dc, the channel input sequences $(+1)$ and (-1) with a spectral null at dc should be eliminated. The notation (S) indicates the sequence obtained by periodically repeating the string S infinitely many times, e.g., $(ab) = ababab\cdots$. The k-constraint presented above limits the maximum length of 0s at the input of the $1/(1 \oplus D)$ precoder to k or, equivalently, the length of channel input patterns (after precoding) of type $(+1)$, (-1) to $k+1$. Similarly, it can readily be seen that for an arbitrary channel-shaping polynomial with spectral nulls at both the dc and the Nyquist frequency, the channel input sequences $(+1-1)$ and $(-1+1)$ with a spectral null at the Nyquist frequency should also be eliminated. For example, for channel-shaping polynomials of the form $(1-D^2)(1+p(D))$, the channel input sequences $(+1)$, (-1), with a spectral line at dc, as well as the sequences $(+1-1)$, $(-1+1)$, with a spectral line at the Nyquist frequency, should be eliminated. The G-constraint mentioned above limits the maximum length of 0s at the input of the $1/(1 \oplus D^2)$ precoder to G or, equivalently, the maximum length of channel input patterns of all four types $(+1)$, (-1), $(+1-1)$ and $(-1+1)$ to $G+2$ [22].

In connection with the PR4 shaping polynomial above, we have seen that another desirable code property is the elimination of quasi-catastrophic error propagation, which is inherent in maximum likelihood sequence detection of any generalized PR channel with spectral nulls [43]. This property allows the path memory size of the sequence detector to be reduced without degrading its bit error rate performance. Quasi-catastrophic error propagation is prevented by eliminating all channel-input error sequences $\{\varepsilon_i\} = \{a_i - \hat{a}_i\}$ that have spectral energy at those frequencies where the channel has spectral nulls. For generalized PR polynominals of the form $(1-D)^m(1+p(D)), m \geq 1$, the k-constraint at the input of a $1/(1 \oplus D)$ precoder is sufficient to eliminate quasi-catastrophic error propagation, because it limits the maximum length of channel-input error patterns of type $(+2)$, (-2) to $k+1x$. Conversely, for shaping polynominals that also exhibit spectral nulls at the Nyquist frequency, such as for example the shaping polynominal $(1-D^2)(1+p(D))$, which is very often used in practical systems, more channel-input error patterns need to be eliminated. Specifically, it can readily be seen that in this case it is necessary, and sufficient, to limit the maximum length of channel-input error patterns of the type $(+2)$, (-2), $(+2\,-2)$, $(-2\,+2)$, $(+2\,0)$, $(0\,+2)$, $(-2\,0)$, $(0\,-2)$. In general, these channel-input patterns exhaustively characterize the undesired quasi-catastrophic sequences for arbitrary shaping polynomials of the form $(1-D)^m(1+D)^n(1+p(D))$, where $n, m \geq 1$ (see also [44]). The I-constraint discussed above at the input of a $1/(1 \oplus D^2)$ precoder limits the maximum length of channel-input error patterns of type $(+2)$, (-2), $(+2\,-2)$, $(-2\,+2)$ to $2I+2$, and of type $(+2\,0)$, $(0\,+2)$, $(-2\,0)$, $(0\,-2)$ to $2I+3$. Note that an additional G-constraint, $G \leq 2I$, further reduces the maximum length of error patterns of type $(+2)$, (-2), $(+2\,-2)$, $(-2\,+2)$ [22].

The most widely used code rates for PRML(G, I) codes in the industry have been 8/9 and 16/17. The optimal block lists of length 9 for constructing block-encodable/decodable rate-8/9 PRML$(4, 4)$ and rate-8/9 PRML$(3, 6)$ codes can be found in [45]. By slightly relaxing the G and I parameters, it was also possible to construct a rate-16/17 PRML$(6, 6)$ code [46].

3.3. MTR Codes

The MTR(j, k) codes introduced in [21] result in a direct coding gain or improved performance by eliminating some of the predominant error events in conjunction with sequence detection. In addition to the benefits of an enhanced coding gain, MTR codes are useful in controlling nonlinear phenomena associated with the fast switching of the write head. More specifically, the MTR constraints are characterized by the parameters j, the maximum number of consecutive 1s that can occur, and k, the maximum number of zeros that can occur. Figure 10 shows the FSTDs generating sequences according to the MTR$(j = 2, k = \infty)$ and MTR$(j = 3, k = \infty)$ constraints. The FSTD of the MTR$(j = 2, k = 5)$ constraint is illustrated in Fig. 11. It can readily be verified that in all sequences obtained by traversing the paths of the diagram in Fig. 11, the consecutive runs of 1s are limited to two. Also, the runs of consecutive 0s are limited to five for timing-recovery purposes. The capacities of these FSTDs are Cap(MTR$(j = 2, k = 5)) = 0.8578$, Cap(MTR$(j = 2, k = \infty)) = 0.8791$, and Cap(MTR$(j = 3, k = \infty)) = 0.9468$.

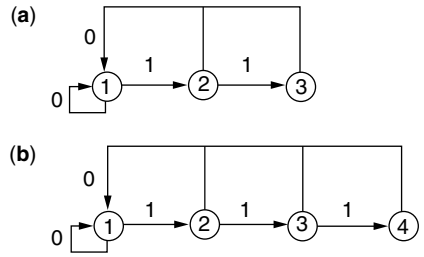


Figure 10. FSTDs for (**a**) MTR($j = 2, k = \infty$) and (**b**) MTR($j = 3$, $k = \infty$) constrained sequences.

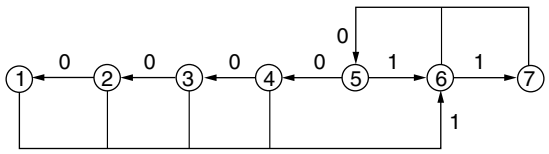


Figure 11. FSTD for MTR($j = 2, k = 5$) constrained sequences.

In general, MTR(j, k) codes do not avoid quasi-catastrophic error propagation in sequence detectors for PR polynomials with spectral nulls at both dc and the Nyquist frequency. The k-constraint avoids channel-input error sequences that have spectral energy only at dc, whereas the j-constraint avoids channel-input error sequences that have spectral energy only at the Nyquist frequency. Therefore, an additional constraint is needed to limit the maximum length of channel-input error sequences of type $(+2\,0)$, $(0\,+2)$, $(-2\,0)$, $(0\,-2)$ that have spectral energy at both dc and the Nyquist frequency. This new constraint for MTR codes, known as the *twins* or t-constraint, has been introduced in [22]. In simple terms the MTR encoder satisfies a t-constraint if it does not allow $t + 1$ consecutive pairs of 0s or 1s (twins). For example if $t = 8$ then nine consecutive twins are not allowed, that is, the string 00 00 11 00 00 00 11 00 11 is not allowed. Sequences that satisfy the t-constraint and at the same time are j- and k-constrained are denoted by MTR(j, k, t) [22]. The derivation of the deterministic FSTD that describes the MTR(j, k, t) constraint is presented in [22]. Figure 12 illustrates a 30-state labeled directed graph for the MTR($j = 2, k = 7, t = 4$) constraint. The capacity of this FSTD is Cap(MTR($j = 2, k = 7, t = 4$) = 0.8591. Tables 5 and 6 list the capacities Cap(MTR($j = 2, k, t$)) and Cap(MTR($j = 3, k, t$)) for various values of the parameters k and t. Codes based on the MTR(j, k, t) constraints eliminate the problem of error propagation in sequence detection for all PR shaping polynomials of the form $(1 - D)^m(1 + D)^n(1 + p(D))$, $m, n \geq 1$. Note that this class of PR polynomials includes PR4, EPR4 (extended PR4), and E^{2}PR4, (extended-square PR4) corresponding to $(1 - D^2)$, $(1 - D^2)(1 + D)$, and $(1 - D^2)(1 + D)^2$ shaping polynomials, respectively, which have been very important in practical systems. For shaping polynomials with memory greater than $j + 1$, the j-constraint can readily be incorporated into the detector to reduce the number of states or branches in the trellis, and to increase the capacity by allowing new potential sequences which were forbidden in MTR(j, k, t) [22].

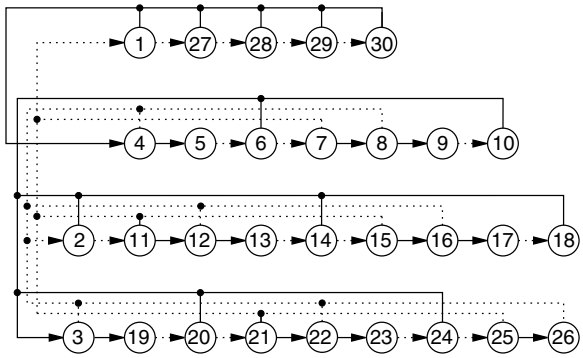


Figure 12. FSTD for MTR($j = 2, k = 7, t = 4$) constrained sequences. Solid line: 1, dotted line: 0.

Table 5. Capacity of MTR($j = 2, k, t$) Constraints

t	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
4	0.85514	0.85915	0.86052	0.86127	
5	0.86259	0.86732	0.86920	0.87026	0.87063
6	0.86567	0.87069	0.87296	0.87424	0.87476
7	0.86699	0.87213	0.87461	0.87599	0.87663
8	0.86756	0.87275	0.87536	0.87678	0.87748
9	0.86782	0.87302	0.87569	0.87714	0.87788
10	0.86792	0.87313	0.87584	0.87730	0.87806
11	0.86797	0.87319	0.87591	0.87737	0.87814
12	0.86799	0.87321	0.87594	0.87740	0.87818

Table 6. Capacity of MTR($j = 3, k, t$) Constraints

t	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
4	0.92830	0.93151	0.93223	0.93260	
5	0.93476	0.93834	0.93969	0.94047	0.94065
6	0.93720	0.94100	0.94265	0.94355	0.94390
7	0.93816	0.94203	0.94385	0.94481	0.94524
8	0.93853	0.94243	0.94434	0.94533	0.94581
9	0.93868	0.94259	0.94454	0.94554	0.94604
10	0.93874	0.94266	0.94462	0.94563	0.94614
11	0.93877	0.94268	0.94466	0.94567	0.94618
12	0.93878	0.94269	0.94467	0.94568	0.94620

In a NRZI format the j-constraint imposes a limit on the maximum number of consecutive transitions in the write current. In particular, the original MTR$(j = 2, k)$ codes that do not allow three consecutive transitions to appear in any encoded sequence have the interesting property of eliminating bit patterns that cause the most common error events in sequence detection. Figure 13 shows typical error patterns that are eliminated by these codes. These error patterns correspond to the NRZ error events of the form $\{\pm 2, \mp 2, \pm 2\}$. It can easily be seen that error bursts of the form $\{\pm 2, \mp 2, \pm 2, \ldots\}$ and length greater than three are also eliminated by the $j = 2$ constraint. These error events correspond to mistaking the polarity of an alternating write current for three or more channel symbol intervals. However, the distance gain realized by eliminating these error events is offset by a significant rate loss penalty. In particular, the maximum possible code rates for the original MTR$(j = 2, k)$ constraint [21]

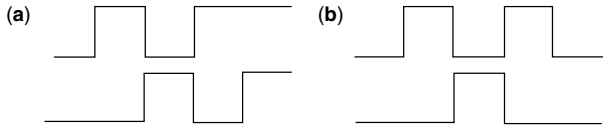


Figure 13. Error patterns eliminated by the $j = 2$ constraint. (**a**) Tribit shift, (**b**) quadbit to dibit.

and the MTR$(j = 2, k, t)$ constraint [22] are less than 8/9, leading to an unacceptable loss of performance due to the low code rate. Time-varying constraints, in which $j = 2$ is observed only at even bit positions (referred to as MTR $j = 2, 3$ constraints), permit the design of higher-rate MTR codes while maintaining their distance gain. The rates of the codes that have been designed to satisfy time-varying MTR$(j = 2, 3, k)$ constraints are 8/9 [47], and slightly higher [48,49]. These rates are still not adequately high for magnetic-recording applications.

An alternative coding strategy is to allow the error event $\{\pm 2, \mp 2, \pm 2\}$ to occur partially and to eliminate all other error events that correspond to mistaking the polarity of an alternating write current for four or more channel symbol intervals. In this way higher-rate codes can be constructed, thereby reducing the signal-to-noise ratio (SNR) penalty due to rate loss. For example, by using a $j = 3$ constraint in conjunction with a j-constrained trellis, all error events of the type $\{\pm 2, \mp 2, \pm 2, \mp 2, \}$, $\{\pm 2, \mp 2, \pm 2, \mp 2, \pm 2\}$, and so on are eliminated. Moreover, also the error event $\{\pm 2, \mp 2, \pm 2\}$ is partially eliminated because the quadbit-to-dibit error cannot occur (see Fig. 13). The rate-16/17 MTR$(j = 3, k = 12, t = 16)$ look-ahead code and the time-varying MTR$(j = 3, 4, k = 18, t = 14)$ block code, described in [22], have this interesting performance-enhancing property and have been implemented in various magnetic-recording systems. The construction of the rate-16/17 block-encodable/decodable MTR$(j = 3, 4, k = 18, t = 14)$ code, as described in [22], will be outlined briefly because of its practical significance. It can be verified that there are in total $65753 > 2^{16}$ potential codewords that can be generated by starting in state two in Fig. 10b, making 17 transitions and terminating in states one, two, or three. Among these codewords there are 199 codewords that begin or end with 10 zeros. After discarding these codewords and 17 more codewords that start with the first 15 bits of one of the strings (1001), (0110), (0011), (1100), or end with the first 16 bits of one of the strings (1001), (0110), (0011), or (1100), a set of 65537 codewords is obtained. These 17-bit codewords can be freely concatenated without violating the $j = 3$ constraint except at the border of two codewords, where the constraint is relaxed to $j = 4$. Furthermore, $k = 18$ and $t = 14$ is obtained.

In general, the performance-enhancing features of the class of high-rate MTR codes render them very attractive compared with conventional PRML(G, I) codes. Of course, ultimately, the coding gain of an MTR code is determined by the tradeoff among various factors such as rate loss, error propagation at the MTR decoder, and performance-improving properties of the MTR code. Finally, it is worthwhile mentioning that for a long time the conventional (G, I) and MTR constraints were discussed separately in the literature, and no connection was made between them. Very recently a theoretical result was presented in [22] showing that the precoded (G, I) constraints are a subclass of the precoded MTR constraints. This very interesting property allows an alternative code-construction methodology for (G, I) codes that is based on employing the $1/(1 \oplus D)$ precoder used by MTR codes.

3.4. Combined Modulation/Parity Codes

As discussed above, PR shaping for the magnetic-recording channel was normally used in conjunction with byte-oriented PRML(G, I) or MTR(j, k, t) modulation codes to aid timing recovery and gain control, to limit the path memory length, and to enhance the performance of the sequence detector. Until recently, these codes have been widely used in the disk-drive industry since the introduction of PR4 and maximum likelihood sequence detection more than ten years ago [10].

Because of the channel memory and the slight noise coloration, sequence detection produces some error patterns more frequently than others. These predominant error patterns or error events at the sequence detector output depend on the generalized PR shaping polynomial, the noise blend, and the normalized linear recording density D_c. For example, Table 7 shows the error events at the 16-state NPML detector output in the case of electronics noise only and $D_c = 3.4$. A shorthand notation to represent ternary error events has been used. For example, $\{+, -, +\}$ is used to denote the events $\{+2, -2, +2\}$ and $\{-2, +2, -2\}$. At high-recording linear densities, the predominant error event is of the type $\{+, -, +\}$. Simulation and experimental data have shown that at low and moderate linear densities the error event $\{+\}$ predominates. Note that the relative percentage of the various error events also depends on the noise conditions, that is, the noise blend between electronics, transition, and colored stationary media noise. These observations suggested the use of modulation codes combined with multiparity block codes to improve performance even further, but at the expense of a slight decrease in code rate. The basic concept is to use parity to detect the presence of an error event from a list of predominant error events. Decoding is achieved by a technique that combines syndrome and soft-decision decoding and is known in the industry as parity-based post-processing [23–28]. A combined modulation/parity code is constructed from a

Table 7. Error Events at NPML Detector Output (PW50/T = 3.4)

Error Events	Relative Frequency
$+-+$	62%
$+$	15%
$+-$	10%
$+-+-+$	4%
$+-+-$	2%
$+\,0\,0+$	1%
$+-+-+-$	1%
other	5%

(G, I) or MTR code by adding parity bits to the (G, I) or MTR codewords. Specifically, the design aims at single or double error-event detection from a prespecified list of error events so that the probability of miscorrection is minimum. Practical reasons, such as complexity and decoding delay, dictate the use of a short list of error events and short codes. Therefore, to keep the code-rate as high as possible, only a small number of parity bits may be used. Typically, the (G, I) codewords are extended by one to four parity bits. In [27] an elaborate approach to designing error-event detection codes is proposed. The code construction methodology is based on a recursive approach for building the parity-check matrix by adding one column at a time. A new column is added to the parity-check matrix if and only if the prespecified error-event detection capability is satisfied [27].

The approach adopted in [23,25,26,28] is based on simple polynomial codes characterized by their generator polynomial $g_c(D)$, whose degree specifies the total number of parity bits. The single-parity code $g_c(D) = 1 \oplus D$ ensures that there is an even number of 1s in the codeword. It can therefore detect error events of odd length. The double-parity code with $g_c(D) = 1 \oplus D^2$ adds two parity bits such that both odd and even interleaves of the codeword have an even number of 1s. The polynomial codes with primitive generator polynomials $g_c(D) = 1 \oplus D \oplus D^3$ and $g_c(D) = 1 \oplus D \oplus D^4$ add three and four parity bits, respectively. In all cases the combined modulation/parity code satisfies the prespecified (G, I) constraint. Examples of combined modulation/parity codes proposed in the literature include rate-32/35 PRML$(G = 7, I = 7)$ single-parity bit code [23], and rate-64/66, -64/67, -64/68, and -64/69 PRML(G, I) single-parity, dual-parity, triple-parity, and quadruple-parity codes, respectively, [25,26,28]. Moreover, longer block-length rate-96/101 and -96/102 PRML(G, I) triple-parity and quadruple-parity codes, respectively, have been discussed in [24–26].

An alternative design approach is to use performance-enhancing MTR modulation codes instead of the simple (G, I) codes. For example, the time-varying MTR$(j = 3, 4, k = 18, t = 14)$ block code presented above eliminates all error events of the type $\{\pm 2, \mp 2, \pm 2, \mp 2\}$, $\{\pm 2, \mp 2, \pm 2, \mp 2, \pm 2\}$, and so on. More important, this code also eliminates almost 50% of the error events of the type $\{\pm 2, \mp 2, \pm 2\}$. Therefore, two parity bits are adequate to detect the remaining predominant error events $\{\pm 2, \mp 2, \pm 2\}$, $\{\pm 2\}$, $\{\pm 2, \mp 2\}$, and $\{\pm 2, 0, 0, \pm 2\}$. By concatenating six 17-bit MTR$(j = 3, 4, k = 18, t = 14)$ codewords and inserting two parity bits in appropriate locations a rate-96/104 MTR$(j = 3, 4, k = 18, t = 14)$ dual-parity code is obtained.

4. EQUALIZATION AND DETECTION TECHNIQUES

The role of advanced signal processing has been crucial in increasing the areal density in magnetic storage devices. Until the beginning of the past decade, all commercially available disk-drive systems used analog peak detection. A peak detector operates on the analog readback signal to detect the presence of a pulse within a sliding observation window. At low linear densities, the location of peaks in the readback waveform will closely correspond to the location of the transitions in the input current waveform. Therefore, with an accurate clock signal, the recorded data pattern can in principle be reconstructed by correctly determining the pulse positions in the readback waveform. For a long time, peak detectors combined with RLL(d, k) modulation codes, which improve missing-bit errors by spacing transitions further apart, provided a low-cost detection strategy.

Recent advances in VLSI technologies and the need for ever higher areal densities paved the way for advanced signal-processing techniques, including maximum-likelihood sequence detection (MLSD) or near-MLSD schemes. The noise in hard-disk drive systems is a combination of electronics noise, colored stationary media noise, and transition noise. Although the first two noise sources can be treated as additive and Gaussian sources, the third is data-dependent, non-Gaussian and nonadditive.

For the purpose of introducing the optimum detection scheme for the readback signal, we will first consider the linear ISI model for the magnetic-recording system shown in Fig. 5. In this case, it can readily be shown [50,51] that the optimum signal processing and detection method consists of a filter matched to the pulse response, $h(t)$, a symbol-rate sampler, a whitening filter, and a MLSD as shown in Fig. 14. The output of the symbol-rate sampler will then be

$$u_i = \sum_{\ell} a_\ell R_h(i - \ell) + \nu_i \tag{13}$$

where $R_h(i)$ is the sampled autocorrelation function of $h(t)$ defined by

$$R_h(i) = \int_{-\infty}^{+\infty} h(t)h(t + iT)\,dt \tag{14}$$

$1/T$ is the rate at which the encoded date are written in the magnetic medium, and $\{\nu_i\}$ is the additive noise sequence given by

$$\nu_i = \int_{-\infty}^{+\infty} \eta(t)h(t - iT)\,dt \tag{15}$$

In Fig. 14, the binary modulation-encoded data sequence $\{a_i\}$ is represented by the D-transform $a(D) = \sum_i a_i D^i$ and

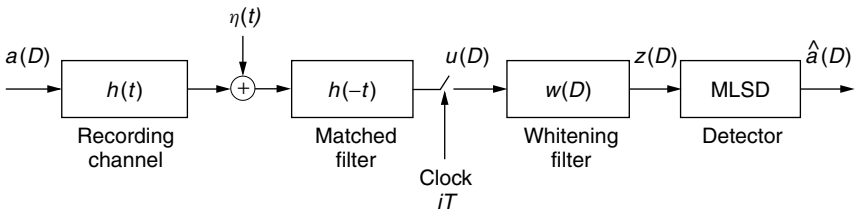


Figure 14. Magnetic-recording system with optimum signal processing and MLSD detection.

has power spectral density $S_a(D)$. Also the D-transform of the symbol-rate sampler output sequence $\{u_i\}$ is denoted by

$$u(D) = a(D)S_h(D) + \nu(D) \tag{16}$$

where $S_h(D)$ indicates the D-transform of the sampled autocorrelation function $R_h(i)$. By observing Eq. (15) it can readily be seen that the power spectral density of the noise sequence $\{\nu_i\}$ is $S_\nu(D) = S_\eta(D)S_h(D)$, where $S_\eta(D)$ denotes the D-transform of the sampled autocorrelation function $R_\eta(i)$ of the additive Gaussian process $\eta(t)$, that is, $R_\eta(i) = \int_{-\infty}^{+\infty} \eta(t)\eta(t+iT)\,dt$. Note that if the only noise source is the electronics noise due to the read head and preamplifier, then $S_\nu(D) = N_0\, S_h(D)$.

After symbol-rate sampling, the sequence at the output of the matched filter enters the whitening filter with transfer function $w(D)$. The output of the whitening filter is given by

$$z(D) = a(D)S_h(D)w(D) + \nu(D)w(D) = a(D)g(D) + e(D) \tag{17}$$

Observe that the desired signal component is affected by ISI through the overall transfer function $g(D) = S_h(D)w(D)$. Although it appears that the ISI may affect an infinite number of previously recorded symbols, in practice only a finite number of terms of $g(D)$ play a role in the sequence-detection process. This allows modeling the digital magnetic-recording system by an equivalent discrete-time finite-impulse-response (FIR) model with a minimum-phase transfer function $g(D)$ [52], where

$$g(D) = \sum_{\ell=0}^{M} g_\ell D^\ell \tag{18}$$

The MLSD that follows the whitening filter determines the most likely recorded sequence $\hat{a}(D)$ based on the observed sequence $z(D)$. It is well known that the MLSD is efficiently implemented with the Viterbi algorithm, whose complexity increases exponentially with the memory length. For an ISI span of M symbols, the state complexity of the optimum MLSD is 2^M. At low-linear recording densities, M may extend to only a few symbols. However, for $D_c > 3$, the ISI may extend to eight or even more symbols.

The optimum detector in the presence of ISI and additive Gaussian noise could therefore be prohibitively complex to implement because of its excessive state complexity. To circumvent this problem, a variety of suboptimal lower-complexity schemes have been developed aiming at reducing the span of ISI and consequently the complexity of the resulting MLSD. PR shaping to prescribed ISI targets has been very early recognized as an affective way to reduce the channel memory and also match the overall channel frequency response. In the following sections the most important suboptimum detection schemes will be reviewed, with emphasis on the ones that have extensively been used in high-performance direct-access storage devices.

4.1. PRML Detection

Partial response techniques for the magnetic-recording channel are similar to those that have been used in digital communication systems [50]. According to these techniques, the overall channel is shaped to some prescribed ISI pattern for which simple detection methods are known. The similarity between the pulse response in a magnetic-recording system and certain PR shapes was first noted in [53]. In particular, it was observed that at recording densities corresponding to $D_c \approx 2$, the frequency-domain representation of the Lorentzian pulse response closely resembles that of a linear filter with impulse response given by

$$f(t) = \operatorname{sinc}\left(\frac{t}{T}\right) - \operatorname{sinc}\left(\frac{t-2T}{T}\right) \tag{19}$$

where $\operatorname{sinc}(t) = \sin(\pi t)/\pi t$. It can readily be seen that at all times that are multiples of T, the value of the function $f(t)$ is zero except at times $t = 0$ and $t = 2T$, where the function takes the values $+1$ and -1, respectively. The discrete-time representation of this recording channel model corresponds to an overall noiseless input–output relationship given by

$$x_i = a_i - a_{i-2} \tag{20}$$

In D-transform notation, the input–output relationship becomes

$$x(D) = a(D)f(D) = a(D)(1 - D^2) \tag{21}$$

where the overall transfer function $f(D) = 1 - D^2$ is known in the literature as partial-response class-4 or PR4 shape. For higher linear recording densities, PR polynomials of the form

$$f_N(D) = (1 - D)(1 + D)^N \qquad N \geq 1 \tag{22}$$

are more suitable because their spectral characteristics match the spectral characteristics of the magnetic-recording channel better [54]. Clearly, the frequency response corresponding to $f_N(D)$ exhibits a spectral null at zero frequency and an N-th order null at the Nyquist frequency. Moreover, the PR4 polynomial corresponds to $N = 1$, whereas the PR polynomials with $N > 1$ are referred to as *extended* PR4 polynomials and are denoted as $\mathrm{E^{N-1}PR4}$. The PR4 and EPR4 shaping polynomials have been used extensively in hard-disk drives.

Simple linear equalization techniques may be employed to yield a prescribed PR shape. After symbol-rate sampling at the output of the matched filter, the samples enter a PR equalizer with transfer function $c(D)$, as shown in Fig. 15. The coefficients of the PR equalizer $\{c_\ell\}$ are optimized such that the overall transfer function, including channel and matched filter, closely matches a desired PR polynomial, that is,

$$c(D) = \frac{f_N(D)}{S_h(D)} \tag{23}$$

This equalizer is called zero-forcing linear equalizer, and shapes the incoming sequence $\{u_\ell\}$ according to any of the PR polynomials given by Eq. (22). If the magnetic-recording channel with additive Gaussian noise

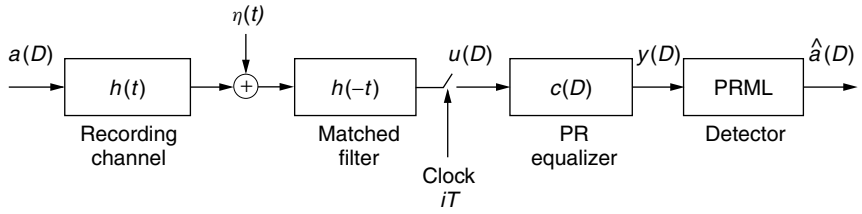


Figure 15. Magnetic-recording system with PRML detection.

is equalized to a PR target $f_N(D)$, the signal at the input of the detector becomes

$$\begin{aligned} y(D) &= a(D)S_h(D)c(D) + \nu(D)c(D) \\ &= a(D)f_N(D) + n(D) = x(D) + n(D) \end{aligned} \tag{24}$$

Thus, the zero-forcing linear equalizer limits the span of ISI to a small number of symbols, allowing an implementation of the MLSD for the desired PR target $f_N(D)$ with a small number of states.

After zero-forcing linear equalization, the noise sequence $n(D)$ is a discrete-time filtered version of the additive Gaussian noise $\eta(t)$. The power spectral density of $n(D)$ is

$$S_n(D) = \frac{f_N(D)f_N(D^{-1})}{S_h(D)} S_\eta(D) \tag{25}$$

Therefore, the variance of the noise sequence $\{n_i\}$ is

$$\sigma_n^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_n(e^{jw})\,dw = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|f_N(e^{jw})|^2}{S_h(e^{jw})} S_\eta(e^{jw})\,dw \tag{26}$$

This variance is a measure of the noise power at the input to the detector. Given that the noise sequence $n(D)$ is Gaussian and assuming that it also is an i.i.d. sequence of random variables (a reasonable assumption provided the selected PR target closely matches the magnetic-recording channel characteristics at the specified linear recording density), the most likely recorded sequence $\hat{a}(D)$ is the one that minimizes the squared Euclidean distance between the equalized sequence $y(D)$ and the noiseless sequence $x(D) = a(D)f_N(D)$. Equivalently,

$$\begin{aligned} \hat{a}(D) &= \arg\min_{a(D)} \|y(D) - x(D)\|^2 \\ &= \arg\min_{a(D)} \|y(D) - a(D)f_N(D)\|^2 \end{aligned} \tag{27}$$

The minimization in Eq. (27) can be efficiently achieved using the Viterbi algorithm. The combination of PR equalization techniques with MLSD is known in the industry as PRML detection [10]. The Viterbi algorithm is usually described via a finite-state transition diagram that evolves in time, known as the trellis diagram. The implementation complexity of the Viterbi algorithm depends on the number of states of the corresponding trellis. A $E^{N-1}PR4$ shaping polynomial gives rise to a trellis with 2^{N+1} states. In the case of PR4 and EPR4, the number of states is four and eight, respectively.

The application of the Viterbi algorithm for a PR4 shaped magnetic-recording channel was first proposed in [55], where it was also shown that a potential gain of 3 dB could be achieved using this dynamic programming procedure. The state complexity of a PR4-based detector can be further reduced by noting that the Euclidean metric in Eq. (27) can be split into two terms: one involving the odd indices and the other the even indices. Thus, in this special case, the Viterbi algorithm can be applied to two independent 2-state trellises operating on the even and odd PR4-equalized subsequences $\{y_{2i}\}$ and $\{y_{2i+1}\}$, respectively [10,56]. The Viterbi algorithm on each of the two 2-state trellises shown in Fig. 16 can be further simplified substantially by considering only the difference between the two survivor metrics and thus transforming the algorithm into a dynamic-threshold computation scheme ([10,56], and references therein). All these simplifications inherent in PR4 shaping led to the incorporation of the PR4-based PRML detection technology into magnetic hard-disk drives as well as magnetic tape systems. It was first introduced in the early 1990s in a commercial 5.25-inch IBM disk drive. The PR4-based PRML technology had a tremendous impact in the hard-disk drive industry and became the de-facto standard. Analytical studies, simulation results, and experimental data have shown that PR4-based PRML systems offer a 30–50% increase in linear recording density over RLL(2, 7) or RLL(1, 7) peak detection systems.

At higher linear recording densities, that is, $D_c > 2$, the linear PR4 equalizer leads to substantial noise enhancement, which increasingly affects the performance of the PRML system. EPR4 shaping alleviates this problem, to a certain extent, but increases the memory of the shaping polynomial to three and hence the number of states in the corresponding trellis to eight. Like PR4 shaping, the EPR4 polynomial exhibits spectral nulls at zero and the Nyquist frequencies; therefore, the magnetic-recording system would require similar PRML(G, I) codes

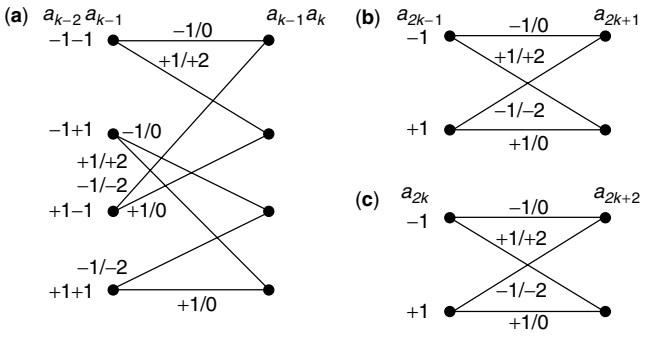


Figure 16. (**a**) 4-state PR4 trellis diagram. (**b**) 2-state odd and (**c**) 2-state even interleaved $1 - D$ trellis.

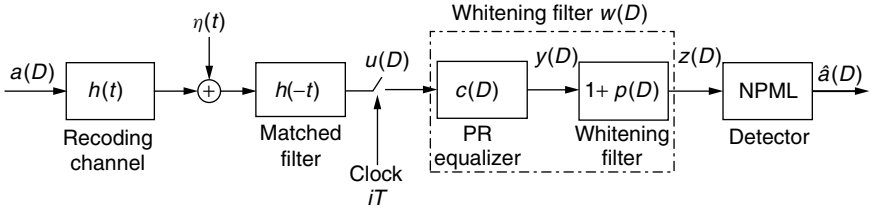


Figure 17. Magnetic-recording system with NPML detection.

to eliminate quasi-catastrophic error propagation, limit the path memory length, and facilitate timing and gain control. EPR4-based PRML systems achieve an additional 15% increase in linear recording density over PR4-based PRML systems, and the scheme has been used by some disk-drive manufactures for a limited period of time. Higher-order PR polynomials, that is, $N > 2$, although suitable for magnetic recording from the shaping characteristics point of view, do not achieve the matched filter bound and thus are of less interest. Hence, generalized PR polynomials in which the coefficients of the desired target response are nonintegers became significant in practical systems.

4.2. NPML Detection

In the absence of noise enhancement and noise correlation, the PRML sequence detector performs maximum-likelihood sequence detection. But there is an obvious loss of optimality associated with linear PR equalization as the operating point moves to higher linear recording densities. Clearly, a very close match between the desired target polynomial and the physical channel will guarantee that this loss will be minimal. An effective way to achieve near optimal performance independent of the operating point — in terms of linear recording density — and the noise conditions is via noise prediction. In particular, the power of the noise sequence $n(D)$ at the output of the PR linear equalizer can be minimized by using an infinitely long predictor. A linear predictor with coefficients $\{p_\ell\}$ operating on the noise sequence $n(D)$ will produce the estimate $\hat{n}(D)$, where the prediction-error sequence is given by

$$e(D) = n(D) + \hat{n}(D) = n(D)(1 + p(D)) \tag{28}$$

The optimum predictor $p(D) = p_1 D + p_2 D^2 + \cdots$, which minimizes the mean-square error $E\{|e_i|^2\}$ is given by $p(D) = q(D)/q_0 - 1$, where $q(D)$ is the minimum phase causal factor of $1/S_n(D)$ in Eq. (25). Using results from prediction theory, one readily obtains the minimum achievable mean-square error, which is also the variance of the white noise sequence at the output of the whitening filter $1 + p(D)$ shown in Fig. 17 and is given by [11]

$$\sigma_e^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \frac{S_\eta(e^{j\omega})}{S_h(e^{j\omega})}\, d\omega\right\} \tag{29}$$

Assuming that the additive noise process before matched-filtering is white and Gaussian with power spectral density $S_\eta(D) = N_0$, that is, the case of electronics noise only, Eq. (29) reduces to an expression that is identical to the minimum mean-square error (MMSE) of an infinitely long zero-forcing decision-feedback equalizer (DFE). Thus, the combination of the PR linear equalizer with the infinitely long noise predictor, shown in Fig. 17, is equivalent to the forward filter of a zero-forcing DFE, which in turn is the optimum whitening filter $w(D)$ of the MLSD [57]. Hence,

$$w(D) = c(D)(1 + p(D)) \tag{30}$$

and the sequence at the input of the detector becomes

$$z(D) = a(D)f(D)(1 + p(D)) + e(D) \tag{31}$$

where $e(D)$ is an AWGN sequence with variance given by Eq. (29). Note that in general the linear PR equalizer $\{c_\ell\}$ can be optimized so that the coefficients of the desired target $\{f_k\}$ can take any arbitrary value, that is,

$$f(D) = \sum_{\ell=0}^{K} f_\ell D^\ell \tag{32}$$

If $f_0 = 1$ and $f(D)$ is minimum phase, then the variance of the noise at the output of the whitening filter is still given by Eq. (29) [11].

So far, the zero-forcing criterion for obtaining the equalizer coefficients $\{c_\ell\}$ has been considered for the development of the NPML detection technique. Similar results can be obtained by using the MMSE criterion for optimizing the linear PR equalizer. For more details on this topic the reader is referred to [11].

The important result of Eq. (31) is that with this method the desired generalized PR target can be factored into two terms. The first factor contains spectral nulls or near nulls at selected frequencies, reflecting the nulls or near nulls of the physical magnetic-recording channel. The second factor, without roots on the unit circle, is optimized depending on the linear recording density and noise conditions. Note that spectral nulls at frequencies $f = 0$ and $f = 1/2T$ play an important role in practical systems because they render the sequence detector insensitive to dc offsets and disturbances around the Nyquist frequency.

An infinitely long predictor filter would lead to a sequence detector structure that requires an unbounded number of states. Finite-length predictors and, in particular, shaping polynomials of the form $g(D) = (1 - D)(1 + p(D))$ with a spectral null at dc and $g(D) = (1 - D^2)(1 + p(D))$ with spectral nulls at both dc and the Nyquist frequency, where $p(D) = \sum_{\ell=1}^{L} p_\ell D^\ell$ is the transfer function of a predictor of finite order L, render the noise at the input of the sequence detector approximately white. This class of generalized PR polynomials, which

is significant in practical applications, as well as any generalized PR shaping polynominal of the form $g(D) = f(D)(1 + p(D))$, when combined with sequence detection, give rise to NPML systems [11,12,14,15]. For a generalized PR target characterized by the polynomial $g(D) = (1 - D)(1 + p(D))$, the effective ISI memory of the system is limited to $M = L + 1$ symbols, and the NPML detector performs maximum-likelihood sequence detection using the 2^{L+1}-state trellis corresponding to $g(D)$. The same holds for generalized PR targets of the form $g(D) = (1 - D^2)(1 + p(D))$. In this case the effective ISI memory of the system is limited to $M = L + 2$ symbols, and the NPML detector performs maximum-likelihood sequence detection using the 2^{L+2}-state trellis corresponding to $g(D)$. Finally, if $g(D) = f(D)(1 + p(D))$, then the effective memory of the system is limited to $M = L + K$, giving rise to a 2^{L+K}-state NPML detector. In any case, the NPML detector is efficiently implemented by using the Viterbi algorithm, which recursively computes

$$\hat{a}(D) = \arg \min_{a(D)} \|z(D) - a(D)g(D)\|^2 \tag{33}$$

For large values of L, which implies white Gaussian noise at the output of the predictor filter, the performance of NPML may be determined by the technique described in [52]. At high SNR, the probability of error for NPML detection can be approximated by

$$P_b \approx K_0 Q\left(\frac{d_{\min}}{2\sigma_e}\right) \tag{34}$$

where $d^2_{\min}$ is the minimum Euclidean distance of an error event, σ_e^2 is the variance of the AWGN at the output of the predictor given in Eq. (29), and K_0 is a constant.

The variance σ_e^2 of the noise is a measure of performance because it represents the effective SNR at the input to the NPML detector. It is plotted in Fig. 18 as a function of the normalized linear density D_c for selected generalized PR shaping polynomials. The system is only affected by electronics noise, and the SNR is assumed constant at 25 dB. Curve 1 corresponds to the case of an infinitely long whitening filter $w(D)$. Curves 2 and 3 correspond to the memory-four shaping polynomials $g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2)$ and $g(D) = (1 - D)(1 + 0.75D)(1 + p_1 D + p_2 D^2)$, respectively. Finally, curve 4 corresponds to a memory-six shaping polynomial assuming a PR4 equalizer followed by a four-coefficient predictor. In all cases a simple 5-pole Butterworth filter instead of a matched filter has been assumed. Furthermore, the results for curves 2 and 4 have been obtained by using a linear PR4 equalizer with 10 coefficients, whereas those for curve 3 have been obtained by using a $(1 - D)(1 + 0.75D)$ 10-coefficient PR linear equalizer. As can be seen a memory 4-target giving rise to a 16-state NPML detector is uniformly good for $2 \le D_c \le 3.6$. In fact, the variance of the noise at the input of the 16-state NPML detector is at most 1.2 dB worse than the variance of an infinitely long whitening filter. The class of 16-state NPML detectors associated with order $L = 3$ predictor filters $(g(D) = (1 - D)(1 + p_1 D + p_2 D^2 + p_3 D^3))$ or order $L = 2$ predictor filters $(g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2))$ provide significant performance gains over the E^2PR4 shaping polynomial and is currently the state-of-the-art detection scheme in the hard-disk drive industry.

In practical applications a simple low-pass filter is used instead of a matched filter. Furthermore, the PR equalizer can be implemented as a FIR digital filter or a continuous-time filter before sampling. There are many commercial products that have used the latter approach, but today most of the hard-disk drives use discrete-time FIR equalization techniques. For known channel characteristics, the coefficients of the finite-length PR equalizer and predictor can be obtained by solving two sets of equations separately. In a first step the optimum coefficients of a finite-length PR equalizer according to the MMSE or zero-forcing criterion are obtained [50]. The predictor coefficients are then the solution of the well-known normal equations. An alternative approach is to combine the PR equalizer and the predictor into a single whitening filter and obtain its coefficients together with the NPML detector target by using a joint optimization procedure. Such an approach is reminiscent of the computation of the coefficients of a PR DFE [11,58]. To cope with the slow variations of the magnetic-recording channel as well as with the need for different sets of coefficients depending on whether the outer or the inner tracks are read back, standard adaptation algorithms can be applied. In some commercial hard-disk drives, the whitening-filter

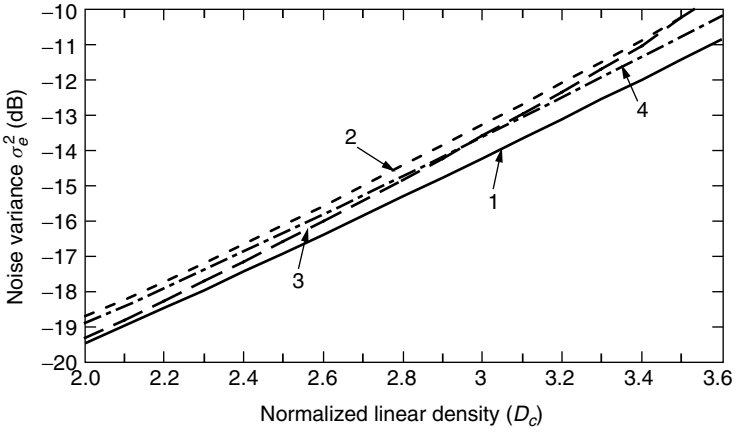


Figure 18. Noise variance at NPML detector input as a function of D_c for a Lorentzian channel with AWGN and an SNR of 25 dB. Curve 1: infinitely long whitening filter $w(D)$; curve 2: $g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2)$; curve 3: $g(D) = (1 - D)(1 + 0.75D)(1 + p_1 D + p_2 D^2)$, and curve 4: $g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2 + p_3 D^3 + p_4 D^4)$.

(PR equalizer/predictor) coefficients are trained during manufacturing, and large sets of coefficients are obtained reflecting the various operating conditions. These sets of coefficients are stored and frozen before shipping the products. Other commercial products have the capability to retrain the whitening filter and NPML detector target adaptively in real time.

4.3. Other Detection Techniques

Several other techniques to increase the areal recording density have been proposed. One of these, which is well known in the field of digital communications, is DFE [50]. The DFE does not force the overall transfer function to a prescribed target, and thus suffers from less noise enhancement than PRML techniques. In fact, the response at the output of the forward section of a DFE could be viewed as a particular form of generalized PR polynomial that matches the magnetic-recording channel characteristics very well. It has already been pointed out above that the forward section of an infinitely long zero-forcing DFE is equivalent to the whitening filter of the optimum MLSD. The feedback section of the DFE using past decisions cancels the causal ISI and a simple threshold symbol-by-symbol detector provides the final decision. Application of DFE to storage channels has been studied in [59]. An adaptive DFE with look-up table implementation of the feedback section (RAM-DFE) has been proposed in [60,61]. By using a dynamic procedure to update the contents of the RAM, a RAM-DFE can also effectively counter nonlinear distortion. Prototype DFE chips for the magnetic-recording channel have been developed in the past [62] but this approach did not enjoy commercial success primarily because of the problem of error propagation.

Another technique that has attracted considerable attention is the fixed-delay tree-search approach with decision feedback (FDTS/DF) [31]. The FDTS/DF employs the forward section of a DFE to create a minimum phase causal response with most of the energy of the causal ISI concentrated in the first few terms. If M is the total span of ISI, the last $M-\tau$ ISI terms are canceled by a feedback mechanism similar to a DFE arrangement, whereas the first τ ISI terms are being processed by a detector structure that searches into a tree with $2^{\tau+1}$ branches. Clearly, the feedback cancellation scheme works properly provided that the FDTS/DF detector releases decisions with a delay of τ symbols. Thus, the detector looks ahead τ steps into the tree, and computes the $2^{\tau+1}$ Euclidean metrics associated with all the look-ahead paths in the tree. These metrics are then used to decide whether the symbol at the root of the tree should be $+1$ or -1. This scheme is a derivative of the delay-constrained optimal detector approach presented in [63]. The application of noise-predictive PR equalization schemes in conjunction with FDTS has been studied in [12].

Finally, reduced-state sequence-detection schemes [64–66] have also extensively been studied for application in the magnetic-recording channel ([12,25] and references therein). For example, it can readily be seen that the NPML detectors can be viewed as a family of reduce-state detectors with imbedded feedback. They also exist in a form in which the decision-feedback path can be realized by simple table look-up operations, whereby the contents of these tables can be updated as a function of the operating conditions [12]. Analytical and experimental studies have shown that a judicious tradeoff between performance and state complexity leads to practical schemes with considerable performance gains. Thus, reduced-state approaches appear promising for increasing the linear density even further.

5. PARITY-BASED POST-PROCESSING

In generalized PR systems combined with sequence detection a short list of error events dominates. In general, post-processors are suboptimum reduced-complexity receiver structures that improve error-rate performance by correcting the most likely error events at the output of the sequence detector. The use of a post-processor requires a modest increase in implementation complexity. Based on the short list of preselected error events the post-processor computes the log-likelihood ratio (LLR) of each of these selected error events at each point in time. These LLRs are then used to decide the type and the location of the most likely error event to be corrected. The LLR corresponding to the i-th error event in the list, $\varepsilon_i(D)$, is computed as the difference of the following Euclidean metrics

$$\begin{aligned}\mathrm{LRR}(\varepsilon_i(D)) = \|(z(D) - \hat{a}(D)g(D)\|^2 - \|z(D) \\ - (\hat{a}(D) + \varepsilon_i(D))g(D)\|^2 \end{aligned} \tag{35}$$

where $\|z(D) - \hat{a}(D)g(D)\|^2$ is the Euclidean distance between the noisy sequence at the input of the sequence detector, $z(D)$ and the sequence $\hat{a}(D)g(D)$ generated by using final decisions produced by the sequence detector, whereas $\|z(D) - (\hat{a}(D) + \varepsilon_i(D))g(D)\|^2$ is the Euclidean distance between the noisy sequence at the input of the sequence detector, $z(D)$ and the sequence $(\hat{a}(D) + \varepsilon_i(D))g(D)$ generated by an alternative data sequence that includes the specific error pattern.

As can be seen, the LLRs computed according to Eq. (35) use the same metric as the NPML detector does. A post-processor using such a soft metric is also referred to as noise-predictive post-processor. Noise-predictive post-processors are reminiscent of the *one-shot* optimum receiver structures in communications theory [67], and can have threshold-based or parity-based triggering mechanisms [23–28,68,69]. In general, threshold-based post-processing schemes do not suffer from rate loss, whereas parity-based post-processing schemes do. However the parity-based triggering mechanism for initiating error-event correction may be more robust in certain situations.

In a parity-based post-processing scheme typically a single or double error event within a codeword is corrected. After NPML detection and during the symbol-by-symbol processing of each codeword, a short list of the most likely error events is maintained and continuously updated together with the associated error type, polarity, and location. Once the entire codeword has been received and the short list finalized, the LLRs and the syndromes of all combinations of error events are computed. Finally,

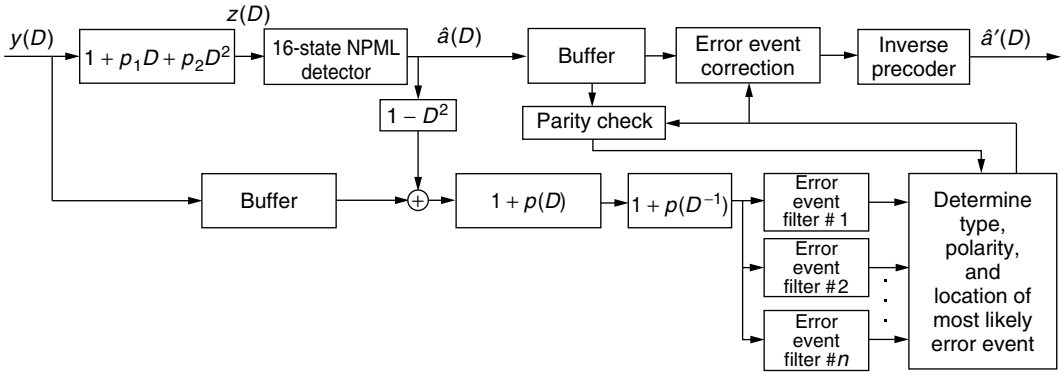


Figure 19. 16-state NPML detector and noise-predictive parity-based post-processor. From [26], © IEEE 2001.

after rejecting error-event candidates that violate certain conditions, such as for example those that do not produce valid syndromes, the single (double) error event with the minimum LLRs is (are) corrected. Figure 19 shows the block diagram of the 16-state NPML detector in tandem with a parity-based post-processor.

Figure 20 illustrates the performance of various modulation/parity codes after 16-state NPML detection and parity-based post-processing. These results have been obtained via computer simulations assuming the Lorentzian channel model and electronics noise. The user linear density is set to $\mathrm{PW50}/Tu = 3.2$. The channel density D_c depends on the rate of the code. The front-end filter is a five-pole low-pass Butterworth filter with 3-dB cutoff frequency at 55% of the channel symbol rate. The equalizer is a 10-coefficient PR4-shaping zero-forcing equalizer. Curve 1 shows the performance of a 16-state NPML detector in conjunction with a rate-32/34 PRML(G, I) single-parity code and a post-processor that detects and corrects the three types of error events $\{\pm 2\}$, $\{\pm 2, \mp 2, \pm 2\}$, and $\{\pm 2, \mp 2, \pm 2, \mp 2, \pm 2\}$. Curve 2 indicates the performance of the 16-state NPML detector in conjunction with a rate-96/102 PRML(G, I) code with quadruple-parity and a post-processor that detects and corrects seven types of error events. Finally, curve 3 corresponds to the performance of the 16-state NPML detector in conjunction with rate-96/104 MTR$(j = 3, 4, k = 18, t = 14)$ code with dual-parity and a post-processor that detects and corrects the four types of error events $\{\pm 2\}$, $\{\pm 2, \mp 2, \pm 2\}$, and $\{\pm 2, 0, 0, \pm 2\}$, and $\{\pm 2, \mp 2, \pm 2, \mp 2\}$. As can be seen, the rate-96/104 MTR-based code outperforms the rate-32/34 and -96/102 (G, I)-based codes by 1.3 and 0.7 dB, respectively, although it is approximately 2% less efficient. The difference in performance between the rate-96/104 MTR-based and rate-96/102 (G, I)-based codes is less pronounced under data-dependent medium noise conditions. This is attributed to the fact that in such a case, even at very high linear densities, the error $\{\pm 2\}$ is the predominant error event and the MTR constraints are not very effective.

Figure 21 shows the evolution of the various architectures that have been used over the past decade by various disk-drive manufactures. In particular, it shows the SNR requirements for achieving a symbol-error rate of 10^{-6} after the parity-based post-processor as a function of the normalized linear density D_c. Curve 1 corresponds to the conventional PRML architecture. It shows the performance of the 2-state interleaved PR4-based PRML

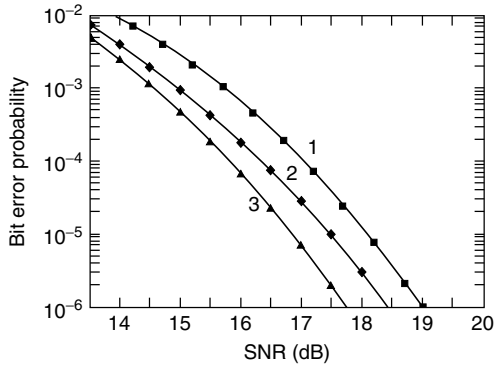


Figure 20. Bit-error rate performance after post-processing for a Lorentzian channel with $\mathrm{PW50}/Tu = 3.2$, AWGN, and a 16-state NPML detector. Curve 1: rate-32/34 PRML(G, I), 1 parity bit; curve 2: rate-96/102 PRML(G, I), 4 parity bits, and curve 3: rate-96/104 MTR(j, k, t), 2 parity bits.

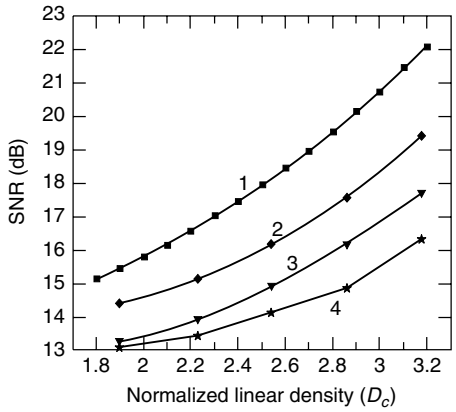


Figure 21. SNR requirements as a function of D_c for a Lorentzian channel with AWGN. Error probability $P_b = 10^{-6}$. Curve 1: PRML architecture, rate-16/17 PRML(G, I); curve 2: EPRML architecture, rate-32/34 PRML(G, I), 1 parity bit; curve 3: NPML architecture, rate-32/34 PRML(G, I), 1 parity bit, and curve 4: NPML architecture, rate-96/104 MTR(j, k, t), 2 parity bits.

detector in conjunction with a rate-16/17 PRML(G, I) constrained code. Curve 2 corresponds to the EPRML architecture. In this architecture the detector is based on the EPR4 shaping polynomial giving rise to an 8-state trellis. The inner code is a rate-32/34 PRML(G, I) constrained code and includes a single parity bit. Curve 2 demonstrates the benefit of the higher-order shaping polynomial as well as of the combined constrained/parity inner code, in particular at high linear recording densities, where a gain of approx. 2.5 dB over the conventional PRML scheme is obtained. Finally, curves 3 and 4 correspond to the NPML architecture, which has only recently been introduced in hard-disk drive products. The shaping polynomial is based on a 2-coefficient predictor and has spectral nulls at both the dc and the Nyquist frequency. Curve 3 shows the performance of the 16-state NPML detector in conjunction with the single-parity rate-32/34 PRML(G, I) constrained code, whereas curve 4 illustrates the performance of the 16-state NPML detector in conjunction with the dual-parity rate-96/104 MTR$(j = 3, 4, k = 18, t = 14)$ code. As can be seen, the 16-state NPML detector in conjunction with the rate-96/104 MTR-based dual-parity code, provides a 5.5 dB gain over the conventional PR4-based PRML detection with the rate-16/17 PRML(G, I) code at $D_c = 3.2$. Alternatively, the current NPML system architecture provides a 55% linear recording density increase over the conventional PR4-based PRML architecture.

6. DATA-DEPENDENT NPML DETECTION

Today's hard-disk drive devices employ thin-film media that appear primarily to exhibit nonstationary data-dependent transition or medium noise as opposed to colored stationary medium noise. Improvements on the quality of the readback head as well as the incorporation of low-noise preamplifiers may render the data-dependent medium noise a significant component of the total noise affecting the performance of the magnetic-recording system. Because the medium noise is correlated and data-dependent, information about the noise and data patterns in past samples can provide information about the noise in the current sample. Thus, the concept of noise prediction for stationary Gaussian noise sources developed in [11] can be naturally extended to the case where the noise characteristics depend highly on the local data patterns [16,18].

By modeling the data-dependent noise as a finite-order Markov process, the optimum MLSD for channels with ISI has been derived in [17,19]. In particular, it has been shown that when the data-dependent noise is conditionally Gauss–Markov, the branch metrics can be computed from the conditional second-order statistics of the noise process. In other words, the optimum MLSD can be implemented efficiently by using the Viterbi algorithm, where the branch-metric computation involves data-dependent noise prediction. Because both predictor coefficients and prediction error depend on the local data pattern, the resulting structure has been called data-dependent NPML detector [20]. In real systems, the data-dependent medium noise is not a strictly Markov noise process, and clearly the data-dependent NPML detector is only a near-MLSD structure. Nevertheless, physical models for data-dependent medium noise such as the one developed in [35] or the more accurate microtrack model [18] can be used to investigate the impact of the model parameters on the data-dependent Markov assumption and the performance of the NPML detector.

Let $\{y_i\}$ be the sequence of samples at the output of the PR linear equalizer with target $f(D)$. Then

$$y_i = x_i + n_i(a) = a_i + \sum_{\ell=1}^{K} f_\ell a_{i-\ell} + n_i(a) \tag{36}$$

where the noise sample $n_i(a)$ at time instance iT is assumed to be a zero-mean Gaussian random variable with statistics depending on the data sequence $a \triangleq \{a_i\}$. Figure 22 shows an example of a channel with ISI that spans K symbols and data-dependent Gauss–Markov noise. This model corresponds to Eq. (36), where the additive noise is generated by a L-order autoregressive filter whose coefficients depend on the last $K + 1$ recorded data symbols $a_i^{i-K} = \{a_i, a_{i-1}, \ldots, a_{i-K}\}$. In this case the correlated and data-dependent noise sample is given by

$$n_i = \sigma_e(a_i^{i-K})\upsilon_i - \sum_{\ell=1}^{L} p_\ell(a_i^{i-K}) n_{i-\ell} \tag{37}$$

where $\{\upsilon_i\}$ denotes a zero-mean unit-variance white Gaussian noise sequence, $\sigma_e(a_i^{i-K})$ indicates the data-dependent standard deviation, and $\{p_\ell(a_i^{i-K})\}$ represents the data-dependent coefficients of the autoregressive filter. Clearly, there are 2^{K+1} sets of filter coefficients, and the total memory of the model is $K + L$, giving rise to a trellis with 2^{K+L} states. A Viterbi algorithm, whose branch metrics have been modified to account for the data-dependent correlated noise, can operate on this trellis to estimate recursively the most likely recorded data sequence $\{a_i\}$. The implementation of this algorithm is very similar to the standard Viterbi algorithm for NPML detection described above. The main difference is that in the computation of the branch metrics, a window of observed values $y_i^{i-L} = \{y_i, y_{i-1}, \ldots, y_{i-L}\}$ is used, instead of just one sample z_i, which is the output of a "global"

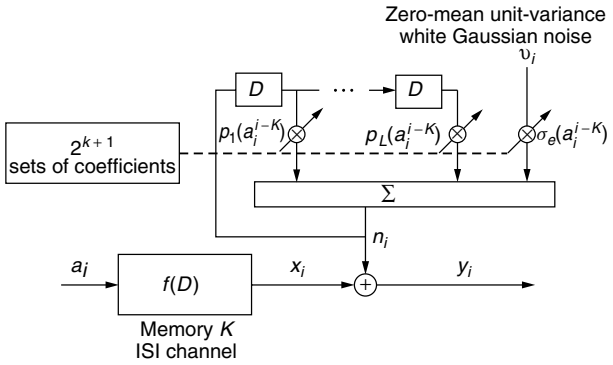


Figure 22. Block diagram of ISI channel with Gauss–Markov data-dependent noise.

whitening/prediction filter. The optimality of this approach for the Gauss–Markov noise case was shown in [19].

Let $\xi_i \triangleq a_i^{i-K-L} = \{a_i, a_{i-1}, \ldots, a_{i-K-L}\}$ denote a transition on the 2^{K+L}-state trellis. Then, assuming Gaussian data-dependent noise, the branch metric associated with transition ξ_i is given by

$$\gamma_i(\xi_i) = \ln \sigma_e^2(a_i^{i-K}) + \frac{[y_i - x_i + \hat{n}_i(\xi_i)]^2}{2\sigma_e^2(a_i^{i-K})} \tag{38}$$

where $\hat{n}_i(\xi_i)$ is a data-dependent predicted value of the current noise sample n_i, based on the past L noise samples $\{n_{i-1}, n_{i-2}, \ldots, n_{i-L}\}$, that is,

$$\hat{n}_i(\xi_i) = \sum_{\ell=1}^{L} p_\ell(a_i^{i-K})(y_{i-\ell} - x_{i-\ell}) = \sum_{\ell=1}^{L} p_\ell(a_i^{i-K}) n_{i-\ell} \tag{39}$$

and $\sigma_e^2(a_i^{i-K})$ represents the data-dependent variance of the prediction error. If the additive noise is correlated and Gaussian and does not depend on the data pattern, then $\{p_l(a_i^{i-K})\}$ and $\sigma_e^2(a_i^{i-K}) = \sigma_e^2$. In such a case, the data-dependent NPML detection structure reduces to the NPML detection technique presented above.

A finite-length predictor filter with coefficients that depend on a particular data pattern can be obtained by applying the normal equations separately and conditioned on each of the 2^{K+1} possible data patterns a_i^{i-K} that affect the noise process. For this purpose, the noise process can be characterized by 2^{K+1} autocorrelation functions of the form

$$R_n(l \mid a_i^{i-K}) = E\{n_i n_{i-\ell} \mid a_i^{i-K}\} \tag{40}$$

where the expectation is taken not only with respect to the noise statistics but also with respect to the data symbols that are not included in the conditioning. Using the autocorrelation functions described by Eq. (40), a set of 2^{K+1} L-th order prediction filters $\{p_l(a_i^{i-K})\}$ can be obtained together with their corresponding prediction errors $\sigma_e^2(a_i^{i-K})$, which can then be used in the branch metric computation of Eq. (38).

In practical systems the noise autocorrelation function can be estimated based on training patterns and statistical averaging, which can be performed during the manufacturing process of the disk drive. To avoid matrix inversions and the numerical problems associated with ill-conditioned matrices, an alternative approach is to use standard adaptation procedures to learn the set of noise predictor coefficients and noise prediction error-variances conditioned on the local data patterns [20]. Finally, note that the reduced-state sequence detection schemes discussed in connection with NPML detection can also be applied to data-dependent NPML, providing a significant reduction of implementation complexity.

Figure 23 illustrates the performance of two 16-state NPML detection schemes in the presence of nonstationary data-dependent medium noise. These results have been obtained via computer simulations assuming the microtrack channel model with a transition-width parameter of $a = 0.15$, and no electronics or any other source of stationary noise. In this case the system is affected by 100% data-dependent medium noise. The user linear recording density is set to PW50/$Tu = 3.6$. In both cases a rate-96/102 PRML(G, I) has been used, resulting in a normalized channel linear density of $D_c = 3.825$. The front-end filter is a five-pole low-pass Butterworth filter with 3-dB cutoff frequency at 55% of the channel symbol rate. The equalizer is a 10-coefficient PR4-shaping zero-forcing equalizer. Curve 1 shows the performance of a 16-state NPML detector whose branch metric computation uses the same set of predictor coefficients optimized as if the system were affected by stationary Gaussian noise only. Curve 2 indicates the performance of the data-dependent 16-state NPML detector whose branch metrics have been modified according to Eq. (38) to account for the presence of 100% data-dependent noise. As can be seen, the data-dependent NPML detector yields a gain of 1.3 dB at a symbol error rate of 10^{-5}. Note however that 100% data-dependent medium noise is not a realistic scenario in today's hard-disk drives, and therefore the gains expected by using a more complex data-dependent NPML detection scheme are less pronounced.

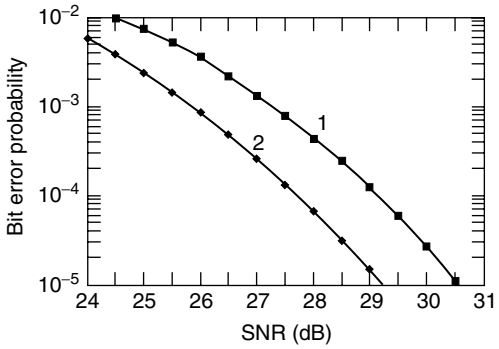


Figure 23. Bit-error rate performance for microtrack channel with PW50/$Tu = 3.6$, $a = 0.15$, and 100% data-dependent medium noise. Curve 1: 16-state NPML detector, and curve 2: 16-state data-dependent NPML detector.

7. FUTURE TRENDS

Currently, MTR or (G, I) codes combined with multiparity block codes, and 16-state NPML detection for generalized PR shaping channels, followed by a post-processor for soft decoding of the combined multiparity/constrained code, represent the state of the art in the industry. Outer error-correction coding has also played an important role in achieving high data integrity in magnetic-recording systems. In hard-disk drives, interleaved byte-oriented RS coding is currently the standard outer coding scheme. In the future, RS symbols with more than eight bits and sector sizes having a length of more than 512 8-bit bytes will have a significant impact on efforts to improve performance and push areal density even further.

The recent advances in coding theory and in particular the introduction of Turbo codes in the mid-1990s and the rediscovery of the powerful low-density parity check (LDPC), hold the promise to push the areal density to the ultimate limit for given magnetic-recording components. In spite of the current limit of the sector-size in a hard-disk drive to 512 bytes, which constrains the block size of an

outer code, and the high code-rate requirements, it has been shown that LDPC or Turbo codes with rather simple iterative decoding schemes can achieve a gain of more than 2 dB over existing systems and bring performance to within 1.5 dB of the ultimate information-theoretic limit: the capacity.

However, despite progress in the area of reduced-complexity detection and decoding algorithms, Turbo equalization structures with iterative detectors/decoders have not yet found their way into digital recording systems because of the still unfavorable tradeoff between performance, implementation complexity, and latency. The design of high-rate, short-block-length Turbo-like codes for recording systems remains an area of active research.

BIOGRAPHY

Evangelos S. Eleftheriou received as B.S. degree in electrical engineering from the University of Patras, Greece, in 1979, and M.Eng. and Ph.D. degrees in electrical engineering from Carleton University, Ottawa, Canada, in 1981 and 1985, respectively. He joined the IBM Zurich Research Laboratory in Rüschlikon, Switzerland, in 1986, where he has been working in the areas of high-speed voice-band data modems, wireless communications, and coding and signal processing for the magnetic recording channel. Since 1998, he has managed the magnetic recording and wired transmission activities at the IBM Zurich Research Laboratory.

His primary research interests lie in the areas of communications and information theory, particularly signal processing and coding for recording and transmission systems. He holds more than 30 patents (granted and pending applications) in the areas of coding and detection for transmission and digital recording systems. He was editor of the IEEE Transactions on Communications from 1994 to 1999 in the area of Equalization and Coding. He was guest editor of the *IEEE Journal on Selected Areas of Communications* special issue, "The Turbo Principle: From Theory to Practice."

BIBLIOGRAPHY

1. D. A. Thompson and J. S. Best, The future of magnetic data storage technology, *IBM J. Res. Develop.* **44**: 311–322 (2000).
2. S. Iwasaki and Y. Nakamura, An analysis for the magnetization mode for high density magnetic recording, *IEEE Trans. Magn.* **MAG-13**: 1272–1277 (1977).
3. S. Iwasaki and K. Ouchi, Co-Cr recording films with perpendicular magnetic anisotropy, *IEEE Trans. Magn.* **MAG-14**: 849–851 (1978).
4. H. Takano et al., Realization of 52.5 Gb/in.^{2} perpendicular recording, *J. Magn. Magn. Mater.* **235**: 241–244 (2001).
5. H. N. Bertram and M. Williams, SNR and density limit estimates: A comparison of longitudinal and perpendicular recording, *IEEE Trans. Magn.* **36**: 4–9 (2000).
6. R. Wood, The feasibility of magnetic recording at 1 terabit per square inch, *IEEE Trans. Magn.* **36**: 36–42 (2000).
7. R. Cideciyan, E. Eleftheriou, and T. Mittelholzer, Perpendicular and longitudinal recording: A signal-processing and coding perspective, *IEEE Trans. Magn.* **38**: 1698–1704 (2002).
8. A. Dholakia, E. Eleftheriou, and T. Mittelholzer, On iterative decoding for magnetic recording channels, in *Proc. 2nd Intl. Symp. on Turbo Codes & Related Topics*, Brest, France, 219–226, 2000.
9. D. Arnold and E. Eleftheriou, On the information-theoretic capacity of magnetic recording systems in the presence of medium noise, *IEEE Trans. Magn.* **38**: (Sept. 2002) (in press).
10. R. D. Cideciyan et al., A PRML system for digital magnetic recording, *IEEE J. Sel. Areas Commun.* **10**: 38–56 (1992).
11. P. R. Chevillat, E. Eleftheriou, and D. Maiwald, Noise predictive partial-response equalizers and applications, *Proc. IEEE Intl. Conf. Commun.* 942–947 (1992).
12. E. Eleftheriou and W. Hirt, Noise-predictive maximum-likelihood (NPML) detection for the magnetic recording channel, *Proc. IEEE Intl. Conf. Commun.* 556–560 (1996).
13. E. Eleftheriou and W. Hirt, Improving performance of PRML/EPRML through noise prediction, *IEEE Trans. Magn.* **32**(part1): 3968–3970 (1996).
14. R. Karabed and N. Nazari, Trellis-coded noise predictive Viterbi detection for magnetic recording channels, in *Dig. The Magnetic Recording Conf.* (*TMRC*), Minneapolis, MN, Aug. 1997.
15. J. D. Coker, E. Eleftheriou, R. L. Galbraith, and W. Hirt, Noise-predictive maximum likelihood (NPML) detection, *IEEE Trans. Magn.* **34**(part1): 110–117 (1998).
16. J. Caroselli, S. A. Altekar, P. McEwen, and J. K. Wolf, Improved detection for magnetic recording systems with media noise, *IEEE Trans. Magn.* **33**: 2779–2781 (1997).
17. A. Kavcic and J. M. F. Moura, Correlation-sensitive adaptive sequence detection, *IEEE Trans. Magn.* **34**: 763–771 (1998).
18. J. P. Caroselli, Modeling, analysis, and mitigation of medium noise in thin film magnetic recording channels, Ph.D. dissertation, University of California, San Diego, 1998.
19. A. Kavcic and J. M. F. Moura, The Viterbi algorithm and Markov noise memory, *IEEE Trans. Inform. Theory* **46**: 291–301 (2000).
20. J. Moon and J. Park, Pattern-dependent noise prediction in signal-dependent noise, *IEEE J. Sel. Areas Commun.* **19**: 730–743 (2001).
21. B. Brickner and J. Moon, Design of a rate 6/7 maximum transition run code, *IEEE Trans. Magn.* **33**(part1): 2749–2751 (1997).
22. R. D. Cideciyan, E. Eleftheriou, B. H. Marcus and D. S. Modha, Maximum transition run codes for generalized partial response channels, *IEEE J. Sel. Areas Commun.* **19**: 619–634 (2001).
23. T. Conway, A new target response with parity coding for high density magnetic recording channels, *IEEE Trans. Magn.* **34**: 2382–2486 (1998).
24. J. L. Sonntag and B. Vasic, Implementation and bench characterization of a read channel with parity check post-processor, in *Dig. The Magnetic Recording Conf.* (*TMRC*), Santa Clara, CA, Aug. 2000.
25. R. D. Cideciyan, J. D. Coker, E. Eleftheriou, and R. L. Galbraith, NPML detection combined with parity-based post-processing, in *Dig. The Magnetic Recording Conf.* (*TMRC 2000*), Santa Clara, CA, Aug. 2000.

26. R. D. Cideciyan, J. D. Coker, E. Eleftheriou, and R. L. Galbraith, NPML detection combined with parity-based postprocessing, *IEEE Trans. Magn.* **37**: 714–720 (2001).
27. B. Vasic, A graph based construction of high-rate soft decodable codes for partial response channels, in *Proc. IEEE ICC'01*, Helsinki, Finland, 2716–2720, 2001.
28. W. Feng, A. Vityaev, G. Burd, and N. Nazari, On the performance of parity codes in magnetic recording systems, in *Proc. IEEE Global Telecommun. Conf.*, 1877–1881, 2000.
29. R. I. Potter, Digital magnetic recording theory, *IEEE Trans. Magn.* **10**(part1): 502–508 (1974).
30. K. G. Ashar, *Magnetic Disk Technology*, IEEE Press, New York, 1997.
31. J. Moon and L. R. Carley, Performance comparison of detection methods in magnetic recording, *IEEE Trans. Magn.* **26**: 3155–3172 (1990).
32. J. Moon, The role of signal processing in data-storage systems, *IEEE Signal Proc. Mag.* 54–72 (July 1998).
33. L. L. Nunnelley, D. E. Heim, and T. C. Arnoldussen, Flux noise in particulate media: Measurement and interpretation, *IEEE Trans. Magn.* **23**: 1767–1775 (1987).
34. H. N. Bertram, *Theory of Magnetic Recording*, Cambridge University Press, Cambridge, UK, 1994.
35. S. K. Nair, H. Shafiee, and J. Moon, Modeling and simulation of advanced read channels, *IEEE Trans. Magn.* **29**: 4056–4058, (1993).
36. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (1948).
37. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorders, *IEEE Trans. Inform. Theory* **44**: 2260–2299 (1998).
38. R. L. Adler, D. Coppersmith, and M. Hassner, Algorithms for sliding-block codes: An application of symbolic dynamics to information theory, *IEEE Trans. Inform. Theory* **29**: 5–22 (1983).
39. B. H. Marcus, P. H. Siegel, and J. K. Wolf, Finite-state modulation codes for data storage, *IEEE J. Sel. Areas Commun.* **10**: 5–37 (1992).
40. K. A. S. Immink, *Coding Techniques for Digital Recorders*, Prentice-Hall International, UK, 1991.
41. P. H. Siegel and J. Wolf, Modulation and coding for information storage, *IEEE Commun. Mag.* **29**: 68–86.
42. G. Jacoby, A new look-ahead code for increased data density, *IEEE Trans. Magn.* **13**: 1202–1204 (1977).
43. G. D. Forney and A. R. Calderbank, Coset codes for partial-response channels; or, Coset codes with spectral nulls, *IEEE Trans. Inform. Theory* **35**: 925–943 (1989).
44. R. Karabed, P. H. Siegel, and E. Soljanin, Constrained coding for binary channels with high intersymbol interference, *IEEE Trans. Inform. Theory* **45**: 1777–1797 (1999).
45. Method and Apparatus for Implementing Optimum PRML Codes, U. S. Patent 4, 707, 681 (1987) J. Eggenberger and A. M. Patel.
46. A. M. Patel, Rate 16/17 (0,6/6) Code, *IBM Tech. Discl. Bull.* **31**(8): 21–23 (1989).
47. W. G. Bliss, An 8/9 rate time-varying trellis code for high density magnetic recording, *IEEE Trans. Magn.* **33**: 2746–2748 (1997).
48. K. K. Fitzpatrick and C. S. Modlin, Time-varying MTR codes for high density magnetic recording, in *Proc. IEEE Global Telecommun. Conf.*, 1250–1253, 1997.
49. B. E. Moision and P. H. Siegel, Distance enhancing constraints for noise predictive maximum likelihood detectors, in *Proc. IEEE Global Telecommun. Conf.*, 2730–2735, 1998.
50. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
51. J. G. Proakis, Equalization techniques for high-density magnetic recording, *IEEE Signal Proc. Mag.* 73–82 (July 1998).
52. G. D. Forney, Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 363–378 (1972).
53. H. Kobayashi and D. T. Tang, Application of partial-response channel coding to magnetic recording systems, *IBM J. Res. Develop.* **14**: 368–375 (1970).
54. H. K. Thapar and A. M. Patel, A class of partial response systems for increasing storage density in magnetic recording, *IEEE Trans. Magn.* **MAG-23**: 3666–3668 (1987).
55. H. Kobayashi, Application of probabilistic decoding to digital magnetic recording, *IBM J. Res. Develop.* **15**: 65–74 (1971).
56. R. W. Wood and D. A. Peterson, Viterbi detection of class IV partial response on a magnetic recording channel, *IEEE Trans. Commun.* **COM-34**: 454–461 (1986).
57. R. Price, Nonlinearly feedback equalized PAM vs. capacity for noisy filter channels, in *Proc. IEEE Intl. Conf. on Commun.*, 22.12–22.17, 1972.
58. J. W. M. Bergmans, Partial response equalization, *Philips J. Res.* **42**: 209–245 (1987).
59. J. W. Bergmans, Density improvements in digital magnetic recording by decision feedback equalization, *IEEE Trans. Magn.* **22**: 157–162 (1986).
60. K. D. Fisher et al., An adaptive RAM-DFE for storage channels, *IEEE Trans. Commun.* **39**: 1559–1568, (1991).
61. J. Cioffi et al., Adaptive equalization in magnetic-disk storage channels, *IEEE Commun. Mag.* **28**: 14–29 (1990).
62. J. W. M. Bergmans et al. Dual-DFE read/write channel IC for hard-disk drives, *IEEE Trans. Magn.* **34**: 172–177 (1998).
63. K. Abend and B. D. Fritchman, Statistical detection for communication channels with intersymbol interference, *Proc. IEEE* **58**: 779–785 (1970).
64. V. M. Eyuboglu and S. U. Qureshi, Reduced-state sequence estimation with set partitioning and decision feedback, *IEEE Trans. Commun.* **COM-36**: 13–20 (1988).
65. A. Duell-Hallen and C. Heegard, Delayed decision-feedback sequence estimation, *IEEE Trans. Commun.* **COM-37**: 428–436 (1989).
66. P. R. Chevillat and E. Eleftheriou, Decoding of trellis-encoded signals in the presence of intersymbol interference and noise, *IEEE Trans. Commun.* **COM-37**: 669–676 (1989).
67. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons Inc., New York, 1965.
68. R. W. Wood, Turbo-PRML: A compromise EPRML detector, *IEEE Trans. Magn.* **29**: 4018–4020 (1993).
69. H. Sawaguchi and S. Mita, Soft-output decoding for concatenated error correction in high-order PRML channels, in *Proc. IEEE Intl. Conf. on Commun.*, 632–1637, 1992.

SIGNAL QUALITY MONITORING IN OPTICAL NETWORKS

IPPEI SHAKE
NTT Corporation
Kanagawa, Japan

1. INTRODUCTION

Signal quality monitoring is an important issue in relation to the design, operation, and maintenance of optical transport networks. From a network operator's point of view, monitoring techniques are required to provide connections, undertake protection and/or restoration, perform maintenance, and establish service level agreements. To realize these functions, the monitoring techniques should be able to offer the following: in-service (nonintrusive) measurement, signal deterioration detection [both signal-to-noise ratio (SNR) degradation and waveform distortion], fault isolation (locate faulty sections or nodes), transparency and scalability (irrespective of the signal bit rate and signal format), and simplicity (small size and low cost).

There are several approaches, including both digital and analog techniques, that make it possible to detect various types of impairment. Bit error rate (BER) measurement is a fundamental method for evaluating end-to-end signal quality, but it is not a practical solution for optical networks because it requires clock and data synchronization between transmitters and receivers. Error block detection or error counting utilizing a SONET/SDH frame or other frame is currently a practical solution for optical networks. However, these techniques fail to satisfy the aforementioned requirements for performance monitoring, for instance, transparency in an optical transport network (OTN). Several approaches have recently been proposed to overcome this problem. These techniques include optical SNR or power evaluation with optical/electrical spectrum measurement, pilot tone detection, pseudo-BER estimation or error monitoring using variable decision circuits, and a statistical method using histogram evaluation accompanied by synchronous or asynchronous eye diagram measurement. A fundamental performance monitoring parameter of any digital transmission system is its end-to-end BER. However, the BER can be correctly evaluated only with an out-of-service BER measurement by using a known test bit pattern in place of the real signal. By contrast, in-service measurement can provide only rough estimates through the measurement of digital parameters [e.g., BER estimation, error block detection, and error count in forward error correction (FEC) or analog parameters (e.g., optical SNR, optical/electrical spectrum, and Q-factor)].

2. PERFORMANCE MONITORING UTILIZING DIGITAL PARAMETERS

2.1. Digital Frame Based Technique

2.1.1. Bit Interleaved Parity (SONET/SDH). Error block detection is a digital technique for the end-to-end performance monitoring of optical channels. The error block is detected by means of an error detection code, for example, bit interleaved parity (BIP) in a SONET/SDH frame [1]. A BIP code that consists of X bits is called a BIP-X code. The BIP-X code is written as part of the overhead of the following payload. The data sequence to be monitored is divided into sequences, each of which is X bits long, the even parity of the nth bit (n is an integer between 1 and X) of all the X bits sequences is written in the nth bit of the BIP-X code. These procedures are undertaken at the transmitter end first. Then, at the receiver end, the BIP code is calculated again over each frame using the same procedure and compared with the transmitted code. This technique can only be used to monitor odd numbers of bit errors because it uses a BIP code in which an even parity is written over long sequences of data. This means that BIP-X is valid when the BER is low enough to cause only 1 bit error in the nth bit of all the X bit sequences.

2.1.2. Digital Wrapper. The digital wrapper is a SONET/SDH-based frame format technology in OTN [1]. This frame includes not only data and a channel header but also the forward error correction (FEC) code. The FEC makes it possible to correct bit errors that occur in the transmission line at the receiver by sending an additional bit sequence with data and channel header sequences. FEC is a powerful technology for long-haul high-speed optical transmission because of this real-time error correction capability. From the viewpoint of performance monitoring, the FEC procedure can provide us with the number of errors. This frame typically is used on a link per link basis, so it should not be used for end-to-end performance monitoring.

2.1.3. Others. Other frame format technologies have been developed for wide area networks (WAN), although most are also SONET-based. For example, 10 Gbit/s Ethernet (IEEE802.3ae) for WAN (WAN-PHY) has an operation, administration, management, and provisioning function based on SONET technology. Another extension of the 10 Gbit/s Ethernet frame for WAN has also been proposed.

2.2. Pseudo BER Estimation Using Variable Decision Circuit

When the system BER is too low to be measured within a reasonable amount of time, it is necessary to estimate the BER or other parameters representing signal quality. This subsection introduces one BER estimation approach that uses a variable decision circuit [2]. The decision level is changed step by step from a low amplitude to a high amplitude through the whole eye opening. The BER is measured only when the decision levels are relatively low or high. This is because the decision level is directly related to the BER, as shown in the following equation (with a Gaussian assumption),

$$\mathrm{BER} = (1/2)\mathrm{erfc}(Q/\sqrt{2}),\ Q = (\mu_1 - V_t)/\sigma_1 = (V_t - \mu_0)/\sigma_0 \quad (1)$$

where μ_i and σ_i are the mean and the standard deviation of mark ($i = 1$) and space ($i = 0$) noise levels, respectively, and V_t is the decision level. The BER becomes relatively high when the decision level is lower or higher than the

optimum level, and the measurement can be finished in a sufficiently short time. BER data are then plotted on a graph, where the vertical axis is the BER and the horizontal axis is the decision level. There are two linear fitting curves in the figure. Finally, the lowest BER is estimated by extracting the intersection of these two fitting curves.

This method is useful because a very low BER can be easily estimated within a reasonable amount of time. However, its use of BER measurement means that it requires data synchronization between the transmitter and receiver.

Recently, a new approach has been proposed as an application of this method, which resolves the problem of data synchronization by using two decision circuits [3,4]. The receiver equipment uses these two decision circuits. The first is used for a working channel with a fixed decision level and the second is used for measurement with a variable decision level. The received signal is divided into two, launched into these two decision circuits, and then the two logical outputs from the two circuits are combined with an exclusive OR gate (EXOR). The EXOR output is interpreted as the BER (Fig. 1). This provides a pseudo-BER estimation because the result from the fixed decision level is used as the reference data instead of the bit sequence of a real data code. Using these estimated BER values, we can estimate the lowest BER value by employing the same procedure as [2] (extracting the intersection of the fitting curve). The merit of this method is that it requires no knowledge of the transmitted bit sequences, and the performance of the data regeneration process in the master decision circuit is not degraded by the monitoring function.

By counting the bit error number, we also can estimate the amplitude histograms [3]. When the decision level of the second decision circuit is changed step by step through the whole eye opening, the number of "1" events counted at EXOR is recognized as the pseudo-error probability, which is the number of events whose amplitude is between the decision levels of the first and second decision circuits. The amplitude histograms are given by the absolute value of the derivation of this distribution, which is the same as that obtained by the synchronous sampling measurement. The Q-factor is evaluated using the amplitude histograms.

EXOR can be eliminated to realize the same function [3]. By counting "1" events at both the first (fixed decision level) and second (variable decision level) decision circuits while counting the clock period, and subtracting the number of "1" events, we can recognize the difference between a "1" event of the first and second decision circuits. Then the deviation of the probability curve of the difference also provides the amplitude histograms.

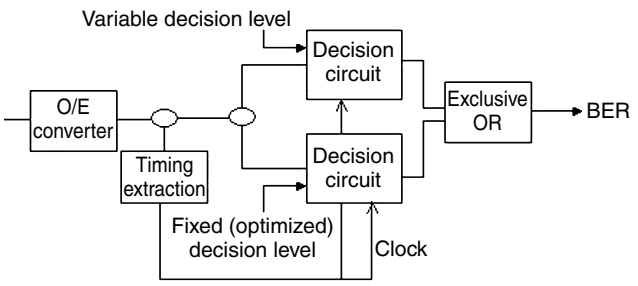


Figure 1. Typical configuration for dual decision circuits technique.

If the profile of the amplitude histograms does not change in the measurement time, the amplitude histograms can be estimated by using one decision circuit and a clock counter [5]. As a single decision circuit with a variable decision level does not have a reference, the error counter cannot reflect the influence of the amplitude fluctuation of each bit. This method is particularly advantageous with regard to BER measurement or error block detection, although it still requires timing extraction, which might depend on the signal format, bit rate, and/or modulation format.

3. ANALOG PARAMETERS

3.1. Statistical Method Using Histogram Evaluation

A method for signal quality monitoring that will provide a good measure of signal quality without the complexity of termination has long been desired and studied. Amplitude histogram estimation using a variable decision level (Section 2) can be recognized as a statistical method, but in this section, I focus on amplitude histogram evaluation using a sampling technique. Q-factor evaluation using synchronous sampling is more of a statistical method than a variable decision circuit technique because the sampling rate typically is low compared with the signal bit rate. However, Q-factor evaluation using the sampling technique is also a useful method when the system BER is too low to be measured within a reasonable amount of time, because it can be used for high bit rate signals, for which no decision circuit is used.

The amplitude histogram of an optical binary signal is obtained from an eye-diagram measured by the synchronous sampling technique. An amplitude histogram exhibits amplitude distributions at both mark and space levels at a fixed timing phase where the eye opening is at its maximum (Fig. 2a). The Q-factor at a fixed timing phase t ($Q(t_0)$) is estimated from the amplitude histogram, which is generally generated at timing phase t in the pattern as opposed to the data eye. It is defined as

$$Q(t_0) = |\mu_1(t) - \mu_0(t)|/(\sigma_1(t) + \sigma_0(t)) \tag{2}$$

where $\mu_i(t)$ and $\sigma_i(t)$ are the mean and standard deviation of the mark ($i = 1$) and space ($i = 0$) levels at t, respectively. The Q-factor (Q) is analytically related to the BER on a Gaussian assumption due to the equation defined by (1).

When the BER is low (e.g., $< 10^{-7}$), the theoretical value of Q is almost the same as the measured $Q(t_0)$ value:

$$Q(t_0) \sim Q \tag{3}$$

Eye monitoring using synchronous electrical sampling is conventionally undertaken with a digital sampling oscilloscope. However, the signal bit rate is limited by the O/E conversion bandwidth. Recently, an optical sampling

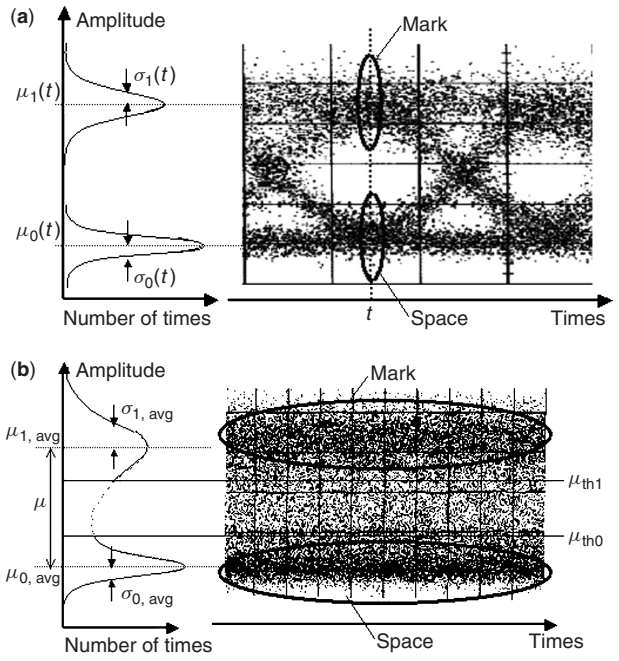


Figure 2. (**a**) Amplitude histograms at a fixed timing phase t of eye diagrams obtained by synchronous sampling; (**b**) amplitude histograms obtained by asynchronous sampling.

technique has been developed that has a temporal resolution of less than 1 ps and which overcomes the bit rate limitation [6]. Signal quality evaluation using eye diagrams obtained by means of optical sampling has also been reported [7,8]. The noise characteristics of optical amplifier systems do not have a Gaussian distribution. The analysis of non-Gaussian distributions is discussed in Ref. 9. Evaluations of crosstalk due to chromatic dispersion using histograms obtained by electrical sampling are presented in Ref. 10.

As mentioned previously, a statistical method using synchronous sampling is a useful technique for Q-factor evaluation. However, all sampling-based methods require synchronization and then some analysis, which makes them similar to protocol-aware termination in terms of cost and complexity. In fact, synchronous sampling also requires timing extraction using complex equipment that is specific to each BER and each format. Recently, the situation has begun to change. A simple, asynchronous histogram method was developed for Q-factor measurement [11]. Performance can be monitored at different monitoring points such as optical line repeaters, regenerators, or optical switching nodes (requires premeasurement). In other words, this method is expected to be applied to monitoring points where electrical termination is impossible. If we consider the all-optical network of the future, an optical switching node that has an all-optical switch will require performance monitoring without electrical termination.

Here, the averaged Q-factor (Q_{avg}) measurement obtained through asynchronous sampling is presented [11]. The asynchronous amplitude histogram is obtained from an eye-diagram measured by asynchronous optical sampling (Fig. 2b). Of the sampling points that constituted the histogram, it is determined that a set of points whose level is higher than a predetermined threshold level, μ_{th1}, belongs to level "mark" (i.e., "1"), while a set of points whose level is lower than a predetermined threshold level, μ_{th0} belongs to level "space" (i.e., "0"). Q_{avg} is defined by

$$Q_{avg} = |\mu_{1,avg} - \mu_{0,avg}|/(\sigma_{1,avg} + \sigma_{0,avg}) \tag{4}$$

where $\mu_{i,avg}$ and $\sigma_{i,avg}$ are the mean and standard deviation of the mark ($i = 1$) and space ($i = 0$) level distributions, respectively. The data obtained by asynchronous sampling include unwanted cross-point data in the eye-diagram, which reduces the measured value of the averaged Q-factor. Thus, it is necessary to remove the cross-point data. In this way, the two threshold levels were set at $\mu_{th1} = \mu_{1,avg} - \alpha\mu$ and $\mu_{th0} = \mu_{0,avg} + \alpha\mu$, and the coefficient α was defined as falling between 0 and 0.5.

In another approach, the peak levels of both mark and space level distribution (μ_1 and μ_0) are extracted, the data between μ_1 and μ_0 are eliminated and residual data at the mark level (larger than μ_1) and space level (smaller than μ_0) are symmetrically the reverse of the μ_1 and μ_0, respectively [12], and then the same evaluation is performed.

The essence of this method is that it does not use timing extraction or evaluate asynchronous eye diagrams. That is why this method provides signal format, modulation format, and bit rate flexibility. However, this technique is not timing jitter-sensitive despite the fact that jitter impairs the BER. This is the tradeoff with this method. Some analysis of bit rate flexibility and chromatic dispersion dependence has been provided [8].

3.2. Optical Power, Wavelength, and Optical SNR Monitor with Spectrum Measurement

Many methods using spectrum measurement have been published with a view to optical power, wavelength, and/or optical SNR (OSNR) monitoring. A simple approach for monitoring such kinds of analog parameter involves measuring the optical power spectrum of a tapped optical signal. However, this technique had two main problems. One is that the experimental equipment is large and expensive. The progress on DWDM networks makes it more difficult to monitor the optical power/ OSNR and wavelength of all channels. The second problem is that an optical spectrum monitor can measure the optical signal to out-of-band noise ratio, but it cannot monitor the optical SNR including in-band noise.

Recently, compact and stable equipment has been proposed by several groups to deal with the first problem. One approach employs an optical power and frequency monitor with a grating and a photo detector (PD) array [13], while another uses an arrayed waveguide grating (AWG) filter. With the latter technique, the AWG filtering wavelengths are controlled with the wavelength monitor and its feedback, using a reference light, thus realizing a precise wavelength and power monitoring. With yet another technique, a tunable Fabry-Perot etalon filter is used to separate the different spectral components in the spatial domain using temperature control [14]. An OSNR and optical wavelength monitor using an acousto-optic tunable filter have also been reported. The monolithic

integration of PD modules into an AWG filter is another approach for realizing an optical power and wavelength monitor in WDM networks. A point of interest as regards the abovementioned optical spectrum monitoring is its applicability to DWDM networks where the channel spacing is less than 50 GHz [15].

Further approaches have also been developed with a view to achieving precise OSNR monitoring. One technique monitors the OSNR by using the polarization-nulling technique, which employs the different polarization properties of optical signals and amplified spontaneous emission (ASE) noise [16]. The most recent approach to use this technique is reported in Ref. 16, where the polarization of an optical signal with ASE noise is controlled so that it is linear. The optical signal is split into a signal +ASE component and an ASE only component with polarization beam splitter (PBS). Then the ASE only component is divided into two with a 3 dB coupler. One of these ASE only components is optically filtered with a band-pass filter (BPF), and the powers of these three components are measured. As a result, OSNR is written as an equation of these three powers.

Another technique monitors OSNR by analyzing the low-frequency noise characteristics at the receiver [17]. With this technique, the total received power and received noise power are measured and compared. The noise power density is measured using an analog to digital converter and an FFT unit operating in the 40–50 kHz range. This technique is ineffective when the pattern length of the optical signal becomes longer than PRBS15. However, this limitation has been relaxed recently. In the latest result [18], the optical signal was split into orthogonally polarized lights by a polarization beam splitter and recombined after an optical delay $\Delta\tau$ in one of the paths. Then the measured electrical power P after the orthogonal delayed-homodyne module is written as

$$P = \{1 - 4\gamma(1-\gamma)\sin 2(\pi f \Delta\tau)\}S + N \qquad (5)$$

where γ is the power ratio at the polarization beam splitter, S and N are the electrical powers of the signal and noise, respectively, and f is the measured frequency. By setting $\Delta\tau$, f, and γ at adequate values, the electrical power component of the signal becomes zero, and the receiver noise component can be measured.

3.3. Pilot Tone Technique, Sub-Carrier Multiplexing

The pilot tone technique has been discussed for several years, and various approaches have been adopted. Sub-carrier multiplexing is used for the pilot tone and there are two main kinds of sub-carrier usage. One approach is to use the sub-carrier for monitoring the signal power, wavelength crosstalk, or OSNR. The other approach involves using the sub-carrier for an optical signal header. This means that the sub-carrier signal, which is modulated by header information data, is multiplexed into an optical signal (payload data) as a packet header or supervisory channel. Moreover, a variety of sub-carrier frequencies are used in these approaches (Fig. 3). The keys to these techniques relate to the methods used for combining and

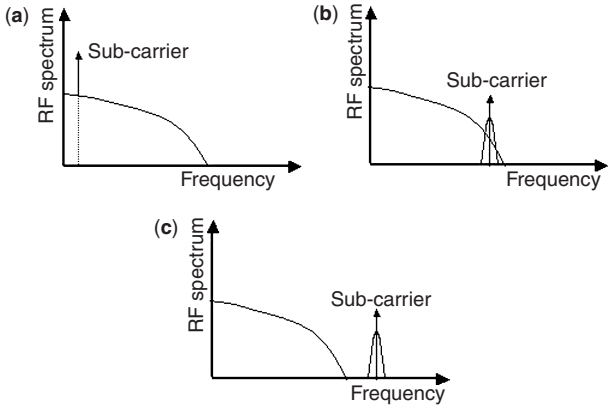


Figure 3. Typical RF spectrum of a data signal and a sub-carrier (pilot tone), when the sub-carrier frequency is (**a**) low (~ MHz), (**b**) near the signal bit rate (modulated by header information), and (**c**) higher than the signal bit rate (modulated by header information).

detecting the sub-carrier, the monitoring accuracy, and the influence of sub-carrier multiplexing on the signal.

The most effective approach for the pilot tone technique is to add a specific sinusoidal tone with a small amplitude to each optical WDM channel. In Ref. 19, the pilot tone is simply added to the laser bias current at the transmitter and is used to supervise individual wavelength channels along the optical path. At the transmitter an electrical pilot tone in the kHz regime is added to the signal, each wavelength channel being coded with a different tone frequency. The pilot tones are extracted at nodes between the transmitter and receiver by tapping a small portion of the signal power into a monitor module. The tapped optical power is detected, the pilot tones are filtered out electronically, and their levels are registered. The tones will provide signal identification and power level information for fault management. The tone amplitude is not more than 10% of the data level, and the tone frequency is below 100 kHz. This ensures that interference with the low-frequency components of the data results in negligible sensitivity degradation. Another approach uses a 10% peak-to-peak pilot tone modulation with a 50-kHz frequency and heterodyne detection with 20-Hz IF electrical filtering [20]. In Ref. 20, OSNR values measured by an optical spectrum analyzer and the pilot tone technique are compared, and good agreement in the OSNR region below 30 dB is reported. The OSNR sensitivity is limited to 30 dB with this technique because the electrical noise in the detector becomes nonnegligible.

Another approach designed simply for accurate frequency monitoring uses an AWG. With this approach, pilot tones with different RF frequencies are used for different wavelength channels. The pilot tones are split by the AWG where each center wavelength of each filter channel is set at the ITU grids using a temperature controller, so one AWG filter channel can split two pilot tones. The ratio of these two pilot tones with the same frequency split by adjacent AWG filters indicates the frequency of the optical channel [21].

Deciding on an adequate frequency range is also an important issue with respect to the accuracy of this

technique. Values ranging from several tens of kHz to the sub MHz level typically are used to minimize the penalty on a payload optical signal (Fig. 3a). However, a recent report states that when pilot tone–based monitoring techniques are used in amplified networks, their performance is deteriorated by the cross-gain modulation (XGM) of erbium-doped fiber amplifiers (EDFAs). The XGM problem is solved by using high-frequency tones in the 1-MHz range, but even when the pilot tone frequencies are in the few MHz range, the performance is limited by the Raman effect. As a result, tone frequencies higher than 100 MHz are recommended [22]. Another analysis suggests that, with regard to the pilot tone carrier to noise ratio, the tone detection sensitivity is limited by the power spectral density of the payload signal, and a pilot tone higher than 1 GHz is recommended for a 2.5-Gbit/s payload signal [23]. Some other approaches use pilot tone frequencies around the bit rate frequency, some of which is modulated by header information (Fig. 3b,c). The subject of pilot tone frequency remains an open issue.

In relation to the influence of the sub-carrier on the signal, the modulation intensity of a sub-carrier influences the payload signal, but detection sensitivity is reduced when the modulation intensity is low. This is a trade off problem and, for example, the number of payload signal wavelengths is limited because of this problem [23].

Wavelength conversion at intermediate nodes will become an important function in future DWDM networks. Some pilot tone–based approaches focusing on this function have already been reported [24]. These papers discuss the influence of the pilot tone in an interferometric wavelength converter using SOA, pilot tone frequency conversion with wavelength conversion using an semiconductor optical amplifier (SOA) + DFB laser, and the pilot tone technique when three SOA-interferometers are cascaded for wavelength conversion.

Some approaches use the sub-carrier technique for an optical signal header. Sub-carrier frequencies higher than the bit rate are sometimes used. The sub-carrier is encoded as a packet header using amplitude shift keying (ASK) or phase shift keying (PSK). The use of frequencies higher than the signal bit rate is advantageous in terms of the crosstalk between the sub-carrier and the payload signal. However, it is necessary to install high-speed, high-bandwidth electrical equipment at the transmitter and receiver. A sub-carrier frequency lower than the signal bit rate is used in [25] as a 50-Mbit/s NRZ channel overhead. A frequency of 9.73 GHz is used when the signal bit rate is 10 Gbit/s, using a differentially driven Mach-Zehnder (MZ) interferometer modulator. The sub-carrier is directly detected and filtered by an LPF. Another sub-carrier detection technique is proposed in Ref. 26. Here, the sub-carrier is encoded on the optical carrier by means of a dual-arm MZ $LiNbO_3$ (LN)-modulator, one arm is used for 10-Gbit/s data, the other is used for a 16.7-GHz RF tone modulated with a 100-Mbit/s ASK. A fiber loop mirror using polarization maintaining (PM) fiber birefringence is used to separate the baseband signal from the sub-carrier.

Other techniques focus on chromatic dispersion monitoring [27]. These techniques employ a chromatic dispersion monitor using the sub-carrier ratio method between higher and lower frequencies. The sub-carrier power measured at the receiver decreases due to the phase delay that the sub-carrier experiences in the dispersive fiber. A chromatic dispersion monitor using the optical side-band suppression method, which uses a sub-carrier frequency higher than the bit rate, measures the relative phase (time) delay between sub-carrier sidebands.

BIOGRAPHY

Ippei Shake was born in Kobe, Japan, in 1970. He received the B.S. and M.S. degrees in physics from Kyoto University, Kyoto, in 1994 and 1996, respectively. In 1996, he joined NTT Optical Network System Laboratories, NTT Corporation, Yokosuka, Japan. Since then he has been engaged in research and development of high-speed optical signal processing and high-speed optical transmission systems. He is now with NTT Network Innovation Laboratories, Yokosuka, Japan. His research interests also include optical networks, optical performance monitoring, and optical time-division multiplexing/demultiplexing circuits. He is a member of the Institute of Electrical and Electronics Engineers and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

BIBLIOGRAPHY

1. ITU-T Recommendation G.709.
2. N. S. Bergano, F. W. Kerfoot, and C. R. Davidson, Margin measurements in optical amplifier systems, *IEEE Photonics Tech. Lett.* **3**: 304–306 (1993).
3. R. Weismann, O. Bleck, and H. Heppner, Cost effective performance monitoring in WDM systems, *Optical Fiber Communication Conference 2000* (*OFC2000*), WK2, 2000.
4. M. Fregolent, S. Herbst, H. Soehnle, and B. Wedding, Adaptive optical receiver for performance monitoring and electronic mitigation of transmission impairments, *26th European Conference on Optical Communication* (*ECOC2000*), S2.1, 2000.
5. S. Ohteru and N. Takachio, Optical signal quality monitor using direct Q-factor measurement, *IEEE Photonics Tech. Lett.* **11**(10): 1307–1309 (1999).
6. H. Takara et al., 100 Gbit/s optical signal eye-diagram measurement with optical sampling using organic nonlinear optical crystal, *Electron. Lett.* **24**: 2256–2258 (1996).
7. C. Schmidt et al., Optical Q-factor monitoring at 160 Gb/s using an optical sampling system in an 80km transmission experiment, *Conference on Lasers and Electro-Optics 2002* (*CLEO 2002*), 579–580, 2002.
8. I. Shake and H. Takara, Transparent and flexible performance monitoring using amplitude histogram method, *OFC2002*, TuE1.
9. S. Norimatsu and M. Maruoka, Accurate Q-factor estimation of optically amplified systems in the presence of waveform distortion, *J. Lightwave, Tech.* **20**(1): 19–29 (2002).

10. C. M. Weinert, C. Caspar, M. Konitzer, and M. Rohde, Histogram method for identification and evaluation of crosstalk, *Electron. Lett.* **36**(6): 2000.

11. I. Shake, H. Takara, S. Kawanishi, and Y. Yamabayashi, Optical signal quality monitoring method based on optical sampling, *Electron. Lett.* **34**(22): 2152–2154 (1998).

12. M. Rasztovits-Wiech, K. Studer, and W. R. Leeb, Bit error probability estimation algorithm for signal supervision in all-optical networks, *Electron. Lett.* **35**(20): 1754–1755 (1999).

13. K. Otsuka et al., A high-performance optical spectrum monitor with high-speed measuring time for WDM optical networks, *23rd European Conference on Optical Communication* (*ECOC'97*), 147–150, 1997.

14. S. K. Shin, C. H. Lee, and Y. C. Chung, A novel frequency and power monitoring method for WDM network, *Optical Fiber Communication Conference '98* (*OFC'98*), WJ7, 1998.

15. H. Suzuki and N. Takachio, Optical signal quality monitor built into WDM linear repeaters using semiconductor arrayed waveguide grating filter monolithically integrated with eight photo diode, *Electron. Lett.* **35**(10): 836–837 (1999).

16. J. H. Lee and Y. C. Chung, Improved OSNR monitoring technique based on polarization-nulling method, *Electron. Lett.* **37**(15): (2001).

17. S. K. Shin, K. J. Park, and Y. C. Chung, A novel optical signal-to-noise ratio monitoring technique for WDM networks, *Optical Fiber Communication Conference 2000* (*OFC2000*), WK6, 2000.

18. C. J. Youn, K. J. Park, J. H. Lee, and Y. C. Chung, OSNR monitoring technique based on orthogonal delayed-homodyne method, *Optical Fiber Communication Conference 2002* (*OFC2002*), TuE3, 2002.

19. G. R. Hill et al., A transport network layer based on optical network elements, *J. Lightwave, Tech.* **11**(5): 667–679 (1993).

20. G. Bendelli, C. Cavazzoni, R. Girardi, and R. Lano, Optical performance monitoring technique, *26th European Conference on Optical Communication* (*ECOC2000*) **4**: 113–116 (2000).

21. C. J. Youn, S. K. Shin, K. J. Park, and Y. C. Chung, Optical frequency monitoring technique using arrayed-waveguide grating and pilot tones, *Electron. Lett.* **37**(16): 2001.

22. H. S. Chung et al., Effects of stimulated Raman scattering on pilot-tone-based WDM supervisory technique, *IEEE Photonics Tech. Lett.* **12**(6): 731–733 (2000).

23. Y. Hamazumi and M. Koga, Transmission capacity of optical path overhead transfer scheme using pilot tone for optical path network, *J. Lightwave, Tech.* **15**(12): 2197–2205 (1997).

24. A. Bissons et al., Analysis of evolution of over-modulated supervisory data in a cascade of all-optical wavelength converters, *Optical Fiber Communication Conference 2000* (*OFC2000*), ThD4, 2000.

25. M. Rohde et al., Control modulation technique for client independent optical performance monitoring and transport of channel overhead, *Optical Fiber Communication Conference 2002* (*OFC2002*), TuE2, 21–22, 2002.

26. G. Rossi, O. Jerphagnon, B.-E. Olsson, and D. J. Blumenthal, Optical SCM data extraction using a fiber-loop mirror for WDM network system, *IEEE Photonics Tech. Lett.* **12**(7): 897–899 (2000).

27. M. N. Petersen et al., Online chromatic dispersion monitoring and compensation using a single inband subcarrier tone, *IEEE Photonics Tech. Lett.* **14**(4): 570–572 (2002).

SIGNATURE SEQUENCES FOR CDMA COMMUNICATIONS

Tony Ottosson
Erik Ström
Arne Svensson
Chalmers University of Technology
Göteborg, Sweden

1. INTRODUCTION

A multiple access method is a method for allowing several users to share a common physical channel, such as, a coaxial cable, an optical fiber, or a band of radio frequencies. Common multiple access methods are frequency-division multiple access (FDMA) and time-division multiple access (TDMA) [1–3]. Strictly speaking, the term "multiple access" is applicable for the case when the users are not at the same geographic location and the term "multiplexing" is applicable when they are; however, we use "multiple access" for both cases. In FDMA, the users' signals are selected such that they do not overlap in the frequency domain (an example is FM or AM broadcast radio), and in TDMA, the users' signals do not overlap in time (i.e., the users take turns using the channel). Ideally, this means that the users do not disturb each other, and the signals are said to be orthogonal. (In practice, there will be some interuser interference due to imperfections in the implementation and nonideal channels, but we will ignore these complications here.)

A third multiple access method that has become increasingly popular is code-division multiple access (CDMA) [4,5]. In CDMA, the users signals are overlapping in both time and frequency. This does not, however, imply that CDMA signals are necessarily nonorthogonal. As we will see shortly it is indeed possible, under certain strong restrictions, to find orthogonal signals that overlap in time and frequency.

Let us consider a simple example to illustrate the main idea behind the most common type of CDMA: direct-sequence CDMA (DS-CDMA). The name DS-CDMA stems from the fact that the users' signals are direct-sequence spread spectrum signals (DS-SS signals) [6–8]. Suppose that we have two users who each want to transmit data at a rate of $1/T$ bits/second. Let $b_1(t)$ and $b_2(t)$ be the data waveform of user 1 and 2, respectively, where bits are coded as ± 1 amplitudes of the waveforms. The kth user is assigned a signature waveform, $c_k(t)$, and the transmitted signal, $s_k(t)$, is formed as $s_k(t) = b_k(t)c_k(t)$, see Fig. 1. The data waveform is defined by the user's information bit sequence: in this case $b_1[n] = \{1, -1, \ldots\}$ for user 1 and $b_2[n] = \{-1, 1, \ldots\}$ for user 2. Similarly, the signature waveforms are defined by the users' signature sequences: $c_1[n] = \{1, -1, 1, 1, 1, -1, 1, 1, \ldots\}$ for user 1 and $c_2[n] = \{1, 1, -1, 1, 1, 1, -1, 1, \ldots\}$ for user 2. We note that the data waveforms can change polarity every T seconds and the signature waveforms can change polarity every T_c seconds. The ratio $N = T/T_c$ is called the spreading factor and also processing gain.

It is the signature waveform (also known as the code waveform or spreading waveform) that enables the

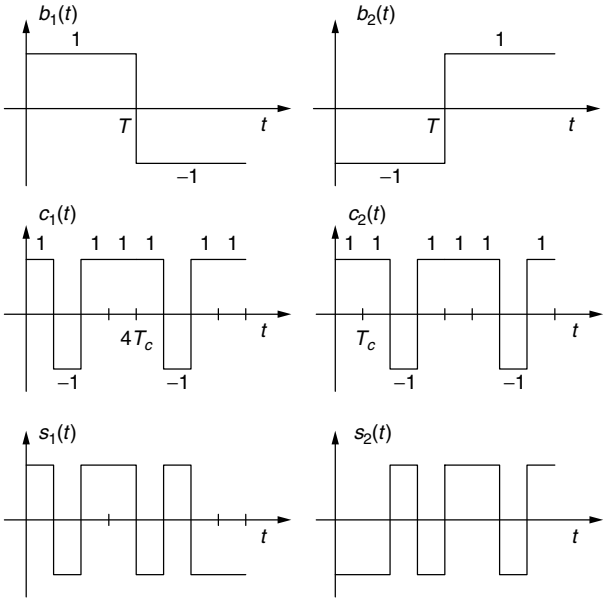


Figure 1. Data waveforms, signature waveforms, and transmitted waveforms for two users in a direct-sequence CDMA system.

receiver to separate users. Hence, the waveforms must be unique among the users. The process of multiplying the data waveform with the signature waveform is called spreading, since the bandwidth of the transmitted signal is significantly larger than the bandwidth of the data waveform. As a matter of fact, the bandwidth of the transmitted signal is approximately N times larger than the data waveform bandwidth (which justifies the terminology "spreading factor" for N).

The transmitted signals are added by the channel[1] and the received waveform is $r(t) = s_1(t) + s_2(t)$ (we ignore channel noise and other nonideal channel effects here). Now suppose that the receiver is interested in detecting the data from user 1. A block diagram of the system under consideration is found in Fig. 2. The receiver multiplies the received signal with the signature waveform of user 1, $c_1(t)$, and this results in the signal

$$\begin{aligned} z_1(t) &= r(t)c_1(t) \\ &= [b_1(t)c_1(t) + b_2(t)c_2(t)]c_1(t) \\ &= b_1(t)\underbrace{c_1^2(t)}_{=1} + b_2(t)c_2(t)c_1(t) \\ &= b_1(t) + b_2(t)c_2(t)c_1(t) \end{aligned}$$

We see that the signal $z_1(t)$ is the sum of two terms: the desired data waveform of user 1 and an interference term that is due to user 2. Since we have reduced the bandwidth of the desired part of the signal, the process of multiplying the received signal with the signature waveform is called despreading.

The standard method for recovering the data from a signal disturbed by additive noise or interference is to process the noisy signal with a matched filter. In this case, the filter should be matched to a rectangular pulse of length T seconds. Hence, the matched filter can be implemented as an integrator over T seconds. That is, the output of the matched filter is

$$y_1(t) = \int_{t-T}^{t} z_1(u)\,du$$

To decided on the nth bit, we sample the matched filter at time $(n+1)T$. Hence, to recover $b_1[0]$, we sample the filter at time T, which results in

$$\begin{aligned} y_1(T) &= \int_0^T z_1(t)\,dt \\ &= \int_0^T b_1(t)\,dt + \int_0^T b_2(t)c_1(t)c_2(t)\,dt \\ &= \int_0^T b_1[0]\,dt + b_2[0]\underbrace{\int_0^T c_1(t)c_2(t)\,dt}_{=0} \\ &= b_1[0]T \end{aligned}$$

where $b_1[0] = 1$ and $b_2[0] = -1$ are the transmitted bits. We see that the data bit from user 1 can be recovered as $\text{sgn}[y_1(T)]$, regardless of the value of $b_2[0]$. The condition

[1] There are situations when the transmitted signals are added already in the transmitter. This is the case when several users' signals are transmitted from the same geographical position. One example of this is a downlink in a cellular CDMA system, where all the users' signals in one cell are transmitted from the same base station [1–3].

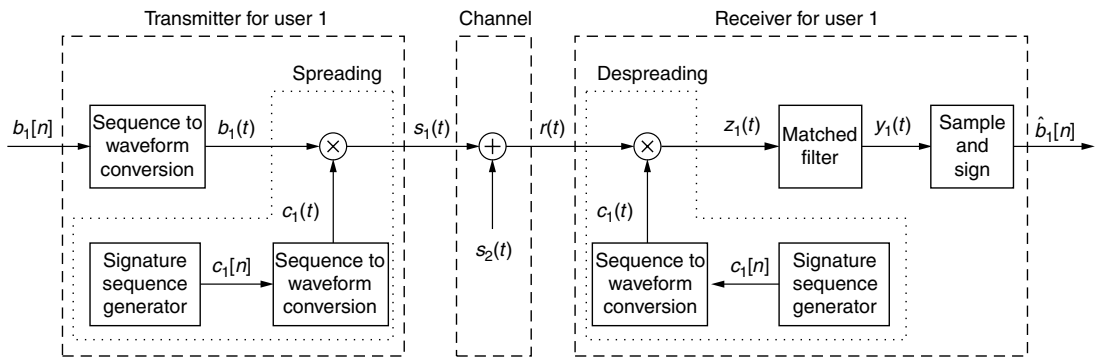


Figure 2. A block diagram for the transmitter and receiver for user 1. The channel is assumed to be noise-free and the only interference added is the signal from user 2.

that the cross-correlation between $c_1(t)$ and $c_2(t)$ is zero, that is,

$$\int_0^T c_1(t)c_2(t)\,dt = 0$$

proves that the signature waveforms $c_1(t)$ and $c_2(t)$ are orthogonal (over the interval $0 \le t \le T$). As long as the signature waveforms are orthogonal, users will not interfere with each other after despreading and matched filtering, and the data can be recovered without error (if we ignore channel noise). However, it is not necessary for the signature sequences to be completely orthogonal for a CDMA system to work. As a matter of fact, as long as the cross-correlation between the signature waveforms is small, the interference alone cannot cause errors (although the resistance against noise can be reduced).

It should now be clear that we need to choose signature sequences carefully to avoid excessive interference between users. In short, we want to select sequences that have low cross-correlations. As is discussed in more detail later, it is also important to select sequences that have good partial cross-correlation and autocorrelation properties. The autocorrelation of a sequence is the cross-correlation between the sequence and a time-shifted version of the same sequence. Ideally, we want the autocorrelation to be small for all nonzero time-shifts. The partial cross-correlation is the correlation between partial segments of the sequences. Ideally, the partial cross-correlation should be small for all choices of sequence segments.

This definition of "good correlation properties" is applicable for DS-CDMA, but not necessarily for other types of CDMA. Although many forms of CDMA exist, the most common type apart from DS-CDMA is frequency-hopping CDMA (FH-CDMA). FH-CDMA is similar to FDMA in that the users are assigned different frequency channels that are meant to be for exclusive use by the users. In FH-CDMA, the frequency channel allocation for a user is constantly changing. That is, the user hops among frequencies, and the hopping is done according to the user's signature sequence. Every now and then, two or more users will be assigned the same frequency. This is known as a collision and is a highly undesirable event, and FH-CDMA signature sequences are therefore primarily designed to avoid collisions. Hence, the signature sequences used for DS-CDMA and FH-CDMA are different. We will not cover FH-CDMA sequences further in this article; the interested reader is referred to Ref. [9] for an overview. More details on FH sequences (and DS sequences) can also be found in Refs. [6–8].

In addition to good cross-correlation properties (or good collision properties), it is desirable that the sequences are easily generated, that is, with as little hardware and software complexity as possible. Furthermore, in some applications, we also are interested in making the signature sequence a secret for all but the intended receiver. It is then important that it will be difficult for an unauthorized receiver to predict future values of the signature sequence by observing its past values. It should be noted that in most current CDMA systems, the signature sequences have not been designed with this last requirement in mind. Hence, security in these systems usually is obtained by other cryptographic methods.

We know that completely random sequences have, on average, good correlation properties and are impossible to predict. However, we cannot use true random sequences since both the transmitter and the receiver must be able to produce the same sequences. Instead, most sequences used in practice are derived from pseudorandom sequences (PN-sequences). The theory of PN sequences involves finite-field algebra; however, we will leave out all details here and rather give a brief description of how PN-sequences can be generated with digital hardware (or software). Most sequences in practical use are binary sequences (also known as bi-phase sequences) or 4-ary sequences (quadriphase sequences). The latter type is useful for quadrature modulated systems (e.g., phase-shift keying or quadrature amplitude modulation).

2. SPREAD SPECTRUM AND CDMA

In the introduction, it is noted by a simple example that, the signature sequences should be unique and have low partial cross-correlations for all time-shifts. To quantify these statements, let us go into more details.

As before, assume that a user k transmits the signal

$$s_k(t) = \sqrt{E_k/T}\,b_k(t)c_k(t) \tag{1}$$

where

$$b_k(t) = \sum_{i=0}^{\infty} b_k[i]h(t - iT) \tag{2}$$

is the data waveform

$$c_k(t) = \sum_{i=0}^{\infty} c_k[i]g(t - iT_c) \tag{3}$$

the signature waveform, and $h(t)$ a rectangular pulse shape given by

$$h(t) = \begin{cases} 1 & 0 \le t \le T \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The chip-pulse-shape $g(t)$ can in general take any form. To simplify the presentation here, only the rectangular chip-pulse-shape is considered, such that

$$g(t) = \begin{cases} 1 & 0 \le t \le T_c \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In practice, more spectrally efficient waveforms are used, but for the purpose of this article this only makes the mathematics untrackable. The properties and results that are discussed here hold with small changes for all pulse shapes used in practice. The data sequence and the chip sequence are denoted $b_k[i]$ and $c_k[i]$, respectively. In general, both of these sequences are complex-valued and take values from limited sets. Again for simplicity, we will concentrate here on real-valued data symbols and binary chips. The only difference in this model compared to the model presented in the introduction is that we have normalized the signal so its energy is

$$\int_0^T s_k^2(t)\,dt = \frac{E_k}{T}\int_0^T b_k^2(t)c_k^2(t)\,dt = \frac{E_k}{T}T = E_k \tag{6}$$

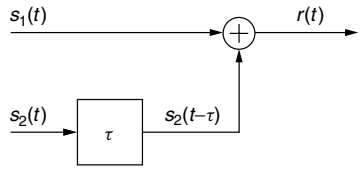


Figure 3. Block diagram of a two-user asynchronous noise-free channel with relative delay τ.

Let us consider a two-user case where the users are asynchronous (i.e., not time-aligned); see Fig. 3. This is the typical situation in a true multiple access scenario (such as the uplink in a cellular system). The received waveform can be expressed as

$$r(t) = s_1(t) + s_2(t-\tau) \tag{7}$$

where τ is the relative time-offset between the signals of users 1 and 2 depending on the propagation times of the user signals, and hence their distances to the receiver. Without loss of generality, we assume that $0 \le \tau < T$. Here we have, for simplicity, neglected the receiver noise that is always present, since it will not have any influence on the properties of the signature waveforms.

Assuming that we are interested in detecting user 1, we first despread the signal by multiplying with the waveform of user 1 as

$$\begin{aligned} z_1(t) &= r(t)c_1(t) \\ &= \left[\sqrt{E_1/T}b_1(t)c_1(t) + \sqrt{E_2/T}b_2(t-\tau)c_2(t-\tau)\right]c_1(t) \\ &= \sqrt{E_1/T}b_1(t) + \sqrt{E_2/T}b_2(t-\tau)c_2(t-\tau)c_1(t) \end{aligned} \tag{8}$$

The despread signal consists of the desired signal part and an interfering part from user 2. To recover $b_1[0]$ we calculate the output of the normalized filter matched to $h(t)$ at time T, which becomes

$$\begin{aligned} y_1(T) &= \frac{1}{\sqrt{T}}\int_0^T z_1(t)\,dt \\ &= \sqrt{E_1}\frac{1}{T}\int_0^T b_1(t)\,dt \\ &\quad + \sqrt{E_2}\frac{1}{T}\int_0^T b_2(t-\tau)c_2(t-\tau)c_1(t)\,dt \\ &= \sqrt{E_1}b_1[0]\frac{T}{T} + \sqrt{E_2}b_2[-1]\frac{1}{T}\int_0^\tau c_2(t-\tau)c_1(t)\,dt \\ &\quad + \sqrt{E_2}b_2[0]\frac{1}{T}\int_\tau^T c_2(t-\tau)c_1(t)\,dt \end{aligned} \tag{9}$$

The integrals in this expression are partial cross-correlations between the waveforms $c_1(t)$ and a time-shifted version of $c_2(t)$. Assuming that τ is a multiple of the chip time given as $\tau = pT_c$ $(0 \le p < N)$, we can calculate the individual partial cross-correlations as function of the discrete signature sequences as

$$\frac{1}{T}\int_0^\tau c_2(t-\tau)c_1(t)\,dt = \begin{cases} \dfrac{T_c}{T}\displaystyle\sum_{n=0}^{p-1} c_2[n-p]c_1[n] & 0 < p < N \\ 0 & p = 0 \end{cases} \tag{10}$$

$$\frac{1}{T}\int_\tau^T c_2(t-\tau)c_1(t)\,dt = \frac{T_c}{T}\sum_{n=p}^{N-1} c_2[n-p]c_1[n] \tag{11}$$

These derivations can easily be extended to more than two users with user i as the desired user. We now introduce the discrete partial cross-correlations as

$$X_{k,i}(p) = \sum_{n=p}^{N-1} c_k[n-p]c_i[n] \quad 0 \le p < N \tag{12}$$

and

$$\overline{X}_{k,i}(p) = \begin{cases} \displaystyle\sum_{n=0}^{p-1} c_k[n-p]c_i[n] & 0 < p < N \\ 0 & p = 0 \end{cases} \tag{13}$$

With these discrete partial cross-correlations the matched filter output in Eq. (9) can be rewritten as

$$\begin{aligned} y_1(T) &= \sqrt{E_1}b_1[0] + \sqrt{E_2}b_2[-1]\overline{X}_{2,1}(p)/N \\ &\quad + \sqrt{E_2}b_2[0]X_{2,1}(p)/N \end{aligned} \tag{14}$$

We clearly see that we get the desired signal $\sqrt{E_1}b_1[0]$ but also two multiple-access interference (MAI) terms that depend on the partial cross-correlations between the signature sequences of the two users. For ideal output (only the desired part), the sum of the two MAI terms should be zero for any combination of data symbols $b_2[-1]$ and $b_2[0]$, and for all values of the time-shift p, since p may vary. These equations can easily be extended to more users. The MAI term from user k on user i is obtained by replacing index 2 by index k and index 1 by index i in Eq. (14). Ideally, the sum of all these MAI terms should be zero for all time-shifts p. This is guaranteed when $\overline{X}_{k,i}(p) = 0$ and $X_{k,i}(p) = 0$ for all $p \in [0, N-1]$ and for all $k \ne i$, although in some cases this is an unnecessarily strong requirement.

For signature sequences with period N the cross-correlations can be expressed as

$$\begin{aligned} X_{k,i}(p) &= C_{k,i}(p) \\ \overline{X}_{k,i}(p) &= C_{k,i}(p-N) \end{aligned} \tag{15}$$

where we have introduced the aperiodic cross-correlation parameter $C_{k,i}(m)$ defined as

$$C_{k,i}(m) = \begin{cases} \displaystyle\sum_{n=0}^{N-1-m} c_k[n]c_i[n+m] & \text{if } 0 \le m \le N-1, \\ \displaystyle\sum_{n=0}^{N-1+m} c_k[n-m]c_i[n] & \text{if } 1-N \le m < 0, \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

which is commonly used in the literature [10].

In many systems, the channel for the kth user can be modeled as a channel with L distinct propagation paths, where the lth path has complex gain $g_k(l)$ and delay $\tau_k(l)$.

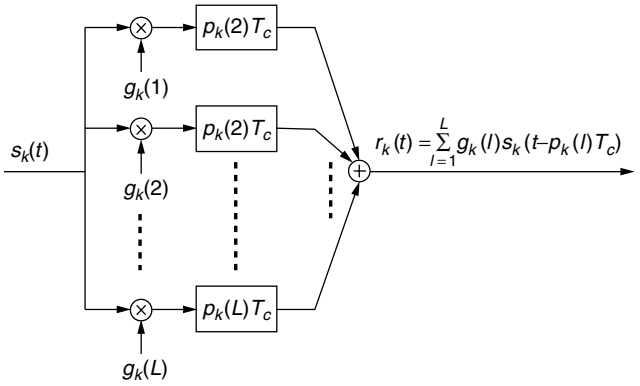


Figure 4. Noise-free L-path channel for the kth user. The lth path has complex gain $g_k(l)$ and delay $p_k(l)T_c$.

For simplicity, we assume that the delays can be written as $\tau_k(l) = p_k(l)T_c$, where $p_k(l)$ is an integer in $[0, N-1]$ and $p_1(1) = 0$. The multipath channel for the kth user is shown in Fig. 4.

Generalizing the given expressions for the case of a uplink CDMA system with K users experiencing delay spread, it can be shown that the matched filter output for user 1 at time T (path 1) can be written as

$$\begin{aligned} y_1(T) = {} & \sqrt{E_1}g_1(1)b_1[0] + \\ & \underbrace{\sum_{l=2}^{L}\sqrt{E_1}g_1(l)\Big(b_1[-1]\overline{X}_{1,1}(p_1(l))/N + b_1[0]X_{1,1}(p_1(l))/N\Big)}_{\text{ISI(+ICI)}} \\ & + \underbrace{\sum_{k=2}^{K}\sum_{l=1}^{L}\sqrt{E_k}g_k(l)\Big(b_k[-1]\overline{X}_{k,1}(p_k(l))/N + b_k[0]X_{k,1}(p_k(l))/N\Big)}_{\text{MAI}} \end{aligned} \tag{17}$$

The desired part is $\sqrt{E_1}g_1(1)b_1[0]$ in Eq. 17. The second part is the intersymbol interference (ISI), which is a weighted sum of aperiodic autocorrelations (partial cross-correlations between two equal sequences with different time-shifts) $\overline{X}_{1,1}(p_1(l))$ and $X_{1,1}(p_1(l))$. The part of the ISI that depends on $b_1[0]$ is sometimes referred to as interchip interference (ICI), but we prefer to refer to all of it as ISI. The third part is the MAI part, which is a weighted sum of the partial cross-correlations $\overline{X}_{k,1}(p_k(l))$ and $X_{k,1}(p_k(l))$. The ideal signature sequences should make the matched filter output as close as possible to the desired component. Again, this is obtained when all partial cross-correlations are zero for all time-shifts and the partial autocorrelation is zero for all time-shifts except zero. In practice, it may be enough to require that these partial correlations are close to zero.

From Eq. (17) it is seen that some information about the desired data symbol $b_1[0]$ is not used to form the desired part of the decision variable but instead becomes interference. The performance of a spread spectrum and CDMA receiver can be improved by using all the information about $b_1[0]$, which is available in the received signal (also that in the multipath components of the desired signal) [6–8]. Such a receiver is not matched to the transmitted waveform of the desired user only, but to that waveform convolved with the channel impulse response, and is commonly referred to as a RAKE receiver [6,8,11]. Now the decision variable becomes the weighted sum

$$y_{\text{rake}} = \sum_{l=1}^{L} g_1^*(l)y_1(T + p_1(l)T_c) \tag{18}$$

This expression can be worked out in the same way as Eq. (17). Again, it consists of three terms: the desired term, an ISI term, and a MAI term. The ISI and MAI terms are guaranteed to becomes zero under the same requirements as above, that is, the partial autocorrelation should be zero at all time-shifts except 0 and the partial cross-correlations should be zero at all time-shifts.

So far we have considered only the decision variable for data detection. The normalized matched filter output

$$y_1(t) = \frac{1}{\sqrt{T}}\int_{t-T}^{t} z_1(u)\,du \tag{19}$$

may also be used for acquisition and synchronization. Note that we have assumed that the sampling point T used to obtain a decision variable for $b_1[0]$ was known in the preceding derivation. In practice, this sampling point must be estimated. One common way of doing this is to find the maximum over a symbol interval of the matched filter output, since this maximum most likely corresponds to the time-shift where the timing of the received signal and the regenerated signal in the receiver are aligned. This normally is implemented as a two-stage procedure, where the matched filter output $y_1(t)$ is compared to a threshold in the first stage. The time where the threshold is exceeded is taken as a rough estimate of the timing (aquisition). In the next stage, the maximum of $y_1(t)$ in the vicinity of the point found in the first stage is found (synchronization) [6,8].

The contribution from the desired signal in $y_1(t)$ becomes very similar to the MAI term in Eq. (9), except that the user index is the same on both signature waveforms. After some derivations,

$$y_1(pT_c) = \sqrt{E_1}b_1[-1]\overline{X}_{1,1}(p)/N + \sqrt{E_1}b_1[0]X_{1,1}(p)/N \tag{20}$$

when $r(t) = s_1(t)$. On a multipath channel, there will be more terms of the type $\overline{X}_{1,1}(p + p_1(l))$ and $X_{1,1}(p + p_1(l))$. To have a distinct peak to use for timing estimation, it is clear that the requirements on the signature sequence used for synchronization is the same as the requirement to obtain low ISI in data detection.

3. COMMON SIGNATURE SEQUENCES

In this section, we describe some commonly used signature sequences and briefly discuss their properties.

The signature waveform of user k will, as before, be denoted as $c_k(t)$. This waveform is obtained as given in Eq. (3) where $c_k[i]$ takes values from $\{\pm 1\}$. The signature sequences are commonly defined by arithmetics in the binary field GF(2) where the elements are denoted as 0 and 1. In the following, we use uppercase letters to denote variables in GF(2) and lowercase letters to denote the corresponding antipodal variables. This means that the antipodal signature sequence is obtained from the corresponding signature sequence in GF(2) from

$$c_k[i] = 1 - 2\,C_k[i] \tag{21}$$

We do not use different symbols for addition and multiplication, because we believe it is clear from the expression whether it is operations in GF(2) or in the field of real numbers. The user index k will be suppressed when there is no reason to distinguish between several users.

Periodic sequences, like the ones discussed in this section, can be used in several ways to form signature waveforms for spread spectrum and CDMA. One way is to map the complete binary sequence to an antipodal sequence using Eq. (21), and then form the signature waveform as described in Eq. (3). When the spreading factor N is identical to the period P, this is referred to as short codes, while the case of $P > N$ corresponds to long codes.[2] With short codes, the signature waveform becomes identical for each transmitted symbol, while it changes over time for long codes. However, sometimes instead the signature waveform is obtained from a periodic repetition of part of the full period of the original sequence. Mathematically, this means that the signature sequence used to form the signature waveform is given by

$$\{C[0], C[1], \ldots, C[Q], C[0], C[1], \ldots, C[Q], \ldots\} \tag{22}$$

where $Q < P$. Different users may use different phase shifts of the same periodic sequence, or they may use different periodic sequences.

3.1. Maximal Length Sequences

Maximal length sequences (m-sequences) are generated by a shift register with feedback and have the maximum period that can be obtained with the given shift register [12]. The signature sequence $\{C[0], C[1], \ldots\}$ is generated by the recursive formula

$$\begin{aligned} C[i] &= G[1]C[i-1] + G[2]C[i-2] + \cdots + G[m]C[i-m] \\ &= \sum_{k=1}^{m} G[k]C[i-k] \end{aligned} \tag{23}$$

where $i \geq m$ and m is the length (memory) of the shift register and is commonly also referred to as the degree of the sequence. The coefficients $\{G[1], G[2], \ldots, G[m]\}$ and the initial state $\{C[0], C[1], \ldots, C[m-1]\}$ specify the sequence. The coefficient $G[m]$ is always 1 for binary m-sequences. The maximum period of this signature sequence is $P = 2^m - 1$ and is obtained when the characteristic polynomial $G(D) = G[m]D^m + G[m-1]D^{m-1} + \cdots + D + 1$ is an irreducible and primitive polynomial that divides $D^P + 1$ [7,8,12,13]. An irreducible polynomial cannot be factored and a primitive polynomial $G(D)$ of degree m is one for which the period of the coefficients of $1/G(D)$ is P.

M-sequences exist for all $m > 1$ and the number of characteristic polynomials is

$$N_p(m) = \frac{2^m - 1}{m} \prod_{i=1}^{k} \frac{P_i - 1}{P_i}$$

where $\{P_1, P_2, \ldots, P_k\}$ are prime numbers such that

$$2^m - 1 = \prod_{i=1}^{k} P_i^{m_i}$$

where $\{m_1, m_2, \ldots, m_k\}$ are integers.[3] It should be noted that a characteristic polynomial $G(D)$ can be reversed as $G'(D) = D^m + G[1]D^{m-1} + \cdots + G[m-1]D + G[m]$ to give a reversed m-sequence. These reversed polynomials are included in $N_p(m)$. For a given characteristic polynomial, different initial states give a different phase shift of the same sequence. In Table 1, the number of characteristic polynomials and the maximum period are given for degree m m-sequences. It is clear that the number of characteristic polynomials and thus sequences increases very fast when m increases. Characteristic polynomials for many periods can be found in Refs. [7,8,11,14,15] and we summarize some of them in Table 2 (the reversed polynomials are not included in this table).

Table 1. The Maximum Period and the Number of Characteristic Polynomials for m-Sequences of Degree *m*

m	$P = 2^m - 1$	$N_p(m)$
2	3	1
3	7	2
4	15	2
5	31	6
6	63	6
7	127	18
8	255	16
9	511	48
10	1023	60
11	2047	176
12	4095	144
13	8191	630
14	16383	756
15	32767	1800
16	65535	2048
17	131071	7710
18	262143	8064
19	524287	27594
20	1048575	24000

[2] The case $N > P$ should be avoided and is not discussed further here.

[3] This is a so-called prime decomposition.

Table 2. Characteristic Polynomials of m-Sequences. The Polynomials Are Given in Octal Notation. After Converting the Octal Numbers to Binary Numbers, the Left-Most Binary 1 Corresponds to $G[m]$. As an Example, $23_{\text{octal}} = 010011_{\text{binary}}$, Which Corresponds to $G(D) = D^4 + D + 1$

m	Characteristic Polynomials in Octal Form: $\{G[m], G[m-1], \ldots, G[1], 1\}$
2	7
3	13
4	23
5	45, 75, 67
6	103, 147 155
7	211, 217, 235, 367, 277, 325, 203, 313, 345
8	435, 551, 747, 453, 545, 537, 703, 543
9	1021, 1131, 1461, 1423, 1055, 1167, 1541, 1333, 1605, 1751, 1743, 1617, 1553, 1157
10	2011, 2415, 3771, 2157, 3515, 2773, 2033, 2443, 2461, 3023, 3543, 2745, 2431, 3177

In every period of length P of a m-sequence, the number of zeros is $2^{m-1} - 1$ and the number of ones is 2^{m-1}. Moreover, half of the number of runs of ones and zeros have length 1, 1/4 have length 2, 1/8 have length 3, and in general $1/2^k$ have length k with $k < m$.

With short codes using the full period of the m-sequences, that is, $Q = P = N$, it is well known that the periodic autocorrelation function is given by Refs. [7,8]

$$\phi(p) = \overline{X}_{k,k}(p) + X_{k,k}(p) = \begin{cases} N & \text{when } p = 0 \\ -1 & 1 \le p < N \end{cases} \tag{24}$$

For large N, $\phi(p)/\phi(0) = -1/N$, where $p \neq 0$, becomes insignificant, and the autocorrelation function approaches the autocorrelation function of an uncorrelated sequence. Therefore, m-sequences have excellent properties for synchronization in direct sequence spread spectrum when the data waveform is constant and the channel flat, since it contains well-defined autocorrelation peaks. They also lead to very small ISI under the same assumption. However, it is also well known that the periodic cross-correlation given by

$$\Phi_{k,i}(p) = \overline{X}_{k,i}(p) + X_{k,i}(p) \tag{25}$$

may have significant values [10,11]. The maximum periodic cross-correlation

$$\Phi_{\max} = \max_{p,k,i \neq k} |\Phi_{k,i}(p)| \tag{26}$$

between any pair of m-sequences is given in Table 3 for some short m-sequences. From table 3 it is clear that the maximum cross-correlation peak may be more than one-third of the length of the sequence. The maximum MAI is proportional to the maximum cross-correlation values when $b_k[-1] = b_k[0]$. Much less appears to be known about ISI and MAI in the general case when $b_k[-1] \neq b_k[0]$ and also about the synchronization properties when $b_i[-1] \neq b_i[0]$.

As an example of short codes based on m-sequences, we choose the length 31 sequence generated by using $G(D) = D^5 + D^4 + D^3 + D^2 + 1$ (75 in octal as seen in Table 2). When the initial values of the signature sequence are all ones, the first period becomes {1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0}. In Fig. 5 the aperiodic and periodic discrete autocorrelation sequences ($\phi(p)$) are shown as well as $X_{1,1}(p) - \overline{X}_{1,1}(p)$, which is proportional to the MAI when $b_k[-1] = -b_k[0]$. With binary antipodal modulation, the absolute value of the maximum MAI between two users using different shift of the same m-sequence is proportional to the maximum of $|X_{1,1}(p) - \overline{X}_{1,1}(p)|$ and

Table 3. The Maximum Periodic Cross-Correlation $\Phi_{\max}$ for m-Sequences of Degree m

m	$\Phi_{\max}$	$\Phi_{\max}/P$
3	5	0.71
4	9	0.60
5	11	0.35
6	23	0.36
7	41	0.32
8	95	0.37
9	113	0.22
10	383	0.37
11	287	0.14
12	1407	0.34

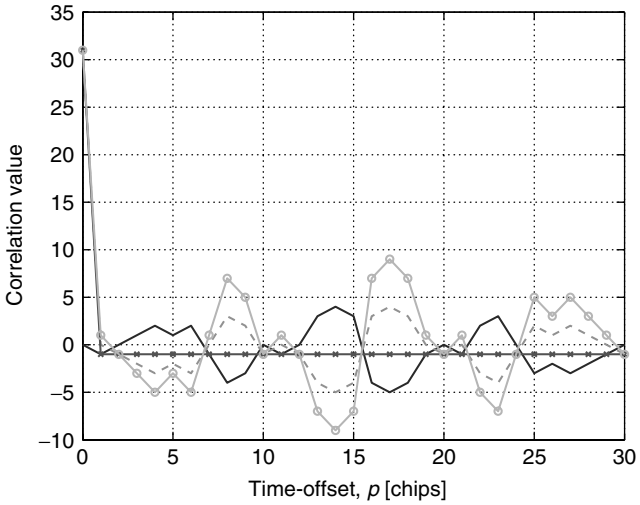


Figure 5. $\overline{X}_{1,1}(p)$ (solid line), $X_{1,1}(p)$ (dashed line), $\phi(p)$ (cross), and $X_{1,1}(p) - \overline{X}_{1,1}(p)$ (circle) for the m-sequence based on characteristic polynomial $G(D) = D^5 + D^4 + D^3 + D^2 + 1$.

$|\phi(p)|$ for $p \neq 0$. From Fig. 5 we see that this maximum MAI in this example is not larger than the maximum periodic cross-correlation $\Phi_{\max}$ as shown in Table 3. However, the maximum cross-correlation $|\Phi_{k,i}(p)|$ between the m-sequences generated by $G(D) = D^5 + D^4 + D^3 + D^2 + 1$ and $G(D) = D^5 + D^2 + 1$ (45 in octal) is 13, so the maximum correlation is not guaranteed to be at most the values given in Table 3. The conclusion is that $\Phi_{\max}$ should not be taken as a guarantee of the maximum MAI.

With long codes, much less is known about the autocorrelation and cross-correlation properties of signature sequences based on m-sequences. In this case, it is the partial correlations over $N < P$ chips that are important. The same is true for both long and short codes on multipath fading channels. In this case, the correlation properties also depend on whether a simple matched filter receiver is used or a RAKE receiver is used (or any other receiver). In the first case, it is the partial correlation between the transmitted signature waveform of user k convolved with the channel impulse response of user k and the signature waveform of the desired used that should have small values to reduce MAI. In the second case, it is the partial correlation between the transmitted signature waveform of user k convolved with the channel impulse response of user k and the signature waveform of the desired used convolved with the channel impulse response of the desired user that influence MAI. It seems that very little is known about these correlation properties on multipath channels in general for m-sequences.

M-sequences are used in the IS-95 CDMA system developed by Qualcomm [16]. In the uplink, a long m-sequence with period $2^{42} - 1$ is used to distinguish different channels (channelization). In both the uplink and the downlink, m-sequences with period $2^{15} - 1$ are used to separate mobiles (uplink) and base stations (downlink). Separate m-sequences are used on the I and Q channels in both directions. In the downlink, the data are also scrambled by a decimated long m-sequence.

3.2. Gold Sequences

M-sequences lead to the excessive amount of MAI in CDMA systems as seen previously. Another family of periodic signature sequences with somewhat better properties are Gold sequences [7,8,11,17,18]. A Gold sequence is obtained as the sum of two so-called preferred pairs of m-sequences, that is, $C[i] = C'[i] + C''[i]$, where $C'[i]$ and $C''[i]$ are the ith chips of two different m-sequences. In fact, each preferred pair of m-sequences generates a whole family of Gold sequences, namely, each of the two m-sequences alone and the sum of one of them with any shift of the other. Therefore, every preferred pair generates $P + 2$ Gold sequences, where P as before is the period of the m-sequence. A limited set of characteristic polynomials for preferred pairs of m-sequences is given in Table 4. A more complete table can be found in Ref. [7, page 502].

When $N = P$, the periodic autocorrelation $\phi(p)$ with $p \neq 0$ and the periodic cross-correlation $\Phi_{k,i}(p)$ can be shown to take at most three values, which are $\{-1, -t(m), t(m) - 2\}$, where

$$t(m) = \begin{cases} 2^{(m+1)/2} + 1 & m \text{ odd} \\ 2^{(m+2)/2} + 1 & m \text{ even} \end{cases} \tag{27}$$

Table 4. Characteristic Polynomial for Preferred Pairs of m-Sequences That Are Used to Form Gold Sequences. A More Complete Table can be found in Ref. [7, page 502]. The Octal Notation is Explained in Table 2

m	$P = 2^m - 1$	Preferred Pairs of Generator Polynomials in Octal Form
5	31	[45,75]
6	63	None
7	127	[211,217], [211,277]
8	255	[747,703]
9	511	[1021,1131], [1131,1423]

Table 5. The Maximum Periodic Autocorrelation and Cross-Correlation $\Phi_{\max}$ for Gold Sequences of Degree m

m	$t(m) = \Phi_{\max}$	$\Phi_{\max}/P$
5	9	0.29
6	17	0.27
7	17	0.13
8	33	0.13
9	33	0.06
10	65	0.06
11	65	0.03
12	129	0.03

These values are given in Table 5 together with the corresponding normalized values. Here it is clearly seen that these values are much improved compared with the corresponding values for m-sequences (see Table 3).

Thus, the MAI on a flat channel is reduced when $b_k[-1] = b_k[0]$. However, for other combinations of data, the MAI may be larger. The properties for synchronization on a flat channel have become poorer as compared with m-sequences, since the periodic autocorrelation function is larger for $0 < p < N$. This is illustrated in Figs. 6 and 7 for

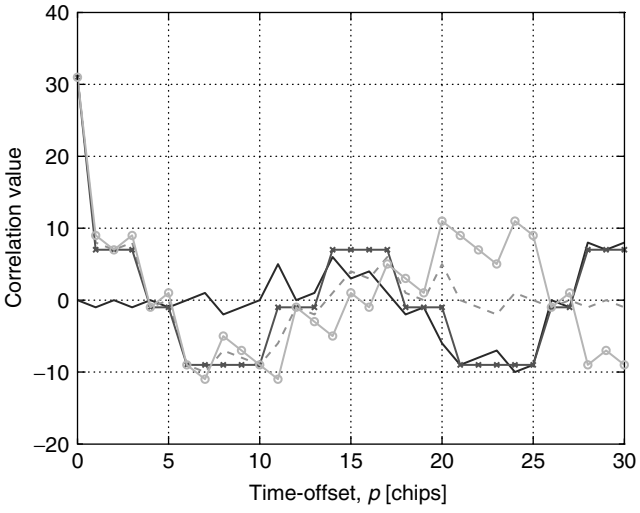


Figure 6. $\overline{X}_{1,1}(p)$ (solid line), $X_{1,1}(p)$ (dashed line), $\phi(p)$ (cross), and $X_{1,1}(p) - \overline{X}_{1,1}(p)$ (circle) for Gold sequence.

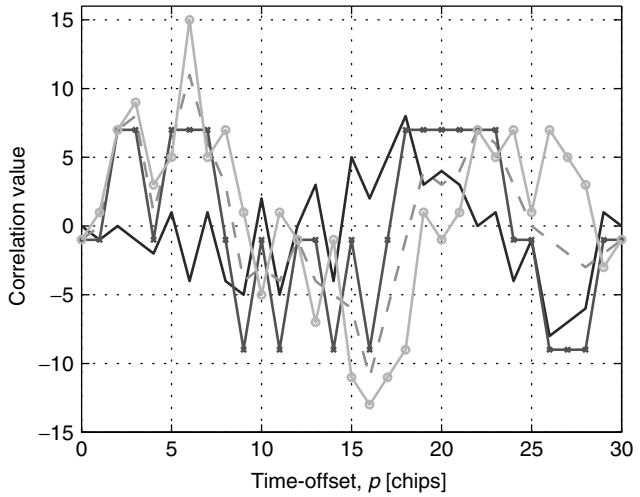


Figure 7. $\overline{X}_{2,1}(p)$ (solid line), $X_{2,1}(p)$ (dashed line), $\Phi_{2,1}(p)$ (cross), and $X_{2,1}(p) - \overline{X}_{2,1}(p)$ (circle) for two Gold sequences.

Gold sequences with period 31. The preferred pairs used are those given in Table 4 for $m = 5$. Both m-sequences have been generated starting from all ones. Figure 6 shows $\overline{X}_{1,1}(p)$ (solid line), $X_{1,1}(p)$ (dashed line), $\phi(p)$ (cross), and $X_{1,1}(p) - \overline{X}_{1,1}(p)$ (circle) for the Gold sequence obtained by adding the two sequences with a shift of 2 chips on the one generated from characteristic polynomial 45. Figure 7 shows $\overline{X}_{2,1}(p)$ (solid line), $X_{2,1}(p)$ (dashed line), $\Phi_{2,1}(p)$ (cross), and $X_{2,1}(p) - \overline{X}_{2,1}(p)$ (circle) where user two uses the sequence above and user 1 the sequence obtained by shifting 15 chips instead of 2. Both figures verify the three valued behavior of $\phi(p)$ and $\Phi_{k,i}(p)$ as discussed previously (curves shown with cross). However, in both cases, it is also clear that larger MAI may occur when $b_2[-1] = -b_2[0]$ (curves with circle).

Again, much less is known about the correlation properties when long codes based on Gold sequences are used and in general on multipath channels. In the Wideband CDMA (WCDMA) system used for third-generation mobile communications, Gold sequences are used to distinguish cells in the downlink and to distinguish mobiles in the uplink when simple receivers are used in the base station [4,5]. In both cases, long codes are used. The Gold sequence used in the downlink has period $2^{18} - 1$, while the Gold sequence used in the uplink has period $2^{41} - 1$. In both cases, only 38,400 chips are used to form a periodic sequence with period 38,400. The spreading factor is variable between 4 and 512.

3.3. Kasami Sequences

The *small set* of Kasami sequences can be obtained in a way similar to the Gold sequences [11,13,19]. Again, two m-sequences are added, but in this case, these two m-sequences have different periods. To obtain a period of $P = 2^m - 1$, with m even, of the Kasami sequence, one starts with an m-sequence with the same period. This m-sequence is then decimated by $2^{m/2} + 1$, which results in another m-sequence of period $2^{m/2} - 1$. By adding these two sequences, a Kasami sequence is obtained. Yet more Kasami sequences can be obtained by adding the original m-sequence with the other $2^{m/2} - 2$ shifts of the decimated sequence. By also including the original m-sequence in the set, $2^{m/2}$ Kasami sequences with period $2^m - 1$ have been obtained. It turns out that another way to generate all these sequences is by using the characteristic polynomial $G(D)G'(D)$ in (23), where $G(D)$ is the characteristic polynomial of the original m-sequence and $G'(D)$ is the characteristic polynomial of the decimated m-sequence.

The periodic discrete autocorrelation and cross-correlation is also three-valued for the *small set* of Kasami sequences. The values are from the set $\{-1, -(2^{m/2} + 1), 2^{m/2} - 1\}$. The *small set* of Kasami sequences therefore satisfies the Welch lower bound [11,20], which states that the maximum cross-correlation $\Phi_{\max}$ between any two sequences in a set of M sequences is bounded as

$$\Phi_{\max} \geq P\sqrt{\frac{M-1}{MP-1}} \approx \sqrt{P} \tag{28}$$

where the approximation is valid for large values of P and M. This set of sequences is therefore considered as optimal.

The *large set* of Kasami sequences with period $P = 2^m - 1$, with m even, contains both the Gold sequences and the *small set* of Kasami sequences and is obtained in the following way. Three different sequences are added. One is an m-sequence of period P. The other two are obtained by decimating the original m-sequence by $2^{m/2} + 1$ and $2^{(m+2)/2} + 1$. The two can be used in any possible shift. The number of sequences obtained are $2^{3m/2}$ when $m = 0 \bmod 2$ and $2^{3m/2} + 2^{m/2}$ when $m = 2 \bmod 2$. All the values of the periodic autocorrelation and cross-correlation are from the set $\{-1, -1 \pm 2^{m/2}, -1 \pm 2^{m/2+1}\}$. The Welch bound is not asymptotically approached with this larger set, but the packing of signal space is more efficient than for the Gold sequences. This set can be generated directly by using the characteristic polynomial $G(D)G'(D)G''(D)$, where $G(D)$ is the characteristic polynomial of the original m-sequence, $G'(D)$ is the characteristic polynomial of the first decimated m-sequence, and $G''(D)$ is the characteristic polynomial of the second decimated m-sequence.

However, still little is known about the performance of the sets of Kasami sequences when they are used as short or long signature sequences on multipath channels in CDMA. The reason, as before, is that it is not only the periodic autocorrelation and periodic cross-correlation between the full period of the signature sequences that matters.

3.4. Walsh–Hadamard Sequences

Two orthogonal signature sequences (on a flat channel) over the full period cannot be obtained by any of the sequence families discussed previously since $\Phi_{k,i}(p) \neq 0$. Complete orthogonality would make MAI to disappear and is therefore interesting. In the introduction, we indicated that under certain restrictions, it is in fact possible to obtain complete orthogonality. The family of orthogonal sequences is referred to as Walsh–Hadamard sequences

(sometimes Hadamard codes or Walsh functions). A Hadamard matrix of length $P = 2^m$ is defined as

$$\mathbf{H}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\mathbf{H}_P = \begin{bmatrix} \mathbf{H}_{P/2} & \mathbf{H}_{P/2} \\ \mathbf{H}_{P/2} & \overline{\mathbf{H}}_{P/2} \end{bmatrix} \tag{29}$$

where $\overline{\boldsymbol{H}}_{P/2}$ denotes the complement of $\mathbf{H}_{P/2}$. The rows of such a matrix are orthogonal. Thus, a user can form its signature sequence as a periodic repetition of one row in the Hadamard matrix of length $P = N$. If another user forms its signature sequence in the same way, but based on another row in the same matrix, and these two users are synchronized such that the periods of the signature sequences completely overlap, the periodic cross-correlation $\Phi_{k,i}(0)$ becomes zero. These two users will therefore not interfere with each other on a channel without multipath propagation. However, $\Phi_{k,i}(p)$ and $\phi(p)$ for $0 < p < N$ are in general not zero and are in many cases quite large (can in fact be as large as $N - 1$). This means that the system must be synchronized and that Walsh-Hadamard sequences are not good for synchronization purposes. The orthogonality is also lost on multipath channels.

Signature sequences based on rows or parts of rows of a Hadamard matrix can also be used when different users have different spreading factors as in WCDMA [4,5]. From Eq. (29), it is seen that each Hadamard matrix can in fact be decomposed into four Hadamard matrices of half the size. Since the rows in Hadamard matrices of all sizes have orthogonal rows, it is possible to allow a user with spreading factor $P/2$ to form its signature sequence as a repetition of one of the rows in $\mathbf{H}_{P/2}$. To keep orthogonality between all signature sequences also over the period $P/2$, users with spreading factor P are now restricted to use one of the remaining rows not starting with the sequence of length $P/2$ already used. This can be generalized to any spreading factor between 2 and P, when the Hadamard matrix of length P is used.

In WCDMA, the Hadamard matrix of length 256 is used to allocate signature waveforms to different channels in a base station or in a mobile (referred to as channelization codes in WCDMA). This matrix can be used for spreading factors from 4 to 256 (only powers of two are used) and the orthogonality is always preserved as long as the channel is flat. A user with spreading factor 4, is allocated the first 4 bits on one row in the matrix. It forms its signature sequence by periodically repeating these four bits. Since 1/4 of the rows in the Hadamard matrix start with the same 4 bits, these cannot be used for other users, because this would not make them orthogonal over a 4 chip period. A user with spreading factor 8 can be allocated the 8 first bits on the one of the remaining (allowed) rows for its signature sequence. This means that another 1/8 of all the rows are not allowed for the rest of the users. This scheme can now be continued as long as there are rows available to allocate. These sequences are referred to as Orthogonal Variable Spreading Factor (OVSF) sequences in WCDMA.

Walsh–Hadamard sequences are also use in IS-95 CDMA [16]. In the downlink, different channels are separated by different Walsh–Hadamard sequences of length 64. Walsh–Hadamard sequences are also used in the uplink, not as signature sequences but to obtain orthogonal 64-level modulation.

4. SOME CONCLUDING REMARKS ON SIGNATURE SEQUENCES

Although we have not covered all known signature sequences, it is clear that families of signature sequences with $\overline{X}_{k,i}(p) = 0$ except for $p = 0$ and $k = i$ do not exist. This means that there will be interference (the sum of ISI, MAI, and receiver noise) in the decision variable of each user. The choice of signature waveforms can to a certain extent reduce the interference, but in practice channel coding is also used in the system to reduce the effects of the interference. Since most known channel codes are designed for independent errors, it is important that interference be white (or almost white), such that the interference in the decision variable is independent from one symbol interval to the next. Furthermore, the power of the total interference should be small. The average interference power (AIP) for user 1 in the output at time T of a filter matched to $h(t)$ is given by

$$\begin{aligned} \mathrm{AIP}_1(T) &= \mathrm{E}\,[y_1(T) - \sqrt{E_1}g_1(1)b_1[0]]^2 \\ &= \sum_{l=2}^{L} E_1 \mathrm{E}\,|g_1(l)|^2 \left(\overline{X}_{1,1}^2(p_1(l))/N^2 + X_{1,1}^2(p_1(l))/N^2\right) \\ &\quad + \sum_{k=2}^{K}\sum_{l=1}^{L} E_k\,\mathrm{E}\,|g_k(l)|^2 \left(\overline{X}_{k,1}^2(p_k(l))/N^2 \right. \\ &\quad \left. + X_{k,1}^2(p_k(l))/N^2\right) \end{aligned} \tag{30}$$

where the same assumptions as in Eq. (17) about the channels have been used. Here, we have used the fact that bits from different users and different times are independent and hence the expected value of the cross-terms are zero. This average depends on the channel (both the complex coefficients and the relative delays) so in order for signature sequences to be good, they must lead to low AIP for all reasonable channels. The average also depends on the received energy from the different users through E_k and $|g_k(l)|^2$. These can to a certain extent be reduced by power control, such that a nearby user is not received at much higher power than a distant user. Results exist that seem to show that all the different families of signature sequences discussed lead to approximately the same AIP with practical channels in asynchronous DS-CDMA [21]. There also is some evidence that Gold sequences lead to reasonably good performance on land mobile radio channels [22].

With short signature sequences, the correlation time of the interference only depends on the correlation time of the channel coefficients. Thus, a large interference value due to high cross-correlation between two users, may remain for a significant time, and this makes channel coding less efficient. Long signature sequences are a means to reduce the correlation between the interference in neighboring symbol intervals, since different segments of the signature

sequences are used in neighboring symbol intervals, leading to different cross-correlations. This significantly improves the performance of channel coding.

There also exist many different receivers that may be used with DS-CDMA [23–25]. Many of these attempt to remove some or all interference before the decision variable is formed. For such receivers, the actual choice of signature sequences may be somewhat less important. The processes of mapping data symbols $b_k(t)$ to a transmitted waveform in DS-CDMA can also be described as repetition coding followed by scrambling, where the scrambling sequences plays the role of the signature sequence above [26]. In such a system, the outer channel coding and the repetition coding can be combined into one encoding process and interleaving can be done either on symbols or on chips. For all these different kind of systems, signature sequences should be designed jointly with several other things like channel coding, interleaving scheme, detection algorithm, and so on. For such more elaborate schemes, it remains an open issue how to design signature sequences and what the actual performance of the system will be on different channels.

BIOGRAPHIES

Tony Ottosson received the M.Sc. in electrical engineering from Chalmers University of Technology, Göteborg, Sweden, in 1993, and the Lic. Eng. and Ph.D. degrees from the Department of Information Theory, Chalmers University of Technology, in 1995 and 1997, respectively. Currently, he is an associate professor at the Communication Systems Group, Department of Signals and Systems, Chalmers University of Technology. During 1999 he was also working as a Research Consultant at Ericsson Inc., Research Triangle Park, North Carolina.

Professor Ottosson's research interests are in communication systems and information theory and are targeted mainly to CDMA systems. Specific topics are modulation, coding, multirate schemes, multiuser detection, combined source-channel coding, joint decoding techniques, and synchronization.

Erik G. Ström received his M.Sc. in electrical engineering in 1990 from the Royal Institute of Technology (KTH), Stockholm, Sweden, and Ph.D. degree in electrical engineering from the University of Florida, Gainesville, in 1994. He joined the Department of Signals, Sensors, and Systems at KTH as a postdoc in 1995 and was appointed assistant professor (forskarassistent) in 1996. Later that year, Ström joined Chalmers University of Technology, Göteborg, Sweden, where he is now an associate professor (högskolelektor/docent). He received the Chalmers Teacher's Prize in 1998. Since 1990 Dr. Ström acted as a consultant for the Educational Group for Individual Development, Stockholm, Sweden. He is a contributing author and associate editor for the Royal Admiralty Publishers' FesGas-series. Ström is a member of the board of the IEEE VT/COM chapter of the Swedish section and was a coguest editor for the special issue IEEE *Journal on Selected Areas in Communications* on "Signal Synchronization in Digital Transmission Systems." His research interests include code-division multiple access, synchronization, and wireless communications. He has published approximately 40 journal and conference papers.

Arne Svensson received the M.Sc. degree in electrical engineering in 1979, and the Dr. Ing. (Tekniska Licentiat) and the Dr. Techn. (Ph.D.) in telecommunication theory in 1982 and 1984, respectively, from University of Lund, Lund, Sweden. He joined Ericsson in 1987 as a research engineer and became later a specialist in communications. At Ericsson he worked on the design and analysis of a communication systems for the Swedish air force and the personal digital cellular system for mobile communications in Japan. Since 1993, he has been professor in communication systems at Chalmers University of Technology, Göteborg, Sweden, where he has been working on design and analysis of air interfaces for mobile communications. Dr. Svensson has published more than 150 papers in international journals and conference proceedings, and he is a fellow of IEEE. In 1986, he received the paper of the year award from the IEEE Vehicular Technology Society. His areas of interest include channel coding/decoding, digital modulation methods, channel estimation, data detection, multiuser detection, digital satellite systems, wireless IP-based systems, CDMA and spread spectrum systems, personal communication networks, and ultra wideband systems.

BIBLIOGRAPHY

1. T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.
2. M. D. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press, Boca Raton, Fl, 1993.
3. G. L. Stüber, *Principles of Mobile Communication*, 2nd ed., Kluwer, Boston, 2001.
4. H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, Wiley, New York, 2000.
5. T. Ojanperä and R. Prasad, *Wideband CDMA for Third Generation Mobile Communications*, Artech House, Boston, 1998.
6. M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook*, revised edition, McGraw-Hill, New York, 1994.
7. R. C. Dixon, *Spread Spectrum Systems with Commercial Applications*, 3rd ed., Wiley, New York, 1994.
8. R. L. Peterson, R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
9. D. V. Sarwate, Optimum PN sequences for CDMA systems. In *Proc. IEEE Third International Symposium on Spread Spectrum Techniques and Applications*, 27–35, Oulu, Finland, July 1994.
10. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudorandom and related sequences, *IEEE Proceedings* **68**(5): 593–619 (May 1980).
11. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.

12. S. W. Golomb, *Shift Register Sequences*, revised edition, Aegean Park Press, Laguna Hills, CA, 1982.
13. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* **36**(9): 48–54 (Sept. 1998).
14. P. Fan and M. Darnell, *Sequence Design for Communication Applications*, UK Research Studies Press, 1996.
15. W. Stahnke, Primitive binary polynomial, *Math. Comp.* **27**: 977–980 (Oct. 1976).
16. S. C. Yang, *CDMA RF System Engineering*, Artech House, Boston, 1998.
17. R. Gold, Optimal binary sequences for spread spectrum multiplexing, *IEEE Trans. Inform. Theory* **IT-13**: 619–621 (Oct. 1967).
18. R. Gold, Maximal recursive sequences with 3-valued recursive cross correlation functions, *IEEE Trans. Inform. Theory*, **IT-14**: 154–156 (Jan. 1968).
19. T. Kasami, Weight distribution formula for some class of cyclic codes, Technical Report R-285, Coordinated Science Laboratory, University of Illinois, Urbana, IL, April 1966.
20. L. R. Welch, Lower bounds on the maximum cross correlation of signals, *IEEE Trans. Inform. Theory* **IT-20**: 397–399 (1974).
21. K. H.A Kärkkäinen and P. A. Leppänen, Comparison of the performance of some linear spreading code families for asynchronous DS/SSMA systems, *Proc. IEEE Military Communications Conference*, 784–790, November 1991.
22. H. Elders-Boll, The optimization of spreading sequences for CDMA system in the presence of frequency-selective fading, *Proc. IEEE 6th Symp. on Spread Spectrum Techniques and Applications*, 414–418, New Jersey Institute of Technology, New Jersey, USA, September 2000.
23. S. Verdú, *Multiuser Detection*, Cambridge University Press, UK, 1998.
24. S. Moshavi, Multi-user detection for DS-CDMA communications, *IEEE Commun. Mag.* **34**(10): 124–136 (Oct. 1996).
25. A. Duel-Hallen, J. Holtzman, and Z. Zvonar, Multiuser detection for CDMA systems, *IEEE Personal Commun. Mag.* **2**(2): 46–58 (April 1995).
26. P. Frenger, P. Orten, and T. Ottosson, Code-spread CDMA using maximum free distance low-rate convolutional codes, *IEEE Trans. Commun.* **48**(1): 135–144 (Jan. 2000).

SIMULATION OF COMMUNICATION SYSTEMS

K. SAM SHANMUGAN
University of Kansas
Lawrence, Kansas

1. INTRODUCTION

Modeling and simulation of communication systems can be viewed in a hierarchical manner starting at the network layer and progressing down to transmission systems level and then on to implementation details. The performance issues and tradeoffs addressed in each layer, and the modeling and simulation methods and the tools used at the various layers differ significantly. At the network layer, the simulation model will consist of processors, routers, traffic sources, buffers, transmission links, network topology, and protocols. The flow of packets and messages over the network will be simulated using an event-driven simulation framework, and system performance metrics such as network throughput, latency, resource utilization, and quality of service will be estimated from simulations. On the other hand, at the bottom level in the hierarchy dealing with implementation details, simulation of digital hardware, for example, will be done using hardware description language (HDL) simulators at the gate level. Performance metrics and design tradeoffs at this level may include power, speed, and chip area.

The focus of this article is on waveform level simulation of transmission systems or communication links, an example of which is shown in Fig. 1.

The primary simulation technique used at the link level is time-driven Monte Carlo simulation of the flow of waveforms or signals over the transmission link (Refs. 1–3 cover this topic in great detail). Waveform-level simulation of communication systems involves the following steps:

1. Modeling the communication system in a block diagram form in which each functional block performs a specific signal processing operation such as modulation, and filtering
2. Generating sampled values of the input signals, noise, and interference
3. Letting the functional blocks in the simulation model operate on the input samples
4. Gathering the samples generated during the simulation and estimating performance measures such as bit error rates as a function of E_b/N_0 and other design parameters

Details of these steps are presented in the following sections.

2. DISCRETE TIME REPRESENTATION OF SIGNALS AND SYSTEMS

In the simulation domain, all systems and signals are represented by their discrete-time equivalents. The simulation models and simulation algorithms draw heavily from DSP (digital signal processing) concepts [4,5]. The fundamental concept that permits us to go

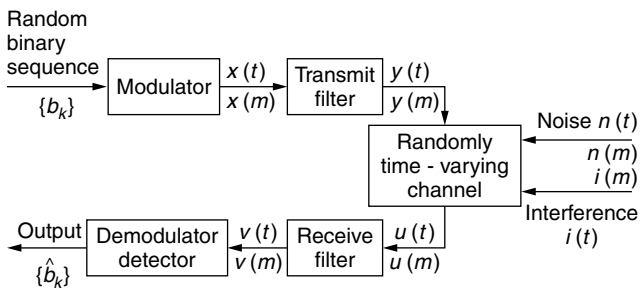


Figure 1. Waveform level simulation model of a communication system.

back and forth between discrete- and continuous-time representation of signals and systems is the sampling theorem. Implications of the sampling theorem as it applies to waveform-level simulation of communication systems are presented below.

2.1. Discrete-Time Representation of Lowpass Signals and Systems

The uniform sampling theorem in its simplest form states that a deterministic signal that is lowpass and band-limited to B Hz in the continuous-time domain can be represented exactly, *without loss of information*, in terms of its sampled values as long as the sampling rate f_s is greater than the *Nyquist rate* of $2B$ samples per second [4,5]. The relationship between the continuous time signal $x(t)$ and its sampled values $x(kT_s)$, where T_s is the time between samples, is given by the following equations:

$$x_s(t) = \sum_{k=-\infty}^{\infty} x(kT_s)\delta(t - kT_s) \tag{1}$$

$$X_s(f) = f_s \sum_{k=-\infty}^{\infty} X(f - kf_s) \tag{2}$$

$$x(t) = x_s(t)^* \frac{\sin(\pi f_s t)}{\pi f_s t} = \sum_{k=-\infty}^{\infty} x(kT_s) \frac{\sin(\pi f_s(t - kT_s))}{(\pi f_s(t - kT_s))} \tag{3}$$

Equation (3) represents an interpolation formula in the time domain that yields the values of $x(t)$ for *all* values of t in terms of the sampled values $x(kT_s)$. This interpolation operation in the time domain is equivalent to filtering of $x_s(t)$ in the frequency domain using an ideal lowpass filter with a bandwidth of $f_s/2$ Hz. If the signal is not strictly band-limited and/or the sampling rate is less than $2B$, then it will not be possible to reconstruct $x(t)$ exactly from $x_s(t)$ due to aliasing, which is a result of spectral overlap that occurs in the frequency domain [see Eq. (2)].

The lowpass sampling theorem applies to the discrete-time representation of both lowpass *signals* and lowpass *systems*. The impulse response $h(t)$ of a lowpass system in the continuous-time domain can also be represented in the discrete-time domain using the sampling theorem.

2.2. Sampling of Bandpass (Deterministic) Signals and Systems

Both lowpass and bandpass signals and components are usually present in communication systems. Information-bearing signals are usually lowpass in nature prior to modulation, after which they become bandpass in nature. Components such as filters can be lowpass or bandpass. Hence we need to represent both lowpass and bandpass signals and systems in the discrete time domain for simulation.

Bandpass signals and systems can be sampled directly using the bandpass version of the sampling theorem, or they can be sampled using the principle of the lowpass sampling theorem and the concept of the lowpass equivalent representation [1]. In the time domain, modulated signals can be expressed in the form

$$x(t) = x_c(t)\cos(2\pi f_c t) - x_s(t)\sin(2\pi f_c t) \tag{4}$$

$$= \text{Re } al\{\tilde{x}(t)\exp(j2\pi f_c t)\} \tag{5}$$

$$\tilde{x}(t) = x_c(t) + jx_s(t) \tag{6}$$

where $x_c(t)$ and $x_s(t)$ are the in-phase and quadrature phase modulating signals, which are usually lowpass; f_c is the carrier frequency; and $\tilde{x}(t)$ is the *complex envelope* of the bandpass signal. $\tilde{x}(t)$ will be lowpass since the modulating signals $x_c(t)$ and $x_s(t)$ are lowpass. Since the bandwidth of the modulating signals will be small compared to f_c, fewer samples will be needed to represent the lowpass complex envelope $\tilde{x}(t)$, which contains all the information in the modulated bandpass signal. Equation (5) clearly shows that the bandpass signal can be reconstructed from $\tilde{x}(t)$ by simply multiplying it with the complex carrier and taking the real part.

The lowpass equivalent representation of a deterministic bandpass signal can also be derived in the frequency domain using the Hilbert transform. When the bandwidth of the signal is small compared to the carrier frequency, the lowpass equivalent representation in the frequency domain is given by [1]

$$\begin{aligned} X_{\text{LP}}(f) &= \{2X_{\text{BP}}(f + f_c)\}_{\text{lowpass}} \\ &= 2X_{\text{BP}}(f + f_c)U(f + f_c) \end{aligned} \tag{7}$$

where $U(f)$ is the unit step function, $U(f) = 1$ for $f > 0$ and zero for $f < 0$.

Note that the lowpass equivalent can be asymmetric around $f = 0$ if the bandpass spectrum is not symmetric around the carrier frequency. This leads to a complex time-domain function as shown in Eq. (6). The minimum sampling rate for the lowpass equivalent representation is B *complex* samples per second or $2B$ real samples per second.

It should be noted that the complex envelope representation is with respect to a single-carrier frequency f_c. This representation can also be used for simulating multicarrier FDM (frequency-division multiplex) systems by using one of the mid-band carriers as the reference and representing the total composite complex envelope of the sum of the FDM carries with respect to the reference carrier according to

$$\sum_{i=1}^{n} A_i e^{j2\pi f_i t + j\phi_i} = e^{j2\pi f_c t} \sum_{i=1}^{n} A_i e^{j2\pi (f_i - f_c)t + j\phi_i} \tag{8}$$

where f_c is the carrier frequency chosen as the reference.

If multiple signals with widely differing bandwidths are present in a system, then a multirate sampling strategy will be useful for simulating such systems. Each signal is sampled at a rate consistent with its bandwidth, and interpolation and decimation techniques are used to upconvert or downconvert the sampling rates as necessary.

2.3. Sampling of Lowpass and Bandpass Random Processes

Signals, noise, and interference in communication systems are usually modeled as stationary random processes

characterized in the frequency domain by power spectral density functions. Frequency domain parameters such as bandwidth, and properties such as lowpass and bandpass are based on power spectral densities (PSDs). The sampling principle and the concept of lowpass equivalent representation also apply to random processes in terms of their PSDs [6].

A lowpass random process in the continuous-time domain can be represented in the discrete-time domain in terms of its sampled values, and it is possible to recover the continuous-time random signal (with a mean-squared error approaching zero) from its sampled values as long as the process being sampled is band-limited to B Hz and the sampling rate is greater than $2B$. The concept of lowpass equivalent representation applies for bandpass random processes also. For example, a bandpass Gaussian process $n(t)$ can be represented in terms of its lowpass equivalent as

$$\begin{aligned} n(t) &= n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t) \\ &= \text{Real}\,\{\tilde{n}(t)\exp(j2\pi f_c t)\}, \tilde{n}(t) = n_c(t) + jn_s(t) \end{aligned} \tag{9}$$

where $n_c(t)$ and $n_s(t)$ are real-valued lowpass Gaussian random processes with the power spectral densities

$$\begin{aligned} S_{n_c n_c}(f) = S_{n_s n_s}(f) &= S_{NN}(f+f_c)U(f+f_c) \\ &\quad + S_{NN}(-f+f_c)U(-f+f_c) \\ jS_{n_s n_c}(f) &= S_{NN}(f+f_c)U(f+f_c) \\ &\quad - S_{NN}(-f+f_c)U(-f+f_c) \end{aligned} \tag{10}$$

If the bandpass process is nonsymmetric around the carrier frequency, then $n_c(t)$ and $n_s(t)$ will be correlated with the cross-PSD given above. For simulation purposes, bandpass random processes are sampled using their lowpass equivalent representations.

2.4. Simulation of Bandpass Systems with Bandpass Inputs

In order to minimize the computational burden, bandpass systems with bandpass inputs are simulated using sampled values of lowpass equivalent representations. It should be noted that while the lowpass equivalent representation of bandpass *signals* in the frequency domain contains a factor of 2 in Eq. (7), the lowpass equivalent representation of bandpass *systems* does not include the factor of 2.

The input–output relationship for the bandpass and lowpass equivalent representations are given by

$$\tilde{y}_{\text{LP}}(t) = \tilde{h}_{\text{LP}}(t)^* \tilde{x}_{\text{LP}}(t); \quad y_{\text{BP}}(t) = \text{Real}\,\{\tilde{y}_{\text{LP}}(t)e^{j2\pi f_c t}\} \tag{11}$$

2.5. Factors Influencing the Sampling Rate

The most important factor in determining an appropriate sampling rate for simulations is the amount of aliasing that can be tolerated. Other factors that have to be considered in setting the sampling rate for simulations include the effect of sampling on modeling functional blocks such as filters (frequency warping due to bilinear z transform), nonlinearities (bandwidth expansion), and feedback loops (delay in feedback loops). All the deleterious effects of sampling on simulation accuracy can be minimized by increasing the sampling rate. However, increasing the sampling rate will increase the computational burden.

This tradeoff between sampling rate and the accuracy of simulations is a very important one in simulations. A practical value for sampling rate that offers a good tradeoff between simulation accuracy and computational burden is 16–32 samples per hertz or symbol [1]. It is most convenient to use an integer number of samples per symbol for simulating digital transmission systems, and it is computationally most efficient to choose 16 or 32 samples per hertz so that the fastest version of the discrete Fourier transform (DFT) algorithm can be used during simulations.

3. MODELING AND SIMULATION OF FUNCTIONAL BLOCKS IN COMMUNICATION SYSTEMS

At the waveform level simulation of a communication system, each functional block in the simulation model performs a specific signal processing operation. If the signal processing operation performed by the functional block is discrete time, algorithmic, and at the symbol level (e.g., convolutional encoder/decoder), then there is very little modeling per se for such functional blocks: the simulation model of the functional block is the algorithm itself (e.g., a Viterbi decoder). On the other hand, there are a number of other functional blocks that perform (analog) waveform processing operations such as filtering and amplification. The signal processing operations performed by these blocks as well as the communication channel have to be *modeled* for simulation purposes. Examples of such *models*, which are described below, include infinite and finite impulse response filters and AM-to-AM and AM-to-PM models for memoryless nonlinearities.

3.1. Modeling and Simulation of Linear Time-Invariant (LTIV) Components

Many components in communication systems such as filters, optical fibers, cables, and other guided channels are linear and time-invariant. Such components are described in the time domain in terms of the impulse responses $h(t)$ or transfer functions $H(f)$. If these components are bandpass in nature, then for simulation purposes we will use the lowpass equivalent representations described in Section 2. The DSP literature contains a wide array of algorithms for implementing and virgule or simulating filters [4,5]. (We will use the generic term *filters* to represent LTIV components). The choice of simulation algorithm will depend on the nature of the filter specifications, the duration of the impulse response, and the context in which the filter is used in the overall simulation model.

3.1.1. Finite-Impulse Response (FIR) Model — Time-Domain Convolution. If the filter specification is given empirically in terms of the sampled values of the lowpass equivalent impulse response $\tilde{h}(kT_s)$, $k = 0, 1, 2, \ldots, M-1$, then the simplest simulation model is an FIR structure

that implements the input–output relationship as a finite convolution sum of the form

$$\tilde{y}(nT_s) = T_s \sum_{k=0}^{M-1} \tilde{h}(kT_s)\tilde{x}((n-k)T_s) \tag{12}$$

Simulation of this equation requires M complex multiplications and additions for each output sample. This may impose a considerable computational burden if the impulse response is very long. In order to reduce the processing load, impulses responses are truncated to the shortest possible duration without losing a significant amount of energy outside the truncation window and the truncated impulse response is used in Eq. (12). If the filter is specified in the frequency domain in terms of sampled values of the frequency response $H(kf_c)$, the inverse Fourier transform is used to compute the impulse response, which is then truncated and used in Equation (12) to perform time-domain convolutions.

3.1.2. FIR Model — DFT Implementation. The computational burden of time-domain convolution can be reduced by using DFT operators. In this implementation, the input samples and the impulse response are *padded* with enough zeroes and the padded input vector and the impulse response vectors are convolved using the DFTs operators according to

$$Y(kf_0) = \text{DFT}\ (\tilde{y}(kT_s)) = \text{DFT}\ (\tilde{h}(kT_s))\text{DFT}\ (\tilde{x}(kT_s)) \tag{13}$$

$$\tilde{y}(kT_s) = \text{Inv.DFT}\ (Y(kf_0)) \tag{14}$$

The minimum DFT size n is usually chosen to be a power of 2 nearest to the sum of the length of the unpadded impulse response and the input vectors. This will permit the use of fast FFT operators to perform the convolution [4,5].

DFT/FFT implementation will be computationally very efficient (several orders of magnitude faster) compared to direct-time-domain convolution when the length of the impulse response exceeds about 100 samples. One drawback of the DFT/FFT filters is that they introduce a processing delay of n samples (e.g., the output lags the input by n samples), which might cause a problem if the filter being simulated is part of a feedback loop, for example. The one block processing delay, which is strictly an artifact of the DFT/FFT filter, might render the feedback loop unstable and lead to totally incorrect simulation results. If all the blocks in the simulation model are serially cascaded, then this is not a problem.

If the input sequence is very long, then it can be broken up into smaller blocks and the response of the filter to each input block can be computed separately and the results can be added using the superposition principle. Either the input blocks or the output blocks have to be overlapped in order to produce the correct output [1,5,6].

3.1.3. Infinite Impulse Response (IIR) Models. If the filter specifications are given in the form of poles and zeroes or as a ratio of polynomial in the s domain, then the filter can be simulated using recursive computational structures that are very efficient. These recursive structures are derived using either *the impulse-invariant method or the bilinear z transform* [1,4,5]. In the first method the impulse response of the filter is obtained from the inverse transform of the transfer function and the z transform of the infinite-length impulse response (untruncated) is used to derive a recursive computational structure. In the second method, the transfer function in the Laplace transform domain is mapped into the z-transform domain directly using the bilinear z transform defined by

$$H_d(z) = H(s)@s = \frac{2}{T_s}\left[\frac{1-z^{-1}}{1+z^{-1}}\right] \tag{15}$$

The resulting transfer function in both cases can be factored into a product of quadratic factors:

$$H_d(z) = \alpha_0 \prod_{i=1}^{K} H_d^{(i)}(z); \qquad K = \frac{1}{2}(N+1);$$

$$H_d^{(i)}(z) = \frac{1+\alpha_{1i}z^{-1}+\alpha_{2i}z^{-2}}{1+\beta_{1i}z^{-1}+\beta_{2i}z^{-2}} \tag{16}$$

This leads to a simulation structure shown in Fig. 2.

The main source of error in the impulse-invariant IIR simulation model is aliasing, whereas the bilinear z transform suffers from frequency warping introduced by the transform. Both of these effects can be minimized by choosing a sufficiently high sampling rate. While the IIR structures are computationally faster compared to the FIR models, the IIR structures can be easily derived only for filters whose transfer functions are specified in analytical form by a transfer function in the Laplace or Fourier transform domains. The FIR filters can model and simulate filters with any arbitrary frequency or impulse responses specified empirically. Many important filters, such as the square-root raised-cosine filter, cannot be specified in the Laplace transform domain by poles and zeros and hence are not easy to simulate via the IIR methods.

3.2. Modeling and Simulation of Linear Time-Varying (LTV) Components

There are many components in a communication system that exhibit linear time-varying behavior. An important example of a LTV component is the mobile communication channel in which channel characteristics such as attenuation and delay change as a function of time due to relative motion between the transmit and receive antennas. The input–output relationship for LTV components can be expressed in the time domain by the convolution integral

$$\tilde{y}(t) = \int_{-\infty}^{\infty} \tilde{h}(\tau, t)\tilde{x}(t-\tau)d\tau \tag{17}$$

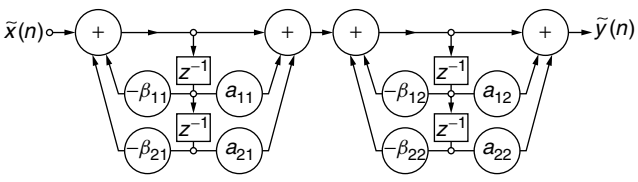


Figure 2. Recursive simulation structure for IIR models [1].

where $\tilde{h}(\tau, t)$ is the time-varying impulse response of the component measured at time t when the impulse is applied to the system input at $t - \tau$. If the input to the system is band-limited to W Hz, then we can derive an FIR model for the time-varying system via the sampling theorem as [7]

$$\begin{aligned}\tilde{y}(kT_s) &\approx \frac{1}{2W}\sum_{n=0}^{\infty}\tilde{h}\left(\frac{n}{2W}, kT_s\right)\tilde{x}\left(kT_s - \frac{n}{2W}\right)\\ &\approx \frac{1}{2W}\sum_{n=0}^{M}\tilde{h}\left(\frac{n}{2W}, kT_s\right)\tilde{x}\left(kT_s - \frac{n}{2W}\right)\\ &= \frac{1}{2W}\sum_{n=0}^{M}\tilde{g}_n(kT_s)\, x\left(kT_s - \frac{n}{2W}\right)\end{aligned} \tag{18}$$

where $\tilde{g}_n(kT_s)$ are called the *tap gain functions* that represent the time-varying impulse response of the system. For a time-*invariant* system, the tap gain functions will be constant and $\tilde{g}_n(kT_s) = \tilde{h}(nT_s)$. The FIR model given above can be implemented as a tapped delay line (TDL) model of the form shown in Fig. 3.

3.3. Modeling and Simulation of Nonlinear Components

Communication systems often contain nonlinear components. Sometimes a nonlinear component is placed in the system to improve performance. For example, a limiter is placed at the front end of a receiver that is subjected to impulsive noise. Also, devices such as high-power amplifiers exhibit an undesirable nonlinear behavior when the input power is high. This introduces nonlinear signal distortion, which might degrade the performance of the communication system significantly. The effects of nonlinearities on system performance are usually difficult to characterize by analytical means but are easy to simulate.

Nonlinearities in communication systems fall into two categories: (1) instantaneous nonlinearities such as limiters and (2) nonlinearities with memory, such as frequency-dependent wideband RF amplifiers. Nonlinearities are also sometimes classified according to whether the input/output signals are baseband or bandpass. Since bandpass nonlinearities are the most common type of nonlinear elements encountered in communication systems, we will concentrate on techniques for modeling and simulation of bandpass nonlinearities in this section.

3.3.1. Lowpass Equivalent Models for Memoryless Nonlinearities. Devices such as bandpass limiters and logarithmic amplifiers can be modeled using the complex envelope representation of the input and output signals. In the bandpass case the input–output relationship of a memoryless nonlinearity can be represented by

$$\begin{aligned}x(t) &= A(t)\cos(\omega_c t + \phi(t))\\ y(t) &= G[x(t)] = G[A\cos(\alpha)];\alpha = \omega_c t + \phi(t)\end{aligned} \tag{19}$$

where $G(\,)$ is a memoryless nonlinearity. If the bandwidth of the input signal is much smaller than f_c, then we can expand $y(t)$ as a Fourier series of the form

$$z = a_0 + \sum_{k=1}^{\infty}(a_k\cos k\alpha + b_k\sin k\alpha) \tag{20}$$

The output of the nonlinearity will contain spectral components in the vicinity of f_c as well as harmonic terms located in the vicinity of kf_c, $k > 1$. If $f_c \gg B$, then the harmonic terms will be far removed from the in-band terms and hence they can be ignored (these components can be easily removed by filtering in a real system). The in-band spectral components or the so-called *first-zone output components* correspond to $k = 1$ in the Fourier series expansion and the first-zone output of the nonlinearity can be expressed in the form

$$y(t) = a_1\cos[\omega_c t + \phi(t)] + b_1\sin[\omega_c t + \phi(t)] \tag{21}$$

where a_1 and b_1 are the Fourier series coefficients defined by

$$\begin{aligned}a_1 &= f_1(A) = \frac{1}{\pi}\int_0^{2\pi} G(A\cos\alpha)\cos\alpha\, d\alpha;\\ b_1 &= f_2(A) = \frac{1}{\pi}\int_0^{2\pi} G(A\cos\alpha)\sin\alpha\, d\alpha\end{aligned} \tag{22}$$

In terms $f_1(A)$ and $f_2(A)$ (which are called the first-order *Chebyshev transforms* of the nonlinearity), the complex lowpass equivalent simulation model for a memoryless nonlinearity is [1]

$$\tilde{y}(t) = f(A)\exp(j\phi + jg(A)) \tag{23}$$

$$f(A)e^{jg(A)} = f_1(A) - jf_2(A) \tag{24}$$

For nonlinearities such as soft and hard limiters, this model, in the form of $f(A)$ and $g(A)$ (which are also called the *AM-to-AM and AM-to-PM* transfer characteristics of

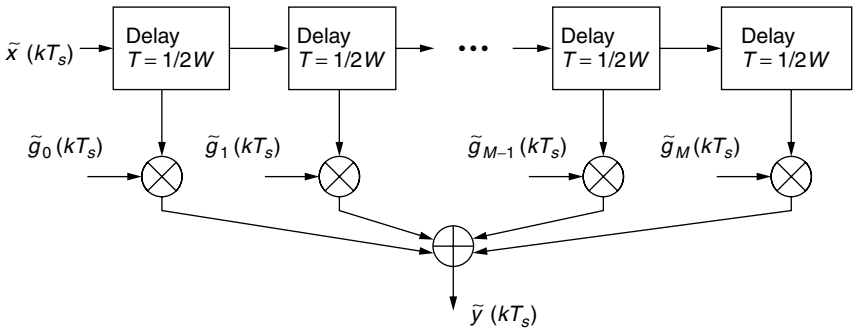


Figure 3. Tapped delay line model for a time-varying component.

the nonlinearity), can be derived in the closed form from the Fourier integrals given in Eq. (22) [1].

The AM–AM and AM–PM characteristics of devices such as high-power amplifiers are usually obtained from *swept-power measurements*, which are made with an input tone of the form $A\cos(\omega_c t + \phi(t))$. The input amplitude and hence the input power $A^2/2$ is varied in steps of 1 dB or so. The AM–AM characteristic is obtained from the input power–output power relationship and the AM–PM characteristic is obtained by measuring the phase offset between the input and output as a function of the input power level. Typical AM–AM and AM–PM characteristics are shown in Fig. 4. The AM–AM and AM–PM model can be simulated using either the empirical AM–AM and AM–PM data or the analytical approximation [8] such as the one shown in Fig. 4.

3.3.2. Lowpass Equivalent Models for Nonlinearities with Memory. When a nonlinear component operates over a wide bandwidth, it might exhibit a frequency selective nonlinear behavior. Models for nonlinearities with memory (or frequency-selective behavior) are difficult to derive analytically, but some useful models can be derived from swept-power/swept-frequency measurements. These measurements are made by probing the device with an unmodulated tone of the form $A\cos(2\pi(f_c t + f_i t))$ and changing both the input power levels and the frequency offset f_i and recording the AM–AM and AM–PM transfer characteristics at different frequencies. If these curves are significantly different at different frequencies, then a model of the form shown in Fig. 5 can be synthesized from the swept-frequency/swept-power measurements to account for the frequency selective behavior of the device. Details of the model synthesis may be found in papers by Saleh and Poza et al. [8,9]. Other simulation models for nonlinearities with memory in the form of Volterra series or nonlinear differential equations may be found in Ref. 1. (The best single source of reference for all the topics addressed in this article is Ref. 1.)

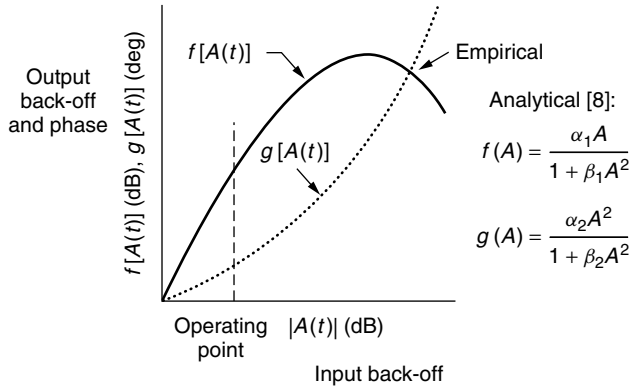


Figure 4. AM–AM and AM–PM characteristics.

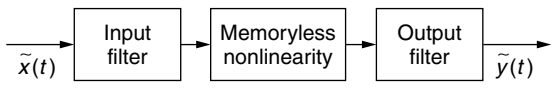


Figure 5. Simulation model for a frequency-selective nonlinearity.

3.4. Modeling and Simulation of Communication Channels

Communication channels introduce a variety of transmission impairments, including attenuation, linear distortion, noise, and interference. Simulation models for communication channels take one of two forms: (1) a transfer function model for time-invariant channels such as optical fibers and cables and (2) a TDL model for time-varying channels such as the mobile radio channel. The transfer function model of a time-invariant channel can be simulated using an FIR or IIR algorithm. Time-varying channels are more difficult to model and simulate. Some of the models and approaches that are used for simulating mobile radio communication channels and other time-varying channels are described below.

3.4.1. Simulation Models for Mobile Communication Channels. In a typical mobile communication environment there will be multiple propagation paths between the transmitter and the receiver due to reflection, refraction, and scattering [10,11]. Also, the path characteristics will be time-varying due to relative motion between the transmit and receive antennas (Fig. 6).

The input–output relationship for the two-ray multipath channel shown in Fig. 6 can be expressed as

$$\tilde{y}(t) = \tilde{a}_1(t)\tilde{x}(t - \tau_1(t)) + \tilde{a}_2(t)\tilde{x}(t - \tau_2(t)) \qquad (25)$$

where $\tau_1(t)$ and $\tau_2(t)$ are the path delays and $\tilde{a}_1(t)$ and $\tilde{a}_2(t)$ are the randomly time-varying complex attenuations of the two multipath components, which causes fluctuations in the received signal power. Changes in the received signal power as a function of time is called *fading*. The terms $\tilde{a}_1(t)$ and $\tilde{a}_2(t)$ are usually modeled as *uncorrelated and stationary* random processes.

Movement of the mobile unit over larger distances ($d \gg \lambda$, where λ is the wavelength), and changes in terrain features affect attenuation and received signal power slowly, and this phenomenon is called *large-scale (or slow) fading*. The received signal in each path in Fig. 6 is made up of a large number of scattered components, and hence the central-limit theorem leads to complex Gaussian process models for the complex attenuation and the complex envelope of the received signal for each path.

Movement over small distances on the order of $\lambda/2$ causes significant phase changes of the scattered components, resulting in rapid fluctuations in signal amplitude and power. This phenomenon is called *small-scale (fast) fading*. Large-scale fading impacts the link

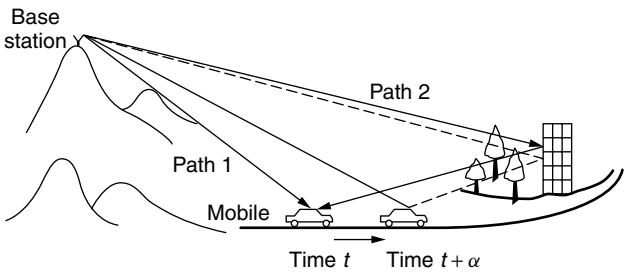


Figure 6. Mobile communication environment.

$S/N + I$ and coverage within a given area. Small-scale fading, on the other hand, impacts all the signal processing operations in the transmitter and receiver, and the effects of small-scale fading are simulated at the waveform level. The basic simulation model for small-scale fading is a TDL model of the form given in Fig. 3.

The tap gains are random processes representing the random variations in the channel characteristics such as the complex attenuation of each multipath component. These are usually modeled as complex Gaussian processes with zero mean and Rayleigh envelope distribution for dense urban environments, and nonzero mean and Rice envelope statistics for rural and suburban environments. The simulation model is specified in terms of the number of multipath components, relative delays, average power received in each path, and the power spectral density of the random process models that describe the random variations of the path characteristics (i.e., the tap gain functions). An example of the multipath model that is used for simulating the mobile radio environments for the design of the third-generation cellular systems is given in Table 1 [12].

In the simulation model, the tap gain functions are generated by filtering uncorrelated Gaussian random processes. The filter transfer function is carefully synthesized to produce the desired Doppler power spectral density. Details of the algorithms used for generating Gaussian sequences with a given power spectral density are discussed in Section 4.

3.4.2. Discrete-Channel Model. Whereas the TDL model is used to simulate the waveform-level distortions introduced by a multipath fading communication channel, a discrete Markov model is often used to characterize the burst errors introduced by fading channels [13–15]. These discrete channels models are computationally more efficient than the waveform-level simulation models for evaluating the performance of error control encoders/decoders, interleaves, and other devices. An example of a simple two-state Markov model for a fading channel is shown in Fig. 7.

Table 1. Example of a TDL Model for an Outdoor Mobile Radio Channel at 2 GHz

Tap Delay (ns)	Average Power (db)
0	0.0
244	−2.4
488	−6.5
732	−9.4
936	−12.7
1220	−13.3
1708	−15.4
1953	−25.4

In the Makov model, the channel is in one of the two states at the beginning of a symbol interval. If the channel is in good state, the transmitted symbol is error-free whereas the probability of transmission error is 0.5 when the channel is in the bad state. While the channel remains in the bad state, it produces bursts of transmission errors. The channel can transition from state i to state j at the end of a symbol interval with a probability of $p_{ij}, i, j = 1, 2$. The rate or the probability of transition between the good and bad states and the error generation probabilities can be derived analytically from the underlying fading channel model or estimated from error sequences obtained using waveform level simulations.

Markov models are also applicable to hard input and soft output channels. Details of the Baum–Welch algorithm used for estimating the structure and parameters of the Markov models and examples of discrete channel models may be found in the literature [16–18].

Simulation of the Markov model is carried out at the symbol rate. For simulating the flow of a symbol through the discrete-channel model, two uniform random numbers are drawn, one to determine the state of the channel (i.e., the transition) and a second random number to decide whether the transmitted symbol suffers a transmission error. Thus, the simulation of the discrete model is very efficient compared to having to generate sampled values of the transmitted waveform and process them through all the functional blocks in the waveform-level simulation model.

4. MONTE CARLO SIMULATION AND RANDOM-NUMBER GENERATION

With respect to the simulation model shown in Fig. 1, the inputs or stimuli that drive the simulation model are sampled values of the input signals, noise, and interference, which are modeled as random processes. Sampled values of random processes are random variables, and hence

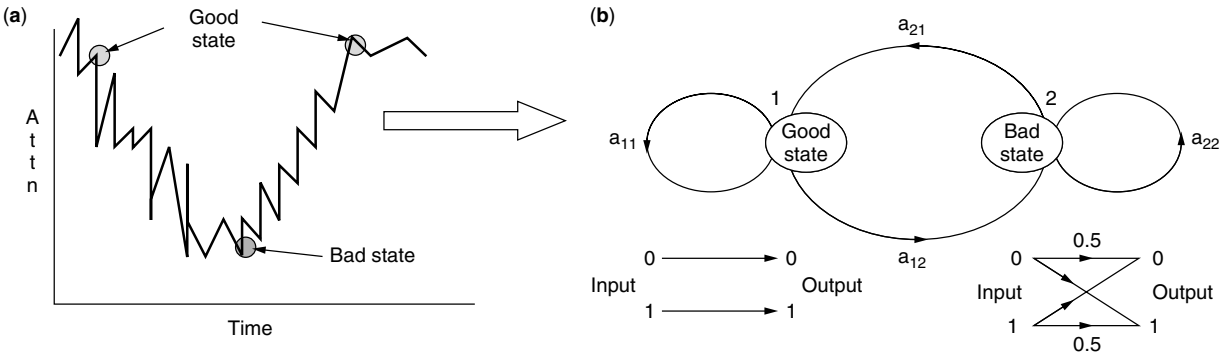


Figure 7. (**a**) Fading channel; (**b**) Markov model for a fading channel.

the input sequences that drive the simulation models are sequences of random numbers from arbitrary distributions and with arbitrary power spectral densities or autocorrelation functions. Hence Monte Carlo simulation involves the generation and processing of random numbers.

4.1. Uniform Random-Number Generator

A typical Monte Carlo simulation might entail the generation and processing of many thousand samples of random variables, and hence we need computationally efficient algorithms for generating random numbers. The starting point of random-number generation is the uniform random-number generator. An independent sequence of uniform random integers in the interval $[0, M-1]$ can be generated using the *linear congruential algorithm* [19–21]

$$X_{j+1} = (aX_j + c) \bmod M \quad \text{(integer arithmetic)} \tag{26}$$

This recursive algorithm is started using an initial random *seed* that is provided by the user and it produces a set of integers uniformly distributed in the interval $[0, M-1]$. A sequence of uniform random numbers in the interval [0,1) can be obtained according to $U_i = float(X_i/M)$.

The output sequence produced by the recursive algorithm given in Eq. (26) will be periodic with a maximum period of M. The values of a, c and M are carefully chosen such that the sequence is independent, is uniformly distributed, and has the maximum period. Two popular algorithms that produce uniform sequences with long periods are the Marsaglia–Zamann algorithm and the Wichmann–Hill algorithm [1].

Since the uniform random-number generators play such a central role in Monte Carlo simulation and random-number generation, they should be thoroughly tested for temporal and distributional properties. Statistical tests for these can be found in the literature [1,6].

4.2. Random-Number Generators for Arbitrary Distributions

An independent sequence of random numbers from arbitrary probability distributions can be generated by applying appropriate memoryless nonlinear transformations to an independent uniform sequence U_i [1]. It can be shown that U_i can be mapped to a sequence of random numbers X_i from an arbitrary distribution $F_X(x)$ according to

$$X_i = F_X^{-1}(U_i) \tag{27}$$

where F_X^{-1} is the inverse cumulative distribution function (CDF) of X. This method is called the *inverse transform method* of generating random numbers, and it can be easily applied to distributions that are analytically tractable. For example, if we want to produce a sequence of random numbers from an exponential probability density function (PDF), the inverse transform method yields the formula $X_i = (-1/\lambda)\ln(1-U_i)$.

If the CDF and/or the inverse of the CDF of X cannot be expressed in closed form, then the inverse transform method can be implemented in empirical form by quantizing the underlying PDF and creating a piecewise linear CDF and applying the inverse transform method empirically. The details of this approach are shown in Fig. 8.

4.3. Gaussian Random-Number Generators

Gaussian random processes are used to model signals, noise, and interference as well as fading in communication channels, and hence it is important to have computationally efficient algorithms for generating Gaussian random numbers. Two algorithms for generating Gaussian random numbers are [20,21]

1. The sum of a large number of uniform random numbers (usually 12), which leads to an approximately Gaussian distribution by virtue of the central-limit theorem
2. The Box–Mueller method, which uses the following algorithm

$$\begin{aligned} X_1 &= [-2\ln(U_1)]^{1/2}\cos 2\pi U_2 \\ X_2 &= [-2\ln(U_1)]^{1/2}\sin 2\pi U_2 \end{aligned} \tag{28}$$

where X_1, X_2 are two independent Gaussian samples derived from two independent uniform random numbers U_1, U_2.

The Box–Mueller method produces a better distribution than does the sum-of-12 method.

4.4. Correlated Gaussian Sequences

Correlated Gaussian sequences with a given power spectral density (PSD) or autocorrelation function can be generated by filtering an uncorrelated Gaussian sequence. The filter transfer function can be synthesized using a number of different approaches. For FIR implementation of the filter, the transfer function of the filter can be chosen according to

$$H(f) = \sqrt{(S_{YY}(f))} \tag{29}$$

where $S_{YY}(f)$ is the desired power spectral density. This method can be used for generating temporally correlated

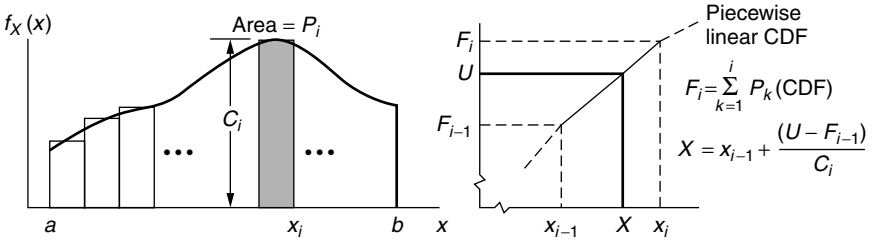


Figure 8. Empirical version of the inverse transform method.

Gaussian processes employed in modeling multipath fading communication channels. An IIR filter transfer function can be synthesized using auotregressive moving-average (ARMA) model. Details of this procedure may be found in the literature [1,6,22].

4.5. Binary and Nonbinary Sequences

The input symbol sequences used in the simulation of digital transmission systems can be generated by mapping the output of a uniform random-number generator into binary and M-ary sequences. Discrete-symbol sequences can also be generated using a linear feedback shift register arrangement in which the feedback tap weights are chosen to be the coefficients of primitive polynomials in Galois field (GF) (2^k), $2^k = M$. An example of this for the binary case is shown in Fig. 9.

These shift register sequences, which are also called *pseudonoise* (PN) *sequences*, have many desirable properties that are useful in the context of simulations. Two of the most important properties of the shift register sequences are that they have the maximum period ($2^m - 1$ in the binary case), and they produce all possible m symbol combinations within one period where m is the number of stages in the shift register structure. This property is very useful for simulating intersymbol interference (ISI) and other forms of linear distortion in digital transmission systems. Although it is possible to do this with random sequences derived from a uniform random-number sequence, the sequence length required to produce all possible m symbol combinations will be much longer than the PN sequences.

Details on PN sequence generation and a list of primitive polynomials can be found in Ref. 1.

5. PERFORMANCE ESTIMATION VIA MONTE CARLO SIMULATION

The primary use of simulation is performance evaluation and tradeoff studies. A number of performance metrics such as power spectral densities, S/N at the output of a receiver and bit error rates (BER) in digital systems can be estimated using Monte Carlo simulations. Just as in measurements, the estimated quantities are subjected to variations that are inherent in estimation of parameters from random observations (or simulation results). The metrics used to judge the quality of the estimators are the bias and variance. While it is easy to construct estimators that are unbiased, the variance of the estimator is not very easy to control since it depends on the type of estimator used as well as the number of observations used to obtain an estimated value. In general, the variance will be inversely proportional to the sample size, and this often leads to long simulation runs, especially in the case of low BER estimation. Since BER estimation is very important in the simulation of digital transmission systems, we present some of the approaches for estimating BERs in the following sections.

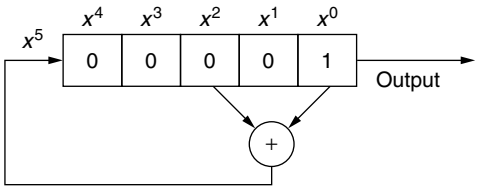


Figure 9. Feedback shift register for generating binary PN sequences; $m = 5$; $g(x) = 1 + x + x^5$.

5.1. MC Techniques for Estimating of BERs in Digital Communication Systems

The Monte Carlo technique is the most general method of estimating BERs and can be applied to any type of communication system with arbitrary distributions for noise and interference. The technique is simple to apply; perform a waveform-level simulation with a long input sequence of length N symbols, count the number of errors between the input symbol stream and the simulated output, and then form a counting estimate for the BER as the ratio of the number of errors counted and the number of symbols simulated (see Fig. 1). If the symbol errors in the system are occurring independently, then the normalized estimation error is given by [1]

$$\text{Normalized error} = \frac{\text{standard deviation of the estimator}}{\text{BER being estimated}} \approx \frac{1}{\sqrt{NP_e}} \tag{30}$$

For estimating a BER on the order of 10^{-6} with a normalized error of, say, 20%, the ordinary MC method requires a sample size on the order of $25(10^6)$ bits to be simulated. Hence the ordinary MC technique is not suitable for estimating very low BERs. A number of alternate methods have been developed in order to reduce sample size requirements for low-BER estimation. An overview of these methods and details may be found in the literature [1,23–25]. Two of these methods are described briefly here.

5.2. Semianalytical (SA) MC Techniques for Low-BER Estimation

This method is applicable to systems in which the effects of noise and interference can be assumed to be additive and Gaussian (of some other known distribution) at the output of the system [1]. In this case the BER in the system can be estimated by running a noiseless simulation to characterize the waveform distortion introduced by all the functional blocks in the system and analytically computing the probability of error due to noise superimposed on the distorted output. A simple example for the binary case with additive noise at the output is shown in Fig. 10. The basic form of the estimator becomes

$$\tilde{P}_e = \frac{1}{N}\sum_{k=1}^{N} P_k; \quad P_k = Q\left(\frac{d_k}{\sigma}\right) \tag{31}$$

where $Q(d_k/\sigma)$ is the analytically computed probability of error for the kth simulated value of the distorted output sample d_k at decision time, N is the number of symbols simulated, and σ is the standard deviation of

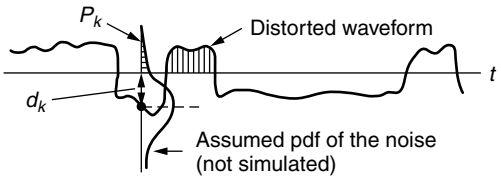


Figure 10. Semianalytical BER (bit error rate) estimator for a binary communication system.

the noise at the output of the system, which is assumed to be Gaussian in this example. With this approach, the simulation length is determined by the number of symbols N needed to simulate the distribution of the waveform distortion accurately. Typically this length will be much smaller than what would be necessary to simulate the effects of noise explicitly, particularly at low BERs. If the primary source of distortion in the system is linear (ISI) and lasts over m symbols, then a PN sequence of length 2^m is all that will be needed in a binary system to simulate all possible distortion values exactly. The variance of the estimator is zero in this case. The SA Monte Carlo methods lead to significant reduction in sample size for low-BER estimation. Indeed, with these methods the sample size requirements are independent of the BER being estimated. The SA methods are applicable to M-PSK and QAM (multiphase shift keying and quadrature amplitude modulation) schemes also [1].

5.3. An Important Sampling Method

This method, which is shown in Fig. 11, is based on biasing the input PDFs such that the important regions of the input PDFs are enhanced to produce a larger number of symbol errors during simulation. The higher error rates can be estimated with smaller sample sizes, and the bias in the estimator, which results from biasing the input PFDs, can be corrected easily at the output of the system where errors are counted. The important sampling method when applied properly has the potential to reduce sample size requirements significantly. Details of the important sampling method can be found in the literature [1,24,26].

6. SUMMARY

Simulation is a very useful tool for the design and analysis of communication systems. In this article we presented the basic principles behind waveform-level simulation of communication systems. The first step is developing a simulation model of the system under study in a block diagram form containing parameterized representations of all the functional blocks in the system that might have a bearing on the design and analysis issues being addressed. The second step is the selection of models for the signal processing operations performed by the functional blocks and representing them appropriately using the lowpass equivalent representation and the sampling theorem. After a simulation model is completely specified in terms of the topology, functional blocks, parameters, and input signals, the next step is to execute the simulations by generating sampled values of all the input signals using appropriate random-number generators and letting the functional blocks in the simulation model operate on the input sequences and produce output sequences. The final step is the estimation of performance metrics of interest such as BERs. This estimation can be done online while the simulation is being executed or at the end of the simulation as an offline postprocessing operation.

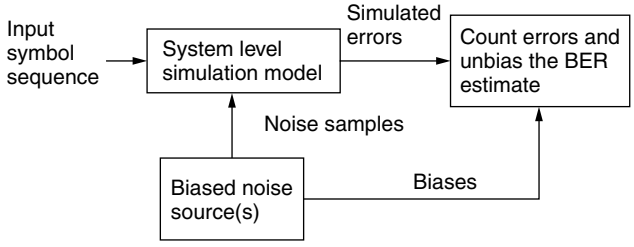


Figure 11. Important sampling method.

The overall simulation accuracy will depend on the modeling assumptions and approximations, accurate representation of signals, and accuracy of the algorithms used for random-number generation, estimation techniques used, and the length of the simulations. All of these issues were addressed in this article.

Simulation of communications is an interdisciplinary activity that requires skills in a broad range of areas, including communication systems, random signal theory, statistics, digital signal processing, and software engineering. A sound understanding of the fundamental principles in these areas as they apply to simulations is essential in order to produce valid and accurate answers to important design and analysis problems that face the communication engineers of today.

BIOGRAPHY

K. Sam Shanmugan received a Ph.D. degree from Oklahoma State University, Stillwater, Oklahoma, in electrical engineering in 1970. He is currently the SW Bell Distinguished Professor of Telecommunication in the Electrical Engineering and Computer Science Departments at the University of Kansas. He has also worked for AT&T Bell Laboratories, TRW, Hughes and Cadence Design Systems. Dr. Shanmugan is the author of over 100 publications in the above areas and is the author/coauthor of three books, *Digital and Analog Communication Systems* (Wiley, 1979), *Random Signals: Detection Estimation and Data Analysis* (Wiley, 1988), and *Simulation of Communication Systems* (Plenum Press, 1992). Dr. Shanmugan is a fellow of the IEEE and is the recipient of many teaching and research awards at the University of Kansas.

BIBLIOGRAPHY

1. M. C. Jeruchim, P. B. Balaban, and K. S. Shanmugan, *Simulation of Communication Systems*, 2nd ed., Kluwer-Plenum Press, New York, 2000 (provides in-depth coverage of all the major topics and also contains several hundred additional references).

2. *IEEE Journal on Selected Areas in Communications*, special issues devoted to computer-aided modeling, analysis and design of communications systems: **SAC-2**(1) (1984); **SAC-6**(1) (1988); **SAC-10**(1) (1992).

3. F. M. Gardner and J. D. Baker, *Simulation Techniques*, Wiley, New York, 1997.

4. A. V. Oppenheim and R. W. Schaffer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

5. A. V. Oppenheim, A. S. Willskey, and L. T. Young, *Signals and Systems*, Prentice-Hall, Englewood-Cliffs, NJ, 1983.

6. K. Sam Shanmugan and A. M. Breipohl, *Random Signals: Detection, Estimation and Data Analysis*, Wiley, New York, 1988.

7. T. Kailath, Channel characterization: Time varying dispersive channels, in *Lecturers in Communication Theory*, McGraw-Hill, New York, 1961.

8. A. A. M. Saleh, Frequency independent and frequency-dependent nonlinear models for TWT amplifiers, *IEEE Trans. Commun.* **Com-29**(11): 1715–1720 (1981).

9. H. M. Poza, Z. A. Sarkozy, and H. L. Berger, A wideband data link computer simulation model, *Proc. NAECON Conf.*, 1975.

10. T. S. Rappaport, *Wireless Communications*, Prentice-Hall, Upper Saddle River, NJ, 1996.

11. B. Sklar, Rayleigh fading channels in mobile digital communications, Parts I and II, *IEEE Commun. Mag.* **35**: 90–110 (July 1997).

12. Modified ITU propagation models for indoor, indoor to pedestrian and vehicular environments, 3GPP Ts.25.101 v.2.1.0 UE Radio Transmission and Reception (FFD), *www.3gpp.org*.

13. B. D. Fritchman, A binary channel characterization using partitioned Markov chains, *IEEE Trans. Inform. Theory* **IT-13**: 221–227 (April 1967).

14. S. Sivaprakasam and K. Sam Shanmugan, An equivalent Markov model for burst errors in digital channels, *IEEE Trans. Commun.* 1347–1356 (April 1995).

15. L. R. Rabiner and B. H. Huang, An introduction to hidden Markov models, *IEEE ASSP Mag.* 4–16 (Jan. 1986).

16. W. Turin, *Performance Analysis of Digital Transmission Systems*, Computer Science Press, Rockville, MD, 1990.

17. W. Turin and P. Balaban, Markov model for burst errors in narrowband CDMA system operating over a fading channel, *Proc. Globecom'98*, Sydney, 1998.

18. A. Beverly and K. Sam Shanmugan, Hidden Markov models for burst errors in GSM and DECT channels, *Proc. Globecom'98*, Sydney, 1998.

19. R. F. W. Coates, G. J. Janacek, and K. V. Lever, Monte Carlo simulation and random number generation, *IEEE J. Select. Areas Commun.* **6**(1): 58–66 (Jan. 1988).

20. D. E. Knuth, *The Art of Computer Programming*, Vol. 2, *Seminumerical Algorithms*, 2nd ed., Addison-Wesley, Reading, MA, 1981.

21. R. Y. Rubinstein, *Simulation and the Monte Carlo Method*, Wiley, New York, 1981.

22. S. L. Marple, Jr., *Digital Spectral Analysis, with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

23. M. C. Jeruchim, Techniques for estimating the bit error rate in the simulation of digital communication systems, *IEEE J. Select. Areas Commun.* **SAC-2**(1): 153–170 (Jan. 1984).

24. K. Sam Shanmugan and P. Balaban, A modified Monte Carlo simulation technique for evaluation of error rate in digital communication systems, *IEEE Trans. Commun.* **COM-28**(11): 1916–1928 (1980).

25. P. M. Hahn and M. C. Jeruchim, Developments in the theory and application of importance sampling, *IEEE Trans. Commun.* **COM-35**(7): 706–714 (July 1987).

SOFT OUTPUT DECODING ALGORITHMS

LANCE C. PÉREZ
University of Nebraska
Lincoln, Nebraska

1. INTRODUCTION

The success of turbo codes and iterative decoding in the field of channel coding for the additive white Gaussian noise (AWGN) and other channels has led to the investigation of iterative techniques in a wide range of disciplines. Indeed, iterative processing is being considered for virtually every component in single and multiuser digital communication systems. Iterative processing typically involves iterative information exchange between system components, such as an equalizer and a channel decoder or an interference canceler and channel decoder, that traditionally operated independently.

The essential element of all iterative processing techniques is some form of a soft-input, soft-output (SISO) module. In this article, a detailed description is given of the most common SISO modules, namely, the soft-output Viterbi algorithm (SOVA) [1] and several versions of the Bahl, Cocke, Jelinek, and Raviv (BCJR) [2] or maximum a posteriori (MAP) algorithm. To make the descriptions concrete, these algorithms are described in the context of their application to the decoding of binary convolutional codes. The extension to other applications is generally straightforward and may be found in the appropriate literature.

The outline of this article is as follows. Section 2 provides a basic system description and introduces the basic concepts and notations of convolutional codes and their trellis representations necessary for the subsequent development of the SISO algorithms. A detailed description of the SOVA is given in Section 3. In Section 4, the MAP algorithm and its max-log variant are described. Section 5 contains some comparisons between these algorithms in the application of iterative decoding of turbo codes. Finally, some concluding remarks and pointers to areas for further reading are given in Section 6.

2. SYSTEM MODEL

For the purposes of this article, a digital communication system with forward error correction (FEC) channel coding can be represented by the block diagram shown in Fig. 1. A detailed derivation of this model and the required assumptions may be found in Ref. [3]. The source is a binary memoryless source that produces a sequence

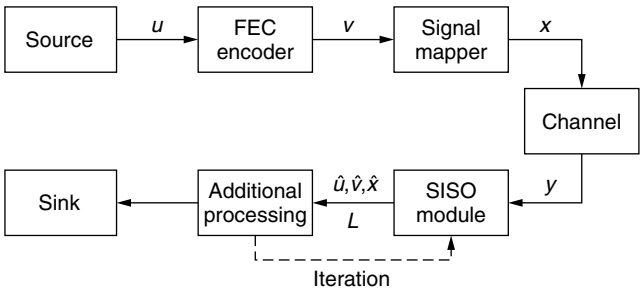


Figure 1. Block diagram of a digital communication system with forward error correction.

$\boldsymbol{u} = [u_0, \ldots, u_r, \ldots, u_{N-1}]$ of independent identically distributed 0's and 1's with equal a priori probabilities of $p_0 = p_1 = 1/2$. The FEC encoder maps the information sequence $\boldsymbol{u}$ to the code sequence $\boldsymbol{v} = [v_0, \ldots, v_r, \ldots, v_{N-1}]$ according to some encoding rule. The signal mapper maps each $n-$tuple v_r to one of 2^n symbols in the signal set. For convenience, the signal mapper is frequently chosen to be a binary phase shift keying (BPSK) or an antipodal modulator that maps each 0 and 1 of the code sequence $\boldsymbol{v}$ to a -1 or $+1$, respectively. The output sequence of the signal mapper, $\boldsymbol{x} = [x_0, \ldots, x_r, \ldots, x_{N-1}]$ is then transmitted across the memoryless AWGN channel, which adds to each transmitted symbol an independent Gaussian random variable with probability density function

$$p_N(n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{n^2}{2\sigma^2}}$$

where $\sigma^2 = \dfrac{N_o}{2}$.

The FEC encoder explicitly considered here is a convolutional encoder of rate $R_c = k/n$ with total encoder memory ν. During each encoding epoch r, the convolutional encoder maps the current $k-$tuple of information bits $u_r = [u_r^{(1)}, \ldots, u_r^{(k)}]$ to an output $n-$tuple of coded bits $v_r = [v_r^{(1)}, \ldots, v_r^{(n)}]$ based on the current input and the current encoder state s_r. The $n-$tuple of coded bits v_r and its corresponding signal mapper output x_r are referred to as a symbol. Although not required, the subsequent decoder descriptions will assume finite length information sequences and thus $r = 0, \ldots, N-1$. The decoding algorithms considered in this article all require the notion of the trellis representation of a convolutional code, which is simply the time expansion of the state transition diagram. In this case, the trellis has a total of 2^ν distinct states $s_r = j$, $j = 0, \ldots, 2^\nu - 1$, at epoch r with 2^k state transitions, or branches, leaving and entering each state.

For example, Fig. 2 depicts a rate $R_c = 1/2$ convolutional encoder with total encoder memory $\nu = 2$ realized in nonsystematic feedforward form [4]. The code is specified by its two generator polynomials, $g_0 = 1 + D + D^2 = 111$ and $g_1 = 1 + D^2 = 101$, which are frequently represented in right justified octal format as $g_0 = 7$ and $g_1 = 5$. The equivalent recursive systematic encoder realization [5] is shown in Fig. 3. Assuming that the encoder is initialized to the all zero state, the trellis diagram with $N = 7$ sections for this encoder is shown in Fig. 4. Each full section of

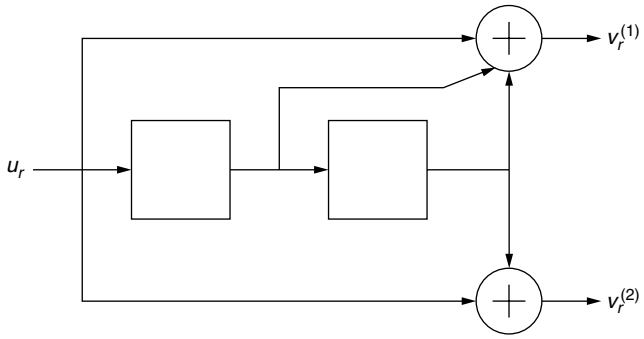


Figure 2. A rate $R_c = 1/2$ convolutional encoder with $\nu = 2$ realized in nonsystematic feedforward form with generator polynomials $g_0 = 1 + D + D^2$ and $g_1 = 1 + D^2$.

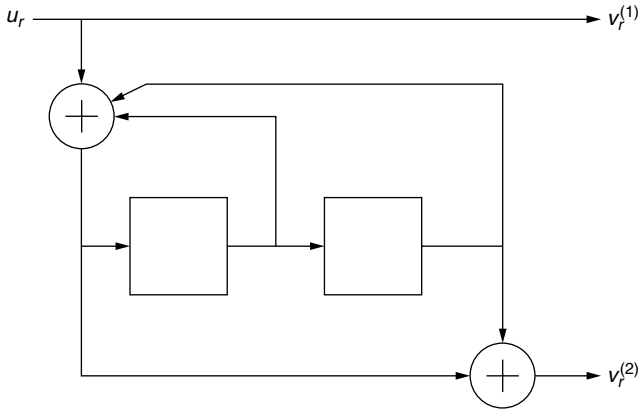


Figure 3. A rate $R_c = 1/2$ convolutional encoder with $\nu = 2$ realized in recursive systematic form with generator polynomials $g_0 = 1 + D + D^2$ and $g_1 = 1 + D^2$.

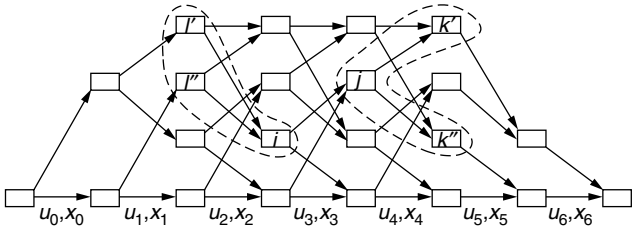


Figure 4. Trellis diagram corresponding to the encoders of Figs. 2 and 3 with information sequences of length 5 and two tail bits.

the trellis has $2^2 = 4$ states and there are $2^1 = 2$ branches leaving and entering each state. In this example, the trellis ends in the all zero state as well. This is referred to as *terminating* the trellis and is accomplished for rate one-half feedforward encoder realizations by appending a *tail* of ν zeroes to the end of the information sequence. Thus, this trellis represents information sequences of length 5 with a tail of 2 zeroes. For recursive systematic and systematic feedback encoder realizations, trellis termination is accomplished by a nonzero, state-dependent tail of ν bits [5]. Although not strictly required by the decoding algorithms described in this article, the subsequent development assumes that the code trellis is terminated.

The decoder operates on the noisy received sequence y. In traditional decoding algorithms for convolutional codes, such as the Viterbi algorithm [6] or sequential decoding, the primary goal of the decoder is to produce an estimate of the transmitted symbol sequence $\hat{\mathbf{x}}$, or equivalently an estimate of the information sequence $\hat{\mathbf{u}}$, consistent with minimizing, or nearly minimizing, an appropriate cost function, such as the probability of an information bit error P_b or the probability of a frame or sequence error P_f. The purpose of the SISO algorithms is to produce an estimate of the transmitted symbol sequence or information sequence along with a sequence $L = [L_0, \ldots, L_r, \ldots, L_{N-1}]$ of soft reliability information indicating the confidence of each component of the sequence estimate. In the context of this article, the reliability information is based on the calculation, or approximation, of the a posteriori probabilities of either the information bits $\Pr[u_r \mid \boldsymbol{y}]$, or the a posteriori probabilities of the transmitted symbols $\Pr[x_r \mid \boldsymbol{y}]$.

In traditional decoding, maximizing the a posteriori probabilities of the information bits, although optimum in terms of P_b, leads to only minor improvements compared to the Viterbi algorithm, which minimizes the sequence error rate $\Pr[\hat{\mathbf{v}} \neq \mathbf{v}] \mid \boldsymbol{y}]$. Thus, even though the MAP algorithm was originally formulated for convolutional codes by Bahl, Cocke, Jelinek, and Raviv [2] in 1972, it was not widely used because it provided no significant improvement over maximum-likelihood decoding and is significantly more complex. Interest in decoding algorithms with soft outputs was rekindled by the work of Hagenauer and Hoeher [1] on the soft output Viterbi algorithm (SOVA) in 1989 and pushed to its current ubiquity with the discovery of turbo codes and iterative decoding in 1993 [7]. For this reason, the discussion of the SISO algorithms begins with the SOVA and then progresses through the MAP algorithm and its variants.

The SISO algorithms described in this chapter all attempt to provide soft reliability information about each bit, either $u_r^{(i)}$ or $v_r^{(i)}$, or each symbol x_r. The basic approach to this is to partition the set of code sequences into two sets Ω_r, which contains sequences where the r^{th} bit or symbol takes on the desired value, and Ω_r^c, which contains sequences where the r^{th} bit or symbol differs from the desired value. The metrics associated with paths in each set may then be used to generate a reliability value for the current bit or symbol. The algorithms described here differ in two fundamental ways: (1) the number of paths used in each set Ω_r and Ω_r^c, and (2) the method for finding these paths.

3. SOFT OUTPUT VITERBI ALGORITHM (SOVA)

As its name suggests, the SOVA is a version of the classical hard output Viterbi algorithm modified to provide soft output reliability information. The Viterbi algorithm minimizes the probability of a sequence error by decoding that sequence $\boldsymbol{v}$ with the largest likelihood $\Pr[\boldsymbol{x} \mid \boldsymbol{y}]$ given the received sequence $\boldsymbol{y}$. For equally likely inputs, this is equivalent to maximizing $\Pr[\boldsymbol{y} \mid \boldsymbol{x}]$. (Thus, the Viterbi algorithm is a *sequence* MAP decoder.) The basic idea of the SOVA is to derive bit or symbol level reliability information from the sequence a posteriori probabilities. The SOVA described here is the algorithm discovered by Hagenauer and Hoeher [1] as modified by Fossorier et al. [8].

To see how this is done, we begin with a brief description of the Viterbi algorithm. For the AWGN channel with BPSK modulation, it is straightforward to show that

$$\Pr[\boldsymbol{y} \mid \boldsymbol{x}] = \prod_{r=0}^{N-1} \prod_{i=1}^{n} \frac{1}{\sqrt{\pi N_o}} \exp\left\{-\frac{(y_r^{(i)} - x_r^{(i)})^2}{N_o}\right\} \sim \prod_{r=0}^{N-1} \prod_{i=1}^{n} \exp\left\{-\frac{(y_r^{(i)} - x_r^{(i)})^2}{N_o}\right\}$$

Using logarithms, this becomes

$$\log \Pr[\boldsymbol{y} \mid \boldsymbol{x}] \sim -\sum_{r=0}^{N-1} \sum_{i=1}^{n} (y_r^{(i)} - x_r^{(i)})^2 \tag{1}$$

and maximizing $\Pr[\boldsymbol{y} \mid \boldsymbol{x}]$ is equivalent to finding the symbol sequence $\boldsymbol{x}$ that is closest to the received sequence in terms of squared Euclidean distance. The Viterbi algorithm is an efficient technique based on the code trellis for finding the closest sequence.

To formulate the Viterbi algorithm, notice that the inner summation in Eq. (1) corresponds to the squared Euclidean distance between the r^{th} symbol in the transmitted sequence $\boldsymbol{x}$ and the r^{th} received symbol in $\boldsymbol{y}$. Since each code sequence is represented by a path through the code trellis, the r^{th} transmitted symbol corresponds to a state transition or branch from a state $s_r = i$ to state $s_{r+1} = j$ and the inner summation in Eq. (1) is referred to as a branch metric. Formally, the branch metric associated with transmitted symbol x_r is defined to be

$$\beta_r(x_r = x) = \sum_{i=1}^{n} (y_r^{(i)} - x^{(i)})^2 \tag{2}$$

Finally, define the *partial path metric* for path l at epoch R and state j as

$$M_{R,l}(j) = \sum_{r=0}^{R-1} \beta_r(x_{r,l} = x) \tag{3}$$

where $j = 0, \ldots, 2^v - 1$ and $x_{r,l}$ is the r^{th} symbol on the l^{th} path.

With these definitions the Viterbi algorithm may now be stated as

Step 1: Initialize $M_0(0) = 0$ and $M_0(i) = \infty$ for $i \neq 0$. Set $R = 0$.

Step 2: Increment $R = R + 1$. For each state $s_R = j$, compute the branch metrics for the 2^k branches entering that state and compute the 2^k partial path metrics

$$M_{R,l}(j) = M_{R-1}(i) + \beta_{R-1}(x_{R-1,l} = x), \tag{4}$$

where $s_{R-1} = i$ is a state at epoch $R-1$ with a branch leading to state $s_R = j$ with branch label $x_{R-1,l}$.

Step 3: For each state $s_R = j$, compare the 2^k partial path metrics $M_{R,l}(j)$ and choose the minimum, that is,

$$M_R(j) = \min_l M_{R,l}(j).$$

Store the corresponding partial sequence $\tilde{\boldsymbol{u}}^j = [\hat{u}_0, \ldots, \hat{u}_{R-1}]$. This partial sequence is known as the survivor for state j at epoch R.

Step 4: If $R < N$, then return to Step 2. If $R = N$, then the survivor to state $s_N = 0$ with path metric $M_N(0)$ is the decoded sequence.

Note that this statement assumes that the encoder started in the all zero state and that the trellis is terminated and ends in the all zero state. Steps 2 and 3 are commonly referred to as the add-compare-select (ACS) operation of the Viterbi algorithm.

The output of the Viterbi algorithm is simply a decoded sequence with no explicit reliability information. The goal of a SISO algorithm is to provide some reliability information, usually in the form of an a posteriori probability, in addition to the decoded sequence. The basic observation behind the SOVA is that the real valued path metric, $M_N(0)$, of the decoded maximum likelihood sequence, $\hat{\mathbf{u}}$ or $\hat{\mathbf{x}}$, provides some reliability information about each decoded information bit $\hat{u}_r$. Assume for the moment that $\hat{u}_r = 1$. If the Viterbi algorithm could be modified to provide the path metric $M_N^c(0)$ of a path with $\hat{u}_r = 0$, then the path metric difference

$$\Delta = |M_N(0) - M_N^c(0)| \tag{5}$$

could be used as a reliability value for the r^{th} bit.

To provide the best reliability measure, the metric $M_N^c(0)$ should correspond to the *best* path with $\hat{u}_r = 0$. That is, $M_N^c(0)$ should correspond to the most likely path in the complementary set of sequences Ω_r^c. The question remains as to how this path can easily be found for each $\hat{u}_r$ of the decoded sequence. One solution is a modified Viterbi algorithm, called the SOVA, that computes reliability information for each bit via a traceback operation during the normal forward recursion of the Viterbi algorithm. These reliability values are stored for every bit in the partial sequence $\tilde{\mathbf{u}}^j$ for every state and the SOVA requires additional memory compared to the standard Viterbi algorithm.

For clarity of exposition and notation, the SOVA will be described for rate $R_c = 1/n$ codes realized in recursive systematic form. The latter assumption incurs no loss of generality, but simplifies the notation by ensuring that the information bit on the last branch of the two competing paths entering each state is different. That is, path 0 entering state $s_R = j$ has $\hat{u}_{R-1,0} = 0$ and path 1 entering $s_R = j$ has $\hat{u}_{R-1,1} = 1$. Consequently, each ACS operation at epoch R immediately generates a reliability value $L_{R-1}^j = \Delta = |M_{R,0} - M_{R,1}|$ for the current bit $\hat{u}_{R-1}$ at the current state $s_R = j$. Since this is the first time that the bit u_{R-1} is decoded, this is the best possible reliability value available at the moment. The situation for the remaining bits in the partial path is more complicated.

As the SOVA progresses through the trellis, it must update these reliability values such that when $R = N$ the reliability values for each bit $\hat{u}_r$ in the decoded sequence is generated from the best path with the complementary bit value in the r^{th} position. This is accomplished through careful updating of the reliability values L_r^j for $r = 0, \ldots, R-2$ following the ACS operation. To understand how this is done, two distinct cases must be considered for each ACS operation. A graphical representation of the ACS operation at state $s_R = 0$ is shown in Fig. 5. Without loss of generality, assume that path 0 is chosen as the survivor.

The first case occurs when the r^{th} bit, for some $0 \le r < R-1$, on path 0, denoted by $\hat{u}_{r,0}$, differs from the corresponding bit $\hat{u}_{r,1}$ on path 1. There are three reliability values to be considered when updating L_r^0:

1. $\Delta = |M_{R,0}(0) - M_{R,1}(0)|$, the difference in the partial path metrics of path 0 and path 1.
2. $L_r^0 = L_{r,0}$, the current reliability value of $u_r^0 = \hat{u}_{r,0}$ on path 0.
3. $L_{r,1}$, the current reliability value of $\hat{u}_{r,1}$ on path 1.

$L_{r,0}$ comes from an ACS operation prior to epoch R between path 0 and, say, path m in which $\hat{u}_{r,0} \neq \hat{u}_{r,m}$. That is, path m belongs to the complementary set Ω_r^c for $\hat{u}_{r,0}$. $L_{r,1}$ comes from an ACS operation prior to epoch R between path 1 and, say, path l in which $\hat{u}_{r,1} \neq \hat{u}_{r,l}$. It follows that $\hat{u}_{r,l} = \hat{u}_{r,0}$ and path l does not belong to the complementary set Ω_r^c for $\hat{u}_{r,0}$. Thus, the update equation for this first case is

$$L_r^j = \min\{\Delta, L_{r,0}\} \tag{6}$$

The second case occurs when the r^{th} bit, for some $0 \le r < R-1$, on path 0, denoted by $\hat{u}_{r,0}$, agrees with the corresponding bit $\hat{u}_{r,1}$ on path 1. Arguing as above, we find that in this case path l is now in the complementary set Ω_r^c with a reliability value of $\Delta + L_{r,1}$ and the update equation is

$$L_r^j = \min\{\Delta + L_{r,1}, L_{r,0}\} \tag{7}$$

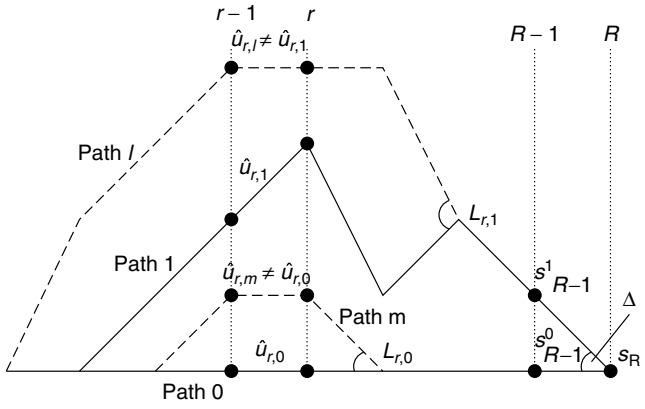


Figure 5. Figure illustrating the paths used in updating the reliability information at state 0 in the SOVA.

With these two update equations, the SOVA may now be stated as follows for rate $R_c = 1/n$ codes.

Step 1: Initialize $M_0(0) = 0$ and $M_0(i) = \infty$ for $i \neq 0$. Set $R = 0$.

Step 2: Increment $R = R + 1$. For each state $s_R = j$, compute the branch metrics for the 2^k branches entering that state and compute the 2^k partial path metrics

$$M_{R,l}(j) = M_{R-1}(i) + \beta_{R-1}(x_{R-1,l} = x) \tag{8}$$

where $s_{R-1} = i$ is a state at epoch $R-1$ with a branch leading to state $s_R = j$ with branch label $x_{R-1,l}$.

Step 3: For each state $s_R = j$, compare the 2^k partial path metrics $M_{R,l}(j)$ and choose the minimum, that is,

$$M_R(j) = \min_l M_{R,l}(j)$$

Store the corresponding partial sequence $\tilde{\boldsymbol{u}}^j = [\hat{u}_0, \ldots, \hat{u}_{R-1}]$. This partial sequence is known as the survivor for state j at epoch R.

Step 3a: For each state $s_R = j$, compute Δ, trace back the survivor and update the reliability values L_R^j of each bit according to Eqs. (6) and (7)

Step 4: If $R < N$, then return to Step 2. If $R = N$, then the survivor to state $s_N = 0$ with path metric $M_N(0)$ is the decoded sequence and the reliability values are L_N^0.

Note that this statement assumes that the encoder started in the all zero state and that the trellis is terminated and ends in the all zero state.

This version of the SOVA computes reliability values based on a single path from each of the sets Ω_r and Ω_r^c. The path used in each set is optimum in the maximum likelihood sequence sense of the Viterbi algorithm. The traceback operation required in this version of the SOVA for updating the reliability values introduces considerable complexity and may not be suitable for some applications. Several alternative algorithms may be found in the literature [9,10]. In the sequel, algorithms that attempt to compute or approximate the a posteriori probabilities directly are discussed.

4. A POSTERIORI PROBABILITY ALGORITHMS

4.1. BCJR or MAP Decoding

The ultimate purpose of this algorithm is the calculation of a posteriori probabilities, such as $\Pr[u_r \mid \boldsymbol{y}]$, or $\Pr[x_r \mid \boldsymbol{y}]$, where $\boldsymbol{y}$ is the received sequence observed at the output of a channel, whose input is the transmitted sequence $\boldsymbol{x}$. Following Ref. 2, it is convenient to calculate the probability that the encoder traversed a specific branch in the trellis, that is, $\Pr[s_r = i, s_{r+1} = j \mid \boldsymbol{y}]$, where s_r is the state at epoch r, and s_{r+1} is the state at epoch $r+1$. The BCJR or MAP algorithm computes this probability as

$$\begin{aligned}\Pr[s_r = i, s_{r+1} = j \mid \boldsymbol{y}] &= \frac{1}{\Pr(\boldsymbol{y})}\Pr[s_r = i, s_{r+1} = j, \boldsymbol{y}] \\ &= \frac{1}{\Pr(\boldsymbol{y})}\alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j) \end{aligned} \tag{9}$$

The α-values are internal variables of the algorithm and are computed by the *forward recursion*

$$\alpha_{r-1}(i) = \sum_{\text{states } l} \alpha_{r-2}(l)\gamma_{r-1}(i,l) \tag{10}$$

This forward recursion evaluates α-values at time $r-1$ from previously calculated α-values at time $r-2$, and the sum is over all states l at time $r-2$ that connect with state i at time $r-1$. The forward recursion for the trellis of Fig. 4 is illustrated in Fig. 6. To enforce the boundary condition that the encoder begins in state 0, the α-values are initialized as $\alpha_0(0) = 1, \alpha_0(1) = \alpha_0(2) = \alpha_0(3) = 0$.

The β-values are calculated in a similar manner, called the *backward recursion*

$$\beta_r(j) = \sum_{\text{states } k} \beta_{r+1}(k)\gamma_{r+1}(k,j) \tag{11}$$

with initial values of $\beta_N(0) = 1, \beta_N(1) = \beta_N(2) = \beta_N(3) = 0$ to enforce the terminating condition of the trellis code. The sum is over all states k at time $r+1$ to which state j at time r connects. The backward recursion is illustrated in Fig. 7.

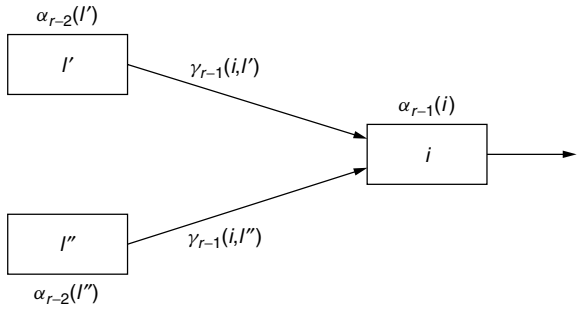


Figure 6. Illustration of the forward recursion of the MAP algorithm.

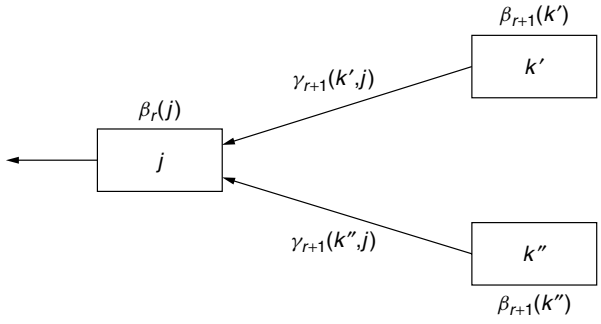


Figure 7. Illustration of the backward recursion of the MAP algorithm.

The γ-values are conditional transition probabilities and are the inputs to the algorithm based on the received sequence. Specifically, $\gamma_r(j,i)$ is the joint probability that the state at time $r+1$ is $s_{r+1}=j$ and that y_r is received, given $s_r=i$. It is calculated using the expression

$$\begin{aligned}\gamma_r(j,i) &= \Pr(s_{r+1}=j, y_r \mid s_r=i)\\ &= \Pr[s_{r+1}=j \mid s_r=i]\Pr(y_r \mid x_r)\end{aligned} \tag{12}$$

where $\Pr[s_{r+1}=j \mid s_r=i]$ is the a priori transition probability and is related to the probability of u_r. For feedforward encoders, the top transition in the trellis diagram of Fig. 4 is associated with $u_r=1$ and the bottom transition with $u_r=0$. This term is used to account for a priori probability information on the bits u_r. To simplify notation, this transition probability will be denoted as

$$p_{ij} = \Pr[s_{r+1}=j \mid s_r=i] = \Pr[u_r] \tag{13}$$

The second term, $\Pr(y_r \mid x_r)$, is simply the conditional channel transition probability, given that symbol x_r is transmitted. Note that x_r is the symbol associated with the transition from state $i \to j$.

The a posteriori symbol probabilities $\Pr[u_r \mid \boldsymbol{y}]$ can now be calculated from the a posteriori transition probabilities (9) by summing over all transitions corresponding to $u_r=1$, and, separately, by summing over all transitions corresponding to $u_r=0$, to obtain

$$p[u_r=1 \mid \boldsymbol{y}] = \frac{1}{\Pr(\boldsymbol{y})}\sum_{u_r=1}\Pr[s_r=i, s_{r+1}=j, \boldsymbol{y}] \tag{14}$$

$$p[u_r=0 \mid \boldsymbol{y}] = \frac{1}{\Pr(\boldsymbol{y})}\sum_{u_r=0}\Pr[s_r=i, s_{r+1}=j, \boldsymbol{y}] \tag{15}$$

From these equations it is clear that the MAP algorithm computes the a posteriori probabilities using all the sequences in the sets Ω_r and Ω_r^c.

The derivation of the MAP algorithm requires the probability

$$q_{ij}(x) = \Pr(\tau(u_r, s_r)=x \mid s_r=i, s_{r+1}=j) \tag{16}$$

that is, the a priori probability that the output x_r at time r assumes the value x on the transition from state i to state j. For convolutional codes, as opposed to coded modulation schemes, this probability is a deterministic function of i and j.

To begin, define the internal variables α and β by their probabilistic meaning. These are

$$\alpha_r(j) = \Pr(s_{r+1}=j, \tilde{\boldsymbol{y}}) \tag{17}$$

the joint probability of the partial sequence $\tilde{\boldsymbol{y}}=(y_{-l},\ldots,y_r)$ up to and including time epoch r and state $s_{r+1}=j$; and

$$\beta_r(j) = \Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j) \tag{18}$$

the conditional probability of the remainder of the received sequence $\boldsymbol{y}$ given that the state at time $r+1$ is j.

It is now possible to calculate

$$\begin{aligned}\Pr(s_{r+1}=j, \boldsymbol{y}) &= \Pr(s_{r+1}=j, \tilde{\boldsymbol{y}}, (y_{r+1},\ldots,y_l))\\ &= \Pr(s_{r+1}=j, \tilde{\boldsymbol{y}})\Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j, \tilde{\boldsymbol{y}})\\ &= \alpha_r(j)\beta_r(j)\end{aligned} \tag{19}$$

where we have used the fact that $\Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j, \tilde{\boldsymbol{y}}) = \Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j)$, that is, if $s_{r+1}=j$ is known, events after time r are independent of the history $\tilde{\boldsymbol{y}}$ up to s_{r+1}.

In the same way we calculate via Bayes' expansion

$$\begin{aligned}\Pr(s_r=i, s_{r+1}=j, \boldsymbol{y}) &= \Pr(s_r=i, s_{r+1}=j, (y_{-l},\ldots,y_{r-1}),\\ &\quad \times y_r, (y_{r+1},\ldots,y_l))\\ &= \Pr(s_r=i, (y_{-l},\ldots,y_{r-1}))\\ &\quad \times \Pr(s_{r+1}=j, y_r \mid s_r=i)\\ &\quad \times \Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j)\\ &= \alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j)\end{aligned} \tag{20}$$

Now, again applying Bayes' rule and $\sum_b p(a,b) = p(a)$, we obtain

$$\begin{aligned}\alpha_r(j) &= \sum_{\text{states } i}\Pr(s_r=i, s_{r+1}=j, \tilde{\boldsymbol{y}})\\ &= \sum_{\text{states } i}\Pr(s_r=i, (y_{-l},\ldots,y_{r-1}))\Pr(s_{r+1}=j, y_r \mid s_r=i)\\ &= \sum_{\text{states } i}\alpha_{r-1}(i)\gamma_r(j,i)\end{aligned} \tag{21}$$

For a trellis code started in the zero state at time $r=-l$ we have the starting conditions

$$\alpha_{-l-1}(0)=1, \alpha_{-l-1}(j)=0;\ j\neq 0 \tag{22}$$

As above, we similarly develop an expression for $\beta_r(j)$, that is,

$$\begin{aligned}\beta_r(j) &= \sum_{\text{states } i}\Pr(s_{r+2}=i, (y_{r+1},\ldots,y_l) \mid s_{r+1}=j)\\ &= \sum_{\text{states } i}\Pr(s_{r+2}=i, y_{r+1} \mid s_{r+1}=j)\\ &\quad \times \Pr((y_{r+2},\ldots,y_l) \mid s_{r+2}=i)\\ &= \sum_{\text{states } i}\beta_{r+1}(i)\gamma_{r+1}(i,j)\end{aligned} \tag{23}$$

The boundary condition for $\beta_r(j)$ is

$$\beta_l(0)=1, \beta_l(j)=0; \quad j\neq 0 \tag{24}$$

for a trellis code which is terminated in the zero state.

Furthermore, the general form of the γ values is given by

$$\begin{aligned}\gamma_r(j,i) &= \sum_{x_r} \Pr(s_{r+1}=j \mid s_r=i) \\ &\quad \times \Pr(x_r \mid s_r=i, s_{r+1}=j)\Pr(y_r \mid x_r) \\ &= \sum_{x_r} p_{ij} q_{ij}(x_r) p_N(y_r - x_r) \end{aligned} \tag{25}$$

Equations (21) and (23) are iterative and the a posteriori state and transition probabilities can now be calculated via the following algorithm.

Step 1: Initialize $\alpha_{-l-1}(0)=1, \alpha_{-l-1}(j)=0$ for all non-zero states ($j \neq 0$) of the encoder, and $\beta_l(0)=1, \beta_l(j)=0, j \neq 0$. Let $r=-l$.

Step 2: For all states j calculate $\gamma_r(j,i)$ and $\alpha_r(j)$ via Eqs. (25) and (21).

Step 3: If $r<l$, let $r=r+1$ and go to Step 2, else $r=l-1$ and go to Step 4.

Step 4: Calculate $\beta_r(j)$ using Eq. (23). Calculate $\Pr(s_{r+1}=j, \boldsymbol{y})$ from Eq. (19), and $\Pr(s_r=i, s_{r+1}=j; \boldsymbol{y})$ from Eq. (9).

Step 5: If $r>-l$, let $r=r-1$ and go to Step 4.

Step 6: Terminate the algorithm and output all the values $\Pr(s_{r+1}=j, \boldsymbol{y})$ and $\Pr(s_r=i, s_{r+1}=j, \boldsymbol{y})$.

The a posteriori state and transition probabilities produced by this algorithm can now be used to calculate a posteriori information bit probabilities, that is, the probability that the information k-tuple $u_r=u$, where u can vary over all possible binary k-tuples. Starting from the transition probabilities $\Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y})$, we simply sum over all transitions $i \to j$ that are caused by $u_r=u$. Denoting these transitions by $A(u)$, we obtain

$$\Pr(u_r=u) = \sum_{(i,j)\in A(u)} \Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \tag{26}$$

As mentioned previously, another most interesting product of the APP decoder is the a posteriori probability of the transmitted output symbol x_r. Arguing analogously as above, and letting $B(x)$ be the set of transitions on which the output signal x can occur, we obtain

$$\begin{aligned}\Pr(x_r=x) &= \sum_{(i,j)\in B(x)} \Pr(x \mid y_r)\Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \\ &= \sum_{(i,j)\in B(x)} \frac{p_N(y_r-x_r)}{p(y_r)} q_{ij}(x) \\ &\quad \times \Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \end{aligned} \tag{27}$$

where the a priori probability of y_r can be calculated via

$$p(y_r) = \sum_{\substack{x' \\ ((i,j)\in B(x))}} p(y_r \mid x') q_{ij}(x') \tag{28}$$

and the sum extends over all transitions $i \to j$.

Equation (27) can be much simplified if there is only one output symbol on the transition $i \to j$. In this case, the transition automatically determines the output symbol, and

$$\Pr(x_r=x) = \sum_{(i,j)\in B(x)} \Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \tag{29}$$

5. LOG-MAP AND THE MAX-LOG-MAP

5.1. The MAP Algorithm in the Logarithm Domain (Log-MAP)

Although the MAP algorithm is concise and consists only of multiplications and additions, current direct digital hardware implementations of the algorithm lead to complex circuits due to many real-number multiplications involved in the algorithm. To avoid these multiplications, we transform the algorithm into the logarithm-domain. This results in the so-called *log-MAP* algorithm.

First, we transform the forward recursion (10), (21) into the logarithm-domain using the definitions

$$A_r(i) = \log(\alpha_r(i)); \quad \Gamma_r(i,l) = \log(\gamma_r(i,l)) \tag{30}$$

to obtain the *log-domain* forward recursion

$$A_{r-1}(i) = \log\left(\sum_{\text{states } l} \exp\left(A_{r-2}(l) + \Gamma_{r-1}(i,l)\right)\right) \tag{31}$$

Likewise, the backward recursion can be transformed into the logarithm-domain using the analogous definition $B_r(j)=\log(\beta_r(j))$, and we obtain

$$B_r(j) = \log\left(\sum_{\text{states } k} \exp\left(B_{r+1}(k) + \Gamma_{r+1}(k,j)\right)\right) \tag{32}$$

The product in Eqs. (9) and (20) now turns into the simple sum

$$\alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j) \to A_{r-1}(i) + \Gamma_r(j,i) + B_r(j) \tag{33}$$

Unfortunately, Eqs. (31) and (32) contain log() and exp() functions, which are more complex than the original multiplications. However, in most cases of current practical interest, the MAP algorithm is used to decode binary codes with $R_c=1/n$ where there are only two branches involved at each state, and therefore only sums of two terms in Eqs. (31) and (32). The logarithm of such a binary sum can be expanded as

$$\begin{aligned}\log(\exp(a)+\exp(b)) &= \log\left(\exp\left(\max(a,b)\right)\right. \\ &\quad \left. \times\ (1+\exp(-|a-b|))\right) \\ &= \max(a,b) + \log\left((1+\exp(-|a-b|))\right) \end{aligned}$$

The second term is now the only complex operation left and there are a number of methods to approach this including a look-up table.

Finally, for binary codes the algorithm computes the log-likelihood ratio (LLR) $\lambda(u_r)$ of the information bits u_r using the a posteriori probabilities (26) as

$$\begin{aligned}\lambda(u_r) &= \log\left(\frac{\Pr(u_r=1)}{\Pr(u_r=0)}\right)\\ &= \log\left(\frac{\displaystyle\sum_{(i,j)\in A(u=1)} \alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j)}{\displaystyle\sum_{(i,j)\in A(u=0)} \alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j)}\right)\end{aligned}$$

$$\lambda(u_r) = \log\left(\frac{\displaystyle\sum_{(i,j)\in A(u=1)} \exp(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j))}{\displaystyle\sum_{(i,j)\in A(u=0)} \exp(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j))}\right) \quad (34)$$

The range of the LLR is $[-\infty,\infty]$, where a large value signifies a high probability that $u_r = 1$.

5.2. Max-Log-MAP

The complexity of the log-MAP algorithm may be further reduced by approximating the forward and backward recursions by

$$\begin{aligned}A_{r-1}(i) &= \log\left(\sum_{\text{states } l} \exp\left(A_{r-2}(l)+\Gamma_{r-1}(i,l)\right)\right)\\ &\approx \max_{\text{states } l}\left(A_{r-2}(l)+\Gamma_{r-1}(i,l)\right)\end{aligned} \quad (35)$$

and

$$\begin{aligned}B_r(j) &= \log\left(\sum_{\text{states } k} \exp\left(B_{r+1}(l)+\Gamma_{r+1}(k,j)\right)\right)\\ &\approx \max_{\text{states } k}\left(B_{r+1}(l)+\Gamma_{r+1}(k,j)\right)\end{aligned} \quad (36)$$

which results in the max-log-MAP algorithm. The final LLR calculation in Eq. (34) is approximated by

$$\begin{aligned}\lambda(u_r) \approx &\max_{(i,j)\in A(u=1)}\left(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j)\right)\\ &-\max_{(i,j)\in A(u=0)}\left(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j)\right)\end{aligned} \quad (37)$$

The advantage of the max-log-MAP algorithm is that it uses only additions and maximization operations to approximate the LLR of u_r. It is very interesting to note that Eq. (35) is the Viterbi algorithm for maximum-likelihood sequence decoding. Furthermore, Eq. (36) is also a Viterbi algorithm, but it is operated in the reverse direction.

Further insight into the relationship between the log-MAP and its approximation can be gained by expressing the LLR of u_r in the form

$$\lambda(u_r) = \log\left(\frac{\displaystyle\sum_{\boldsymbol{x};(u_r=1)} \exp\left(-\frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0}\right)}{\displaystyle\sum_{\boldsymbol{x};(u_r=0)} \exp\left(-\frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0}\right)}\right) \quad (38)$$

where the sum in the numerator extends over all coded sequences $\boldsymbol{x}$ that correspond to information bit $u_r = 1$, and the denominator sum extends over all $\boldsymbol{x}$ corresponding to $u_r = 0$.

It is quite straightforward to see that the max-log-MAP retains only the path in each sum that has the best metrics, and therefore the max-log-MAP calculates an approximation to the true LLR, given by

$$\lambda(u_r) \approx \min_{\boldsymbol{x};(u_r=0)} \frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0} - \min_{\boldsymbol{x};(u_r=1)} \frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0} \quad (39)$$

that is, the metric difference between the nearest path to $\boldsymbol{y}$ with $u_r = 0$ and the nearest path with $u_r = 1$. That is, like the SOVA described earlier, the max-log-MAP approximates the LLR by using the best sequence in each set Ω_r and Ω_r^c. Fossorier et al. [8] have shown that the max-log-MAP and the SOVA presented earlier are identical.

6. PERFORMANCE IN ITERATIVE DECODING

The SOVA, MAP, and max-log-MAP algorithms offer a myriad of complexity and implementation tradeoffs that are application- and technology-dependent and will be left to the literature. The performance tradeoff is clearer, although still somewhat dependent on the application. From the development presented here, one would expect the MAP and log-MAP algorithms to perform identically (ignoring any issues of numerical stability) and that both would slightly outperform the SOVA and max-log-MAP. The latter two algorithms are again essentially identical. To illustrate this, we consider the use of these SISO algorithms in the iterative decoding of a turbo code. It is worth noting that some modifications are required to enable the SOVA to easily use the extrinsic information of the iterative turbo decoder. Details of these modifications may be found in Refs. 10 and 11.

The performance of a turbo code with iterative decoding is shown in Fig. 8. The turbo code is a rate $R_c = 1/3$ code with identical memory $\nu = 4$ constituent encoders and an interleaver of length 4096 bits. The constituent encoders have feedforward polynomial $g_1(D) = 1 + D^4$ and feedback polynomial $g_0(D) = 1 + D + D^2 + D^3 + D^4$. All of the performance curves shown are for 18 complete decoder iterations. As expected, the performance with the MAP and log-MAP component decoders are identical, as is performance with the SOVA and max-log-MAP component decoders. There is a gap of approximately 0.5 dB between the two curves. This loss is attributable to the poorer quality soft information provided by the SOVA and max-log-MAP algorithms due to the use of only a single sequence in complementary set Ω_r^c.

7. CONCLUSION

This article discussed several SISO algorithms suitable, with minor adaptation, to the majority of applications considering soft-information exchange or iterative processing. They were presented in the context of the decoding of binary convolutional codes with finite block lengths. The extensions to sliding windows, block codes, equalization,

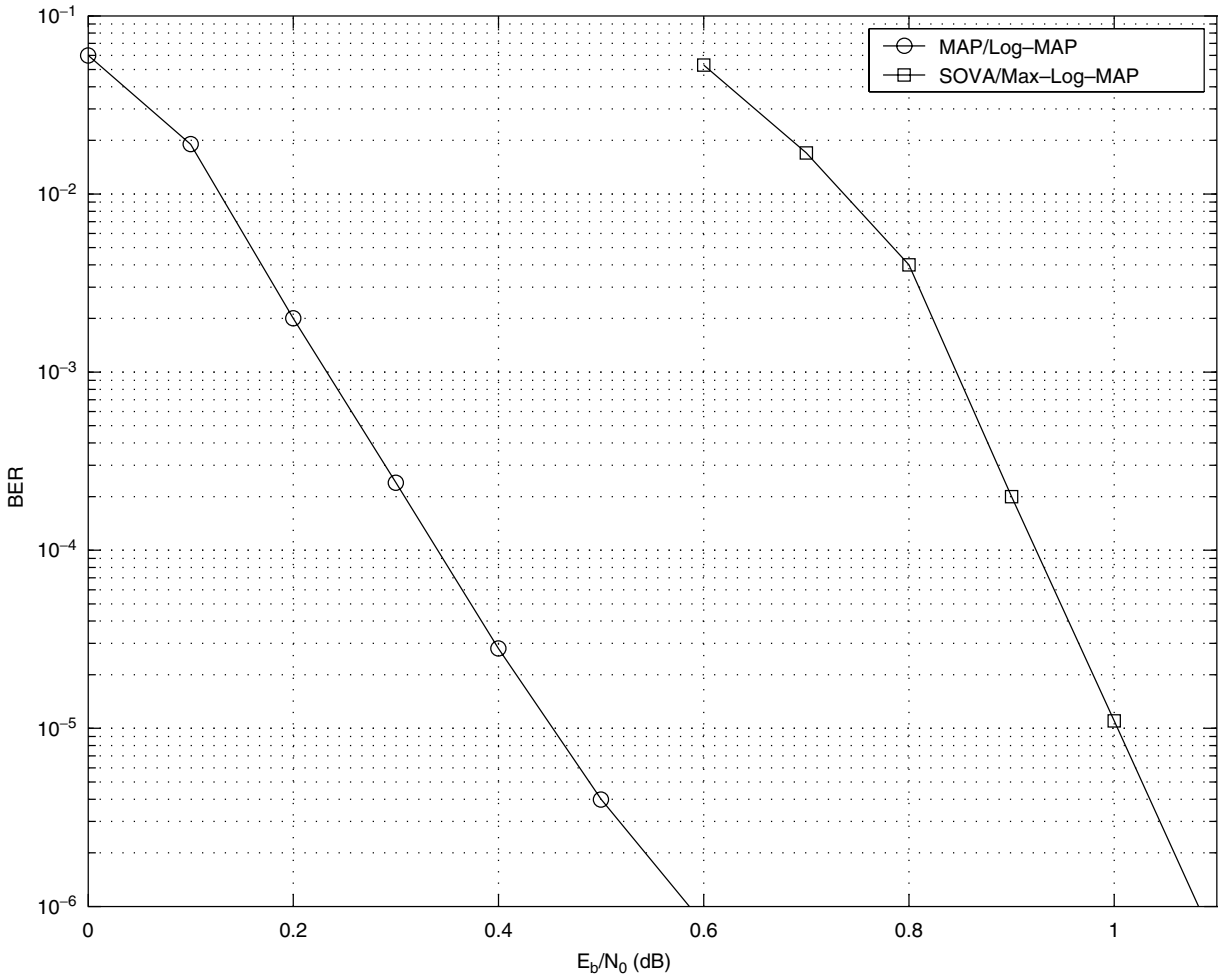


Figure 8. Performance comparison of rate 1/3 turbo code, with interleaver size 4096, and 18 iterations of decoding.

and other applications are relatively straightforward and may be found in the literature.

The virtually universal adoption of SISO algorithms has led to renewed interest in the general theory of decoding algorithms. This has led to several interesting results that unify ideas in fields as diverse as decoding algorithms, graph theory and belief propagation. These results require a degree of mathematical sophistication, but they provide a more general framework in which to understand a broad class of SISO algorithms. The interested reader may find some of these results in Refs. 12–14 and the references therein.

Acknowledgments

The author acknowledge the significant contributions of Christian B. Schlegel to the development of the MAP material in this work and the assistance of Christopher G. Hruby in the preparation of the figures and his valuable comments.

BIOGRAPHY

Lance C. Pérez received the B.S. degree in electrical engineering in 1987 from the University of Virginia, Charlottesville, Virginia, and the M.S. and Ph.D. degrees in electrical engineering from the University of Notre Dame, Notre Dame, Indiana in 1989 and 1995, respectively. From February 1, 1995, to July 31, 1995, Dr. Pérez was also a postdoctoral research associate with a joint appointment from University of Notre Dame and the Institute for Information and Signal Processing at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland. He joined the faculty of the Department of Electrical Engineering at the University of Nebraska, Lincoln, in August 1996 and he is currently an associate professor there. Dr. Pérez is the recipient of a National Science Foundation Career Award and is coauthor with Christian B. Schlegel of the book *Trellis and Turbo Coding* published the IEEE Press/Wiley. His areas of interest are channel coding, digital communications, and engineering education.

BIBLIOGRAPHY

1. J. Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, *Proceedings of GLOBECOM '89*, 1680–1686, Nov. 1989.
2. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).

3. J. G. Proakis, *Digital Communications*, McGraw-Hill, Boston, MA, 1989.
4. S. Lin and D. J. Costello, Jr., *Error Control Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
5. R. Johannesson and K. S. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, Piscataway, NJ, 1999.
6. G. D. Forney, Jr., The Viterbi algorithm, *Proceedings of the IEEE*, PROC-61, 268–278, 1973.
7. Claude Berrou, Alain Glavieux, and Punya Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proceedings of ICC'93*, 1064–1070, May 1993.
8. M. P. C. Fossorier, F. Burkert, S. Lin, and J. Hagenauer, On the equivalence between SOVA and max-log-MAP decodings, *IEEE Comm. Lett.* **COML-2**: 137–139 (May 1998).
9. J. Chen, M. P. C. Fossorier, S. Lin, and C. Xu, Bi-Directional SOVA decoding for turbo-codes, *IEEE Comm. Lett.* **COML-4**: 405–407 (Dec. 2000).
10. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **IT-42**: 429–445 (Mar. 1996).
11. M. Bossert, *Channel Coding for Telecommunications*, Wiley, New York, 1999.
12. R. J. McEliece, On the BCJR trellis for linear block codes, *IEEE Trans. Inform. Theory* **IT-41**: 1072–1092 (July 1996).
13. C. Heegard and S. B. Wicker, *Turbo Codes*, Kluwer, Norwell, 1999.
14. B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.

SOFTWARE RADIO

JOSEPH MITOLA III
Consulting Scientist
Tampa, Florida

1. INTRODUCTION

Software radio (SWR), briefly, is about increasingly wider-bandwidth radiofrequency (RF)-capable digital hardware that is given much or all of its function–personality by software. This chapter provides an overview of the mathematical, engineering, and economics principles of SWR from the theory to practice of radio systems engineering. In the space available, there is little opportunity to address either the larger network architectures, or the software-defined radio (SDR) implementation details. However, the treatment differentiates SWR from SDR. It describes the end-to-end partitioning of SDR requirements. Further, it describes the allocation of critical parameters including dynamic range and processing capacity. Allocation tradeoffs among hardware platforms, firmware and software depend on cost–benefit in the marketplace. In addition, as implementations migrate to software, one must assure that the software is structured well and performs robustly — even when many tasks are competing for processing resources. There are also pointers to relevant industry forums [1] and standards bodies [2].

Some try to "sell" the software radio, but that is not the purpose here. On the contrary, an ideal SWR approach sometimes yields an ineffective product. Pagers, for example, maximize display area and battery life, while minimizing the size and weight of a fixed-function product. Hardware-intensive application-specific integrated circuits (ASICs) are best for that market niche at present. On the other hand, nearly ideal SWR base stations have been deployed since early 2000 because of lower cost of ownership than the baseband–digital signal processor (DSP) base stations that they replace. One therefore must appreciate how analog, digital, and software-intensive approaches complement each other and drive alternative cost–benefit profiles. One may then select the right mix of hardware-intensive and software-defined implementation aspects of a design. This introduction should help the reader appreciate the potential contributions and pitfalls of SWR technology.

SWR is an interdisciplinary technology. It is helpful for software people to understand the RF hardware and air interface standards concepts of an interdisciplinary team. The software-oriented discussion is for people with strong background in RF, analog radio, or DSP but little background in large-scale software. SDRs typically have over one million lines of code (LoC), which is a complex, large-scale software systems. Thus large-scale software tools such as the Unified Modeling Language (UML), the Common Object Request Broker Architecture (CORBA), and the eXtensible Markup Language (XML), typically unfamiliar to RF engineers, loom large in software radio architecture.

The appropriate host platforms for SDR functions change over time. Commercial digital filter ASICs become obsolete as DSP capacity increases, changing the systems level tradeoffs from ASIC to DSP. As needs, technology, and team expertise evolve, the effective choice will also change. Platforms also change with the top–down design constraints, such as market economics, consumer values, and network architecture. For quick time to market, one may procure functions in off-the-shelf code or intellectual property. A sound systems-level architecture facilitates this process, while an inferior architecture inhibits it. The in-depth understanding of SDR therefore includes markets and economics.

One revolutionary aspect of software radio is that knowing how to code a radio algorithm in the programming language C on a DSP chip no longer gives a software engineer the core skills needed to contribute effectively to software radio systems development. In fact, that experience becomes a liability if it causes one to minimize the importance of the new large-scale software engineering methods such as UML, XML, and CORBA.

In addition, European readers will recognize SDL, the ITU-standard Specification and Description Language. In teaching the software radio course on which this article is based, the author finds that Asian and U.S. engineers are less practiced in formal methods for specifying radio functions then their European counterparts. The European Telecommunications Standards Institute (ETSI) emphasis on formal methods and the widespread use of SDL in support of the European standards-setting

process has not permeated U.S. practice, particularly in military, civil, and other non-ITU marketplaces. As a result, U.S. practitioners of radio engineering are doing with generic computer-aided software engineering (CASE) tools what their European counterparts are doing with communications-oriented SDL — defining new radio air interface standards and implementations. This article introduces the formal techniques and software tools needed to effectively develop radios of the next-generation level of complexity.

In addition, software radio has become an industry focus area. Several texts now provide further background reading [3], industry perspectives [4], and in-depth treatment of architecture [5]. This article therefore highlights the numerous approaches with references for the interested reader.

This section introduced SWR and SDR. The next section summarizes the expanding role of SWR in contemporary telecommunications. The subsequent section reviews the fundamental precepts of the ideal SWR: the placement of the analog-to-digital converter (ADC) near the antenna, and the use of software to replace formerly analog or digital hardware. Section 4 introduces implementation-dependent SDR architecture, based on defining functions, components, and design rules that guide SDR product migration. Section 5 examines practical SDR designs, emphasizing implementation constraints. Isochronous performance of the real-time software, for example, is a critical design constraint in the signal processing streams. Section 6 reviews the development parameters and risks associated with choosing designs from the conservative baseband DSP through a variety of SDR alternatives to the ideal SWR. Many SDR projects of the 1990s failed or fell substantially short of requirements because of unanticipated software complexity. Section 7 surveys the broader implications of SWR. Importantly, SWR brings the substantial mitigation of the so-called shortage of radio spectrum, which is an issue more of the economics of spectrum reuse infrastructure than of the laws of physics. Section 8 underscores a few conclusions. A list of acronyms is provided for the reader's convenience, while references are appended in the Bibliography.

2. THE TRANSFORMATION OF RADIO ENGINEERING

We are in the midst of a transformation of radio systems engineering. Throughout the 1970s and 1980s, radio systems migrated from analog to digital in system control, source and channel coding, and baseband signal processing. In the early 1990s, the SWR transformation began to extend these horizons by soft-coding traditionally hardware-defined characteristics of wireless devices, including

- Radiofrequency (RF) band
- RF channel broadband channel coding and bandwidth
- Diversity, intermediate-frequency (IF) combining, beamforming, and antenna characteristics [6]

Today the evolution toward practical SDR is accelerating through a combination of techniques. These include multiband antennas and wideband RF devices. Wideband ADCs and digital-to-analog converters (DACs) now affordably access GHz of spectrum instantaneously. Multiband radios for military, commercial, teleinformatic, intelligent transportation, aircraft, and other applications are increasingly affordable. Processing of IF, baseband, and bit streams is implemented using increasingly general-purpose programmable processors. The complexity of the physical and link-layer protocol software for the many diverse RF bands and physical-layer modes of third- and fourth-generation wireless (3G/4G) now extends to millions of lines of code.

The ideal SWR consists of wideband antenna, wideband ADC/DAC, and general-purpose processor(s). Although SWR configurations are of research interest, they do not meet market constraints. The SDR therefore embraces the evolution of programmable hardware, increasing flexibility via increased programmability within market constraints. The ideal SWR represents and end state of maximum flexibility in this evolution. SDR "futureproofs" practical infrastructure against continually evolving standards and hardware. Strong SWR architecture permits one to insert SDR technology gracefully and affordably. For a clear path for product evolution, one must adapt SWR to specific applications or market niches. The attempts of researchers to build the ideal SWR yield lessons learned from technology pathfinders such as SPEAK*easy* [7,8], FIRST [9,10], and TRUST [11]. Continuing transformation is evident in the creation of the Wireless World Radio Foundation (WWRF) and SDR Forum work in SDR, with global emphasis on SDR as a critical enabler of 3G and 4G wireless. Related work continues to expand in DARPA, the Americas, the ITU, the EC (e.g., GSM MoU and ETSI), and Asia (e.g., ARIB and IEICE).

3. THE IDEAL SOFTWARE RADIO

The top-level components of an ideal SWR handset consist of a power supply, an antenna, a multiband RF converter, a power amplifier, and a single-chip ADC and DAC. These components plus on-chip general-purpose processor and memory perform the radio functions, as illustrated in Fig. 1.

The ideal SWR mobile wireless terminal interfaces directly to the user (e.g., via voice, data, fax, and/or multimedia) and to multiple RF "air" interfaces. Driven by user demands, the mobile unit minimizes dissipated power for long battery life and minimizes manufacturing parts count by maximizing hardware integration for lower unit cost. The generic wireless base station accesses multiple radio air interfaces and the Public Switched Telephone Network (PSTN). With access to the power grid, base stations may employ computationally intensive software designs using modular, open-architecture processing hardware that facilitates technology insertion. Technology insertion futureproofs base-station infrastructure against the continuing evolution of air interfaces. Military base stations ("nodes") need to support multiple networks on

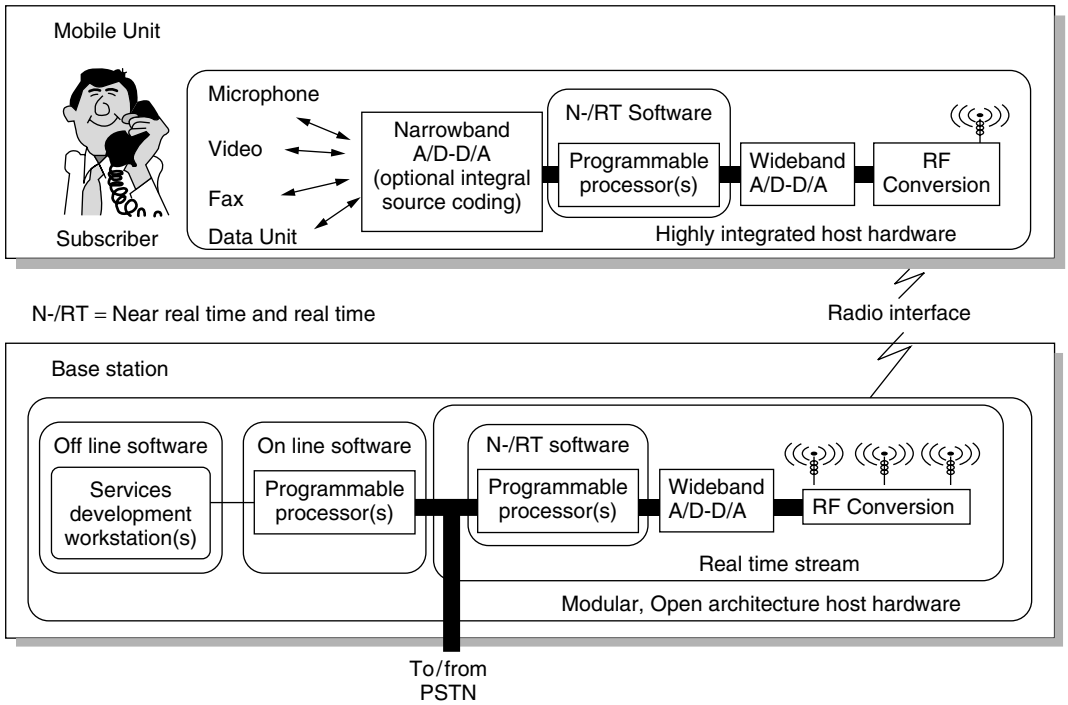


Figure 1. The ideal software radio handset and base station.

multiple RF bands with multiple air interfaces ("modes"). Command-and-control communications (C3) nodes may be formed by the collocation of several radios on mobile vehicles. These configurations often interfere with each other. The military calls this "cosite interference." The SWR basestation attempting to support push-to-talk (PTT) traffic on multiple channels in the same band can also generate self-interference. With software downloads to evolve radio personalities, the management of self-generated interference that was a design issue for the radio engineering laboratory of the 1980s is a deployment-time configuration management issue for SWR.

The placement of the ADC and DAC as close to the antenna as possible and the definition of radio functions in software are the hallmarks of the SWR. Although SWRs use digital techniques, software-controlled digital radios are not ideal SWRs. The essential difference is the total programmability of the SWR, including software-defined RF bands, channel access filters, beamforming, and physical-layer channel modulation. SDR designs, on the other hand, liberally mix analog hardware, digital hardware, and software technologies. SDR has become practical as DSP costs per millions of instructions per second (Mips) have dropped below $10 and continue to plummet. The economics of software radios become increasingly compelling as demands for flexibility increase while these costs continue to drop by a factor of 2 every 1.5–3 years.

By April 2002, SDR technology cost-effectively implemented commercial 1G analog and 2G digital mobile cellular radio air interfaces. 3G SDR base stations were being developed. Over time, wideband 4G air interfaces will also yield to software techniques on wideband RF platforms. In parallel, multiband multimode military radios were being deployed, such as the digital modular radio (DMR) [12], and the Joint Tactical Radio System (JTRS) [13] "clusters."[1] Such SDR implementations require a mix of increasingly sophisticated software technology along with hardware-intensive techniques such as ASICs.[2]

3.1. The Ideal Functional Components

Technology advances have ushered in new radio capabilities that require an expansion of the canonical communications functional model: source coding, the channel, and channel coding. The new aspects are captured in the software radio functional model. First, multiband technology [14], accesses more than one RF band of communications channels at once. The RF channel, then, is generalized to the channel set of Fig. 2. This set includes RF channels, but radio nodes such as personal communications system (PCS) base stations and portable military radios also interconnect to fiber and cable; therefore these are also included in the channel set. The channel encoder

[1] In this article, the conventional notion of cellular radio is extended to embrace the idea that the propagation of RF from any SDR transmitter defines an implicit RF cell. Its size and shape is determined by the physical placement of antenna(s) and the environment. Antenna height, directivity, path loss, diffraction, and multipath loss shape the cell. A multiband, multimode SDR is uniquely suited to turn such implicit cells into explicitly managed ad hoc cellular networks.

[2] In fact, the continuing interplay among military and commercial software radios plays an important role in the evolution of SDR technology. The merger of these market segments around common interest in open architecture SDR platforms is an on-going process, complete with the common interests and discontinuities.

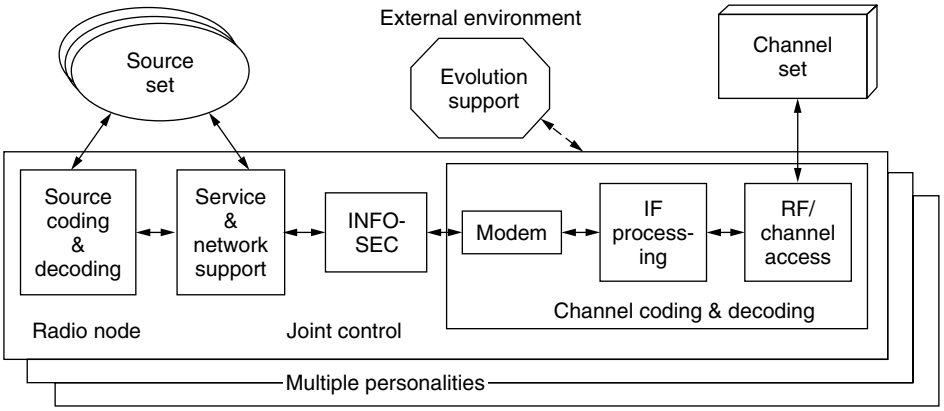


Figure 2. Functional model of a software radiocommunications system.

of a multiband radio includes RF/channel access, IF processing, and modem. The RF/channel access includes wideband antennas, and the multielement arrays of smart antennas [15]. This segment also provides multiple signal paths and RF conversion that span multiple RF bands. IF Processing may include filtering, further frequency translation, space/time diversity processing, beamforming, and related functions. Multimode radios [7] generate multiple air interfaces (also called "waveforms") defined principally in the modem, which is the RF channel modulator–demodulator. These waveforms may operate only in specific RF bands or may span multiple bands. A software-defined personality includes RF band, channel set (e.g., control and traffic channels), air interface waveforms, INFOSEC, network interfaces, and related user interface functions. In this abstraction of information source, the user is just another source (/sink) set.

Although few applications require Information Security (INFOSEC), there are incentives for its use. Authentication reduces fraud. Stream encipherment ensures privacy. Both help assure data integrity. Transmission security (TRANSEC) hides the fact of a communications event (e.g., by spread-spectrum techniques [16]). INFOSEC is therefore included in the functional model, although this function may be null for some applications.

In addition, the source coding/decoding pair of the expanded model includes the data, facsimile, video, and multimedia sources essential for new services. Some sources will be physically remote from the radio node, connected via the synchronous digital hierarchy (SDH), a local-area network (LAN) [17], and other systems, through the Service & Network Support shown in Fig. 2.

These functions may be implemented in multithreaded multiprocessor software orchestrated by a joint control function. Joint control assures system stability, error recovery, timely dataflow, and isochronous streaming of voice and video. As radios become more advanced, joint control becomes more complex, evolving toward autonomous selection of band, mode, and data format in response to implicit user needs. An autonomous software radio capable of machine learning is called a "cognitive radio." Cognitive radio requires a knowledge processing architecture in the joint control function, overlaid on the SWR architecture discussed in this article [18].

Any of the radio node functions may be singleton (e.g., single band vs. multiple bands) or null, further complicating joint control. Agile beamforming supports additional users and enhances quality of service (QoS) [19]. Beamforming today requires dedicated processors, but in the future, these algorithms may timeshare a DSP pool along with, for instance, a rake receiver [20] and other modem functions. Joint source and channel coding [21] also yields computationally intensive waveforms. Dynamic selection of band, mode, and diversity as a function of QoS [22] introduces large variations into demand, potentially causing conflicts for processing resources. These resources may include ASICs, field-programmable gate arrays (FPGAs), DSPs, and general-purpose computers. Channel strapping, adaptive waveform selection, and other forms of data-rate agility [23] further complicate the statistical structure of the computational demand. In addition, processing resources may be unavailable because of equipment failure [24]. Joint control therefore integrates fault modes, personalities, and support functions, mapping the highest priority radio functions onto the available processing resources. to yield a reliable telecommunications object [25].

In a software radio, the user can upload a variety of new air interface personalities [26]. These may modify any aspect of the air interface, including how the carrier is hopped, the spectrum is spread, and beams are formed. The required radio resources are RF access, digitized bandwidth, dynamic range, memory, and processing capacity. Resources used must not exceed those available on the radio platform. Some mechanism for evolution support is therefore needed to define the waveform personalities, to download them (e.g., over the air) and to ensure that each personality is safe before being activated. Evolution support therefore must include a software factory. In addition, the evolution of the radio platform — the analog and digital hardware of the radio node — must also be supported. This may be accomplished via the development of customized hardware/firmware modules, or by the acquisition of commercial off-the-shelf (COTS) modules, or both.

The block diagram of the radio functional model is a partitioning of the blackbox functions of the ideal SWR node into the functional components shown in Fig. 2, characterized further in Table 1.

Table 1. Function Allocation of the Software Radio Functional Model

Functional Component	Allocated Functions	Remarks
Source coding and decoding	Audio, data, video, and fax interfaces	Ubiquitous algorithms (e.g., ITU [27], ETSI [28])
Service and network support	Multiplexing; setup and control; data services; internetworking	Wireline and Internet standards, including mobility [29]
Information security[a]	Transmission security, authentication, nonrepudiation, privacy, data integrity	May be null, but is increasingly essential in wireless applications [30]
Channel coding/decoding: modem[a]	Baseband modem, timing recovery, equalization, channel waveforms, predistortion, black data processing	INFOSEC, modem, and IF interfaces are not yet well standardized
IF processing[a]	Beamforming, diversity combining, characterization of all IF channels	Innovative channel decoding for signal and QoS enhancement
RF access	Antenna, diversity, RF conversion	IF interfaces are not standardized
Channel set(s)	Simultaneity, multiband propagation, wireline interoperability	Automatically employ multiple channels or modes for managed QoS
Multiple personalities[a]	Multiband, multimode, agile services, interoperable with legacy[b] modes	Multiple *simultaneous* personalities may cause considerable RFI[c]
Evolution support[a]	Define & manage personalities	Local or network support
Joint control[a]	Joint source/channel coding, dynamic QoS vs. load control, processing resource management	Integrates user and network interfaces; multiuser; multiband; and multimode capabilities

[a]Interfaces to these functions have historically been internal to the radio, not plug-and-play.
[b]"Legacy" refers to modes that are deployed but may be deprecated.
[c]Radiofrequency interference.

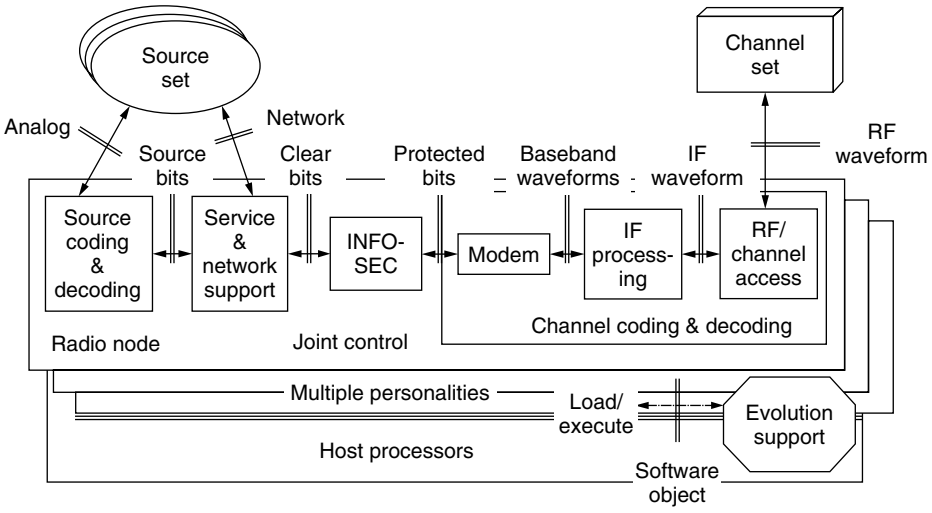


Figure 3. Standard interface points facilitate development, deployment, and evolution.

Not every implementation needs all sub-functions of this functional model. Thus, one may consider the functional model to be a point of departure for the tailoring of SDR implementations.

3.2. The Ideal Functional Interfaces

After identifying the functions to be accomplished in a software radio, one must define the interface points among the functional components. Figure 3 identifies these interfaces. The notation "RF waveform" includes spatial beamforming and the air interface. The IF waveform consists of signals filtered and converted to an intermediate carrier frequency that facilitates analog filtering, signal conditioning, and related analog processing.

In addition, IF processing may include A/D and D/A conversion. If so, some IF processing may be implemented digitally. Baseband waveforms are usually digital streams (e.g., of message packets or coded voice). These digital streams may also be sampled replicas of analog signals, such as digitized FM broadcast waveforms. The modem delivers what may be called *decoded channel bits* that may be encrypted ("black" bits in INFOSEC jargon) to the INFOSEC function if one is present. The modem transforms IF signals to channel bits. INFOSEC then transforms these protected bits into unencrypted ("clear")

bits (also called "red" bits). These bits may be manipulated through a protocol stack in order to yield source or network bit streams. Network bit streams conform to a network protocol, while source bits are appropriate for a source decoder. The interface to local sources of voice, music, video, and other media includes an analog transducer. Access to remote sources is accomplished via the network interface. In addition to these signal processing interfaces, there are control interfaces mediated by the user or network (both of which are source sets).

Personalities are downloaded to the radio via the software object interface. The simplest mechanism for maintaining radio software after deployment is the downloading of a complete binary image of the radio. A more flexible approach allows one to download a specific new function such as a specialized voice coder (vocoder). Such incremental downloads conserve network bandwidth at the expense of increased risk of configuration errors in the download process.

The interfaces thus defined may be thought of as the "horizontal" interfaces of the software radio, since they are concatenated to create the signal and control flows through functional components between sources and channels. These interfaces are further characterized in Table 2. Design-level interfaces ("design to") specify the data to be exchanged, while code-to interfaces directly or indirectly specify the exact format and meaning of each hardware signal, bit, word, byte, and message sequence of the interface.

In traditional radio engineering, these interfaces were addressed primarily in the design and development of the radio. For plug-and-play in software radio, they must be open architecture standards that facilitate the insertion of third-party components in deployment and operations. This is the business model that made the IBM PC a commercial success. How can such interfaces be standardized for industrywide third-party plug-and-play business to grow?

4. SOFTWARE-DEFINED RADIO (SDR) ARCHITECTURE

The *Random House Unabridged Dictionary* defines architecture as "a fundamental underlying design of computer hardware, software, or both" [38]. While this is an agreeable definition, it provides no prescription of what "underlying design" entails. The IEEE prescribes that architecture consists of components and interfaces. This leaves undefined what the components and interfaces are supposed to accomplish. The ideal software radio functional model and interfaces of the previous section begin to specify the functions of the architecture. The Defense Information Systems Agency (DISA) is the U.S. Department of Defense (DoD) agency charged with

Table 2. Top Level Component Interfaces

Interface	Characteristics	Properties
Analog stream	Audio, video, facsimile streams	Continuous, infinite dimensional; filtering constraints are imposed here
Source bit stream	Coded bit streams and packets; ADC, vocoder, text data compression [31]	Includes framing and data structures; finite arithmetic precision defines a coded, Nyquist [32] or oversampled dynamic range[a]
Clear bit stream	Framed, multiplexed, forward-error-controlled (FEC) bit streams and packets	FEC imparts algebraic properties over the Galois fields defined by these bit streams [33]
Protected bit stream	Random challenge, authentication responses; public key; enciphered bit streams [34] and packets	Finite-dimensional; randomized streams, complex message passing for downloads; if null, this interface reverts to clear bits
Baseband waveform	Discrete-time synchronous quantized sample streams (one per carrier)	Digital waveform properties determine fidelity of analytic representation of the signal
IF waveform	Composite, digitally preemphasized waveform ready for upconversion	Analog IF is continuous with infinite dimensions; digital IF may be oversampled
RF waveform	Power level, shape, adjacent channel interference, etc. are controlled	Analog RF: channel impulse response, spatial distributions via beams and smart antennas [35]
Network interface	Packaged bit streams may require asynchronous transfer mode (ATM), SS7, or ISO protocol stack processing	Synchronous digital hierarchy (SDH), ATM, and/or signaling system 7 (SS7)
Joint control	Control interfaces to all hardware and software; initialization; fault recovery	Loads binary images, instantiates waveforms, manipulates control parameters; cognitive joint control learns user needs [36]
Software objects	Download from evolution support systems	Represents binary images, applets; includes self-description of system capabilities [e.g., 37]
Load/execute	Software object encapsulation	Downloads require authentication and integrity

[a]A coded dynamic range is defined by the vocoder. Nyquist dynamic range results when an analog signal is sampled so as to meet the Nyquist criteria for bandwidth recovery of the sampled signal and has been quantized with sufficient bits of sufficient accuracy to represent the two-tone spurious-signal-free dynamic range of the application. Oversampling above the Nyquist rate can yield additional dynamic range through processing gain.

defining architecture. DISA defines architecture in terms of profiles for communications standards [39], defining architecture by analogy to "zoning laws and building codes" that constrain the construction of residential and industrial buildings [40].

4.1. Functions, Components, and Design Rules

None of the many possible definitions of architecture completely suits the architecture needs of the SDR community. One is needed that relates services, systems, technology, and economics. "Architecture" for SDR is therefore defined as a comprehensive, consistent set of *functions, components* and *design rules* according to which radio (and/or "wireless") communications systems may be cost-effectively organized, designed, constructed, deployed, operated, and *evolved over time*. This definition of architecture is consistent with the other definitions, but addresses more clearly the needs for plug-and-play and component reuse. By including functions and design rules, architecture supports component reuse, spanning component migration among hardware and software implementations.

The *design rules* must assure that when SDR hardware and software components are mated, the resulting composite entity accomplishes the intended functions within the performance bounds established by regulatory bodies, service providers, and users. The abstract functions and interfaces of the ideal SWR above constitute a horizontal architecture for software radio, an abstract functional flow from user to antenna and back. The ideal SWR does not specify the physical arrangement of physical components of an actual radio, of an SDR. Vertical levels are also needed to manage SDR hardware platforms and to achieve platform-independence of increasingly complex SDR software.

4.2. Plug-and-Play

In order for SDR architecture to support plug-and-play, design rules must be published that permit hardware and software from different suppliers to work together when plugged into an existing system. Hardware modules will plug-and-play if the physical interfaces and logical structure of the functions supplied by that module are compatible with the physical interfaces, allocation of functions, and related design rules of the host hardware platform. Software modules will plug-and-play if the individual modules and the SDR configuration of modules are computationally stable. For this, there must be a comprehensive interface to the host environment, and the module must describe itself to the host environment so the component can be managed as a radio resource. SDR architecture, then, defines the partitioning of functions at appropriate levels of abstraction so that software functions may be allocated to software components at appropriate levels of abstraction. SDR architecture defines the design rules, including design patterns [41,42] and interface standards. Rules defining logical levels of abstraction hide irrelevant details in the lower layers. These rules comprise a vertical architecture for SDR.

4.3. Vertical Architecture Design Rules

While horizontal architecture applies to any software radio, vertical architecture applies to radio implementations, SDRs. SDR components do not all share the same logical level of abstraction. A DSP module, for example, is part of the SDR platform. CORBA facilities are part of the software infrastructure on which ideal SWR functions are built using practical SDR software modules. In an advanced SDR, a modem may be a software module, a radio application. Bridging from one air interface to another is a service built on air interface radio applications. For example, in a military disaster-relief scenario, a SDR may bridge the Global System for Mobile communications (GSM) [43] to a military air interface like SINCGARS or HAVE QUICK [13]. One must therefore identify the SDR levels of abstraction that naturally partition the hardware and software into radio platforms, middleware,[3] radio applications, and communications services, as illustrated in Fig. 4.

In digital radios of the 1980s, the radio hardware platform ("radio platform") accomplished most of the RF and IF radio functions in hardware. The RF and IF parameters could be set through a microprocessor from a simple

[3] *Middleware* is software that insulates applications from the details of the operating environment (e.g., the hardware).

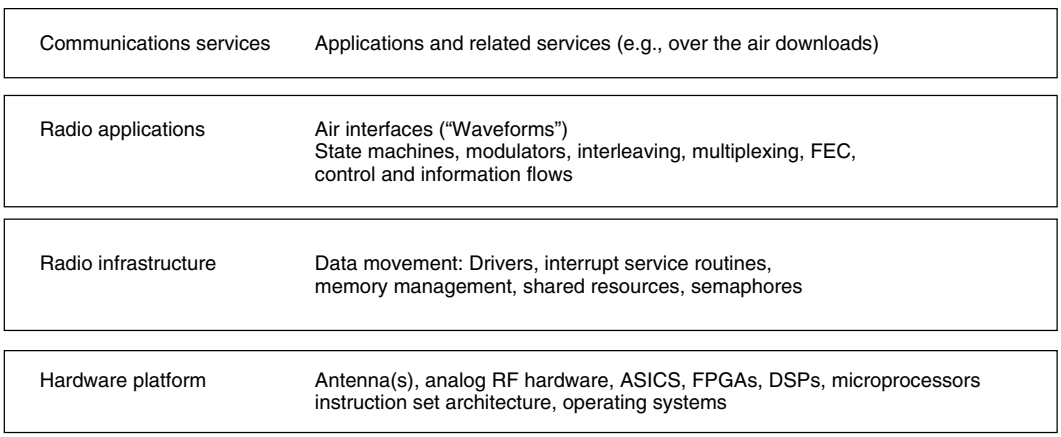


Figure 4. Logical levels of abstraction of the SDR implementations.

user interface or a low-speed data bus. Today's SDR platforms embody GFLOPS of processing capacity that host hundreds of thousands of LoC. This software performs the three top-layer functions of Fig. 4. At the infrastructure level, the code moves data among the distributed multiprocessors of the radio platform. At the applications level, software processes thus distributed cooperate to form radio applications, such as a 3G cellphone (cellular telephone) standard or a military waveform like HAVE QUICK. At the highest level of abstraction, applications software delivers communications services to users. Radio applications may incorporate specialized air interface protocols, and also may employ standard wireline data exchange protocols like TCP/IP.

One must define interfaces among these levels of abstraction, such as using an applications programming interface (API). APIs may map from one horizontal layer to the next. The API calls may be thought of as the vertical interfaces among horizontal layers. This approach has been used with reported success on SWR technology pathfinders [8]. The four layers of abstraction defined above are useful for defining software–software and software–hardware interfaces. Not all API's conform to these four layers. However, they are architecture anchors that help organize the process of evolving SDR implementations.

4.4. Mathematical Structure

Some mathematical principles illuminate the path toward SDR architecture. Some key principles are based on computability and point-set topology [44]. Consider transceiver state, consisting of a set of labels such as "Idle," "Synchronizing," "Receive," "Transmit," and "CarrierFault," asserted by other algorithms such as SquelchDetection (e.g., no squelch means the channel is Idle). SquelchDetection has possible wait states that have topological structure [45]. There is a set (e.g., of state labels, process names) with a family of subsets (e.g., the ones over which the software operations are valid), which has [or fails to have] topological properties. In addition, SDR architecture regarded as a collection of SDR implementations ("instances") has topological structure [46]. Maps over topological spaces define critical mathematical properties such as SDR module composability per the gluing lemma [47]. If the SDR has strong topological structure, then the insertion of plug-and-play components preserves the composability of software modules, including isochronism and controllability [48]. If not, then those critical properties cannot be guaranteed. In the absence of mathematical properties, one must test all possible configurations of SDR modules, a computationally intractable task for merely a few dozen downloadable SDR modules. Industry standards facilitate the application of such mathematical principles to broad classes of SDR.

4.5. Industry-Standard SDR Architectures

An important evolutionary step in the definition of vertical SDR interfaces is use of middleware (e.g., CORBA [49]) in SDR architecture. The Object Management Group (OMG) has defined an Interface Definition Language (IDL) in their Common Object Request Broker Architecture (CORBA). CORBA was developed primarily to define interfaces among software modules that were not originally designed to work together. IDL provides facilities for defining interfaces among software components through the mediation of an Object Request Broker (ORB). Since each new component implements only one interface to the ORB rather than N interfaces to the existing components, the process of integrating a new software component is greatly simplified.

The JTRS JPO began the development of its CORBA-based Software Communications Architecture (SCA) [50] in 1997. Version 2.2 (Fig. 5) includes the architecture specification with supplements on military security,

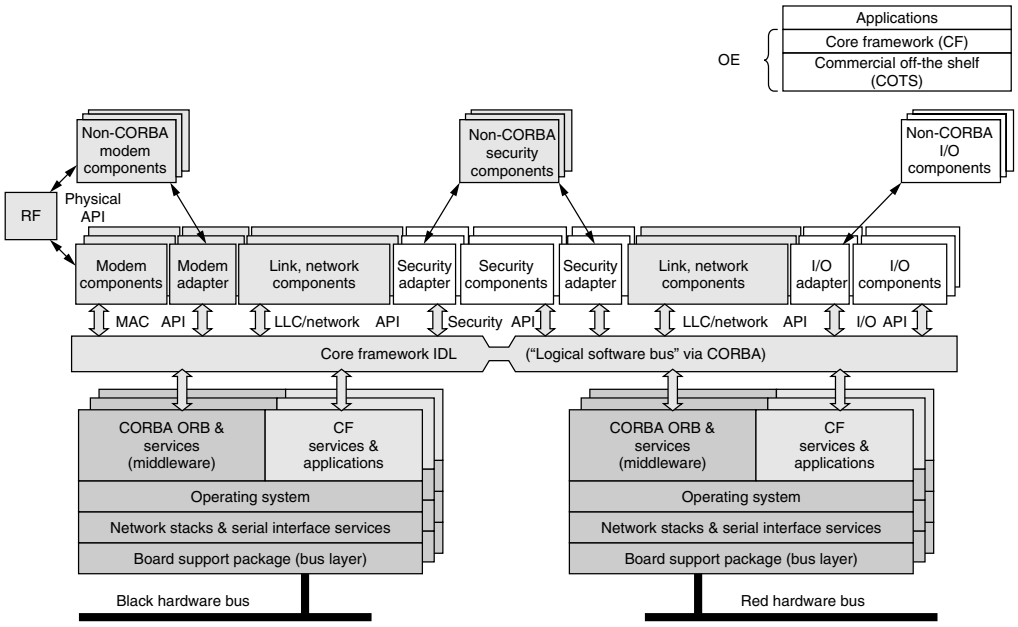


Figure 5. SCA Version 2.2.

APIs, and rationale. The framework applies UML to define hardware devices, software objects, and related interface rules. Its operating environment includes a core framework (CF) consisting of

1. *Base application interfaces* used by all software applications (Port, LifeCycle, TestableObject, PropertySet, PortSupplier, ResourceFactory, and Resource)
2. *Framework control interfaces*, used to control the system (Application, ApplicationFactory, Domain Manager, Device, LoadableDevice, ExecutableDevice, AggregateDevice and DeviceManager)
3. *Framework services interfaces* that support both core-compatible and non-core-compatible (vendor-unique and hardware-defined) applications (File, FileSystem, FileManager, and Timer)
4. A *domain profile* that describes the properties of hardware devices (DeviceProfile) and software components (SoftwareProfile) of the radio system using XML

Since this architecture does not yet map radio communications functions (e.g., 3G or military standards like HAVE QUICK) to its functional model, it provides a framework, an important first step toward plug-and-play architecture. In addition, its code-to level of specification is limited to the XML descriptions of interfaces. Therefore, different "fully compliant" implementations of a given air interface from two different vendors do not necessarily interoperate. For example, one implementation might specify RF in kilohertz while the other specifies it in Hz. With no units-consistency checking or remapping, the different software components would not use the facilities of the RF platform consistently. The government, academic, and industry bodies developing this standard plan to continue to evolve it towards an open-architecture plug-and-play standard.

5. SDR DESIGNS

Design-to and code-to architecture design rules assure that the critical properties of software radios are met as plug-and-play components are configured. Among these is the computational stability of the integrated software, a mathematical property emphasized above. Next is isochronism, the sufficiently precise timing of the real-time signal processing streams. Consider first the signal streams of an ideal SDR, illustrated in Fig. 6. These include a real-time isochronous channel-processing stream, a near-real-time environment management stream, an online control stream to manage the radio's configuration and modes of operation, and radio personalities from offline evolutionary development.

5.1. Real-Time Channel Processing Streams

In practical SDR designs, the real-time channel processing stream is a signal structure within the channel coding/decoding function. In an aggressive SDR design, each broadband channel (e.g., a cellular band) is accessed via a wideband ADC and DAC. This aspect of SDR design merits particular attention. The real-time stream generates subscriber channels in the transmit-path and isolates them in the receive path, such as by

1. Filtering of frequency-division multiple-access (FDMA) [51] waveforms
2. Timing recovery of time-division multiple-access (TDMA) [52] waveforms
3. Despreading military spread-spectrum [53] or commercial code-division multiple-access (CDMA) [54] waveforms

Historically, subscriber channel isolation was allocated to analog IF processing (e.g., in first-generation FDMA cellphones) or ASIC hardware (e.g., a CDMA cellphone).

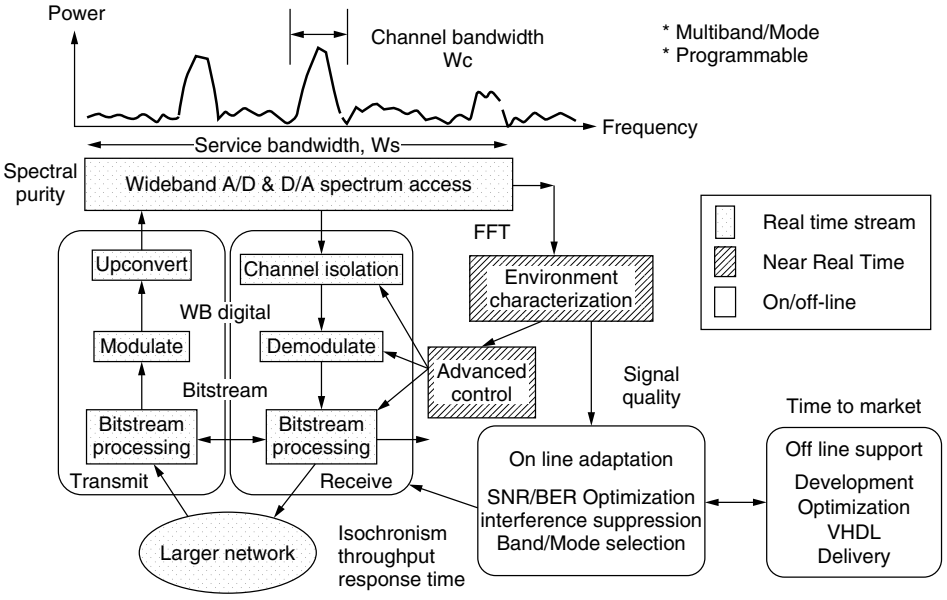


Figure 6. Signal processing streams of software radio.

When implemented in software, the isochronous processing windows for these functions are on the order of the time between DAC/ADC samples: microseconds to tens of nanoseconds. Modulation and demodulation of the channel waveform are also accomplished in the real-time channel-processing stream. Isochronous windows for these functions are on the order of tens to hundreds of microseconds. Once the narrowband subscriber bit streams are recovered, the timing of the isochronous windows increases to milliseconds. INFOSEC encryption and decryption, if applicable, are performed on the subscriber bit streams, with appropriately long isochronous windows — tens to hundreds of milliseconds. For baseband DSP, the time between digital samples (e.g., for baseband voice) is on the order of milliseconds to hundreds of microseconds. This allows plenty of time for processing between samples. In the software radio's IF stream, however, the time between samples is from tens of microseconds to hundreds of nanoseconds. Such point operations require Mflops to Gflops for isochronous performance. For isochronous performance, sampled data values must be computationally produced and consumed within short-buffer timing windows in order to maintain the integrity of the digital signal representation. Subscriber channels may be organized in parallel, resulting in a multiple-instruction multiple-datastream (MIMD) multiprocessing architecture [55]. Input/output (I/O) data rates of this stream approach 200 MB/s (megabytes per second) per IF ADC or DAC. Although these data rates are decimated through processing, to sustain isochronism blocks and events must be timed through I/O interfaces, FPGAS/ASICS and hard real-time embedded software in these streams.

To implement the approaches described above in a practical SDR design requires the integration of the real-time channel processing streams with related radio functions such as local oscillator (LO) signal generation. Figure 7 shows these radio signal flows structured into RF, IF, baseband, bit stream, and source segments, each with order-of-magnitude differences in isochronous windows. This view clarifies the sharing of the power management and low-noise amplifier (LNA) elements with RF conversion and with the RF frequency standard. These RF elements typically need physical proximity to the antenna. The LNA is placed near the antenna in order to set the system sensitivity. The power amplifier is near the antenna for power efficiency. The RF section may be remote from IF processing, such as in diversity architectures.

Digital IF processing in an SDR filters the wideband signal structure from the RF segment to yield the narrower baseband bandwidth. SDR ADCs appear at the IF–RF or RF–antenna interface. The baseband segment performs the modem functions, converting information between channel code and source code. The bit stream segment performs operations on bit streams, including multiplexing, demultiplexing, interleaving, framing, bit stuffing, protocol stack operations, and FEC. SDR system control is included in the bit stream segment because of the digital nature of control messages. The source segment includes the user, the local source and sink of information, and control. Source coding is the transformation of communications signals into bit streams if the design conforms to the ideal SWR functional partitioning described above. The organization of the design of SDR nodes into RF, IF, baseband, bit stream, and source segments promotes the application of a given talent pool and isochronous design discipline within a segment and minimizes the interdependencies between segments.

5.2. The Environment Management Stream

The other shaded boxes in Fig. 6 constitute the near-real-time environment management stream. In an ideal SWR, this stream continuously manages radio environment usage in frequency, time, and space. In a practical SDR design, the message exchanges with the host network are typically defined in specific signaling and multiple-access protocols. These traditionally include the assignment of traffic to clear channels, and the handoff of a mobile subscriber from one cell to the next. This may further include channel identification, equalization, and the estimation of parameters such as multipath time delays and cochannel interference levels. For example, the HF Automatic Link Establishment (ALE) protocol

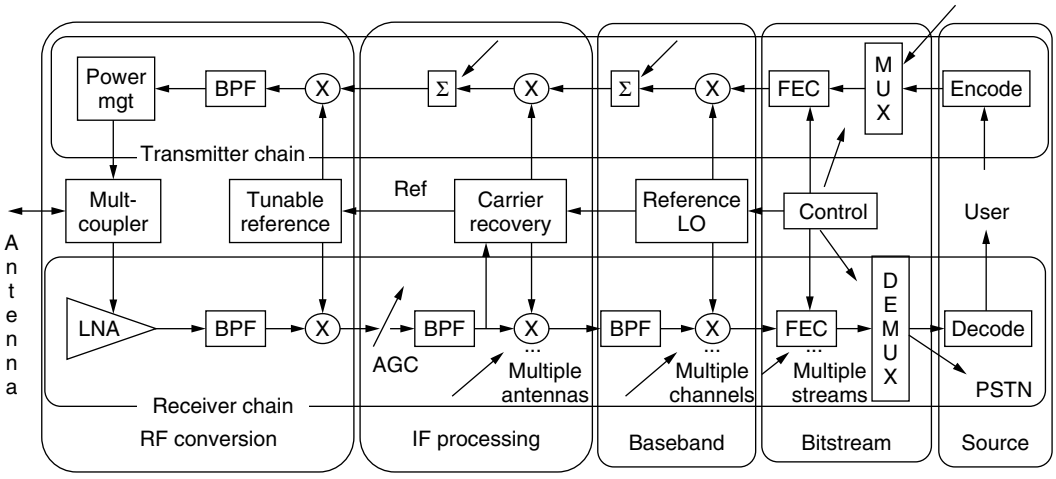


Figure 7. SDR isochronous and interdependent signal flows.

includes probes and responses that characterize several assigned channels [56]. The ALE data are then sent on the channel that is best for the specific subscriber location. In orthogonal FDM (OFDM), narrowband channels with excessive interference may be dynamically deallocated from use [57]. The environment management stream may employ block operations such as matrix multiplications for smart-antenna beamforming. Real-time adaptation by smart antennas may respond to signal parameters computed on every TDMA burst or CDMA symbol. GSM, for example, requires channel identification within 540 μs–2 ms. This establishes top-down limits on the execution time of environment management software. Forgiving operations, such as power-level updates may be refreshed every few frames. Location-aware services (e.g., emergency 911 cellphone location) typically define timing requirements for subscriber emitter location.

5.3. Online Adaptation: Mode Selection and Download Management

Online adaptation includes mode selection, as suggested in Fig. 6, as well as download management. A practical SDR is a multiband/multimode radio. The graceful transition from one RF band or air interface mode to another is called "mode handover." Military radios may change modes as a function of information priority, RF propagation, and other military criteria. An air interface mode typically defines the QoS provided by that mode. 3G air interfaces offer a wide range of data rates. Generally, high data rates require high signal-to-noise ratio (SNR[4]) for a required bit error rate (BER). Online adaptation selects the appropriate air interface mode to satisfy the competing goals of the user and/or of the network. Because of the number and complexity of 3G modes, ITU standards define mode handover in detail.

As modes become more elaborate, users are confronted with an increasing array of choices of QoS versus price. The burden of choosing RF band and mode in the future will be shared among the user, the network, and intelligent wireless appliances [e.g., personal digital assistant (PDA)], SDRs that employ natural-language processing and machine learning to assist the user with mode selection are called "cognitive radios" [18]. Because of Moore's law, cognitive wirelesses PDAs are likely to emerge soon, along with cognitive networks [58]. Cognitive PDAs and networks provide interesting cross-discipline research opportunities, fostering collaboration across natural-language processing, cognitive science, and radio engineering. In the past, mode control was primarily up to the network. As wireless LANs, home wireless networks, and intelligent transportation systems converge with cellular technology, cognitive PDAs will shape offered demand, for example, by delaying a large email attachment until the connection is free.

SDR personality management by over-the-network download also adapts the behavior of practical SDRs. Prior to SDR, the flexibility of RF access of a handheld wireless device was limited to merely choosing one of several predefined air interface modes. With SDR, parameters of the predefined modes may be modified along with higher-level software functions such as the user interface, network applications and air interface protocols.

[4] The SNR may be expressed in terms of unmodulated carrier and interference (CIR), or signal to interference plus noise (SINR).

5.4. Offline Adaptation: The Software Factory

Offline SDR development environments define SWR personalities. Offline functions include radio systems analysis, enhanced algorithms, hardware platform creation, and the rehosting of existing software to new hardware/software platforms. These functions assist in defining incremental service enhancements. For example, an enhanced beamformer, equalizer, and trellis decoder may be developed to increase subscriber density. These enhancements may be prototyped and linked into the channel processing stream in a research, testbed, or evaluation facility [59]. Such an arrangement allows one to debug the algorithm(s) and to experiment with parameter settings prior to deploying new personalities. One may determine the value of the new feature (e.g., in terms of improved subscriber density), as well as its cost. Network traffic to download such features constitutes overhead. Offline adaptation thus includes the definition of personalities [60], download protocols [61], and download traffic [62].

5.5. Software Tools

An advanced SDR does not just transmit a waveform. It characterizes the available transmission channels, probes the available propagation paths, and constructs an appropriate channel waveform. It may also electronically steer its transmit beam in the right direction, select the appropriate power level and pick an appropriate data rate before transmitting. Again, an advanced SDR does not just receive an incoming signal. It characterizes the energy distribution in the channel and in adjacent channels, it recognizes the mode of the incoming transmission, and it creates an appropriate processing stream. With a smart antenna, it also adaptively nulls interfering signals, estimates the dynamics of the multipath, coherently combines desired-signal multipath, and adaptively equalizes this ensemble. It may also trellis decode the channel modulation and then correct residual errors via FEC decoding to receive the signal with the lowest possible bit error rate (BER). Such operations require a family of software components and related tools, including those illustrated in Fig. 8.

Figure 8 organizes software tools according to the time-criticality of the supported software functions. In an SDR, hard real-time software may be delivered as the personality of an ASIC or FPGA. Reduced time-criticality means the function is more compatible with software implementation. The tradeoff among ASIC, FPGA, and software changes with each 18-month Moore's law cycle. The columns labeled C (criticality) and A (availability) identify SDR challenge areas. Bit interleaving, for example, is not challenging in terms of either its criticality to SDR architecture or its availability as a software component. Interference suppression, on the other hand,

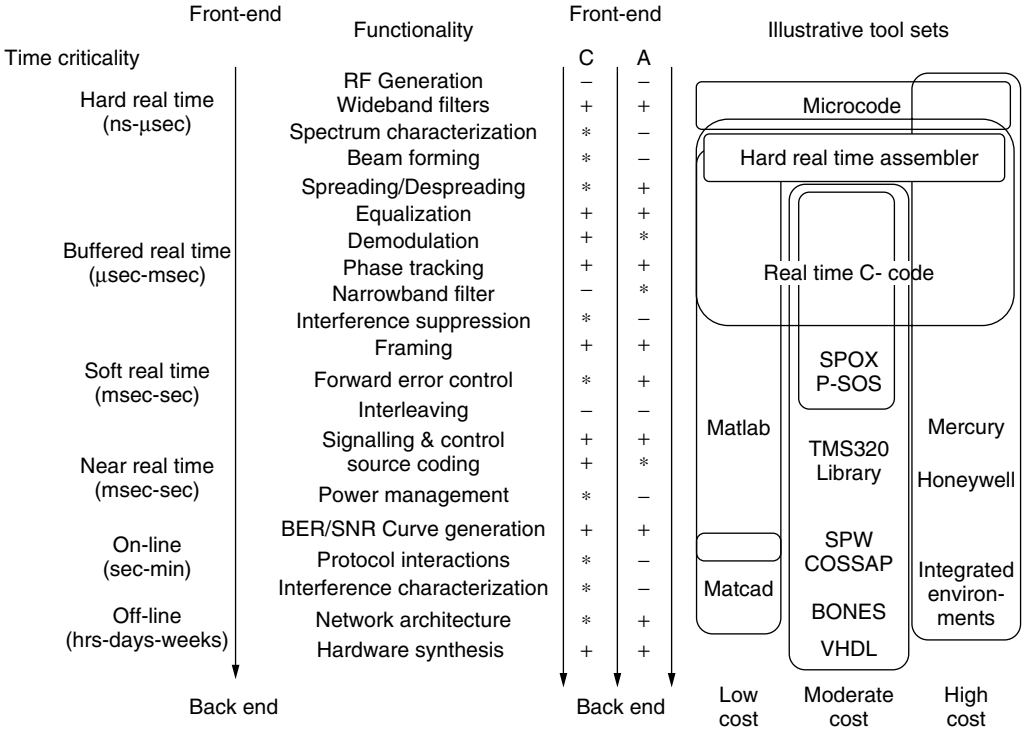


Figure 8. Toolsets of a software factory. (C = criticality; A = availability; $*$ = key performance driver; $+$ = important issue).

is a critical issue differentiating SDR from conventional radios; it also can be a performance driver. To the right are three columns of toolsets that represent the sophistication of the software factory. One may develop software radio products of limited scope [e.g., <40 kloc 40,000 lines of code] using the low-cost tools in the first column. As team size grows, or the mix of ASICs, FPGAs, and DSP hardware in the delivery environment becomes more complex, the investment of tens of thousands of dollars (per design seat) pays off. The largest, most complex systems benefit from the high-cost tool suites costing millions of dollars per system.

5.6. SDR Technology Alternatives

Technology alternatives for digital radios, SDR, and software radios are characterized in the software radio parameter space of Fig. 9. The parameter space compares two critical SDR technology parameters: digital access bandwidth and the flexibility of the processing platform. Digital access bandwidth is approximately half of the sampling rate of the widest bandwidth ADC in the isochronous signal-processing path. Thus, for example, an ideal SDR with 5 GHz conversion rate supports nominally a 2.5-GHz analog bandwidth, based on the Nyquist criterion [32]. Similarly, wideband digital-signal synthesis, digital upconversion, and wideband DAC yield an ideal software radio transmitter.

ADCs with continuous conversion bandwidths of >6 GHz have been built [63], although they are expensive. If all the processing after the ADC were accomplished on a single general-purpose computer, one would have an ideal software radio receiver (the point marked X in the figure). Using a rule of thumb of 100 operations per sample, the digital filtering of a 5-GS/s (gigasamples per second) stream to access a 25-MHz band of RF spectrum requires 500 gigamultiplications (5×10^{11}) per second. This processing capacity is about two orders of magnitude beyond 2002-generation DSPs [64,65] and three or four beyond general-purpose computers. This translates to about 6–10 Moore's law cycles or only 10–15 years of continued exponential development of DSP technology and an additional 5 years beyond that of general purpose computing technology.

Another limitation of the ideal SWR is that no single antenna nor RF stage can sustain the analog bandwidth from 2 MHz to 2.5 GHz RF with reasonable losses or power efficiency. The single wideband RF required for the 5 GHz ADC (and for the transmitter/DAC) is therefore not feasible. Antenna and RF stages depend on properties of materials that have stubbornly resisted pushing bandwidths beyond one RF decade, a 10 : 1 ratio of high to low RF. Thus, the ideal SWR is not possible with today's technology. The ideal properties of such a radio are a useful reference point for measuring progress towards generality and flexibility.

Practical SDR implementations limit RF coverage to medium or narrowband antennas, RF conversion, and IF processing technology. They also use a mix of digital technologies including ASICs, FPGAs, DSP, and general-purpose processors. Examples are illustrated in the figure. The STR-2000 (point *A* in Fig. 9) was an early baseband HF DSP radio developed by Standard Marine AB. This radio digitized its HF IF signal at a 24 kHz sampling

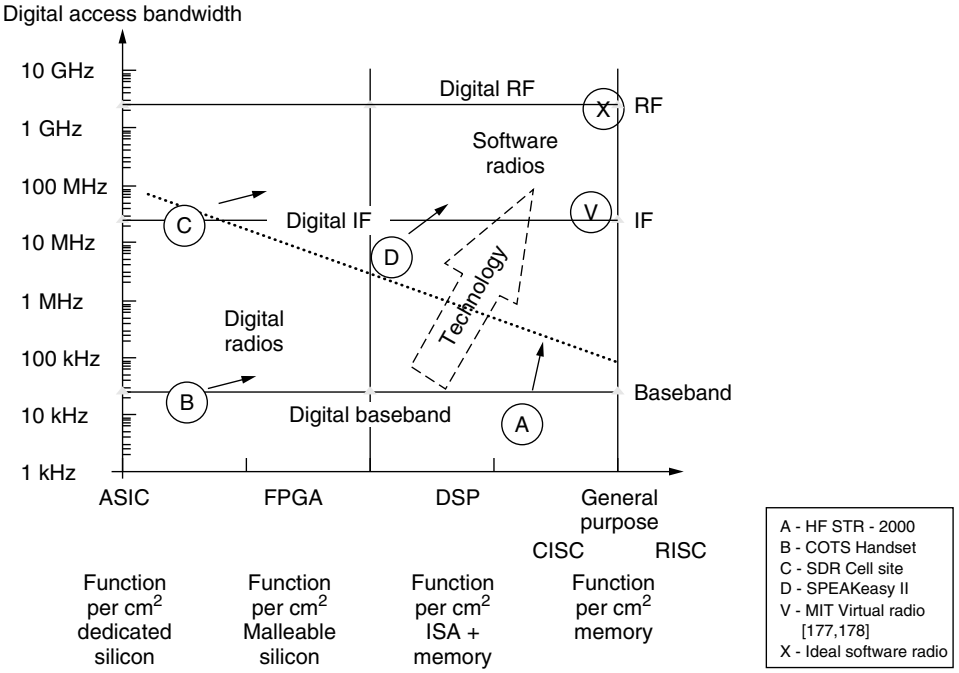


Figure 9. Software radio parameter space.

rate. It used twin Texas Instruments (TI) TMS320C30 DSPs to provide a half-dozen standard HF signal formats digitally. This is readily accomplished by amateur radio operators using general-purpose processors today [66,67]. Second-generation (e.g., GSM) COTS SDR handsets (point *B* in Fig. 9) minimize size, weight, and power using a direct-conversion receiver [68] RF-ASIC [69], and baseband DSP. Combining such ASICs in a handset enclosure for a dual-mode cellphone creates a "Velcro radio" [70]. Aggressive implementations incorporate high-density FPGAs to provide software-driven configurability in a delivery platform that maximizes throughput for a given technology clock rate [71]. 3G CDMA researchers employ FPGAs to replace despreader ASICs [72]. These FPGAs have greater speed and power efficiency for a given clock rate than do DSPs. In addition, the personality of an FPGA can be upgraded in the field by software download. On the other hand, FPGAs lack the silicon area and computational flexibility of the DSP instruction set architecture (ISA). With increasing chip density and area, system-on-chip (SoC) architectures of the near future are likely to include both fixed ISA and variable-personality FPGA coprocessors.

Contemporary software radio cell-site designs (point *C* in Fig. 9) access the allocated unlink[5] RF using a single ADC, such as with 25 MHz of analog bandwidth (viz., 70 MHz conversion rate). These designs employ a bank of digital filter ASICs [73] or parallel digital filters [74] to access a hundred or more subscriber channels in parallel. Research radios like the European ACTS Flexible Interoperable Radio System Technology (FIRST) used Pentek boards [75] with Harris parallel digital filter ASICs. The larger commercial cellular infrastructure suppliers include Alcatel, Ericsson, Fujitsu, Lucent, Motorola, Nippon Electric Corp (NEC), Nokia, NorTel, Siemens, and Toshiba. Although all contribute to open research, few publish the details of their commercial SDR handset or infrastructure products. Research supported by one or more of these industry leaders, explores ASIC [76], DSP [77], and FPGA [78] cell sites [79,80], some with smart antennas [81–82,83]. In addition, emerging products now include ADCs with 200 MHz of bandwidth with >80 dB dynamic range with integrated digital IF processing [84].

Technologically aggressive designs include SPEAK*easy*, the military technology pathfinder. SPEAK*easy* II (point *D* in Fig. 9), which became the baseline for Motorola's WITS 6000 software radio product line [85], incorporated over a Gflops of processing capacity for enhanced flexibility, substantial DSP in 1996–1998. The virtual radio (point *V* in Fig. 9) is the most flexible software radio research implementation reported in the literature [86]. A general purpose DEC Alpha processor running UNIX accesses a wideband IF digitally. Narrowband AM and FM broadcast receivers and an RF LAN were implemented purely in software on this platform. The related SpectrumWare software technology is being commercialized for military and commercial applications [87].

None of the designs *A*–*X* in Fig. 9 is a panacea: the architecture question is the degree of digital RF access and programmability required for the intended market. Contemporary radio designs therefore vary across the dotted line in the phase space. Advancing microelectronics technology moves all implementations inexorably upward and to the right over time. The three fundamental waveform limitations of any SDR implementation, then, are RF access, digital access bandwidth, and digital processing flexibility and capacity.

[5] The *uplink* is the link from mobile to base station. The *downlink* is the reverse link.

5.7. SDR Radio Reference Platform

The definition and use of a radio platform facilitates the evolution of SDR implementations through generations of hardware and software releases. It also enhances the use of UML, CASE tools, and middleware. A radio reference platform is a high-level characterization of the capabilities of the hardware environment of the software radio. Table 3 identifies the critical radio platform parameters that determine the performance of a software radio.

The parameters of Table 3 should be specified with precision. If the platforms in the family are tested for conformance to a well-specified reference platform, then software developed for one member of the family should port readily to another member of the family. The software will not port well (and may not port at all) if special features of the platform beyond the reference set are used. The specification of a minimum level of capability for each parameter defines a reference platform for a family of software radio implementations. Illustrative platforms are suggested in Table 4. The PDAs will have replaced conventional cell phones in this vision of the future. Given the reference platforms, they will have broadband RF, multiple parallel data channels, and wide digital processing bandwidth (BMW).

Devices now in development, mostly in proprietary settings promise to bring such platforms to market in the 2002–2007 timeframe. Such reference platforms closer to the ideal software radio are beginning to make economic and technical sense in infrastructure applications. A reference platform need not have an associated block diagram, but it is often convenient to use such a diagram in the analysis of the feasibility of a reference model.

The reference design of Fig. 10, illustrates the value of a reference platform, but has the following drawbacks. First, it implies that ADCs and DACs are the interface between the digital processing and analog RF sections of the radio. That is often the case, but an ultrawideband (UWB) [88] communications system, for example, uses subnanosecond pulses to spread the communications over 2 GHz or more of bandwidth. These pulses are both transmitted and received with analog circuits, not with DACs and ADCs. The primary value is to associate critical parameters with physical devices in such a way that one may outline an evolutionary path for software radio architecture.

6. DEVELOPMENT PARAMETERS AND RISKS

In 1992 when Mitola [89] introduced the term, almost nobody knew what a "software radio" was. By 1996, 6 months after the publication of the special issue of the *IEEE Communications Magazine* on the software radio, almost every radio vendor claimed to have one. The term had become an industry "buzzword." By 1999, it had become widely understood that nobody even *wanted* to offer an ideal software radio product because one would be unaffordable or inefficient or both. Thus, in 1996, the acronym SDR was introduced as the family affordable, practical implementations of software radio [90]. Today,

Table 3. Software Radio Reference Platform Parameters

Critical Parameter	Remarks
Number of channels	Number of parallel RF, IF, and/or baseband channels
RF access	Continuous coverage from a minimum to a maximum RF
Digital bandwidth	Bandwidth of the maximum ADC for each RF/IF channel
Dynamic range	End to end, including RF, IF, ADC, and processing gain
Interconnect bandwidth	Bandwidth of critical buses, serial ports, backplanes, etc.
Timing accuracy	The precision and stability of system clock(s)
Frequency performance	RF, IF, and local oscillator (LO) accuracy and stability
Processing capacity	Mips, Mflops using standard benchmarks, arithmetic precision (per processor class if appropriate)
Memory capacity	RAM, ROM per processor; mass storage capacity
Hardware acceleration	Parameterize capabilities encapsulated in hardware such as despreader ASICS, FPGAs, and related hybrids.
Operating environment	Operating system and related facilities (including CORBA middleware), interfaces (e.g., APIs), and measured determinism

Table 4. Illustrative Mobile SDR Reference Platforms

Notional Platform	RF Access (MHz)	Channels	Digital Bandwidth (MHz)
Lowband PDA	450–1200	3 (traffic, control, rental)	5
Midband PDA	850–2500	3 (traffic, control, rental)	20
Lowband military	30–500	4 (voice, 2 data, 1 scan)	10
Midband military	88–1200	4 (voice, 2 data, 1 scan)	20
Wideband military	800–4000	6 (4 JTIDS, 1 voice, 1 scan)	250

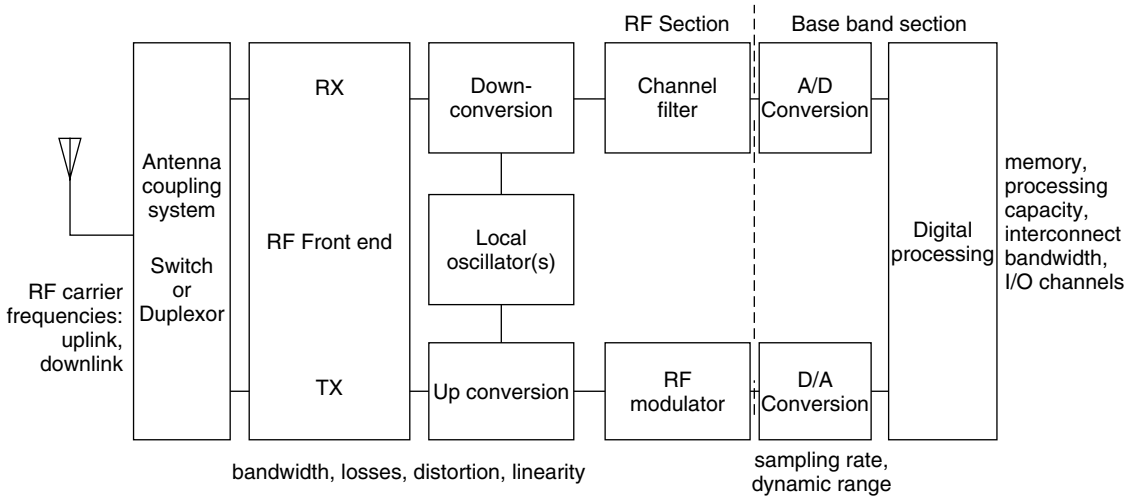


Figure 10. Reference design for an SDR implementation.

SDR is regarded as a ubiquitous technology that is key to the affordable evolution of military and commercial wireless markets. The four key characteristics of SDR from a development or acquisition perspective are

1. The number of air interface channels simultaneously supported (N)
2. The level of programmable digital access (PDA)
3. The degree of hardware modularity (HM)
4. The scope of software flexibility and affordability (SFA)

The number N defines hardware risk. There is low risk with single channel SDRs. With multiple channel (i.e., $N<6$), and full channel access (i.e., N is the full number of subscribers in an allocated RF band), risks increase. Two to four channel nodes are typical of military, civil aviation, and law enforcement hub applications. The full access class is typical of cellular base-station infrastructure. Single-channel SDR provides a baseline of minimum development risk and complexity. Multiple channel nodes require distributed multiprocessing. With small numbers of channels, hardware efficiency per channel is not a major challenge. It becomes a market-discriminator for the full access class, however. This class also carries maximum risk of mismatch between the processing demand offered by the software and the processing capacity deliverable by the hardware. Matching demand to capacity is therefore an essential design issue for practical SDR implementations.

The level of PDA is the point in the software radio functional model at which the conversion to digital occurs. This defines the scope over which the radio's functions are programmable with substantial flexibility. The types of PDA include baseband programmability, IF programmability, and RF programmability.

HM identifies the economic impact of the differences in hardware upgrade paths. Architecture may be based on capability-oriented coarse-grain (possibly programmable) modules such as receivers and exciters that are specific to an air interface. Alternatively, architecture may be based on technology-oriented coarse-grain modules such as COTS ADC and DSP boards. Finer grain modules such as FPGA, ADC and DSP chips are also candidate modules. Finally, the system-on-a-chip approach defines module as a chunk of intellectual property (IP). The granularity of hardware modularity is not prejudicial, but should be determined by the needs of the market segment.

SFA characterizes the service provider's ability to acquire plug-and-play software modules that are driven by a vital, multisupplier marketplace. Software that runs on just one radio platform and is available from only the original manufacturer tends to box the service provider into single-source (sometimes very expensive) maintenance and upgrade paths. If the functionality of the unit will not change over its lifecycle, then this may be a perfectly acceptable path. This would be a rare occurrence in today's fast-moving marketplaces, however. Software that runs on many platforms (e.g., Java) and is available from multiple vendors generally gives the service provider a better software product with more flexibility and at a lower cost over the lifecycle than the alternatives.

7. BROADER IMPLICATIONS

The prospect of a new technology of multiband, multimode software radios — handsets and infrastructure — has social and political implications. Type certification authorities, for example, are charged with administering the equitable use of radio spectrum. Among other things, they certify that radio equipment meets legally imposed constraints. In addition, software radios may operate on any RF band that is within the capabilities of the underlying radio platform, and with any mode for which a software load-image is available. This raises the possibility of truly novel approaches to spectrum management. One of the more interesting is the possibility that software radios could use a spectrum rental protocol to autonomously share spectrum. Another is that by incorporating advanced agent technology, they could evolve their own protocols. As mentioned previously, radios capable of such behavior are called "cognitive radios" [18].

7.1. Type Certification

The prospect of an evolving radio platform raises substantial questions among regulatory bodies about type certification. In remarks before the SDR Forum, the U.S. FCC [91] described type certification of software radios as presenting "regulatory issues." These include the following:

1. To which service(s) is an SDR approved?
2. Is a new approval needed for each "change" to an approved SDR unit?
3. How does the FCC enforce the equipment authorization rules for SDRs?
4. How can an unauthorized use of an SDR be prevented?

Regulators rely on a mix of tactics to achieve their goals. Industry is required to obtain licenses for some uses of spectrum, while others are available without a license, provided the manufacturer complies with the regulations. The FCC relies on legal remedies to motivate manufacturers to comply with the rules. They generally specify license requirements in terms of RF power output, modulation, occupied bandwidth, spurious emissions, and frequency stability (over temperature and voltage supply variations). Analog radios embody these parameters in hardware, so the type certification process has historically focused on the certification of devices. Digital radios, similarly, embody these parameters in a mix of analog and digital hardware, so the process remains valid. Current-generation SDRs with baseband programmable digital access embody these parameters in relatively fixed core images that are tightly coupled to the hardware, and this is compatible with the current process as well.

However, SDRs with at IF programmability embody these parameters in software that is loosely coupled to the hardware. Each combination of band and mode has to be certified separately, according to today's process. Over the air downloads to the SDR complicate the certification process substantially. At present, regulators in the United States are in the process of obtaining the advice of industry through an expected request for comments on proposed rule-making. Industry has the challenge of assisting regulators in defining a certification process that is responsive to the broader social and legal issues, but that does not seriously impede the benefits of SDR technology. Open architecture in some ways exacerbates the certification challenges. A proliferation of software packages enabled by open architecture drives the combinatorial complexity of type certification. Must a service provider certify every possible combination of software modules from every possible vendor? A helpful architecture might have properties that simplify and expedite type certification.

7.2. Incremental Download Stability and Type Certification

In addition to defining a partitioning, an architecture may define principles that assure that plug-and-play with desired properties of controllability and reliability. For example, to type-certify an open-architecture SDR, one must guarantee that the properties specified by the regulatory bodies will be preserved *in spite of the software radio's high degree of flexibility*. The need for such guarantees motivates the study of the mathematical properties of the software radio [92]. For example, one may model the statistical demand for computational resources versus processing capacity using queuing theory [93,94]. Real-time performance can be assured in a fixed architecture using this approach.

The plug-and-play SDR, however, has a *variable architecture* as modules are introduced into the environment and removed. This raises the complexity of the statistics, particularly in complex nodes. In a future 3G cell site, for example, hundreds of users can invoke dozens of variable-bandwidth services via a pool of shared DSP resources. To make this tractable, there should be a predictable relationship of computational demand between plug-and-play software modules and the host processor environment. This calls for a theory of plug-and-play resource bounds for the software radio within which such predictable relationships will exist. The fact that radio software must run to complete in a short, finite time period that can be specified in advance leads to a proof that radio software need not be Turing-computable [92]. The theory translates into a prohibition on unconstrained "while" and "until" loops. These have to be replaced by bounded-while and bounded-until loops that are allowed to run at most n times before generating a protection fault. The related theory of bounded recursion shows how a compiler can calculate n for the programmer so there is no additional programming burden to obtain this protection. Without such protection, while loops may run forever, consuming unacceptable amounts of time and processing power.

This theoretical advance makes it possible for one to provide a software engineering environment that can place tight upper bounds on the computational resources of an arbitrary radio software module. One may therefore prove by induction that a bounded recursive downloaded module will consume resources that are within tightly specified a-priori limits when loaded into a bounded recursive system. This can reduce the combinatorial complexity of the type certification of incremental software downloads. Given, for example, M vocoders and N air interfaces, a bounded recursive software system need test only $M + N$ software configurations, proving the other $MN - (M + N)$ configurations by induction. This supports the incremental download of the M vocoders, reducing download bandwidth on the network. Conventional software has to test all MN integrated load images. Furthermore, a change of vocoder requires the download of a complete load image, with increased network overhead. This article therefore sets forth the technical issues that underlie this tradeoff between network overhead and download complexity.

7.3. Spectrum Management Implications

Given that SDRs will continue to become more capable, one can ask whether they might have some fundamental impact on our approach to the use of the radio spectrum. A new research area, cognitive radio, suggests that this might indeed be the case [18]. Wireless multimedia applications require significant bandwidth, some of which

will be provided by third-generation 3G services. Even with substantial investment in 3G infrastructure, the radio spectrum allocated to 3G will be limited. Cognitive radio is a particular extension of software radio that employs model-based reasoning about users, multimedia content, and communications context. Cognitive radio offers a mechanism for the flexible pooling of radio spectrum using a new class of protocols called "formal radio etiquette." This approach could expand the bandwidth available for conventional uses (e.g., police, fire, and rescue) and extend the spatial coverage of 3G in a novel way. This section characterizes the potential contributions of cognitive radio to spectrum pooling and outlines an initial framework for formal radio etiquette protocols.

Figure 11 illustrates important aspects of spectrum allocation.

Bandwidth that could be made available for the sharing of spectrum, based on current allocations to mobile users, is summarized in Table 5.

The literature describes a protocol for spectrum rental among cognitive radios and infrastructure [95]. The effective use of this new protocol requires software radios that always know where they are (e.g., in latitude, longitude, and altitude above mean sea level), and which embed propagation models that include terrain and buildings. In addition, they must know what their users are doing (e.g., shopping, which is a low precedence use, or in need of emergency assistance, which is a high precedence use). Cognitive radios accomplish this by parsing all incoming and outgoing messages and voice traffic, and analyzing this information to establish the user's priority for use of spectrum. In addition, cognitive infrastructure can offer unused radio spectrum for rent for as little as one second in a microcell. Alternatively, rentals may allow use for minutes to hours in macrocells. The cognitive protocol includes listening for legacy radios to attempt to use the spectrum so that the cognitive radios may politely defer to legacy users. Police, for example, may require the renters to immediately yield the spectrum back to the renting authority. The protocol supports the return of spectrum within 30 ms. Throughput is enhanced if the legacy users can wait for $\geq$0.5 s before being guaranteed clear spectrum.

Table 5. Mobile Spectrum Pools

Band	RF_{min}	RF_{max}	W_c	Remarks
Very low	26.9	399.9	315.21	Long-range vehicular traffic
Low	404	960	533.5	Cellular
Mid	1390	2483	930	PCS
High	2483	5900	1068.5	Indoor and RF LANs

Although cognitive radios may not be practical for years to come, the research points in an interesting direction for spectrum managers. Instead of hard-allocations with primary and secondary users, the spectrum managers at some point in the not-too-distant future should be able to delegate the details of spectrum management to the radios themselves. The spectrum managers would then assume the higher-level task of specifying the rules the radios have to follow to ensure equitable access that conforms to social, political, and legal norms.

This article has provided an overview of software radios. It began top-down by introducing the functional model of the software radio. Next, it introduced the important aspects of software, especially the need for isochronism in multiband multimode radios that share a pool of processing resources among multiple users. A range of hardware implementations were introduced, to differentiated among digital radios, PDRs, SDRs, and ideal software radios. This led to the characterization of acquisition parameters that divide software radios into broad classes. Finally, broader implications were

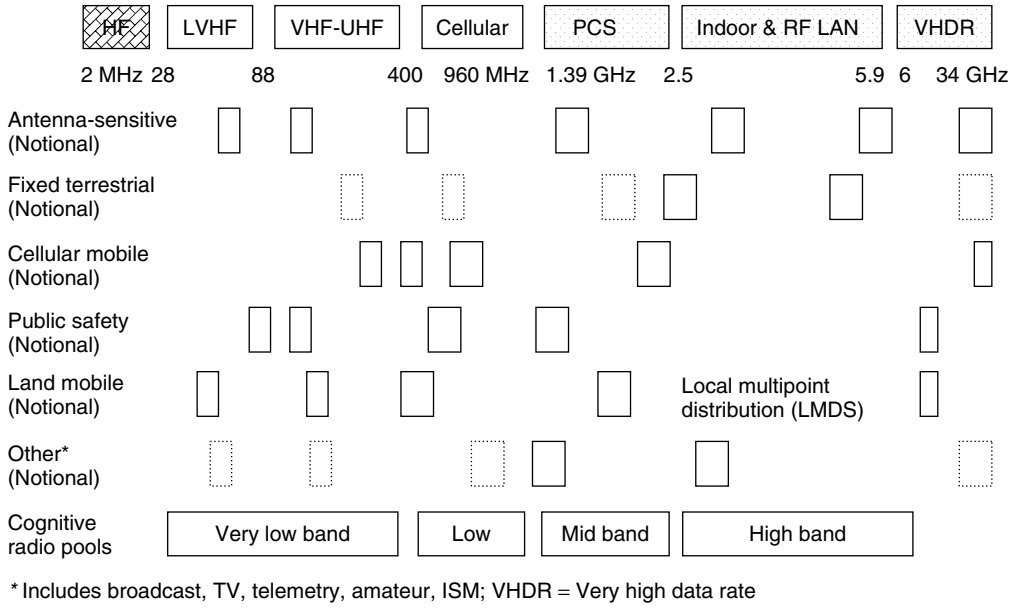


Figure 11. Potential spectrum pools.

presented, including the apparent challenge of type-certifying software radios. The chapter concluded with a view towards the future of the software radio — the evolution toward the cognitive radio.

8. CONCLUSION

Software radio has proved to be a valuable abstraction for the radio science, engineering, and user communities. During 2001, over 65 technical papers published in refereed journals of the IEEE and ACM included software radio or SDR as a theme. Although not a perfect metric, it indicates the degree to which the SWR/SDR abstraction is shaping research and technology development on a global scale. Practical SDR implementations continue to emerge with increasingly wider bandwidths and more tailored air interface capabilities. The challenges to the ideal software radio pose important research challenges, notably in the physics of antennas and RF conversion. The opportunities for innovative SDR implementations continue to attract substantial investments as the potential of this technology to reshape the physical layer of RF communications remains fertile ground.

ACRONYMS

3G/4G	Third- and fourth-generation wireless
ADC	Analog-to-digital-converter
ALE	Automatic Link Establishment, an HF air interface protocol
API	Applications programmer interface
ASICs	Application-specific integrated circuits
BER	Bit error rate
C3	Command-and-control communications
CASE	Computer-aided software engineering
CDMA	Code-division multiple access
CF	Core framework (of the JTRS/SDRF/OMG SCA)
CORBA	The Common Object Request Broker Architecture
DAC	Digital-to-analog converters
DARPA	The Defense Advanced Research Projects Agency
DISA	Defense Information Systems Agency
DMR	Digital modular radio
DoD	Department of Defense (U.S.)
DSP	Digital signal processor (or processing)
EC	The European Community, Brussels, Netherlands
ETSI	European Telecommunications Standards Institute
FDMA	Frequency-division multiple access
FEC	Forward error control
FPGA	Field programmable gate arrays
Gflops	Giga-floating-point operations per second
GHz	Gigahertz, 10^9 Hz
HF	High frequency, nominally 3–30 MHz
Hz	Hertz, cycles per second (e.g., of RF carrier frequency)
I/O	Input/output
IDL	Interface Definition Language
IEEE	Institute of Electrical and Electronics Engineers
IF	Intermediate frequency
INFOSEC	Information Security
IP	Internet Protocol, as in TCP/IP
ISA	Instruction set architecture
ITU	International Telecommunications Union
JTRS	Joint Tactical Radio System
kHz	Kilohertz, 10^3 Hz
LNA	Low-noise amplifier
LO	Local oscillator
loc	Lines of code
LVHF	Low VHF, typically 28–88 MHz
MHz	Megahertz, 10^6 Hz
Mips	Millions of instructions per second
OMG	Object Management Group
ORB	Object Request Broker
PCS	Personal communications system
PDA	Personal digital assistant
PSTN	Public Switched Telephone Network
PTT	Push to talk
QoS	Quality of service
RF	Radiofrequency
SCA	Software Communications Architecture
SDL	Specification and Description Language (ITU Standard Z.100)
SDR	Software-defined radio
SDRF	SDR Forum
SDR Forum	See *www.sdrforum.org*
SNR	Signal-to-noise ratio
SWR	Software radio
TCP	Transmission Control Protocol, usually used with IP, as in TCP/IP
TDMA	Time-division multiple access
UHF	Ultrahigh frequency, typically 300–3000 MHz
UML	The Unified Modeling Language
VHF	Very high frequency, typically 30–300 MHz
XML	The eXtensible Markup Language

BIOGRAPHY

Dr. Joseph Mitola III is an internationally recognized expert on software radio systems and technologies. In addition to having published the first paper on software radio architecture in 1992, he teaches in the United States, Asia, and Europe. He was Founding Chair of the SDR Forum in 1996 and co-chairs the Forum's Technical Symposium, November 2002. He published the first interdisciplinary graduate text on SWR, *Software Radio Architecture* [Wiley Interscience]. His doctoral dissertation, *Cognitive Radio* (KTH, June 2000), created the first teleinformatics framework for autonomous software radios, integrating machine learning and language processing into software radio. He edited the IEEE text *Software Radio Technologies*. With The MITRE Corporation, Dr. Mitola applies his expertise in telecommunications and information processing to the national and tactical needs of DoD. More recently he served as Senior Program Manager at the Defense Advanced Research Projects Agency and General Systems Engineer of the Defense Airborne Reconnaissance Office

(DARO). Prior to MITRE, Dr. Mitola was the Chief Scientist of Electronic Systems, E-Systems Melpar Division, culminating a career at E-Systems that began in 1976. He has also held positions of technical leadership with Harris Corporation, Advanced Decision Systems, and ITT Corporation. He began his career with the U.S. DoD in 1967. Dr. Mitola holds the B.S. in EE (Northeastern University '72); M.S.E. (The Johns Hopkins University, 1974); Licentiate in Engineering (May 1999), and Doctorate in Teleinformatics (The Royal Institute of Technology, KTH, Stockholm, June 2000).

BIBLIOGRAPHY

1. Software-Defined Radio (SDR) Forum (*www.sdrforum.org*).
2. Object Management Group (OMG) (*www.omg.org*).
3. J. Mitola and Z. Zvonar, *Software Radio Technology: Selected Readings*, IEEE Press, New York, 2001.
4. E. Del Re, ed., *Software Radio*, Springer-Verlag, London, 2001.
5. J. Mitola, *Software Radio Architecture*, Wiley, New York, 2000.
6. Kohno, *Software Radio and Software Antenna: Spatial and Temporal Communication Theory Using Software Antenna*, Yokohama National Univ., Yokohama, Japan, 1998.
7. Upmal and Lackey, SPEAKeasy, the military software radio, *IEEE Commun. Mag.* (1995).
8. P. Cook, An architectural overview of the speakeasy system, *IEEE J. Select. Areas Commun.* (April 1999).
9. *ACTS Mobile Communications Summit '98*, European Commission, Rhodes, Greece, June 98.
10. *4th ACTS Mobile Communications Summit '99* (CD-ROM) European Commission, Sorrento, Italy, June 1999.
11. M. Mehta et al., Reconfigurable terminals: An overview of architectural solutions, *IEEE Commun. Mag.* (Aug. 2001).
12. *<http://www.motorola.com/GSS/SSTG/ISSPD/WITS/DMR.html>*.
13. Joint Tactical Radio System homepage, *www.jtrs.saalt.army.mil* (2002).
14. McGarth et al., *RFIC Technology for Wireless Consumer Products — Trends in GaAs*, M/A-COM LOUD & Clear, M/A-COM, Inc., Lowell, MA, 1995.
15. Kennedy and Sullivan, Direction finding and smart antennas using software radio architectures, *IEEE Commun. Mag.* (May 1995).
16. D. Nicholson, *Spread Spectrum Signal Design LPE and AJ Systems*, Computer Science Press, Rockville, MD, 1988.
17. Stallings, *Handbook of Computer-Communications Standards*, Vol. 1, *The Open Systems Interconnection* (*OSI*) *Model and OSI-Related Standards*, Macmillan, New York, 1987.
18. J. Mitola, *Cognitive Radio: Model Based Competence for Software Radios*, Licentiate thesis, KTH (The Royal Institute of Technology), Stockholm, Sweden, Aug. 1999.
19. Pickholtz and Hill, *Adaptive Beamforming for Interference Reduction*, George Washington Univ. PW3312A, Dec. 31, 1990.
20. Zoltowski et al., Blind 2-D rake receivers based on space-time adaptive MVDR processing for IS-95 CDMA system, *Proc. MILCOM 96*, IEEE, New York, Oct. 1996.
21. Belzer et al., *Joint Source Channel Coding of Images with Trellis Coded Quantization and Convolutional Codes*, UCLA, Los Angeles, 1998.
22. Ferguson and Huston, *Quality of Service*, Wiley, New York, 1998.
23. Paradells et al., *DECT Multibearer Channels*, IEEE Press, New York, 1994.
24. Strom and Shaula, Optimistic recovery in distributed systems, *ACM Trans. Comput. Sys.* (1985).
25. Pesonen, *Object-Based Design of Embedded Software Using Real-Time Operating Systems*, IEEE Press, New York, 1994.
26. M. Cummings and S. Heath, Mode switching and software download for software defined radio: The SDR Forum approach, *IEEE Commun. Mag.* (Aug. 1999).
27. ITU Recommendation H.320, *Narrow-band Visual Telephone Systems and Terminal Equipment* (*www.itu.int/publications/itu-t/ituth13.htm*), International Telecommunications Union, 1998.
28. ITU, *Coding of analogue signals by pulse code modulation* (*G.711–G.712*) *and by methods other than PCM* (*G.720–G.729*), International Telecommunications Union, Geneva, Switzerland, 1998.
29. IETF references to internetworking.
30. Mouly and Pautet, Evolution of the GSM system, *IEEE PCS Mag.* (Oct. 1995).
31. J. Storer, *Data Compression*, The Computer Science Press, Rockville, MD, 1988.
32. Ziemer and Peterson, *Digital Communications and Spread Spectrum Systems*, Macmillan, New York, 1985.
33. W. Peterson and E. Weldon, *Error-Correcting Codes*, MIT Press, Cambridge, MA, 1972.
34. G. Simmons, ed., *Contemporary Cryptography*, IEEE Press, New York, 1992.
35. J. Razavilar et al., Software radio architecture with smart antennas: A tutorial on algorithms and complexity, *IEEE J. Select. Areas Commun.* (April 1999).
36. J. Mitola, *Cognitive Radio: An Integrated Agent Architecture for Software-Defined Radio*, doctoral dissertation, KTH (The Royal Institute of Technology), Stockholm, Sweden, June 2000.
37. ITU Recommendation H.320, ITU-T, Geneva, 1998.
38. *Random House Unabridged Webster's Dictionary*, Random House, New York, 1999.
39. *DII Strategic Enterprise Architecture*, DISA, Washington, DC, 1994.
40. *Technical Architecture for Information Management* (*TAFIM*), U.S. DoD, Washington, DC, 1996.
41. E. Gamma et al., *Design Patterns: Elements of Reusable Object-Oriented Software*, Adisson-Wesley, Reading, MA, 1994.
42. K. Gardner et al., *Cognitive Patterns: Problem Solving Frameworks for Object Technology*, Cambridge Univ. Press, Cambridge, UK, 1998.
43. Mouly and Pautet, *The GSM System for Mobile Communications*, (published by the authors), Plaiseau, France, 1992.
44. J. Mitola, Software radio architecture: A mathematical perspective, *IEEE J. Select. Areas Commun.* (April 1999).

45. Hoest and Shavit, Towards a topological characterization of asynchronous complexity, *Proc. PODC'97*, ACM, Santa Barbara, CA, 1997.
46. J. Mitola, Software radios: Technology and prognosis, *Proc. Nat. Telesystems Conf.*, May 1992, IEEE, New York, 1992.
47. Ono, *Introduction to Point Set Topology*, Johns Hopkins Univ. Press, Baltimore, MD, 1974.
48. J. Mitola, Software radio architecture: A mathematical perspective, *IEEE J. Select. Areas Commun.* (April 1999).
49. T. Mowbray and R. Zahavi, *The Essential CORBA*, Wiley, New York, 1995.
50. *Software Communications Architecture Specification*, MSRC-5000SCA V2.2, JTRS Joint Program Office, Rosslyn, VA (online) *www.jtrs.saalt.army.mil/docs/documents/sca.html* (Nov. 17, 2001).
51. W. C. Y. Lee, *Mobile Communications Design Fundamentals*, Sams, Indianapolis, IN, 1986.
52. Mouly and Pautet, Evolution of the GSM system, *IEEE PCS Mag.* (Oct. 1995).
53. D. Nicholson, *Spread Spectrum Signal Design LPE and AJ Systems*, Computer Science Press, Rockville, MD, 1988.
54. Qualcomm, *The Technical Case for Convergence of Third Generation Wireless Systems Based on CDMA* (*www.qualcomm.com*), March 1999.
55. Bensley et al., *Introduction to Parallel Supercomputing*, The MITRE Corp., Bedford, MA, 1988.
56. *Jane's Military Communications 1992–93* Jane's Information Group, Surrey, UK, 1992.
57. K.-C. Chen and S.-T. Wu, A programmable architecture for OFDM-CDMA, *IEEE Commun. Mag.* (Nov. 2000).
58. T. Kanter, *Adaptive Personal Mobile Communication: Service Architecture and Protocols*, doctoral dissertation, KTH (The Royal Institute of Technology), Stockholm, Sweden, Nov. 2001.
59. J. L. Dixon and J. Wilkes, A 'low-cost' software radio testbed, *Proc. IEEE VTS 53rd Vehicular Technology Conf.*, Spring 2001, IEEE Press, New York, 2001.
60. H. Shiba et al., Design and evaluation of software radio prototype with over-the-air download function, *Proc. Vehicular Technology Conf.*, Fall 2001, IEEE Press, New York, 2001.
61. M. Cummings and S. Heath, Mode switching and software download for software defined radio — the SDR Forum approach, *IEEE Commun. Mag.* (Aug 1999).
62. R. Rummler et al., Traffic modeling of software download for reconfigurable terminals, *Proc. 12th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Sept. 2001, IEEE Press, New York, 2001.
63. R. Walden, Analog to digital converter súrvey and analysis, *IEEE J. Select. Areas Commun.* (April 1999).
64. Texas Instruments Corp. homepage, *www.ti.com* (Jan. 2002).
65. Analog Devices homepage, *www.analogdevices.com* (Jan. 2002).
66. R. Dean Straw N6BV, *The ARRL Handbook for Radio Amateurs*, ARRL (National Association for Amateur Radio), Newington, CT, 2000.
67. WinRadio, *www.advdig.com*, Advanced Digital Systems of St. Louis, Saint Louis, MO, Nov. 1999.
68. U. Rhode et al., *Communications Receivers*, McGraw-Hill, New York, 1997.
69. *RF IC Design for Wireless Communication Systems*, Mead Microelectronics, Inc., Corvallis, OR, 1996.
70. *The Software Defined Radio*, BellSouth, Athens, GA, Dec. 1995.
71. C. Dick, Configurable logic for digital communications: Some signal processing perspectives, *IEEE Commun. Mag.* (Aug. 1999).
72. A. Shankiti and M. Leeser, Implementing a RAKE receiver for wireless communications on an FPGA-based computer system, *Proc. FPGA 2000*, Monterey CA, ACM, New York, 2000.
73. *Harris Digital Channelizer Application Note: Channelized Receiver*, Harris Corp. Melbourne, FL, 1990.
74. K. Zangi and R. Koilpillai, Software radio issues in cellular base stations, *IEEE J. Select. Areas Commun.* (April 1999).
75. Pentek homepage, *www.pentek.com* (1999).
76. E. Farag et al., *A Programmable Power-Efficient Decimation Filter for Software Radios*, ACM O-S97913O3-31971O8, ACM, New York, 1997.
77. T. Yokoi et al., Software receiver technology and its applications, *IEICE Trans. Commun.* (Tokyo) (June 2000).
78. X. Reves et al., Software radio implementation of a DS-CDMA indoor subsystem based on FPGA devices, *Proc. 12th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Sept. 2001, IEEE Press, New York, 2001.
79. Y. Suzuki, Software radio base and personal station prototypes, *IEICE Trans. Commun.* (Tokyo) (June 2000).
80. *Proc. Mobile Communications Summit*, Barcelona, Spain, Sept. 2001, European Commission, Brussels, 2001.
81. R. Kohno, Structures and theories of software antennas for software defined radio, *IEICE Trans. Commun.* (Tokyo) (June 2000).
82. *Proc. ACTS Mobile Communications Summit '98*, June 1998, European Commission, Rhodes, Greece, 1998.
83. *Proc. 4th ACTS Mobile Communications Summit '99*, June 1998 (CD-ROM), European Commission, Sorrento, Italy, 1999.
84. J. Rosa, *RF/IF Subsystem of a Commercial SDR Base Station*, SDR Forum Document SDRF-01-I-0012-V0.00, Hypres Corp., White Plains, NY (*www.hypres.com*) Jan. 2001.
85. Motorola homepage, *www.motorola.com*.
86. V. Bose et al., Virtual radios, *IEEE J. Select. Areas Commun.* (April 1999).
87. Vanu, Inc. homepage, *www.vanu.com*.
88. P. Withington, *Impulse Radio Overview* (*www.timedomain.com*), Time Domain, Inc., 1999.
89. J. Mitola, The software radio architecture, *IEEE Commun. Mag.* (May 1995).
90. *The Software Defined Radio Request for Information*, BellSouth, Atlanta, GA, Dec. 1995.
91. Van Tuyl et al., FCC transmitter certification requirements: Issues related to software defined radio, *Proc. SDR Forum*, SDR Forum, Rome, NY, June 1999.
92. J. Mitola, Software radio architecture: A mathematical perspective, *IEEE J. Selec. Areas Commun.* (April 1999).

93. Tebbs & Garfield, *Real Time Systems*, McGraw-Hill, Berkshire, UK, 1977.

94. K. Ellison, *Developing Real-Time Embedded Software in a Market-Driven Company*, Wiley, New York, 1994.

95. J. Mitola, Cognitive radio for mobile multimedia communications, *Proc. IEEE Mobile Multimedia Communications* (MOMUC) Workshop, San Diego, CA, Nov. 1999, IEEE Press, New York, 1999.

SPACE–TIME CODES FOR WIRELESS COMMUNICATIONS

NAOFAL AL-DHAHIR
A. R. CALDERBANK
AT&T Shannon Laboratory
Florham Park, New Jersey

AYMAN F. NAGUIB
Morphics Technology Inc.
Campbell, California

Space–time coding is a communications technique for wireless systems that employ multiple transmit antennas and single or multiple receive antennas. Information theory has been used to demonstrate that multiple antennas have the potential to dramatically increase achievable data rates. Space–time codes realize these gains by introducing temporal and spatial correlation into the signals transmitted from different antennas. There is, in fact, a diversity gain that results from multiple paths between base station and mobile terminal, and a coding gain that results from how symbols are correlated across transmit antennas. Significant increases in throughput are possible with only two antennas at the base station and one or two antennas at the mobile, and with simple receiver structures. The second antenna at the mobile terminal can be used to further increase system capacity through interference suppression. This article provides an overview of space–time coding techniques and the associated signal processing framework for narrowband and broadband wireless communications.

Current cellular standards such as IS136 support circuit data and fax (facsimile) services at a rate of 9.6 kbps (kilobits per second), and a packet data mode is now being standardized. Rapid growth in mobile computing and other wireless data services is inspiring many proposals for high-speed data services in the range of 64–384 kbps for microcellular wide-area and high-mobility applications, and up to 2 Mbps for indoor applications [1].

However, data rates on band–limited wireless channels are limited by multipath fading and interference from other users [2–8]. Deploying multiple antennas at the both the transmitter and the receiver increases the capacity of wireless channels, and information theory provides measures of this increase [9–11]. The standard approach to increasing capacity is to use linear processing at the receiver with optimum linear combining to combat multipath fading and suppress interference [3,4]. Here, the received signals are weighted and combined to maximize the signal-to-interference-plus-noise ratio (SINR) at the receiver. By contrast, transmit diversity schemes use processing at the transmitter to spread the information across multiple transmit antennas. The earliest forms of transmit diversity were proposed by Uddenfeldt and Raith [12] and Wittneben [13]. The latter is a delay diversity scheme where a signal is transmitted from one antenna, then delayed one symbol, and transmitted from a second antenna. Wittneben's work includes the delay diversity scheme of Seshadri and Winters [14] as a special case (see also Refs. 15 and 16). Winters [17] showed that delay diversity is optimal in the sense that the diversity order experienced by an optimal receiver is equal to the number of transmit antennas. Diversity is the link-level advantage (over a single path) obtained from spreading information over multiple independent paths from base station to mobile unit. Note that superposition of fading statistics at the receiver also reduces variation in signal strength, and allows smoother and more efficient power control. This means that the base station can support significantly more users for a given constraint on radiated signal power. Space–time coding [18–25] combines correlation techniques designed for multiple transmit antennas with the appropriate signal processing at the receiver to provide significant gain over the delay diversity schemes in Refs. 13 and 14.

This article is organized as follows. Section 1 describes fundamentals of space–time coding for flat-fading channels. Equalization schemes necessary for implementation of space–time codes over frequency-selective channels are discussed in Section 2, and channel estimation issues are discussed in Section 3. An extensive reference list is provided to help the reader undertake a more detailed study of any of the topics discussed.

1. SPACE–TIME CODING

This section presents a mathematical model of a *narrowband* communications system with N_t transmit antennas and N_r receive antennas and it assumes a flat-fading channel (see Section 2 for frequency-selective channels). As shown in Fig. 1, the space–time encoder transforms the input data at time l into N_t code symbols

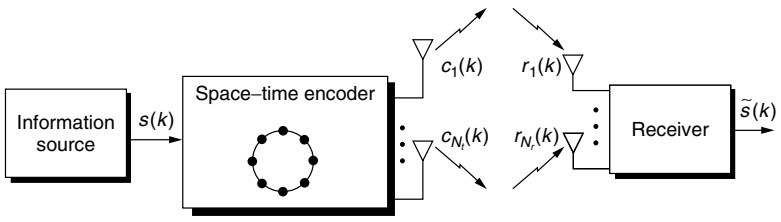


Figure 1. Space–time coding.

$c_1(l), c_2(l), \ldots, c_{N_t}(l)$ that are transmitted *simultaneously* from the different transmit antennas.

Signals arriving at different receive antennas undergo independent fading. The signal at each receive antenna is a noisy superposition of the faded versions of the N_t transmitted signals. Let E_s be the average energy of the signal constellation. The constellation points are scaled by a factor $\sqrt{E_s}$ so that the average energy of the constellation points is 1. Let $r_j(l), j = 1 \cdots N_r$ be the received signal at antenna j after matched filtering. Assuming ideal timing and frequency information, we have

$$r_j(l) = \sqrt{E_s} \cdot \sum_{i=1}^{N_t} \alpha_{ij}(l) c_i(l) + \eta_j(l), \qquad j = 1, \ldots, N_r \tag{1}$$

where $\eta_j(l)$ are independent samples of a zero-mean complex white Gaussian process with two-sided power spectral density $N_0/2$ per dimension. It is also assumed that $\eta_j(l)$ and $\eta_k(l)$ are independent for $j \neq k$, $1 \leq j, k \leq N_r$. The gain $\alpha_{ij}(l)$ models the complex fading channel gain from transmit antenna i to receive antenna j. The channel gain α_{ij} is modeled as a lowpass-filtered complex Gaussian random process with zero mean, variance 1, and autocorrelation function $R_\alpha(\tau) = J_0(2\pi f_d \tau)$, where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind and f_d is the maximum Doppler frequency [26]. It is assumed that signals transmitted from different antennas are subject to independent fades. This can be achieved by separating transmit antennas by more than half the underlying wavelength or by using antennas with different polarizations.

Let $\mathbf{c}_l = [c_1(l), \ldots, c_{N_t}(l)]^T$ be the $N_t \times 1$ *codeword* transmitted from the N_t antennas at time l, $\alpha_j(l) = [\alpha_{1j}(l), \ldots, \alpha_{N_t j}(l)]^T$ be the corresponding $N_t \times 1$ channel vector from the N_t transmit antennas to the jth receive antenna, and $\mathbf{r}(l) = [r_1(l), \ldots, r_{N_r}(l)]^T$ be the $N_r \times 1$ received signal vector. Also, let $\eta(l) = [\eta_1(l), \ldots, \eta_{N_r}(l)]^T$ be the $N_r \times 1$ noise vector at the receive antennas. Furthermore, let us define the $N_r \times N_t$ channel matrix $\mathcal{H}_l$ from the N_t transmit to the N_r receive antennas as $\mathcal{H}(l) = [\alpha_1(l), \ldots, \alpha_{N_r}(l)]^T$. Equation (1) can be rewritten in a matrix form as

$$\mathbf{r}(l) = \sqrt{E_s} \cdot \mathcal{H}(l) \cdot \mathbf{c}_l + \eta(l) \tag{2}$$

We can easily see that the *signal-to-noise ratio* (SNR) *per receive antenna* is given by

$$\text{SNR} = \frac{N_t \cdot E_s}{N_0} \tag{3}$$

1.1. Space–Time Trellis Codes (STTCs)

Suppose that the *codeword* sequence

$$\mathcal{C} = \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L$$

were transmitted and consider the probability that the decoder decides erroneously in favor of another legitimate codeword sequence

$$\tilde{\mathcal{C}} = \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \ldots, \tilde{\mathbf{c}}_L$$

Assume a data frame of length L and define the $N_t \times N_t$ error matrix $\mathcal{A}$ as

$$\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}}) = \sum_{l=1}^{L} (\mathbf{c}_l - \tilde{\mathbf{c}}_l)(\mathbf{c}_l - \tilde{\mathbf{c}}_l)^* \tag{4}$$

The squared distance between $\mathcal{C}$ and $\tilde{\mathcal{C}}$ at the output of the wireless channel turns out to be proportional to $\sum_{j=1}^{N_r} \mathcal{H}_j^* \mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}}) \mathcal{H}_j$, where $\mathcal{H}_j$ is the column vector of path gains from the different transmit antennas to the jth receive antenna. The vector $\mathcal{H}_j$ varies with time, and when it finds the null space of $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$, the jth receive antenna experiences a deep fade. Diversity gain is just the minimum rank of $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$, where the minimization is over all pairs of codewords. Coding gain depends on the product of the nonzero eigenvalues, and again there is a minimization over all pairs of codewords.

If ideal channel state information (CSI) $\mathcal{H}(l), l = 1, \ldots, L$ is available at the receiver, it is straightforward to show that the probability of transmitting $\mathcal{C}$ and deciding in favor of $\tilde{\mathcal{C}}$ is upper-bounded by

$$P(\mathcal{C} \to \tilde{\mathcal{C}}) \leq \left(\prod_{i=1}^{p} \lambda_i \right)^{-N_r} \cdot \left(\frac{E_s}{4N_0} \right)^{-pN_r} \tag{5}$$

where p is the rank of the error matrix $\mathcal{A}$ and $\lambda_i, i = 1, \ldots, p$ are the nonzero eigenvalues of the error matrix $\mathcal{A}$ (see Ref. 27 for details). The bound on probability of error given in Eq. (5) is similar to the probability of error bound for trellis-coded modulation for flat-fading channels. The first term $g_p = (\lambda_1 \lambda_2 \cdots \lambda_p)$ represents the coding gain achieved by the space–time code, and the second term $(E_s/4N_0)^{-pN_r}$ represents a diversity gain of pN_r. This analysis leads to two design criteria for space–time codes. The first is to maximize the rank p of $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$, thereby maximizing diversity gain. The second, for a given diversity gain p, is to maximize the coding gain g_p.

Now consider the problem of decoding space–time codes. Under the assumption that ideal CSI $\mathcal{H}(l), l = 1, \ldots, L$ is available at the receiver, we can derive the maximum-likelihood (ML) decoding rule for the space–time code as follows. Suppose that all codewords are equiprobable, a codeword

$$\mathcal{C} = \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L$$

has been transmitted, and

$$\mathcal{R} = \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_L$$

has been received, where $\mathbf{r}_l$ is given by Eq. (2). At the receiver, optimum decoding amounts to choosing a codeword sequence

$$\tilde{\mathcal{C}} = \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \ldots, \tilde{\mathbf{c}}_L$$

for which the a posteriori probability

$$\Pr(\tilde{\mathcal{C}} | \mathcal{R}, \mathcal{H}(l), l = 1, \ldots, L)$$

is maximized. Since the noise vector is assumed to be a multivariate AWGN (additive white Gaussian noise), it can be easily shown [27] that the optimum decoder is

$$\tilde{\mathcal{C}} = \arg \min_{\tilde{\mathcal{C}} = \tilde{\mathbf{c}}_1, \cdots, \tilde{\mathbf{c}}_L} \sum_{l=1}^{L} \| \mathbf{r}(l) - \sqrt{E_s} \cdot \mathcal{H}(l) \cdot \tilde{\mathbf{c}}_l \|^2 \tag{6}$$

It is obvious that the optimum decoder in (6) can be implemented using the Viterbi algorithm (VA) when the space–time code has a trellis representation.

Figure 2 shows an 8-PSK 8-state space–time code designed for two transmit antennas, where the edge label xy means that symbol x is transmitted from the first antenna and symbol y, from the second antenna. The different symbol pairs in a given row label the transitions (edges) out of a given state, in order, from top to bottom. Observe that labels on edges leaving a given state disagree in the first position. It follows that the rank of the matrix $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$ corresponding to codewords $\mathcal{C}$ and $\tilde{\mathcal{C}}$ (that diverge and then remerge) is equal to 2. The reader may verify that, for odd-numbered states, if the symbol transmitted from the first antenna is negated, the result is the delay diversity scheme proposed by Wittneben. Both schemes provide a diversity gain of 2, but with the space–time code there is an additional coding gain of 2.5 dB.

1.2. Space–Time Block Codes (STBCs)

When the number of antennas is fixed, the decoding complexity of space–time trellis coding (measured by the number of trellis states at the decoder) increases exponentially as a function of the diversity level and transmission rate [22]. In addressing the issue of decoding complexity, Alamouti [18] discovered a remarkable space–time block coding scheme for transmission with two antennas. This scheme supports maximum-likelihood detection based only on linear processing at the receiver. It was later generalized [19] to an arbitrary number of antennas and is able to achieve the full diversity promised by the number of transmit and receive antennas. Here, we will briefly review the basics of space–time block codes [18]. Figure 3 shows the baseband representation for transmit diversity employing space–time block coding with two transmit antennas. The input symbols to the space–time block encoder are divided into pairs. At a given symbol period, the two symbols in each group $\{c_1, c_2\}$ are transmitted simultaneously from the two antennas. The signal transmitted from antenna 1 is c_1, and the signal transmitted from antenna 2 is c_2. In the next symbol period, the signal $-c_2^*$ is transmitted from antenna 1 and the signal c_1^* is transmitted from antenna 2. Let h_1 and h_2 be the channel gains from the first and second transmit antennas to the receive antenna, respectively. The major assumption here is that h_1 and h_2 are constant over two consecutive symbol periods:[1]

$$h_i(nT) = h_i((n+1)T), \qquad i = 1, 2$$

Let r_1 and r_2 be the received signals over two consecutive symbol periods. Then

$$r_1 = h_1 c_1 + h_2 c_2 + \eta_1 \tag{7}$$

$$r_2 = -h_1 c_2^* + h_2 c_1^* + \eta_2 \tag{8}$$

where η_1 and η_2 represent the AWGN and are modeled as i.i.d. complex Gaussian random variables with zero mean and power spectral density $N_0/2$ per dimension. Define the received signal vector $\mathbf{r} = [r_1\ r_2^*]^T$, the codeword vector $\mathbf{c} = [c_1\ c_2]^T$, and the noise vector $\eta = [\eta_1\ \eta_2^*]^{\mathbf{T}}$. Equations (7) and (8) can be rewritten in a matrix form as

$$\mathbf{r} = \mathbf{H} \cdot \mathbf{c} + \eta \tag{9}$$

where the channel matrix $\mathbf{H}$ is defined as

$$H = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \tag{10}$$

The vector η is a complex Gaussian random vector with zero mean and covariance $N_0 \cdot \mathbf{I}$. Define $\mathcal{C}$ as the set of

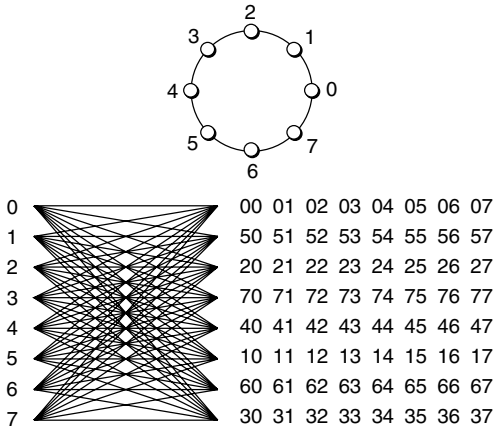


Figure 2. 8-PSK 8-state space–time code with two transmit antennas and a bandwidth efficiency of 3 bits/channel use.

[1] For GSM (Global System for Mobile Communication) mobiles traveling at 60 mph (mi/h) and a carrier frequency of 1 GHz, the channel coherence time is around 11 ms, which is about 3000 GSM symbol durations. Hence, the channel can be safely assumed constant over several hundred symbol durations even at highway speeds. For IS136 mobiles, the symbol duration is around 41 μs; hence, the channel can be assumed constant over few tens of consecutive symbols. Therefore, the assumption of a fixed channel over two consecutive symbols is satisfied for both systems.

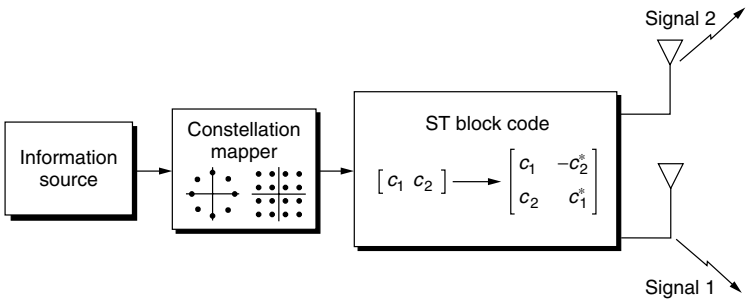


Figure 3. Transmit diversity with space–time block coding.

all possible symbol pairs $\mathbf{c} = \{c_1, c_2\}$ and assume that all symbol pairs are equiprobable. Since the noise vector η is assumed to be a multivariate AWGN, it follows that the optimum ML decoder is

$$\hat{\mathbf{c}} = \arg\min_{\hat{\mathbf{c}} \in \mathcal{C}} \|\mathbf{r} - \mathbf{H} \cdot \hat{\mathbf{c}}\|^2 \tag{11}$$

The ML decoding rule in this equation can be further simplified by realizing that the channel matrix $\mathbf{H}$ is orthogonal ($\mathbf{H}^*\mathbf{H} = (|h_1|^2 + |h_2|^2) \cdot \mathbf{I}$). Consider the modified received signal vector $\tilde{\mathbf{r}}$ given by

$$\tilde{\mathbf{r}} = \mathbf{H}^* \cdot \mathbf{r} = (|h_1|^2 + |h_2|^2) \cdot \mathbf{c} + \tilde{\eta} \tag{12}$$

where $\tilde{\eta} = \mathbf{H}^* \cdot \eta$. In this case the ML decoding rule becomes

$$\hat{\mathbf{c}} = \arg\min_{\hat{\mathbf{c}} \in \mathcal{C}} \|\tilde{\mathbf{r}} - (|h_1|^2 + |h_2|^2) \cdot \hat{\mathbf{c}}\|^2 \tag{13}$$

Since $\mathbf{H}$ is orthogonal, it follows that the noise vector $\tilde{\eta}$ will have a zero mean and covariance $(|h_1|^2 + |h_2|^2) \cdot \mathbf{I}$, that is, the elements of $\tilde{\eta}$ are independent and identically distributed. Hence, it follows immediately that simple linear combining reduces the ML decoding rule in Eq. (13) to two separate, and much simpler, ML decoding rules for c_1 and c_2, as established elsewhere [18]. Assuming that we are using a signaling constellation with 2^b constellation points, this linear combining reduces the number of decoding metrics that have to be computed for ML decoding from 2^{2b} to 2×2^b. It is also straightforward to verify that the SNR for c_1 and c_2 will be

$$\text{SNR} = \frac{(|h_1|^2 + |h_2|^2) \cdot E_s}{N_0} \tag{14}$$

and hence a two-branch diversity performance is obtained at the receiver. When the receiver uses N_r receive antennas, we can write the received signal vector $\mathbf{r}_m$ at receive antenna m and 2, respectively as

$$\mathbf{r}_m = \mathbf{H}_m \cdot \mathbf{c} + \eta_{\mathbf{m}} \tag{15}$$

where $\eta_{\mathbf{m}}$ is the noise vector and $\mathbf{H}_m$ is the channel matrix from the two transmit antennas to the mth receive antenna. In this case the optimum ML decoding rule is

$$\hat{\mathbf{c}} = \arg\min_{\hat{\mathbf{c}} \in \mathcal{C}} \sum_{m=1}^{N_r} \|\mathbf{r}_m - \mathbf{H}_m \cdot \hat{\mathbf{c}}\|^2 \tag{16}$$

As before, in the case of N_r receive antennas, the decoding rule can be further simplified by premultiplying the received signal vector $\mathbf{r}_m$ by $\mathbf{H}_m^*$. In this case, the diversity order provided by this scheme is $2N_r$. Figure 4 shows a simplified block diagram for the receiver with two receive antennas. The properties of the space–time block coding scheme in Fig. 4 and its extension in Ref. 19 can be further exploited to improve wireless capacity and/or throughput. The reader is referred to Ref. 28 for further discussion on this point.

2. EQUALIZATION OF SPACE–TIME CODES ON FREQUENCY-SELECTIVE CHANNELS

As the transmission bandwidth increases beyond the *coherence bandwidth* [29] of the channel, equalization becomes indispensable. Equalization complexity increases with the channel memory (also referred to as the *channel delay spread*), signal constellation size, and the use of multiple transmit and/or receive antennas. The objective of this section is to give an overview of candidate equalization schemes for space–time-coded transmission over broadband wireless channels. In this article, we use the terms *frequency-selective channel*, *broadband channel*, and *intersymbol interference* (ISI) *channel* interchangeably.

We start in Section 2.1 by describing the frequency-selective channel model and assumptions. Effective equalization schemes for STTC and STBC are discussed in Sections 2.2 and 2.3, respectively.

2.1. Channel Model and Assumption

The channel impulse response (CIR) from transmit antenna i to receive antenna j is denoted by the vector $\mathbf{h}_{ij}$. The multiple-input/multiple-output (MIMO) channel memory, denoted by ν, is the maximum memory of all constituent single-input/single-output (SISO) channels. For simplicity, we focus on the case $N_t = 2$ and $N_r = 1$, hence, $\mathbf{h}_{ij}$ will be simply denoted by the vector $\mathbf{h}_i$ or its corresponding D-transform $h_i(D) \stackrel{def}{=} \sum_{k=0}^{\nu} \mathbf{h}_i(k) D^k$. Extension to the general case is straightforward. The CIRs are assumed constant over the transmission block (quasistatic fading) and vary independently from block

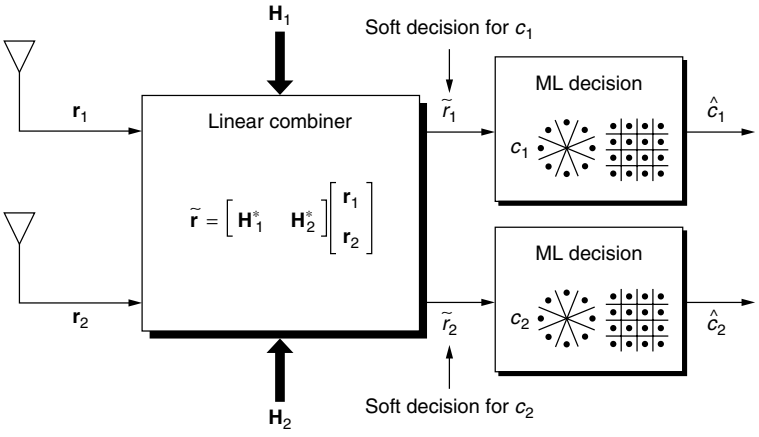


Figure 4. Receiver for space–time block coding.

to block. The input symbols are assumed complex zero-mean and belong to a 2^b signal constellation. The noise is additive white Gaussian and independent of the input.

2.2. Equalization Schemes for Space–Time Trellis Codes

Our focus will be on the 8-state 8-PSK STTC shown in Fig. 2. This code has a rich and transparent structure that can be exploited to simplify equalization.

1. *Turbo Equalization*. While it is possible in theory to model the STTC and the ISI channel, separated by an interleaver,[2] by a single trellis and perform maximum a posteriori (MAP) decoding on this trellis using, for instance, the BCJR (Bahl–Cocke–Jelinek–Raviv) algorithm [30], the complexity would be prohibitive. An alternative lower-complexity decoding scheme views the space–time encoder and the ISI channel, separated by an interleaver, as a serial concatenation of two finite-state machines that can be decoded iteratively using the *Turbo principle* [31]. Using this Turbo equalization scheme, joint space–time equalization and decoding is performed by iteratively exchanging *soft* extrinsic information between the separate BCJR-MAP equalizer and decoder modules. Hard decisions are generated only after the last iteration. The BCJR algorithm consists of a forward and a backward recursion and is usually implemented in the log domain to reduce computational complexity and improve numerical accuracy. Turbo equalization achieves remarkable performance very close to theoretical performance limits [32]. The number of states in the BCJR equalizer module is *exponential* in the channel memory, the number of transmit antennas, and the spectral efficiency (in bps/Hz). The use of MIMO FIR shortening prefilters [33] to reduce the complexity of the BJRC equalizer module and its application to STTC have been studied in [34]. However, for spectrally efficient modulation schemes (such as 8-PSK modulation used in EDGE[3]), the complexity of turbo equalization is still too high [36]. In addition, the long decoding delay might not be acceptable for speech and real-time data applications. An attractive alternative in this case is the M-BCJR equalizer described next.

2. *Prefiltered M-BCJR Equalizer*. The M-BCJR algorithm [37], is a reduced-complexity version of the BCJR algorithm [30] where at each trellis step, only the M active states associated with the highest metrics are retained. An improved version of the M-BCJR algorithm was proposed [38] and applied to the equalization of STTC. Moreover, it was shown [38] that preceding the M–BCJR equalizer with a channel-shortening prefilter improves its performance, especially for small values of M. Even better performance is achieved when a different prefilter is used for the forward and backward recursions of the M-BCJR algorithm. The value of M and the number of prefilter taps can be jointly optimized to achieve the best performance–complexity tradeoffs.

3. *Prefiltered MLSE/DDFSE Equalizer*. Unlike the BJCR-MAP equalizer, which minimizes the *symbol* error probability, maximum-likelihood sequence estimation (MLSE) minimizes the *sequence* error probability assuming equally likely inputs and can be implemented efficiently using the VA. A major advantage of the BCJR-MAP algorithm over the conventional VA[4] is the generation of *soft* information on the decisions. Generalization of the MLSE equalizer to the MIMO case was first reported by Van Etten [40]. For a 2^b signal constellation, N_t transmit antennas, and MIMO channel memory of ν, the MIMO MLSE equalizer has $2^{b \cdot N_t \cdot \nu}$ states in general. The number of equalizer states can be reduced to $2^{b \cdot \nu}$ by using the STTC trellis structure as shown in Ref. 41. However, this complexity is still too high for large signal constellations and long MIMO channel memory. Delayed decision feedback sequence estimation (DDFSE) was introduced [42] as a hybrid scheme between MLSE and decision feedback equalization (DFE) [43] for channels with long memory. Basically, the CIR is divided into a leading part and a tail. Then, an MLSE equalizer is constructed on the basis of the leading part, and the interfering effect of the CIR tail is canceled by feedback using previous (hard) decisions (assumed correct). Like all feedback schemes, DDFSE suffers from error propagation effects. These effects are minimized if most of the channel energy is concentrated in its leading part (as in minimum-phase channels). This is the task of the FIR prefilter designed [44] to improve DDFSE performance in equalizing the 8-state 8-PSK STTC [41].

4. *Orthogonal Frequency Division Multiplexing* (*OFDM*). In OFDM, the high-rate input stream is demultiplexed and transmitted over N low-rate independent frequency subcarriers. This multicarrier transmission scheme is implemented digitally using the efficient fast Fourier transform (FFT) method [45]. OFDM is a block transmission scheme; therefore, a guard sequence (of length at least equal to channel memory) is needed to eliminate interblock interference (IBI). The most popular choice for guard sequence is a *cyclic prefix*, which makes the channel matrix *circulant*; hence, diagonalizable by the FFT. If the FFT size is made large enough such that the width of each frequency bin is less than the *coherence bandwidth* of the channel, then no equalization is needed.[5] A large FFT size (compared to channel memory) also reduces the guard sequence overhead at the expense of increased storage and processing requirements and increased delay which might not be acceptable for delay-sensitive applications.

The OFDM scheme has been extended to the MIMO case [46]. OFDM was successfully applied to equalization of STTC [47].

[2] The randomizing effect of the interleaver is critical to the remarkable performance exhibited by Turbo schemes.

[3] EDGE stands for *enhanced data rates for GSM evolution* and is the proposed third-generation TDMA cellular standard [35].

[4] A modified soft-output Viterbi algorithm (SOVA) was presented by Hagenauer and Hoeher [39]. However, its performance is suboptimal compared to that of BCJR-MAP.

[5] Except for a simple gain and phase adjustment using a single complex tap for each subchannel, assuming negligible intercarrier interference due to Doppler effects or frequency offset errors.

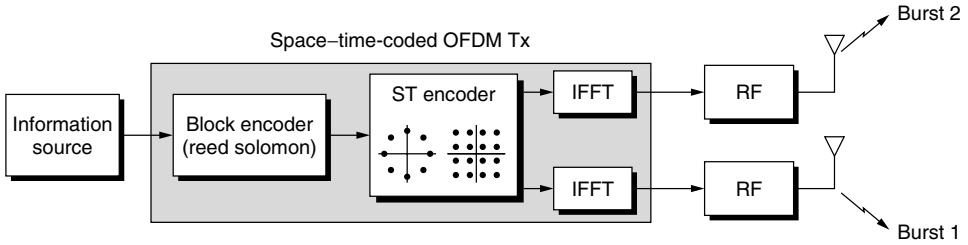


Figure 5. Transmitter for space–time-coded OFDM for broadband applications.

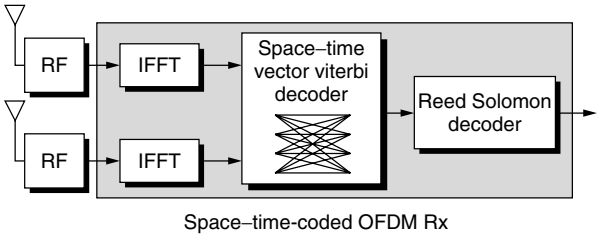


Figure 6. Receiver for space–time-coded OFDM for broadband applications.

Figures 5 and 6 show simplified block diagrams for the transmitter and receiver, respectively, for an OFDM modem with a concatenated space–time coding scheme. The input information symbols are first encoded by an outer conventional channel code. The output of the outer code is then space–time-encoded. Each of the space–time code output streams is then OFDM-modulated and sent over the corresponding antenna. At the receiver, the signal at each receive antenna is OFDM-demodulated. The demodulated signals from the antennas are then fed into the space–time decoder followed by the outer decoder. Figure 7 shows the simulation results for the abovementioned OFDM space–time-coded modem. In this simulation, the available bandwidth is 1 MHz, and the maximum Doppler frequency is 200 Hz. The number of OFDM tones used for modulation is 256. These correspond to a subcarrier separation of 3.9 kHz and OFDM frame duration of 256 μs. To each frame, a cyclic prefix of 40 μs duration is added. Each tone modulates a 4-PSK constellation, although higher-order constellations may be used. We used 16-state 4-PSK space–time code with two transmit and two receive antennas. In addition, an outer (72,64,9) RS code over Galois Field GF(2^7) is used. We plot the frame error probability as function of SNR for different channel delay spreads. From this plot, we can see that an E_b/N_0 between 2.7-4 dB (depending on the delay spread) is needed to achieve a data rate of 1.5 Mbps.

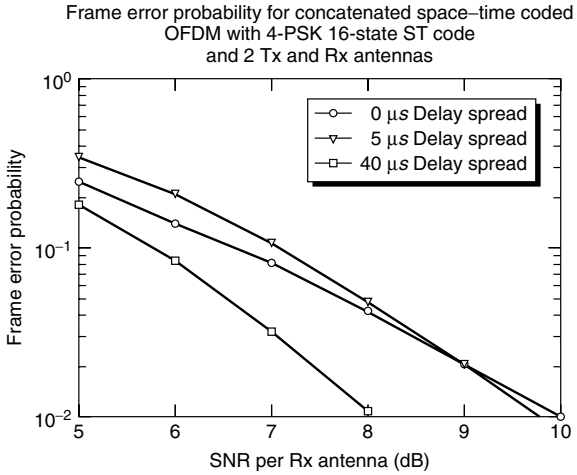


Figure 7. FER of concatenated space–time-coded OFDM with 4-PSK 16-state STC with 2Tx and 2Rx antennas.

2.3. Equalization Schemes for Space–Time Block Codes

Our focus will be on the case of two transmit antennas described in Section 1.2, where a full-rate STBC can be constructed for any signal constellation.

1. *Time-Reversal Space–Time Block Coding* (*TRSTBC*). TRSTBC was introduced [48] as an extension of the Alamouti STBC scheme [18] to frequency-selective channels by imposing the Alamouti orthogonal structure at a *block*, *not symbol*, level as in the flat-fading channel case. At the receiver, TRSTBC employs clever time-domain processing to eliminate the mutual interference effects between the two inputs *while still achieving the maximum diversity gain* of $|\mathbf{h}_1|^2 + |\mathbf{h}_2|^2$. Effectively, TRSTBC converts the two-input/single-output channel to two SISO channels, each with equivalent impulse response $h_{\text{eqv}}(D) = h_1(D)h_1^*(D^{-1}) + h_2(D)h_2^*(D^{-1})$ to which standard SISO equalization schemes such as MLSE [49] or DFE [50] can be applied. TRSTBC assumes that the two channels $h_1(D)$ and $h_2(D)$ are fixed over two consecutive transmission blocks and perfectly known at the receiver and that guard symbols (of length at least equal to channel memory) are inserted between data blocks to eliminate IBI.

2. *Orthogonal Frequency-Division Multiplexing* (*OFDM*). An elegant scheme for combining OFDM and STBC by implementing the Alamouti orthogonal structure at a block level was first reported by Liu et al. [51]. This OFDM–STBC scheme achieves the *full* diversity gain of $|\mathbf{h}_1|^2 + |\mathbf{h}_2|^2$ without bandwidth expansion for two transmit antennas, assuming that the channel is fixed over two consecutive OFDM blocks and known at the receiver and a cyclic prefix is used to eliminate IBI.

3. *Single-Carrier Frequency-Domain Equalization* (*SCFDE*). OFDM has two main drawbacks with respect to single-carrier transmission, namely, a higher peak : average ratio (PAR), which results in larger back-off with nonlinear amplifiers and increased sensitivity to frequency errors and phase noise [52]. An alternative equalization scheme that overcomes these two drawbacks

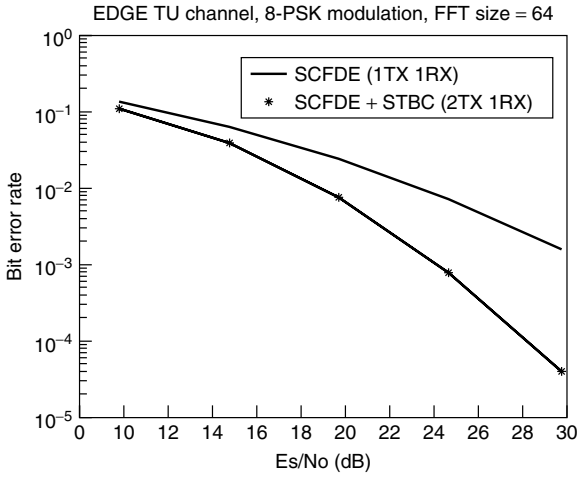


Figure 8. Bit error rate of SCFDE with 1TX and 2TX (STBC) for EDGE TU environment, 8-PSK modulation, and size 64 FFT (1 RX is assumed).

of OFDM while retaining its reduced implementation complexity advantage (due to use of FFT) is single-carrier frequency-domain equalization (SCFDE) [53]. An effective transmit diversity scheme for combining STBC and SCFDE over frequency-selective channels is described in Ref. 54. Figure 8 shows the significant transmit diversity gain achieved as exhibited by the increased slope of the BER curve at high SNR. This simulation assumes a typical urban (TU) EDGE channel with a linearized GMSK transmit pulseshape, 8-PSK modulation, and an FFT size of 64.

3. CHANNEL ESTIMATION ISSUES

Channel estimation for space–time-coded transmissions over flat-fading channels can be performed effectively using orthogonal pilot tones and interpolation as discussed in detail in Ref. 25. Here, we discuss the more challenging frequency-selective channel case.

In single-carrier block transmission systems, a *training sequence* is typically inserted in each block and used to estimate the CIR at the receiver (e.g., using a least-squares algorithm [55]). If the CIR varies within the block, this initial CIR estimate can be tracked using one of various adaptive algorithms.

For single-transmit-antenna scenarios, the training sequence is only required to have a "good" (i.e., impulselike) autocorrelation sequence. However, for the N_t transmit antenna scenarios, the N_t training sequences should, in addition, have "low" (ideally zero) cross-correlation, at least over time lags less than or equal to those of the MIMO channel memory. It can be shown that *perfect root of unity sequences* (PRUS) [56] have these ideal correlation properties. However, PRUS do not belong to standard signal constellations such as PSK. Additional challenges in channel estimation for multiple-transmit-antenna systems over the single-transmit-antenna case are the increased number of channel parameters to be estimated and the reduced transmit power (by a factor of N_t) for each transmit antenna. An obvious transmission scheme that forces the cross-correlation between the N_t training sequences to zero consists of dividing the training interval into N_t subintervals where only one antenna is allowed to transmit its training sequence in each subinterval. This scheme has two major drawbacks: (1) the peak : average ratio (PAR) is increased, which in turn increases amplifier nonlinear distortion; and (2) the effective training period for each transmit antenna is reduced by a factor of N_t.

The rich structure of space–time codes can be used to reduce the number of channel parameters to be estimated. For example, for TRSTBC, decoupling of the two inputs due to the Alamouti orthogonal structure removes the requirement of low cross-correlation between the two training sequences. Similarly, for the 8-state 8-PSK STTC, the special code structure can be exploited to simplify the training sequence design problem while restricting the training symbols to belong to standard signal constellation and incurring negligible performance loss from PRUS [57].

4. CONCLUSION

Space–time coding is a new coding/signal processing framework for wireless communications systems with multiple transmit/receive antennas. This new framework offers the best tradeoff between spectral efficiency and power consumption by optimum combination of modulation, coding, and diversity gains over flat-fading channels.

Space–time trellis codes offer the maximum possible diversity and coding gains without any sacrifice in transmission bandwidth. Their decoding requires a vector Viterbi algorithm. Alamouti-type space–time block codes offer maximum diversity gain, much lower decoding complexity, and full rate transmission but sacrifice coding gain.

For frequency-selective fading channels, we described several competitive equalization schemes for space–time codes that further exploit the temporal diversity of the channel. For space–time trellis codes, prefiltered M-BCJR and OFDM achieve the best performance-complexity tradeoff. For Alamouti's space–time block code, TRSTBC, OFDM–STBC, and FDE–STBC are the most promising candidates. OFDM-based schemes are less attractive when issues of high PAR, frequency errors, and delay become critical. Exploiting the rich structure of space–time codes is critical in simplifying the channel estimation and equalization schemes.

Acknowledgments

We would like to thank the following colleagues (in alphabetical order) for many technical discussions and contributions to this work: G. Bauch, S. N. Diggavi, C. Fragouli, N. Seshadri, A. Stamoulis, V. Tarokh, and W. Younis.

BIOGRAPHIES.

Naofal Al-Dhahir received his M.S. and Ph.D. degrees from Standford University in 1990 and 1994, respectively, in electrical engineering. He was as instructor at Stanford University during Winter 1993. From August 1994 to 1999, he was a member of the technical staff at the Communications Program at GE Corporate R&D Center

in Schenectady, New York, where he worked on various aspects of satellite communication systems design and anti-jam GPS receivers. Since August 1999, he has been a principal member of technical staff at AT&T Shannon Laboratory in Florham Park, New Jersey. His current research intersts include equalization schemes, space-time coding and signal processing, OFDM, and digital subscriber line technology. He has authored more than 40 journal papers and holds 7 U.S. patents in the areas of satellite communications, digital television, and space-time processing. He is a senior member of the IEEE and a member of the IEEE SP4COM technical committee. He is editor for *IEEE Transaction on Signal Processing*, *IEEE Communications Letters*, and *IEEE Transactions on Communications*. He is co-author of the book *Dopler Applications for LEO Satellite System* (Kluwer 2001).

Robert Calderbank is vice president for research at AT&T. He also is responsible for directing the research program in Internet and network systems. This program provides AT&T with technical and industry leadership in all areas of networking technology. These areas include network security, content distribution, operations support, network measurement and management, and end-to-end optical systems.

Dr. Calderbank is an IEEE and AT&T Fellow, and a recipient of the IEEE Third Millennium Medal for his contributions to digital communications. These include the design of high-speed voiceband modems, the development of advanced read channels for magnetic disk storage, and the invention of space-time codes, a breakthrough wireless technology that uses a small number of antennas to significantly improve throughput and reliability.

Ayman Naguib received the B.Sc. Degree (with honors) and the M.S. degree in electrical engineering from Cairo University, Cario, Egypt, in 1987 abd 1990, respectively, and the M.S. degreee in statistics and the Ph.D. degree in electrical engieering from Stanford University, Stanford, California, in 1993 and 1996, respectively.

From 1987 to 1989, he spent his military service at the Signal Processing Laboratory, The Military Technical College, Cairo, Egypt. From 1989 to 1990, he was employed at Cairo University as a research and teaching assistant in the Communication Theory Group, Department of Electrical Engineering. From 1990 to 1995, he was a research and teaching assistant in the Information Systems Laboratories, Stanford University, Stanford California. In 1996, he joined AT&T Labs, Florham Park, New Jersey, as a principal member of technical staff. In September 2000, he joined Morphics Technology Inc. as a technical leader. His current research interests include in general space-time signal processing and coding for high data rate wireless communications (W-CDMA, OFDM, etc.).

BIBLIOGRAPHY

1. Special Issue on the European Path Towards UMTS, *IEEE Pers. Commun. Mag.* **2**: (Feb. 1995).
2. D. J. Goodman, Trends in cellular and cordless communications, *IEEE Commun. Mag.* **29**: 31–40 (June 1991).
3. J. H. Winters, Optimum combining in digital mobile radio with cochannel interference, *IEEE J. Select. Areas Commun.* **JSAC-2**(4): 528–539 (July 1984).
4. J. H. Winters, Optimum combining for indoor radio systems with multiple users, *IEEE Trans. Commun.* **COM-35**(11): 1222–1230 (Nov. 1987).
5. J. H. Winters, On the capacity of radio communication systems with diversity in a Rayleigh fading environment, *IEEE J. Select. Areas Commun.* **JSAC-5**(5): 871–878 (June 1987).
6. P. Balaban and J. Salz, Optimum diversity combining and equalization in digital data transmission with application to cellular mobile radio, *IEEE Trans. Vehic. Technol.* **VT-40**(2): 342–354 (May 1991).
7. P. Balaban and J. Salz, Optimum diversity combining and equalization in data transmission with application to cellular mobile radio — Part I: Theoretical considerations, *IEEE Trans. Commun.* **COM-40**(5): 885–894 (May 1992).
8. P. Balaban and J. Salz, Optimum diversity combining and equalization in data transmission with application to cellular mobile radio — Part II: Numerical results, *IEEE Trans. Commun.* **COM-40**(5): 895–907 (May 1992).
9. G. J. Foschini and M. J. Gans, On limits of wireless communications in a fading environment when using multiple antennas, *Wireless Commun. Mag.* **6**: 311–335 (March 1998).
10. E. Telatar, *Capacity of Multi-Antenna Gaussian Channels*, technical memorandum, AT&T Bell Laboratories, June 1995.
11. G. Foschini, Layered space–time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* **1**: 41–59 (1996).
12. U.S. Patent 5,088,108 (Feb., 1992), J. Uddenfeldt and A. Raith, Cellular digital mobile radio system and method of transmitting information in a digital cellular mobile radio system.
13. A. Wittneben, Base station modulation diversity for digital SIMULCAST, *Proc. IEEE VTC'91*, St. Louis, MO, 1991, Vol. 1, pp. 848–853.
14. N. Seshadri and J. H. Winters, Two schemes for improving the performance of frequency-division duplex (FDD) transmission systems using transmitter antenna diversity, *Int. J. Wireless Inform. Networks* **1**: 49–60 (Jan 1994).
15. A. Wittneben, A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation, *Proc. IEEE ICC'93*, Geneva, Switzerland, 1993, Vol. 3, pp. 1630–1634.
16. J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels, *Proc. IEEE VTC'96*, Atlanta, GA, 1996, Vol. 1, pp. 136–140.
17. J. H. Winters, Diversity gain of transmit diversity in wireless systems with Rayleigh fading, *Proc. IEEE ICC'94*, New Orleans, LA, 1994, Vol. 2, pp. 1121–1125.
18. S. Alamouti, Space block coding: A simple transmitter diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (Oct. 1998).
19. V. Tarokh, H. Jafarkhani, and R. A. Calderbank, Space–time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**: 1456–1467 (July 1999).
20. N. Seshadri, V. Tarokh, and A. R. Calderbank, Space–time codes for high data rate wireless communications: Code

construction, *Proc. IEEE VTC'97*, Phoenix, AZ, 1997, Vol. 2, pp. 637–641.

21. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications: performance criterion and code construction, *Proc. IEEE ICC'97*, Montreal, Canada, 1997, Vol. 1, pp. 299–303.
22. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**: 744–765 (March 1998).
23. V. Tarokh, A. F. Naguib, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications: Mismatch analysis, *Proc. IEEE ICC'97*, Montreal, Canada, 1997, Vol. 1, pp. 309–313.
24. V. Tarokh, A. F. Naguib, N. Seshadri, and A. Calderbank, Space–time codes for high data rate wireless communications: Performance criteria in the presence of channel estimation errors, mobility, and multiple paths, *IEEE Trans. Commun.* **47**: 199–207 (Feb. 1999).
25. A. F. Naguib, V. Tarokh, N. Seshadri, and A. R. Calderbank, A space–time coding based modem for high data rate wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1459–1478 (Oct 1998).
26. W. C. Jakes, *Microwave Mobile Communications*, IEEE Press, 1974.
27. J. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
28. A. F. Naguib and N. Seshadri, Combined interference cancellation and ML decoding of space–time block codes, *IEEE J. Select. Areas Commun.* (in press).
29. T. Rappaport, *Wireless Communications*, IEEE Press, 1996.
30. L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (March 1974).
31. J. Hagenauer, The Turbo principle: Tutorial introduction and state of the art, *Proc. Int. Symp. Turbo Codes*, Sept. 1997, pp. 1–11.
32. C. Douillard et al., Iterative correction of intersymbol interference: Turbo equalization, *Eur. Trans. Telecommun.* 507–511 (Sept.–Oct. 1995).
33. N. Al-Dhahir, FIR channel-shortening equalizers for MIMO ISI channels, *IEEE Trans. Commun.* **50**: 213–218 (Feb. 2001).
34. G. Bauch and N. Al-Dhahir, Iterative equalization and decoding with channel shortening filters for space–time-coded modulation, in *Vehicular Technology Conference Fall*, pp. 1575–1582, 2000.
35. A. Furuskar, S. Mazur, F. Muller, and H. Olofsson, EDGE: Enhanced data rates for GSM and TDMA/136 evolution, *IEEE Pers. Commun. Mag.* 56–66 (June 1999).
36. G. Bauch and A. Naguib, MAP equalization of space–time-coded signals over frequency-selective channels, in *Wireless Communications and Networking Conference*, pp. 261–265, Sept. 1999.
37. V. Franz and J. Anderson, Concatenated decoding with a reduced-search BCJR algorithm, *IEEE J. Select. Areas Commun.* 186–195 (Feb. 1998).
38. C. Fragouli, N. Al-Dhahir, S. Diggavi, and W. Turin, Prefiltered M-BCJR equalizer for frequency-selective channels, in *Conference on Information Sciences and Systems*, March 2001.
39. J. Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, *Global Telecommunications Conf.*, Nov. 1989, pp. 47.1.1–47.1.7.
40. W. V. Etten, Maximum likelihood receiver for multiple channel transmission systems, *IEEE Trans. Commun.* 276–283 (Feb. 1976).
41. A. Naguib and N. Seshadri, MLSE and equalization of space–time-coded signals, in *Vehicular Technology Conference Spring*, pp. 1688–1693, May 2000.
42. A. Duel-Hallen and C. Heegard, Delayed decision-feedback sequence estimation, *IEEE Trans. Commun.* 428–436 (May 1989).
43. N. Al-Dhahir and A. H. Sayed, The finite-length MIMO MMSE-DFE, *IEEE Trans. Signal Process.* 2921–2936 (Oct. 2000).
44. W. Younis and N. Al-Dhahir, FIR prefilter design for MLSE equalization of space–time-coded transmission over multipath fading channels, in *International Symposium on Circuits and Systems*, May 2001, 362–365.
45. S. Weinstein and P. Ebert, Data transmission by frequency-division multiplexing using the discrete fourier transform, *IEEE Trans. Commun.* **19**: 628–634 (Oct. 1971).
46. G. Raleigh and J. Cioffi, Spatio-temporal coding for wireless communication, *IEEE Trans. Commun.* 357–366 (March 1998).
47. D. Agrawal, V. Tarokh, A. Naguib, and N. Seshadri, Space–time-coded OFDM for high data-rate wireless communication over wideband channels, in *Vehicular Technology Conference*, pp. 2232–2236, May 1998.
48. E. Lindskog and A. Paulraj, A transmit diversity scheme for delay spread channels, in *International Conference on Communications*, pp. 307–311, June 2000.
49. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 363–378 (May 1972).
50. J. Salz, Optimum mean-square decision feedback equalization, *Bell Syst. Tech. J.* **52**: 1341–1373 (Oct. 1973).
51. Z. Liu, G. Giannakis, A. Scaglione, and S. Barbarossa, Decoding and equalization of unknown multipath channels based on block precoding and transmit-antenna diversity, *Asilomar Conf. Signals, Systems, and Computers*, 1999, pp. 1557–1561.
52. T. Pollet, M. V. Bladel, and M. Moeneclaey, BER sensitivity of OFDM systems to carrier frequency offset and Wiener phase noise, *IEEE Trans. Commun.* 191–193 (Feb.–April 1995).
53. H. Sari, G. Karam, and I. Jeanclaude, Transmission techniques for digital terrestrial TV broadcasting, *IEEE Commun. Mag.* 100–109 (Feb. 1995).
54. N. Al-Dhahir, *Single-Carrier Frequency-Domain Equalization for Space–Time Block-Coded Transmissions over Frequency-Selective Fading Channels*, IEEE Communications Letters, Vol. 5, July 2001, pp. 304–306.
55. S. Crozier, D. Falconer, and S. Mahmoud, Least sum of squared errors (LSSE) channel estimation, *IEE Proc. Part F*, Aug. 1991, pp. 371–378.
56. D. Chu, Polyphase codes with good periodic correlation properties, *IEEE Trans. Inform. Theory* **IT-18**: 531–532 (July 1972).
57. C. Fragouli, N. Al-Dhahir, and W. Turin, *Channel Estimation for Space–Time-Coded Systems*, AT&T Technical Document 4TKQBS, Feb. 2001.