Broadcasting and Optical Communication Technology

EDITED BY

Richard C. Dorf

The Electrical Engineering Handbook
Third Edition

Broadcasting and Optical Communication Technology

This page intentionally left blank

The Electrical Engineering Handbook Series

Series Editor
Richard C. Dorf
University of California, Davis

Titles Included in the Series

The Electrical Engineering Handbook
Third Edition

Edited by
Richard C. Dorf

Circuits, Signals, and Speech and Image Processing

Electronics, Power Electronics, Optoelectronics, Microwaves, Electromagnetics, and Radar

Sensors, Nanoscience, Biomedical Engineering, and Instruments

Broadcasting and Optical Communication Technology

Computers, Software Engineering, and Digital Devices

Systems, Controls, Embedded Systems, Energy, and Machines

The Electrical Engineering Handbook
Third Edition

Broadcasting and Optical Communication Technology

Edited by

Richard C. Dorf

University of California
Davis, California, U.S.A.

Preface

Purpose

The purpose of *The Electrical Engineering Handbook, 3rd Edition* is to provide a ready reference for the practicing engineer in industry, government, and academia, as well as aid students of engineering. The third edition has a new look and comprises six volumes including:

Circuits, Signals, and Speech and Image Processing
Electronics, Power Electronics, Optoelectronics, Microwaves, Electromagnetics, and Radar
Sensors, Nanoscience, Biomedical Engineering, and Instruments
Broadcasting and Optical Communication Technology
Computers, Software Engineering, and Digital Devices
Systems, Controls, Embedded Systems, Energy, and Machines

Each volume is edited by Richard C. Dorf, and is a comprehensive format that encompasses the many aspects of electrical engineering with articles from internationally recognized contributors. The goal is to provide the most up-to-date information in the classical fields of circuits, signal processing, electronics, electromagnetic fields, energy devices, systems, and electrical effects and devices, while covering the emerging fields of communications, Nanotechnology, biometrics, digital devices, computer engineering, systems, and biomedical engineering. In addition, the final section provides a complete compendium of information regarding physical, chemical, and materials data, as well as widely inclusive information on mathematics. Many articles from this volume and the other five volumes have been completely revised or updated to fit the needs of today, and many new chapters have been added.

The purpose of *Broadcasting and Optical Communication Technology* is to provide a ready reference to subjects in the field of communications, including broadcasting, equalization, optical communication, computer networks, ad hoc wireless networks, information theory, satellites and aerospace, digital video processing, and mobile communications. Here we provide the basic information for understanding these fields. We also provide information about bandwidth modulation, phase-locked loops, telemetry, and computer-aided design and analysis of communication systems.

Organization

The information is organized into two major sections. The first section encompasses 13 chapters and the last section summarizes the applicable mathematics, symbols, and physical constants.

Most articles include three important and useful categories: defining terms, references, and further information. *Defining terms* are key definitions and the first occurrence of each term defined is indicated in boldface in the text. The definitions of these terms are summarized as a list at the end of each chapter or article. The *references* provide a list of useful books and articles for follow-up reading. Finally, *further information* provides some general and useful sources of additional information on the topic.

Locating Your Topic

Numerous avenues of access to information are provided. A complete table of contents is presented at the front of the book. In addition, an individual table of contents precedes both sections. Finally, each chapter

begins with its own table of contents. The reader should look over these tables of contents to become familiar with the structure, organization, and content of the book. For example, see Section I: Communications, then Chapter 1: Broadcasting, and then Chapter 1.1: Modulation and Demodulation. This tree-and-branch table of contents enables the reader to move up the tree to locate information on the topic of interest.

Two indexes have been compiled to provide multiple means of accessing information: the subject index and index of contributing authors. The subject index can also be used to locate key definitions. The page on which the definition appears for each key (defining) term is clearly identified in the subject index.

The Electrical Engineering Handbook, 3rd Edition is designed to provide answers to most inquiries and direct the inquirer to further sources and references. We hope that this handbook will be referred to often and that informational requirements will be satisfied effectively.

Acknowledgments

Richard C. Dorf
Editor-in-Chief

Editor-in-Chief

Richard C. Dorf, professor of electrical and computer engineering at the University of California, Davis, teaches graduate and undergraduate courses in electrical engineering in the fields of circuits and control systems. He earned a Ph.D. in electrical engineering from the U.S. Naval Postgraduate School, an M.S. from the University of Colorado, and a B.S. from Clarkson University. Highly concerned with the discipline of electrical engineering and its wide value to social and economic needs, he has written and lectured internationally on the contributions and advances in electrical engineering.

Professor Dorf has extensive experience with education and industry and is professionally active in the fields of robotics, automation, electric circuits, and communications. He has served as a visiting professor at the University of Edinburgh, Scotland; the Massachusetts Institute of Technology; Stanford University; and the University of California, Berkeley.

Professor Dorf is a Fellow of The Institute of Electrical and Electronics Engineers and a Fellow of the American Society for Engineering Education. Dr. Dorf is widely known to the profession for his *Modern Control Systems, 10th Edition* (Addison-Wesley, 2004) and *The International Encyclopedia of Robotics* (Wiley, 1988). Dr. Dorf is also the co-author of *Circuits, Devices and Systems* (with Ralph Smith), *5th Edition* (Wiley, 1992), and *Electric Circuits, 7th Edition* (Wiley, 2006). He is also the author of *Technology Ventures* (McGraw-Hill, 2005) and *The Engineering Handbook, 2nd Edition* (CRC Press, 2005).

This page intentionally left blank

Advisory Board

This page intentionally left blank

Contributors

Stella N. Batalama
SUNY at Buffalo
Buffalo, New York

Almon H. Clegg
Consultant
Highland, Utah

Thomas M. Cover
Stanford University
Stanford, California

Thomas E. Darcie
AT&T Bell Laboratories
Holmdel, New Jersey

John N. Daigle
University of Mississippi
Oxford, Mississippi

Daniel F. DiFonzo
Planar Communications Corporation
Rockville, Maryland

Richard C. Dorf
University of California
Davis, California

Stephen Horan
New Mexico State University
Las Cruces, New Mexico

Mohammad Ilyas
Florida Atlantic University
Boca Raton, Florida

Dimitri Kazakos
University of Idaho
Moscow, Idaho

Reza Khosravani
Sonoma State University
Rohnert Park, California

Kurt L. Kosbar
University of Missouri
Rolla, Missouri

Carl G. Looney
University of Nevada
Reno, Nevada

Steven L. Maddy
SpectraLink Corporation
Boulder, Colorado

Robert J. Marks II
University of Washington
Seattle, Washington

Stan McClellan
Hewlett Packard Company
Plano, Texas

Sarhan M. Musa
Prairie View A&M University
Prairie View, Texas

Joseph C. Palais
Arizona State University
Tempe, Arizona

Nikos Passas
University of Athens
Athens, Greece

H. Vincent Poor
Princeton University
Princeton, New Jersey

Haoli Qian
Marvell Semiconductor Inc.
Sunnyvale, California

Banmali S. Rawat
University of Nevada
Reno, Nevada

Todd R. Reed
University of Hawaii
Manoa, Hawaii

Richard B. Robrock II
Bell Laboratories Inc.
Bedminster, New Jersey

Martin S. Roden
California State University
Los Angeles, California

Matthew N.O. Sadiku
Prairie View A&M University
Prairie View, Texas

Stanley Salek
Harnett & Edison, Inc.
San Francisco, California

Apostolis K. Salkintzis
Motorola
Athens, Greece

Remzi Seker
University of Arkansas
    at Little Rock
Little Rock, Arkansas

Ronald J. Tallarida
Temple University
Philadelphia, Pennsylvania

Moncef B. Tayahi
University of Nevada
Reno, Nevada

Charles W. Therrien
Naval Postgraduate School
Monterey, California

Joy A. Thomas
Stratify
Mountain View, California

William H. Tranter
Virginia Polytechnic Institute
Blacksburg, Virginia

Sergio Verdú
Princeton University
Princeton, New Jersey

Zhen Wan
University of Texas
Richardson, Texas

Jerry C. Whitaker
Advanced Television Systems
 Committee
Washington, D.C.

Alan E. Willner
University of Southern California
Los Angeles, California

Contents

SECTION II   Mathematics, Symbols, and Physical Constants

Indexes

I

Communications

1
Broadcasting

Richard C. Dorf
University of California

Zhen Wan
University of Texas

Jerry C. Whitaker
*Advanced Television Systems
Committee*

Martin S. Roden
California State University

Stanley Salek
Hammett & Edison, Inc.

Almon H. Clegg
Consultant

1.1 Modulation and Demodulation

Richard C. Dorf and Zhen Wan

Modulation is the process of impressing the source information onto a bandpass signal with a carrier frequency f_c. This bandpass signal is called the modulated signal $s(t)$, and the baseband source signal is called the modulating signal $m(t)$. The modulated signal could be represented by

$$s(t) = \text{Re}\{g(t)e^{j\omega ct}\} \tag{1.1}$$

or, equivalently,

$$s(t) = R(t) \cos\left[\omega_c t + \theta(t)\right] \tag{1.2}$$

and

$$s(t) = x(t) \cos \omega_c t - y(t) \sin \omega_c t \tag{1.3}$$

TABLE 1.1 Complex Envelope Functions for Various Types of Modulation

Type of Modulation	Mapping Functions $g[m]$	Corresponding Quadrature Modulation		Corresponding Amplitude and Phase Modulation		Linearity	Remarks
		$x(t)$	$y(t)$	$R(t)$	$\theta(t)$		
AM	$1 + m(t)$	$1 + m(t)$	0	$\lvert 1 + m(t)\rvert$	$\begin{Bmatrix}0, & m(t) > -1 \\ 180°, & m(t) < -1\end{Bmatrix}$	L[b]	$m(t) > -1$ required for envelope detection
DSB-SC	$m(t)$	$m(t)$	0	$\lvert m(t)\rvert$	$\begin{Bmatrix}0, & m(t) > 0 \\ 180°, & m(t) < 0\end{Bmatrix}$	L	Coherent detection required
PM	$e^{jD_p m(t)}$	$\cos[D_p m(t)]$	$\sin[D_p m(t)]$	1	$D_p m(t)$	NL	D_p is the phase deviation constant (radian/volts)
FM	$e^{jDf \int_{-\infty}^{t} m(\sigma)d\sigma}$	$\cos[D_f \int_{-\infty}^{t} m(\sigma)d\sigma]$	$\sin[D_f \int_{-\infty}^{t} m(\sigma)d\sigma]$	1	$D_f \int_{-\infty}^{t} m(\sigma)d\sigma$	NL	D_f is the frequency deviation constant (radian/volt-sec)
SSB-AM-SC[a]	$m(t) \pm j\hat{m}(t)$	$m(t)$	$\pm\hat{m}(t)$	$\sqrt{[m(t)^2 + [\hat{m}(t)]^2}$	$\tan^{-1}[\pm\hat{m}(t)/m(t)]$	L	Coherent detection required
SSB-PM[a]	$e^{jD_p[m(t)\pm j\hat{m}(t)]}$	$e^{\mp D_p \hat{m}(t)}\cos[D_p m(t)]$	$e^{\mp D_p \hat{m}(t)}\sin[D_p m(t)]$	$e^{\mp D_p \hat{m}(t)}$	$D_p m(t)$	NL	
SSB-FM[a]	$e^{jDf \int_{-\infty}^{t}[m(\sigma)\pm j\hat{m}(\sigma)]d\sigma}$	$e^{\mp D_f \int_{-\infty}^{t} \hat{m}(\sigma)d\sigma}\cos[D_f \int_{-\infty}^{t} m(\sigma)d\sigma]$	$e^{\mp D_f \int_{-\infty}^{t} \hat{m}(\sigma)d\sigma}\sin[D_f \int_{-\infty}^{t} m(\sigma)d\sigma]$	$e^{\mp D_f \int_{-\infty}^{t} \hat{m}(\sigma)d\sigma}$	$D_f \int_{-\infty}^{t} m(\sigma)d\sigma$	NL	
SSB-EV[a]	$e^{\{\ln[1+m(t)]\pm j\widehat{\ln}[1+m(t)]\}}$	$[1+m(t)]\cos\{\widehat{\ln}[1+m(t)]\}$	$\pm[1+m(t)]\sin\{\widehat{\ln}[1+m(t)]\}$	$1+m(t)$	$\pm\widehat{\ln}[1+m(t)]$	NL	$m(t) > -1$ is required so that the ln will have a real value
SSB-SQ[a]	$e^{(1/2)\{\ln[1+m(t)]\pm j\widehat{\ln}[1+m(t)]\}}$	$\sqrt{1+m(t)}\cos\{\frac{1}{2}\widehat{\ln}[1=m(t)]\}$	$\pm\sqrt{1+m(t)}\sin\{\frac{1}{2}\widehat{\ln}[1+m(t)]\}$	$\sqrt{1+m(t)}$	$\pm\frac{1}{2}\widehat{\ln}[1+m(t)]$	NL	$m(t) > -1$ is required so that the ln will have a real value
QM	$m_1(t)+jm_2(t)$	$m_1(t)$	$m_2(t)$	$\sqrt{m_1^2(t)+m_2^2(t)}$	$\tan^{-1}[m_2(t)/m_1(t)]$	L	Used in NTSC color television: requires coherent detection

L = linear, NL = nonlinear, $\hat{[.]}$ is the Hilbert transform (i.e., $-90°$ phase-shifted version) of $[\cdot]$. The Hilbert transform is $\hat{x}(t) \overset{\Delta}{=} x(t) \star \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\lambda)}{t-\lambda}d\lambda$

[a]Use upper signs for upper sideband signals and lower signs for lower sideband signals.

[b]In the strict sense, AM signals are not linear because the carrier term does not satisfy the linearity (superposition) condition.

Source: L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990. With permission.

where $\omega_c = 2\pi f_c$. The complex envelope is

$$g(t) = R(t)e^{j\theta(t)} = x(t) + jy(t) \tag{1.4}$$

and $g(t)$ is a function of the modulating signal $m(t)$. That is,

$$g(t) = g[m(t)]$$

Thus $g[\cdot]$ performs a mapping operation on $m(t)$. The particular relationship that is chosen for $g(t)$ in terms of $m(t)$ defines the type of modulation used.

In Table 1.1, examples of the mapping function $g(m)$ are given for the following types of modulation:

- AM: amplitude modulation
- DSB-SC: double-sideband suppressed-carrier modulation
- PM: phase modulation
- FM: frequency modulation
- SSB-AM-SC: single-sideband AM suppressed-carrier modulation
- SSB-PM: single-sideband PM
- SSB-FM: single-sideband FM
- SSB-EV: single-sideband envelope-detectable modulation
- SSB-SQ: single-sideband square-law-detectable modulation
- QM: quadrature modulation

Modulation

In Table 1.1, a generalized approach may be taken to obtain universal transmitter models that may be reduced to those used for a particular modulation type. We also see that there are equivalent models which correspond to different circuit configurations, yet they may be used to produce the same type of modulated signal at their outputs. It is up to communication engineers to select an implementation method that will optimize performance, yet retain low cost based on the state of the art in circuit development.

There are two canonical forms for the generalized transmitter. Figure 1.1 is an AM-PM type circuit as described in Equation (1.2). In this figure, the baseband signal processing circuit generates $R(t)$ and $\theta(t)$ from $m(t)$. The R and θ are functions of the modulating signal $m(t)$ as given in Table 1.1 for the particular modulation type desired.
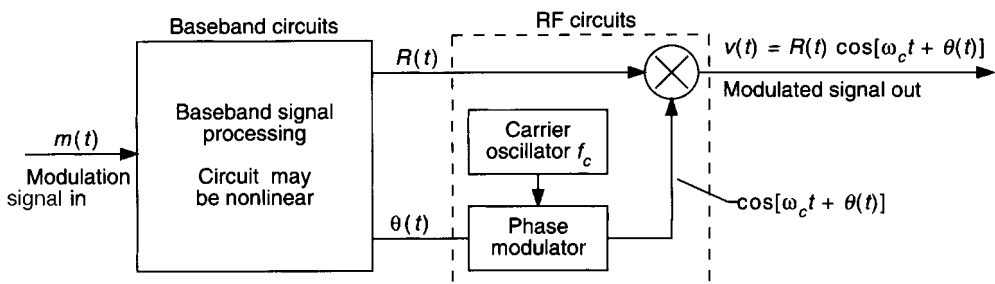
FIGURE 1.1 Generalized transmitter using the AM-PM generation technique. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 279. With permission.)

Figure 1.2 illustrates the second canonical form for the generalized transmitter. This uses in-phase and quadrature-phase (IQ) processing. Similarly, the formulas relating $x(t)$ and $y(t)$ are shown in Table 1.1, and the baseband signal processing may be implemented by using either analog hardware or digital hardware with software. The remainder of the canonical form utilizes radio frequency (RF) circuits as indicated.

Any type of signal modulation (AM, FM, SSB, QPSK, etc.) may be generated by using either of these two canonical forms. Both of these forms conveniently separate baseband processing from RF processing.

Superheterodyne Technique

Most receivers employ the **superheterodyne receiving** technique (see Figure 1.3). This technique consists of either down-converting or up-converting the input signal to some convenient frequency band, called the *intermediate frequency* (IF) band, and then extracting the information (or modulation) by using the appropriate detector. This basic receiver structure is used for the reception of all types of bandpass signals, such as television, FM, AM, satellite, and radar signals.
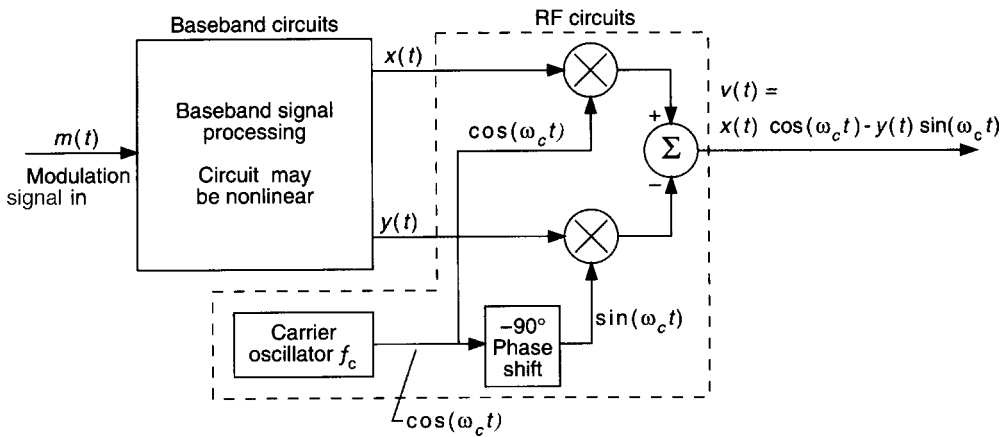
FIGURE 1.2 Generalized transmitter using the quadrature generation technique. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 280. With permission.)
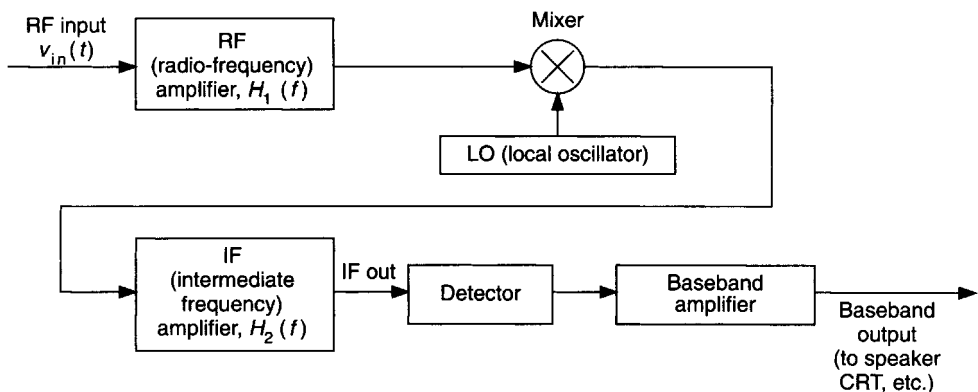
FIGURE 1.3 Superheterodyne receiver. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 281. With permission.)

If the complex envelope $g(t)$ is desired for generalized signal detection or for optimum reception in digital systems, the $x(t)$ and $y(t)$ quadrature components, where $x(t) + jy(t) = g(t)$, may be obtained by using quadrature product detectors, as illustrated in Figure 1.4. $x(t)$ and $y(t)$ could be fed into a signal processor to extract the modulation information. Disregarding the effects of noise, the signal processor could recover $m(t)$ from $x(t)$ and $y(t)$ (and, consequently, demodulate the IF signal) by using the inverse of the complex envelope generation functions given in Table 1.1.

The generalized modulation techniques are shown in Table 1.1. In digital communication systems, discrete modulation techniques are usually used to modulate the source information signal. Discrete modulation includes:

- PCM = pulse-code modulation
- DM = differential modulation
- DPCM = differential pulse-code modulation
- FSK = frequency-shift keying
- PSK = phase-shift keying
- DPSK = differential phase-shift keying
- MPSK = M-ary phase-shift keying
- QAM = quadrature amplitude modulation
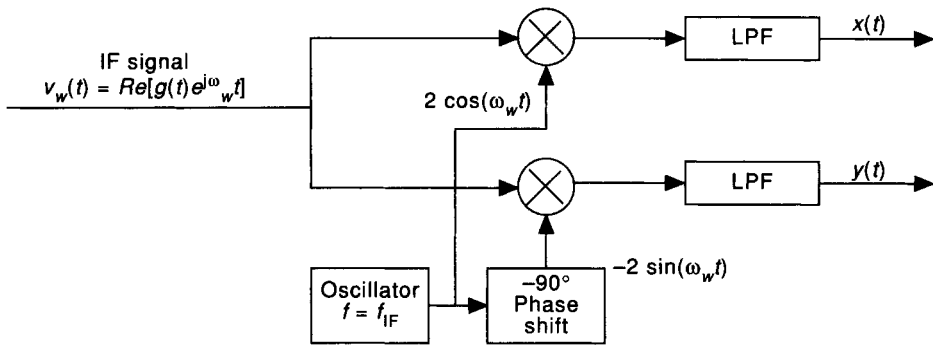
FIGURE 1.4 IQ (in-phase and quadrature-phase) detector. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 284. With permission.)

TABLE 1.2 Performance of a PCM System with Uniform Quantizing and No Channel Noise

Number of Quantizer Levels Used, M	Length of the PCM Word, n (bits)	Bandwidth of PCMSignal (First Null Bandwidth)^n	Recovered Analog Signal Power-to-Quantizing Noise Power Ratios	
			$(S/N)_{pk \text{ out}}$	$(S/N)_{\text{out}}$
2	1	$2B$	10.8	6.0
4	2	$4B$	16.8	12.0
8	3	$6B$	22.8	18.1
16	4	$8B$	28.9	24.1
32	5	$10B$	34.9	30.1
64	6	$12B$	40.9	36.1
128	7	$14B$	46.9	42.1
256	8	$16B$	52.9	48.2
512	9	$18B$	59.0	54.2
1024	10	$20B$	65.0	60.2

[a]B is the absolute bandwidth of the input analog signal.

Pulse-Code Modulation

PCM is essentially analog-to-digital conversion of a special type, where the information contained in the instantaneous samples of an analog signal is represented by digital words in a serial bit stream. The PCM signal is generated by carrying out three basic operations: sampling, quantizing, and encoding (see Figure 1.5). The sampling operation generates a flat-top pulse amplitude modulation (PAM) signal. The quantizing converts the actual sampled value into the nearest of the M amplitude levels. The PCM signal is obtained from the quantized PAM signal by encoding each quantized sample value into a digital word.

Frequency-Shift Keying

The FSK signal can be characterized as one of two different types. One type is called *discontinuous-phase* FSK since $\theta(t)$ is discontinuous at the switching times. The discontinuous-phase FSK signal is represented by

$$s(t) = \begin{cases} A_c\cos(\omega_1 t + \theta_1) & \text{for } t \text{ in time interval when a binary 1 is sent} \\ A_c\cos(\omega_2 t + \theta_2) & \text{for } t \text{ in time interval when a binary 0 is sent} \end{cases} \tag{1.5}$$

where f_1 is called the mark (binary 1) frequency and f_2 is called the space (binary 0) frequency. The other type is continuous-phase FSK. The continuous-phase FSK signal is generated by feeding the data signal into a frequency modulator, as shown in Figure 1.6(b). This FSK signal is represented by

$$s(t) = A_c\cos\left[\omega_c t + D_f \int_{-\infty}^{t} m(\lambda)\mathrm{d}\lambda\right]$$

or

$$s(t) = \mathrm{Re}\{g(t)e^{j\omega ct}\} \tag{1.6}$$
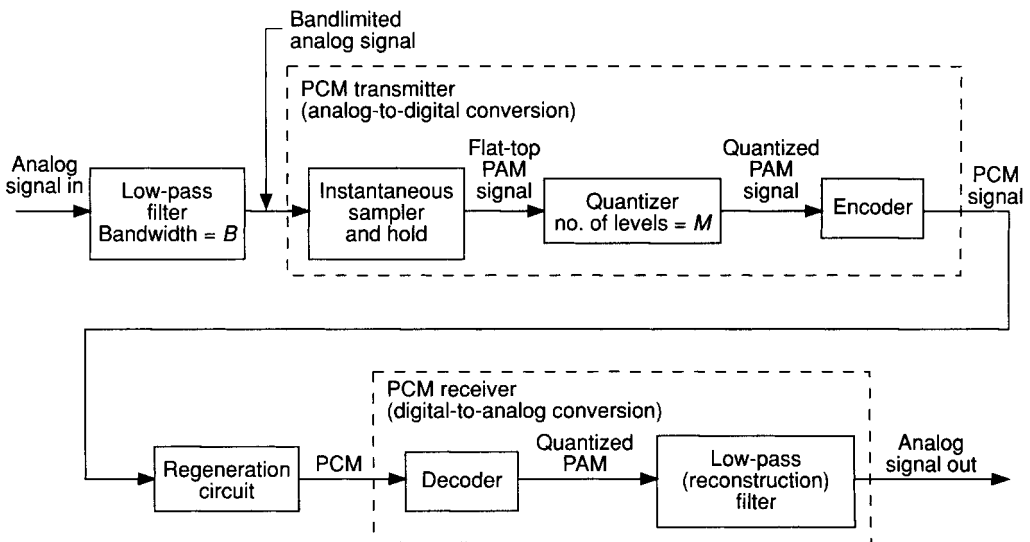
FIGURE 1.5 A PCM transmission system.
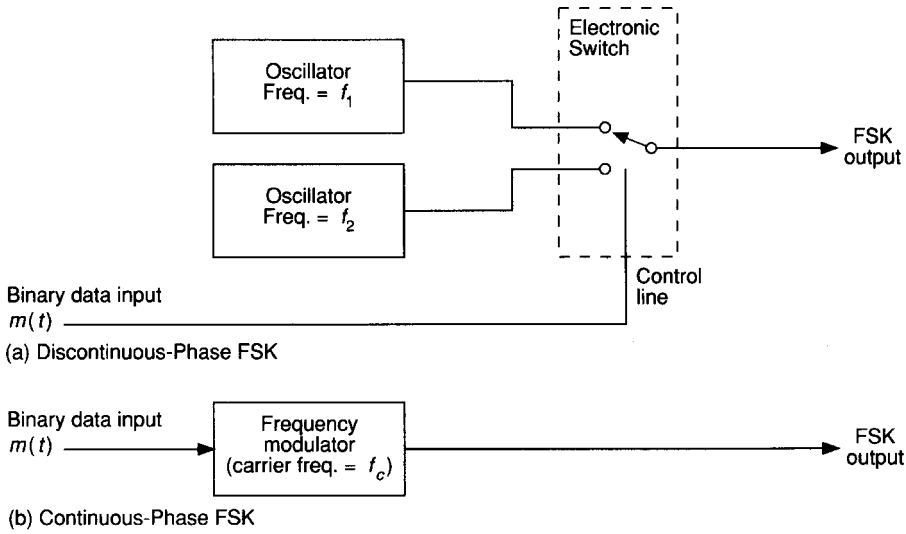
FIGURE 1.6 Generation of FSK. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 337. With permission.)
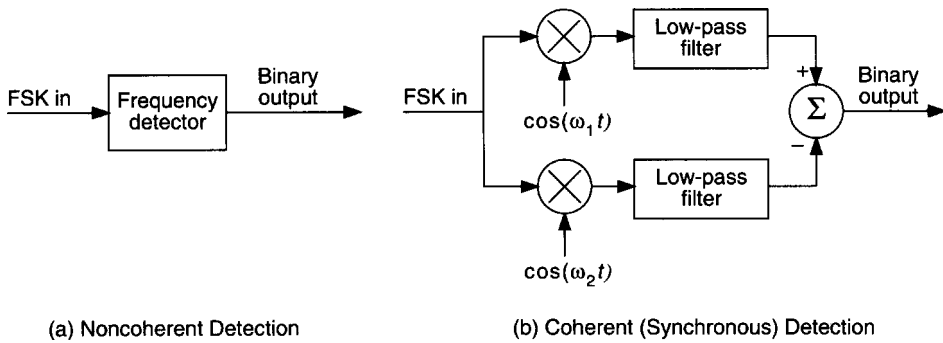
FIGURE 1.7 Detection of FSK. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 344. With permission.)

where

$$g(t) = A_c e^{j\theta(t)} \tag{1.7}$$

$$\theta(t) = D_f \int\limits_{-\infty}^{t} m(\lambda)d\lambda \quad \text{for FSK} \tag{1.8}$$

Detection of FSK is illustrated in Figure 1.7.

M-ary Phase-Shift Keying

If the transmitter is a PM transmitter with an *M*-level digital modulation signal, MPSK is generated at the transmitter output. A plot of the permitted values of the complex envelope, $g(t) = A_c e^{j\theta(t)}$, would contain M

points, one value of g (a complex number in general) for each of the M multilevel values, corresponding to the M phases that θ is permitted to have.

MPSK can also be generated using two quadrature carriers modulated by the x and y components of the complex envelope (instead of using a phase modulator)

$$g(t) = A_c e^{j\theta(t)} = x(t) + jy(t) \qquad (1.9)$$

where the permitted values of x and y are

$$x_i = A_c \cos \theta_i \qquad (1.10)$$

$$y_i = A_c \sin \theta_i \qquad (1.11)$$

for the permitted phase angles θ_i, $i = 1, 2, \ldots, M$, of the MPSK signal. This is illustrated by Figure 1.8, where the signal processing circuit implements Equation (1.10) and Equation (1.11).

MPSK, where $M = 4$, is called quadrature-phase-shift-keyed (QPSK) signaling.

Quadrature Amplitude Modulation

Quadrature carrier signaling is called quadrature amplitude modulation (QAM). In general, QAM signal constellations are not restricted to having permitted signaling points only on a circle (of radius A_c, as was the

(a) Modulator for Generalized Signal Constellation

(b) Modulator for Rectangular Signal Constellation
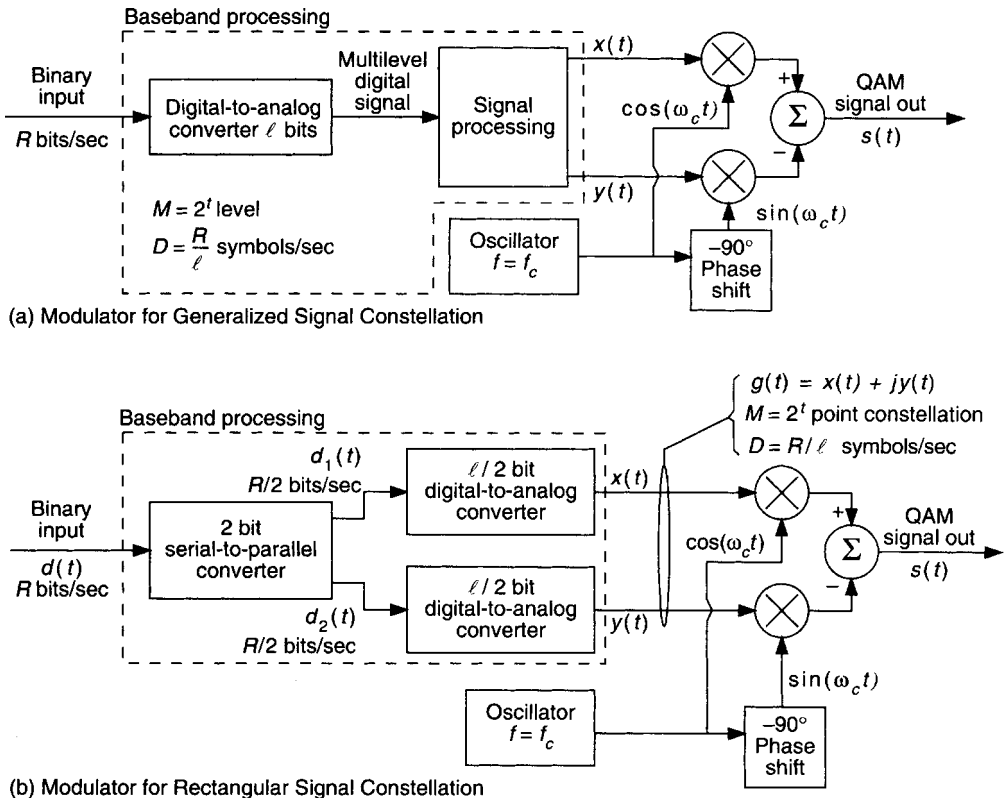
FIGURE 1.8 Generation of QAM signals. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990, p. 346. With permission.)

TABLE 1.3 Spectral Efficiency for QAM Signaling with Raised Cosine-Roll-Off Pulse Shaping

Number of Levels, M (symbols)	Size of DAC, ℓ (bits)	$\eta = \dfrac{R}{B_{\mathrm{T}}} \dfrac{\text{bits/s}}{\text{Hz}}$					
		$r = 0.0$	$r = 0.1$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$
2	1	1.00	0.909	0.800	0.667	0.571	0.500
4	2	2.00	1.82	1.60	1.33	1.14	1.00
8	3	3.00	2.73	2.40	2.00	1.71	1.50
16	4	4.00	3.64	3.20	2.67	2.29	2.00
32	5	5.00	4.55	4.0	3.33	2.86	2.50

DAC = digital-to-analog converter.

$\eta = R/B_{\mathrm{T}} = \ell/2$ bits/s per hertz.

r is the roll-off factor of the filter characteristic.

Source: L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan 1990, p. 350. With permission.

case for MPSK). The general QAM signal is

$$s(t) = x(t) \cos \omega_{\mathrm{c}} t - \gamma(t) \sin \omega_{\mathrm{c}} t \tag{1.12}$$

where

$$g(t) = x(t) + jy(t) = R(t) e^{j\theta(t)} \tag{1.13}$$

The generation of QAM signals is shown in Figure 1.8. The spectral efficiency for QAM signaling is shown in Table 1.3.

Defining Terms

Modulation: The process of impressing the source information onto a bandpass signal with a carrier frequency f_{c}. It can be expressed as

$$s(t) = \mathrm{Re}\{g(t) e^{j\omega ct}\}$$

where $g(t)$ is a function of the modulating signal $m(t)$. That is,

$$g(t) = g[m(t)]$$

$g[\cdot]$ performs a mapping operation on $m(t)$. The particular relationship that is chosen for $g(t)$ in terms of $m(t)$ defines the type of modulation used.

Superheterodyne receiver: Most receivers employ the superheterodyne receiving technique, which consists of either down-converting or up-converting the input signal to some convenient frequency band, called the intermediate frequency band, and then extracting the information (or modulation) by using an appropriate detector. This basic receiver structure is used for the reception of all types of bandpass signals, such as television, FM, AM, satellite, and radar signals.

References

L.W. Couch, *Digital and Analog Communication Systems*, New York: Prentice-Hall, 1995.

F. Dejager, "Delta modulation of PCM transmission using a 1-unit code," Phillips Res. Rep., no. 7, pp. 442–466, Dec. 1952.

J.H. Downing, *Modulation Systems and Noise*, Englewood Cliffs, N.J.: Prentice-Hall, 1964.
J. Dunlop and D.G. Smith, *Telecommunications Engineering*, London: Van Nostrand, 1989.
B.P. Lathi, *Modern Digital and Analog Communication Systems*, New York: CBS College, 1983.
J.H. Park, Jr., "On binary DPSK detection," *IEEE Trans. Commun.*, COM-26, pp. 484–486, 1978.
M. Schwartz, *Information Transmission, Modulation and Noise*, New York: McGraw-Hill, 1980.

Further Information

The monthly journal *IEEE Transactions on Communications* describes telecommunication techniques. The performance of *M*-ary QAM schemes is evaluated in its March 1991 issue, pp. 405–408. The IEEE magazine *IEEE Communications* is a valuable source.

Another source is *IEEE Transactions on Broadcasting*, which is published quarterly by The Institute of Electrical and Electronics Engineers, Inc.

The biweekly magazine *Electronics Letters* investigates the error probability of coherent PSK and FSK systems with multiple co-channel interferences in its April 11, 1991, issue, pp. 640–642. Another relevant source regarding the coherent detection of MSK is described on pp. 623–625 of the same issue. All subscriptions inquiries and orders should be sent to IEE Publication Sales, P.O. Box 96, Stevenage, Herts, SG1 2SD, United Kingdom.

1.2   Radio Broadcasting

Jerry C. Whitaker

Broadcasting has been around for a long time. Amplitude modulation (AM) was the first modulation system that permitted voice communications to take place. This simple modulation system was predominant throughout the 1920s and 1930s. Frequency modulation (FM) came into regular broadcast service during the 1940s. Television broadcasting, which uses amplitude modulation for the visual portion of the signal and frequency modulation for the aural portion of the signal, became available to the public in the mid-1940s. More recently, digital television (DTV) service has been launched in the United States and elsewhere using the conventional television frequency bands and 6-MHz bandwidth of the analog system, but with digital modulation.

AM Radio Broadcasting

AM radio stations operate on 10-kHz channels spaced evenly from 540 to 1600 kHz. Various classes of stations have been established by the Federal Communications Commission (FCC) and agencies in other countries to allocate the available spectrum to given regions and communities. In the United States, the basic classes are *clear*, *regional*, and *local*. Current practice uses the CCIR (international) designations as class A, B, and C, respectively. Operating power levels range from 50 kW for a clear channel station to as little as 250 W for a local station.

High-Level AM Modulation

High-level anode modulation is the oldest and simplest way of generating a high power AM signal. In this system, the modulating signal is amplified and combined with the dc supply source to the anode of the final RF amplifier stage. The RF amplifier is normally operated class C. The final stage of the modulator usually consists of a pair of tubes operating class B in a push–pull configuration. A basic high-level modulator is shown in Figure 1.9.

The RF signal is normally generated in a low-level transistorized oscillator. It is then amplified by one or more solid-state or vacuum tube stages to provide final RF drive at the appropriate frequency to the grid of the
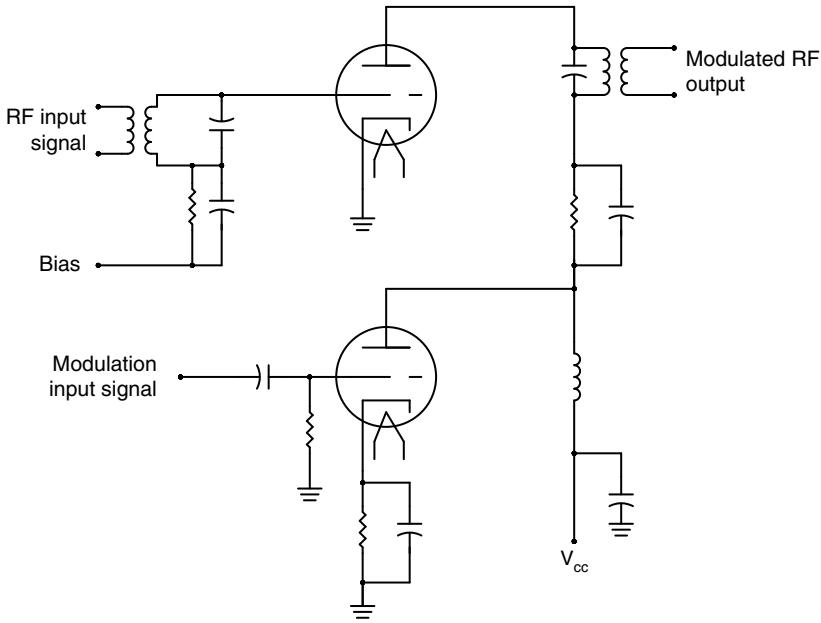
FIGURE 1.9   Simplified diagram of a high-level, amplitude-modulated amplifier.

final class C amplifier. The audio input is applied to an intermediate power amplifier (usually solid state) and used to drive two class B (or class AB) push–pull output devices. The final amplifiers provide the necessary modulating power to drive the final RF stage. For 100% modulation, this modulating power is equal to 50% of the actual carrier power.

The modulation transformer shown in Figure 1.9 does not usually carry the dc supply current for the final RF amplifier. The modulation reactor and capacitor shown provide a means to combine the audio signal voltage from the modulator with the dc supply to the final RF amplifier. This arrangement eliminates the necessity of having dc current flow through the secondary of the modulation transformer, which would result in magnetic losses and saturation effects. In some newer transmitter designs, the modulation reactor has been eliminated from the system, thanks to improvements in transformer technology.

The RF amplifier normally operates class C with grid current drawn during positive peaks of the cycle. Typical stage efficiency is 75 to 83%. An RF tank following the amplifier resonates the output signal at the operating frequency and, with the assistance of a low-pass filter, eliminates harmonics of the amplifier caused by class C operation.

This type of system was popular in AM broadcasting for many years, primarily because of its simplicity. The primary drawback is low overall system efficiency. The class B modulator tubes cannot operate with greater than 50% efficiency. Still, with inexpensive electricity, this was not considered to be a significant problem. As energy costs increased, however, more efficient methods of generating high-power AM signals were developed. Increased efficiency normally came at the expense of added technical complexity.

Pulse-Width Modulation

Pulse-width modulation (PWM), also known as pulse-duration modulation (PDM), is one of the most popular systems developed for modern vacuum tube AM transmitters. Figure 1.10 shows the basic PDM scheme. The PDM system works by utilizing a square-wave switching system, illustrated in Figure 1.11.
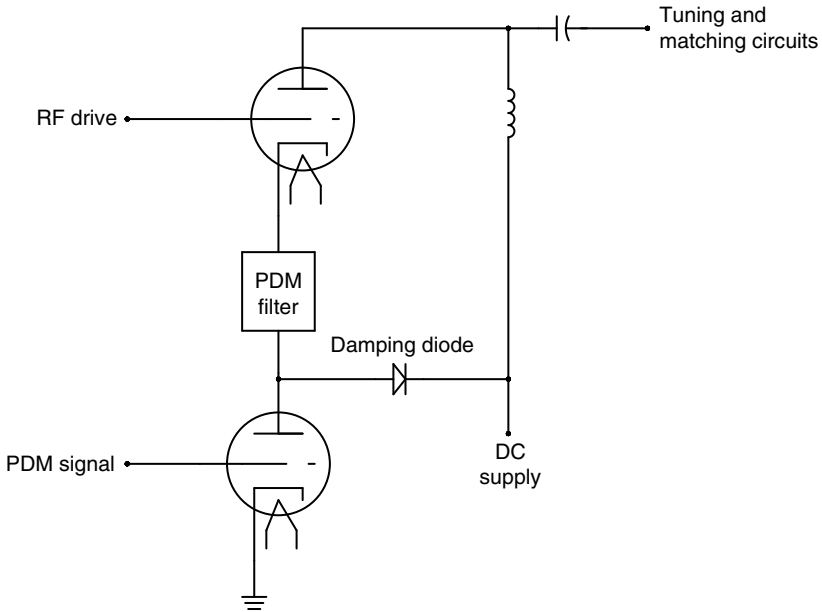
FIGURE 1.10   The pulse-duration modulation (PDM) method of pulse-width modulation.
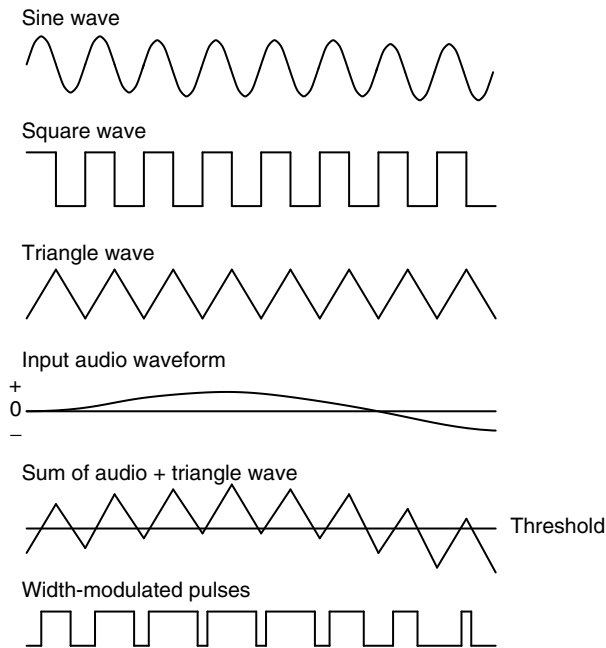
FIGURE 1.11   The principles waveforms of the PDM system.

The PDM process begins with a signal generator (see Figure 1.12). A 75-kHz sine wave is produced by an oscillator and used to drive a square-wave generator, resulting in a simple 75-kHz square wave. The square wave is then integrated, resulting in a triangular waveform that is mixed with the input audio in a summing circuit. The resulting signal is a triangular waveform that rides on the incoming audio. This composite signal is
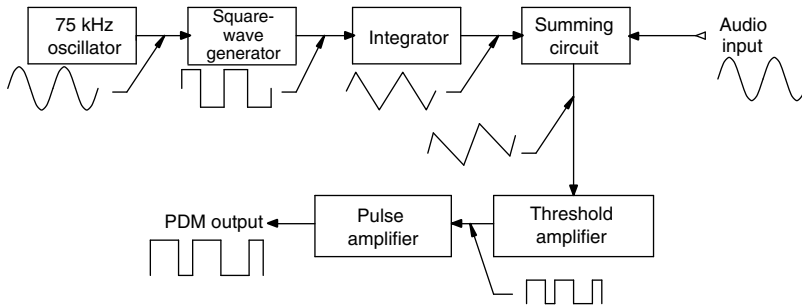
FIGURE 1.12   Block diagram of a PDM waveform generator.

then applied to a threshold amplifier, which functions as a switch that is turned on whenever the value of the input signal exceeds a certain limit. The result is a string of pulses in which the width of the pulse is proportional to the period of time the triangular waveform exceeds the threshold. The pulse output is applied to an amplifier to obtain the necessary power to drive subsequent stages. A filter eliminates whatever transients may exist after the switching process is complete.

The PDM scheme is, in effect, a digital modulation system with the audio information being sampled at a 75-kHz rate. The width of the pulses contains all the audio information. The pulse-width-modulated signal is applied to a *switch* or *modulator tube*. The tube is simply turned *on*, to a fully saturated state, or *off* in accordance with the instantaneous value of the pulse. When the pulse goes positive, the modulator tube is turned on and the voltage across the tube drops to a minimum. When the pulse returns to its minimum value, the modulator tube turns off.

This PDM signal becomes the power supply to the final RF amplifier tube. When the modulator is switched on, the final amplifier will experience current flow and RF will be generated. When the switch or modulator tube goes off, the final amplifier current will cease. This system causes the final amplifier to operate in a highly efficient class D switching mode. A dc offset voltage to the summing amplifier is used to set the carrier (no modulation) level of the transmitter.

A high degree of third-harmonic energy will exist at the output of the final amplifier because of the switching-mode operation. This energy is eliminated by a third-harmonic trap. The result is a stable amplifier that normally operates in excess of 90% efficiency. The power consumed by the modulator and its driver is usually a fraction of a full class B amplifier stage.

The damping diode shown in the previous figure is included to prevent potentially damaging transient overvoltages during the switching process. When the switching tube turns off the supply current during a period when the final amplifier is conducting, the high current through the inductors contained in the PDM filters could cause a large transient voltage to be generated. The energy in the PDM filter is returned to the power supply by the damping diode. If no alternative route is established, the energy will return by arcing through the modulator tube itself.

The PWM system makes it possible to completely eliminate audio frequency transformers in the transmitter. The result is wide frequency response and low distortion. It should be noted that variations on this amplifier and modulation scheme have been used by other manufacturers for both standard broadcast and shortwave service.

Digital Modulation

Current transmitter design work for AM broadcsting has focused almost exclusively on solid-state technology. High-power MOSFET devices and digital modulation techniques have made possible a new generation of energy-efficient systems, with audio performance that easily surpasses vacuum tube designs.

Most solid-state AM systems operate in a highly efficient class D switching mode. Multiple MOSFET driver boards are combined through one of several methods to achieve the required carrier power.

Shortwave Broadcasting

The technologies used in commercial and government-sponsored shortwave broadcasting are closely allied with those used in AM radio. However, shortwave stations usually operate at significantly higher powers than AM stations.

International broadcast stations use frequencies ranging from 5.95 to 26.1 MHz. The transmissions are intended for reception by the general public in foreign countries. Table 1.4 shows the frequencies assigned by the Federal Communications Commission (FCC) for international broadcast shortwave service in the United States. The minimum output power is 50 kW. Assignments are made for specific hours of operation at specific frequencies.

TABLE 1.4   Operating Frequency Bands for Shortwave Broadcasting

Band	Frequnency (kHz)	Meter Band (m)
A	5,950–6,200	49
B	9,500–9,775	32
C	11,700–11,975	25
D	15,100–15,450	19
E	17,700–17,900	16
F	21,450–21,750	14
G	25,600–26,100	11

Very high-power shortwave transmitters have been installed to serve large geographical areas and to overcome jamming efforts by foreign governments. Systems rated for power outputs of 500 kW and more are not uncommon. RF circuits designed specifically for high power operation are utilized.

Most shortwave transmitters have the unique requirement for automatic tuning to one of several preset operating frequencies. A variety of schemes exist to accomplish this task, including multiple exciters (each set to the desired operating frequency) and motor-controlled variable inductors and capacitors. Tune-up at each frequency is performed by the transmitter manufacturer. The settings of all tuning controls are stored in memory. Automatic retuning of a high-power shortwave transmitter can be accomplished in less than 30 seconds in most cases.

Power Amplifier Types

Shortwave technology has advanced significantly within the last 5 years, thanks to improved semiconductor devices. High-power MOSFETs and other components have made solid-state shortwave transmitters operating at 500 kW and more practical. The majority of shortwave systems now in use, however, use vacuum tubes as the power-generating element. The efficiency of a power amplifier/modulator for shortwave applications is of critical importance. Because of the power levels involved, low efficiency translates into higher operating costs.

Older, traditional tube-type shortwave transmitters typically utilize one of the following modulation systems:

- Doherty amplifier
- Chireix outphasing modulated amplifier
- Dome modulated amplifier
- Terman-Woodyard modulated amplifier

FM Radio Broadcasting

FM radio stations operate on 200-kHz channels spaced evenly from 88.1 to 107.9 MHz. In the United States, channels below 92.1 MHz are reserved for noncommercial, educational stations. The FCC has established three classifications for FM stations operating east of the Mississippi River and four classifications for stations west of the Mississippi. Power levels range from a high of 100 kW *effective radiated power* (ERP) to 3 kW or less for lower classifications. The ERP of a station is a function of transmitter power output (TPO) and antenna gain. ERP is determined by multiplying these two quantities together and allowing for line loss.

A transmitting antenna is said to have "gain" if, by design, it concentrates useful energy at low radiation angles, rather than allowing a substantial amount of energy to be radiated above the horizon (and be lost

in space). FM and TV transmitting antennas are designed to provide gain by stacking individual radiating elements vertically.

At first examination, it might seem reasonable and economical to achieve licensed ERP using the lowest transmitter power output possible and highest antenna gain. Other factors, however, come into play that make the most obvious solution not always the best solution. Factors that limit the use of high-gain antennas include:

- Effects of high-gain designs on coverage area and signal penetration
- Limitations on antenna size because of tower restrictions, such as available vertical space, weight, and windloading
- Cost of the antenna

Stereo broadcasting is used almost universally in FM radio today. Introduced in the mid-1960s, stereo has contributed in large part to the success of FM radio. The left and right sum (monophonic) information is transmitted as a standard frequency-modulated signal. Filters restrict this *main channel* signal to a maximum of about 17 kHz. A pilot signal is transmitted at low amplitude at 19 kHz to enable decoding at the receiver. The left and right difference signal is transmitted as an amplitude-modulated subcarrier that frequency-modulates the main FM carrier. The center frequency of the subcarrier is 38 kHz. Decoder circuits in the FM receiver matrix the sum and difference signals to reproduce the left and right audio channels. Figure 1.13 illustrates the baseband signal of a stereo FM station.

Modulation Circuits

Early FM transmitters used *reactance modulators* that operated at low frequency. The output of the modulator was then multiplied to reach the desired output frequency. This approach was acceptable for monaural FM transmission but not for modern stereo systems or other applications that utilize subcarriers on the FM broadcast signal. Modern FM systems all utilize what is referred to as *direct modulation*. That is, the frequency modulation occurs in a modulated oscillator that operates on a center frequency equal to the desired transmitter output frequency. In stereo broadcast systems, a composite FM signal is applied to the FM modulator.

Various techniques have been developed to generate the direct-FM signal. One of the most popular uses a variable-capacity diode as the reactive element in the oscillator. The modulating signal is applied to the diode, which causes the capacitance of the device to vary as a function of the magnitude of the modulating signal. Variations in the capacitance cause the frequency of the oscillator to vary. Again, the magnitude of the frequency shift is proportional to the amplitude of the modulating signal, and the rate of frequency shift is equal to the frequency of the modulating signal.

The direct-FM modulator is one element of an FM transmitter exciter, which generates the composite FM waveform. A block diagram of a complete FM exciter is shown in Figure 1.14. Audio inputs of various types (stereo left and right signals, plus subcarrier programming, if used) are buffered, filtered, and preemphasized before being summed to feed the modulated oscillator. It should be noted that the oscillator is not normally coupled directly to a crystal, but a free-running oscillator adjusted as closely as possible to the carrier frequency of the transmitter. The final operating frequency is carefully maintained by an automatic frequency control system employing a *phase-locked loop* (PLL) tied to a reference crystal oscillator or frequency synthesizer.

A solid-state class C amplifier follows the modulated oscillator and raises the operating power of the FM signal to 20 to 30 W. One or more subsequent amplifiers in the transmitter raise the signal power to several hundred watts for application to the final power amplifier stage. Nearly all current high-power FM transmitters utilize solid-state amplifiers up to the final RF stage, which is generally a vacuum tube for operating powers of 20 kW and above. All stages operate in the class C mode. In contrast to AM systems, each stage in an FM power amplifier can operate class C because no information is lost from the frequency-modulated signal due to amplitude changes. As mentioned previously, FM is a constant-power system.
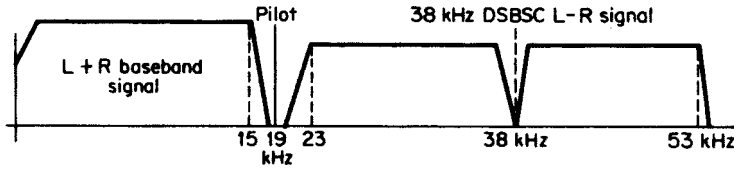
FIGURE 1.13   Composite baseband stereo FM signal.
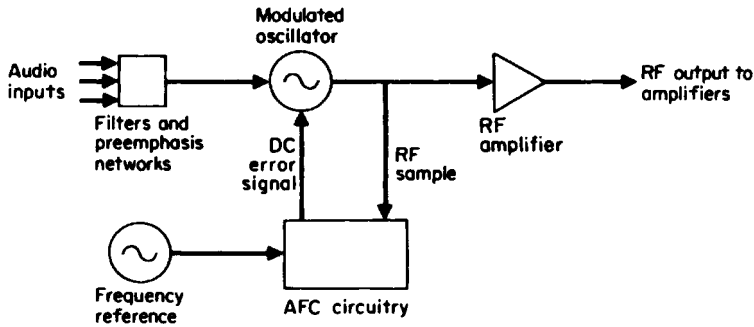
FIGURE 1.14   Block diagram of an FM exciter.

Auxiliary Services

Modern FM broadcast stations are capable of not only broadcasting stereo programming, but one or more subsidiary channels as well. These signals, referred to by the FCC as *Subsidiary Communications Authorization* (SCA) services, are used for the transmission of stock market data, background music, control signals, and other information not normally part of the station's main programming. These services do not provide the same range of coverage or audio fidelity as the main stereo program; however, they perform a public service and can represent a valuable source of income for the broadcaster.

SCA systems provide efficient use of the available spectrum. The most common subcarrier frequency is 67 kHz, although higher subcarrier frequencies may be utilized. Stations that operate subcarrier systems are permitted by the FCC to exceed (by a small amount) the maximum 75-kHz deviation limit under certain conditions. The subcarriers utilize low modulation levels, and the energy produced is maintained essentially within the 200-kHz bandwidth limitation of FM channel radiation.

FM Power Amplifiers

Most high-power FM transmitters manufactured today employ cavity designs. The 1/4-wavelength cavity is the most common. The design is simple and straightforward. A number of variations can be found in different transmitters but the underlying theory of operation is the same. The goal of any cavity amplifier is to simulate a resonant tank circuit at the operating frequency and provide a means to couple the energy in the cavity to the transmission line. Because of the operating frequencies involved (88 to 108 MHz), the elements of the "tank" take on unfamiliar forms.

A typical 1/4-wave cavity is shown in Figure 1.15. The plate of the tube connects directly to the inner section (tube) of the plate-blocking capacitor. The blocking capacitor can be formed in one of several ways. In at least one design, it is made by wrapping the outside surface of the inner tube conductor with multiple layers of insulating film. The exhaust chimney/inner conductor forms the other element of the blocking capacitor. The cavity walls form the outer conductor of the 1/4-wave transmission line circuit. The dc plate voltage is applied to the PA (power amplifier) tube by a cable routed inside the exhaust chimney and inner tube conductor.
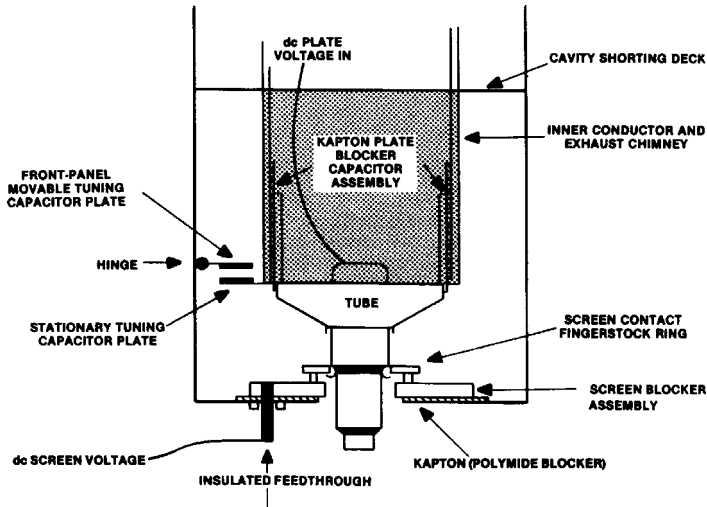
FIGURE 1.15   Physical layout of a common type of 1/4-wave PA cavity for FM broadcast service.

In this design, the screen-contact fingerstock ring mounts on a metal plate that is insulated from the grounded-cavity deck by a blocking capacitor. This hardware makes up the screen-blocker assembly. The dc screen voltage feeds to the fingerstock ring from underneath the cavity deck through an insulated feedthrough.

Some transmitters that employ the 1/4-wave cavity design use a grounded-screen configuration in which the screen contact fingerstock ring is connected directly to the grounded cavity deck. The PA cathode then operates at below ground potential (i.e., at a negative voltage), establishing the required screen voltage for the tube.

Coarse tuning of the cavity is accomplished by adjusting the cavity length. The top of the cavity (the cavity shorting deck) is fastened by screws or clamps and can be raised or lowered to set the length of the assembly for the particular operating frequency. Fine-tuning is accomplished by a variable-capacity plate-tuning control built into the cavity. In the example, one plate of this capacitor, the stationary plate, is fastened to the inner conductor just above the plate-blocking capacitor. The movable tuning plate is fastened to the cavity box, the outer conductor, and is mechanically linked to the front-panel tuning control. This capacity shunts the inner conductor to the outer conductor and varies the electrical length and resonant frequency of the cavity.

AM Broadcast Antenna Systems

Vertical polarization of the transmitted signal is used for AM broadcast stations because of its superior groundwave propagation, and because of the simple antenna designs that it affords. The Federal Communications Commission (FCC) and licensing authorities in other countries have established classifications of AM stations with specified power levels and hours of operation. Protection requirements set forth by the FCC specify that some AM stations (in the United States) reduce their transmitter power at sunset, and return to full power at sunrise. This method of operation is based on the propagation characteristics of AM band frequencies. AM signals propagate further at nighttime than during the day.

The different day/night operating powers are designed to provide each AM station with a specified coverage area that is free from interference. Theory rarely translates into practice insofar as coverage is concerned, however, because of the increased interference that all AM stations suffer at nighttime.

The tower visible at any AM radio station transmitter site is only half of the antenna system. The second element is a buried ground system. Current on a tower does not simply disappear; rather, it returns to Earth

through the capacitance between the Earth and the tower. Ground losses are greatly reduced if the tower has a radial copper ground system. A typical singletower ground system is made up of 120 radial ground wires, each 140 electrical degrees long (at the operating frequency), equally spaced out from the tower base. This is often augmented with an additional 120 interspersed radials 50 ft long.

Directional AM Antenna Design

When a nondirectional antenna with a given power does not radiate enough energy to serve the station's primary service area, or radiates too much energy toward other radio stations on the same or adjacent frequencies, it is necessary to employ a directional antenna system. Rules set out by the FCC and regulatory agencies in other countries specify the protection requirements to be provided by various classes of stations, for both daytime and nighttime hours. These limits tend to define the shape and size of the most desirable antenna pattern.

A directional antenna functions by carefully controlling the amplitude and phase of the RF currents fed to each tower in the system. The directional pattern is a function of the number and spacing of the towers (vertical radiators), and the relative phase and magnitude of their currents. The number of towers in a directional AM array can range from two to six, or even more in a complex system. One tower is defined as the *reference tower*. The amplitude and phase of the other towers are measured relative to this reference.

A complex network of power splitting, phasing, and antenna coupling elements is required to make a directional system work. Figure 1.16 shows a block diagram of a basic two-tower array. A power divider network controls the relative current amplitude in each tower. A phasing network provides control of the phase of each tower current, relative to the reference tower. Matching networks at the base of each tower couple the transmission line impedance to the base operating impedance of the radiating towers.

In practice, the system shown in the figure would not consist of individual elements. Instead, the matching network, power dividing network, and phasing network would all usually be combined into a single unit, referred to as the *phasor*.

Antenna Pattern Design

The pattern of any AM directional antenna system (array) is determined by a number of factors, including:

- Electrical parameters (phase relationship and current ratio for each tower)
- Height of each tower
- Position of each tower with respect to the other towers (particularly with respect to the reference tower)

A *directional array* consists of two or more towers arranged in a specific manner on a plot of land. Figure 1.17 shows a typical three-tower array, as well as the pattern such an array would produce. This is an *in-line array*, meaning that all the elements (towers) are in line with one another. Notice that the *major lobe* is centered on the same line as the line of towers, and that the *pattern nulls* (*minima*) are positioned symmetrically about the line of towers, protecting co-channel stations A and B at true bearings of 315° and 45°, respectively.
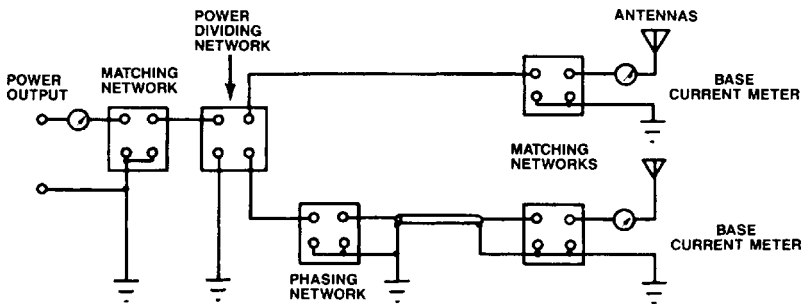
FIGURE 1.16   Block diagram of an AM directional antenna feeder system for a two-tower array.
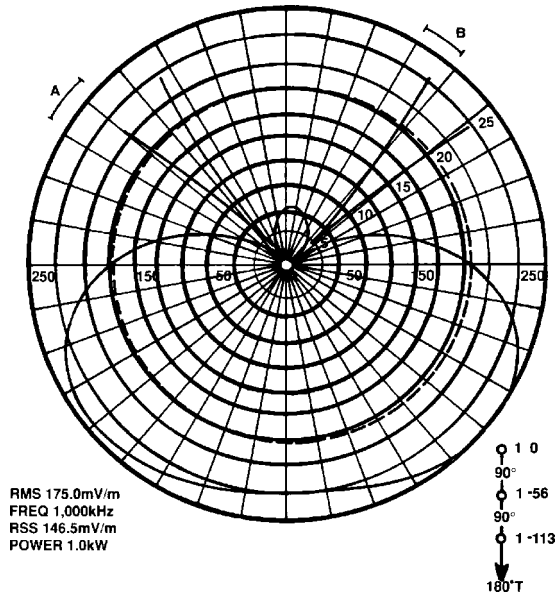
FIGURE 1.17 Radiation pattern generated with a three-tower in-line directional array using the electrical parameters and orientation shown.
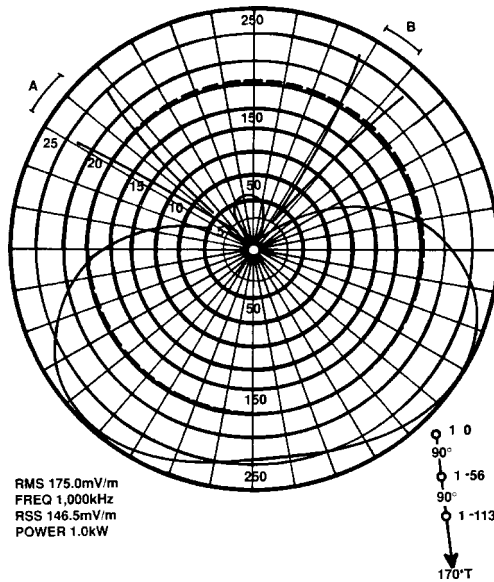
FIGURE 1.18 Radiation pattern produced when the array of Figure 1.13 is rotated to a new orientation.

Figure 1.18 shows the same array, except that it has been rotated by 10°. Notice that the pattern shape is not changed, but the position of the major lobe and the nulls follow the line of towers. Also notice that the nulls are no longer pointed at the stations to be protected.

If this directional antenna system were constructed on a gigantic turntable, the pattern could be rotated without affecting the shape. But, to accomplish the required protections and to have the major lobe(s) oriented in the right direction, there is only one correct position. In most cases, the position of the towers will

be specified with respect to a single reference tower. The location of the other towers will be given in the form of a distance and bearing from that reference. Occasionally, a reference point, usually the center of the array, will be used for a geographic coordinate point.

Bearing

The bearing or azimuth of the towers from the reference tower or point is specified clockwise in degrees from *true north*. The distinction between true and *magnetic north* is vital. The magnetic North Pole is not at the true or geographic North Pole. (In fact, it is in the vicinity of 74° north, 101° west, in the islands of northern Canada.) The difference between magnetic and true bearings is called variation or *magnetic declination*. Declination, a term generally used by surveyors, varies for different locations. It is not a constant. The Earth's magnetic field is subject to a number of changes in intensity and direction. These changes take place over daily, yearly, and long-term (or *secular*) periods. The secular changes result in a relatively constant increase or decrease in declination over a period of many years.

Antenna Monitoring System

Monitoring the operation of an AM directional antenna basically involves measuring the power into the system, the relative value of currents into the towers, their phase relationships, and the levels of radiated signal at certain monitoring points some distance from the antenna. Figure 1.19 shows a block diagram of a typical monitoring system for a three-tower array. For systems with additional towers, the basic layout is extended by adding more pickup elements, sample lines, and a monitor with additional inputs.

Phase/Current Sample Loop

Two types of phase/current sample pickup elements are commonly used: the *sample loop* and *torroidal current transformer* (TCT). The sample loop consists of a single turn unshielded loop of rigid construction, with a fixed gap at the open end for connection of the sample line. The device must be mounted on the tower near the point of maximum current. The loop can be used on towers of both uniform and nonuniform crosssection. It must operate at tower potential, except for towers of less than 130 electrical degrees height, where the loop can be operated at ground potential.

When the sample loop is operated at tower potential, the coax from the loop to the base of the tower is also at tower potential. To bring the sample line across the base of the tower, a sample line isolation coil is used.
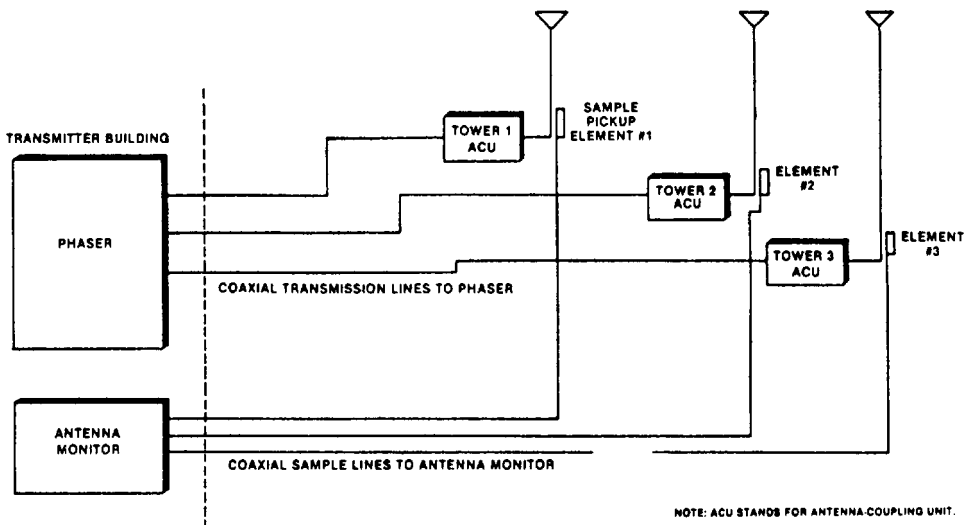
FIGURE 1.19 A typical three-tower directional antenna monitoring system.
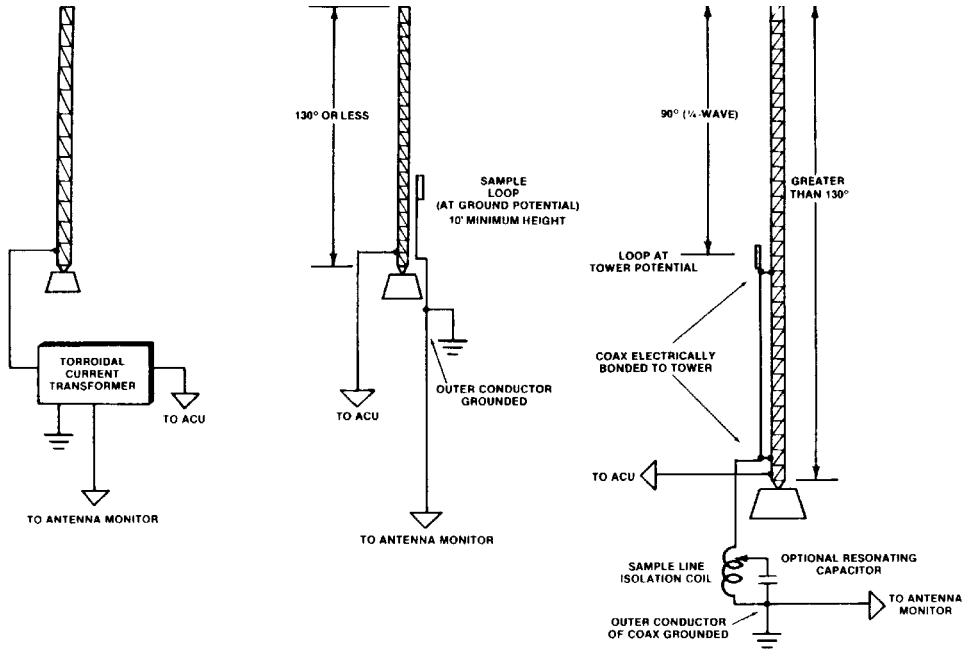
FIGURE 1.20   Three possible circuit configurations for phase sample pickup.

A shielded torroidal current transformer can also be used as the phase/current pickup element. Such devices offer several advantages over the sample loop, including greater stability and reliability. Because they are located inside the tuning unit cabinet or house, TCTs are protected from wind, rain, ice, and vandalism.

Unlike the rigid, fixed sample loop, torroidal current transformers are available in several sensitivities, ranging from 0.25 to 1.0 V per ampere of tower current. Tower currents of up to 40 A can be handled, providing a more usable range of voltages for the antenna monitor. Figure 1.20 shows the various arrangements that can be used for phase/current sample pickup elements.

Sample Lines

The selection and installation of the sampling lines for a directional monitoring system are important factors in the ultimate accuracy of the overall array.

With *critical arrays* (antennas requiring operation within tight limits specified in the station license), all sample lines must be of equal electrical length and installed in such a manner that corresponding lengths of all lines are exposed to equal environmental conditions.

While sample lines may be run above ground on supports (if protected and properly grounded), the most desirable arrangement is direct burial using jacketed cable. Burial of sample line cable is almost a standard practice because proper burial offers good protection against physical damage and a more stable temperature environment.

The Common Point

The power input to a directional antenna is measured at the phasor *common point*. Power is determined by the *direct method*:

$$P = I^2 R$$

where P is the power in watts (W), I is the common point current in amperes (A), and R is the common point resistance in ohms (Ω).

Monitor Points

Routine monitoring of a directional antenna involves the measurement of field intensity at certain locations away from the antenna, called *monitor points*. These points are selected and established during the initial tune-up of the antenna system. Measurements at the monitor points should confirm that radiation in prescribed directions does not exceed a value that would cause interference to other stations operating on the same or adjacent frequencies. The field intensity limits at these points are normally specified in the station license. Measurements at the monitor points may be required on a weekly or a monthly basis, depending on several factors and conditions relating to the particular station. If the system is not a critical array, quarterly measurements may be sufficient.

Folded Unipole Antenna

The *folded unipole* antenna consists of a grounded vertical structure with one or more conductors folded back parallel to the side of the structure. It can be visualized as a half-wave folded dipole perpendicular to the ground and cut in half (see Figure 1.21). This design makes it possible to provide a wide range of resonant radiation resistances by varying the ratio of the diameter of the folded-back conductors in relation to the tower. Top loading can also be used to broaden the antenna bandwidth. A side view of the folded unipole is shown in Figure 1.22.

The folded unipole antenna can be considered a modification of the standard shunt-fed system. Instead of a slant wire that leaves the tower at an approximate 45° angle (as used for shunt-fed systems), the folded unipole antenna has one or more wires attached to the tower at a predetermined height. The wires are supported by standoff insulators and run parallel to the sides of the tower down to the base.

The tower is grounded at the base. The folds, or wires, are joined together at the base and driven through an impedance matching network. Depending on the tuning requirements of the folded unipole, the wires may be connected to the tower at the top and/or at predetermined levels along the tower with shorting stubs.

The folded unipole can be used on tall (130° or greater) towers. However, if the unipole is not divided into two parts, the overall efficiency (unattenuated field intensity) will be considerably lower than the normally expected field for the electrical height of the tower.

FM Broadcast Antenna Systems

The propagation characteristics of VHF FM radio are much different than for MF AM. There is essentially no difference between day and night FM propagation. FM stations have relatively uniform day and night service areas with the same operating power.
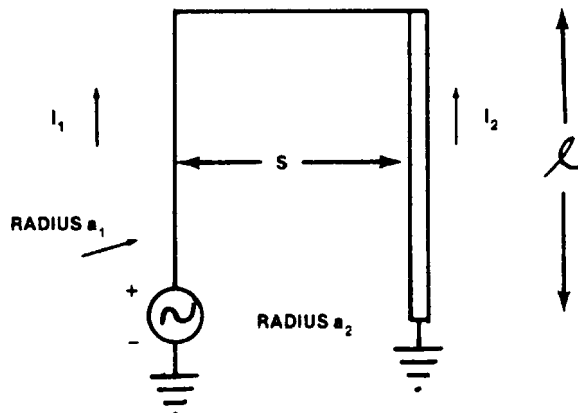
FIGURE 1.21   The folded unipole antenna can be thought of as a 1/2-wave folded dipole antenna perpendicular to the ground and cut in half.

A wide variety of antennas is available for use in the FM broadcast band. Nearly all employ circular polarization. Although antenna designs differ from one manufacturer to another, generalizations can be made that apply to most units.

Antenna Types

There are three basic classes of FM broadcast transmitting antennas in use today: *ring stub* and *twisted ring*, *shunt-* and *series-fed slanted dipole*, and *multi-arm short helix*. While each design is unique, all have the following items in common:

- The antennas are designed for sidemounting to a steel tower or pole.
- Radiating elements are shunted across a common rigid coaxial transmission line.
- Elements are placed along the rigid line every one wavelength.
- Antennas with one to seven bays are fed from the bottom of the coaxial transmission line.
- Antennas with more than seven bays are fed from the center of the array to provide more predictable performance in the field.
- Antennas generally include a means of tuning out reactances after the antenna has been installed through the adjustment of variable capacitive or inductive elements at the feed point.

FIGURE 1.22   The folds of the unipole antenna are arranged either near the legs of the tower or near the faces of the tower.

Figure 1.23 shows a shunt-fed slanted dipole antenna that consists of two half-wave dipoles offset 90°. The two sets of dipoles are rotated 22.5° (from their normal plane) and are *delta-matched* to provide a 50-Ω impedance at the radiator input flange. The lengths of all four dipole arms can be matched to resonance by mechanical adjustment of the end fittings. Shunt-feeding (when properly adjusted) provides equal currents in all four arms.

Wide-band *panel antennas* are a fourth common type of antenna used for FM broadcasting. Panel designs share some of the characteristics listed previously, but are intended primarily for specialized installations in which two or more stations will use the antenna simultaneously. Panel antennas are larger and more complex than other FM antennas, but offer the possibility for shared tower space among several stations and custom coverage patterns that would be difficult or even impossible with more common designs.

The ideal combination of antenna gain and transmitter power for a particular installation involves the analysis of a number of parameters. As shown in Table 1.5, a variety of pairings can be made to achieve the same ERP.

FIGURE 1.23   Mechanical configuration of one bay of a circularly polarized FM transmitting antenna.

TABLE 1.5   Various Combinations of Transmitter Power and Antenna
Gain that Will Produce 100-kW ERP for an FM Station

No. Bays	Antenna Gain	Transmitter Power (kW)
3	1.5888	66.3
4	2.1332	49.3
5	2.7154	38.8
6	3.3028	31.8
7	3.8935	27.0
8	4.4872	23.5
10	5.6800	18.5
12	6.8781	15.3

1.3   Television Systems

Jerry C. Whitaker

The technology of television is based on the conversion of light rays from still or moving scenes and pictures into electronic signals for transmission or storage, and subsequent reconversion into visual images on a screen. A similar function is provided in the production of motion picture film; however, where film records the brightness variations of a complete scene on a single frame in a short exposure no longer than a fraction of a second, the elements of a television picture must be scanned one piece at a time. In the television system, a scene is dissected into a **frame** composed of a mosaic of *picture elements* (pixels). A **pixel** is defined as the smallest area of a television image that can be transmitted within the parameters of the system. This process is accomplished by:

- Analyzing the image with a photoelectric device in a sequence of *horizontal scans* from the top to the bottom of the image to produce an electric signal in which the brightness and color values of the individual picture elements are represented as voltage levels of a video waveform
- Transmitting the values of the picture elements in sequence as voltage levels of a video signal
- Reproducing the image of the original scene in a video signal display of parallel scanning lines on a viewing screen

Scanning Lines and Fields

The image pattern of electrical charges on a camera tube target or CCD, corresponding to the brightness levels of a scene, are converted to a video signal in a sequential order of picture elements in the scanning process. At the end of each horizontal line sweep, the video signal is *blanked* while the beam returns rapidly to the left side of the scene to start scanning the next line. This process continues until the image has been scanned from top to bottom to complete one *field scan*.

After completion of this first field scan, at the midpoint of the last line, the beam again is blanked as it returns to the top center of the target where the process is repeated to provide a second field scan. The spot size of the beam as it impinges upon the target must be fine enough to leave unscanned areas between lines for the second scan. The pattern of scanning lines covering the area of the target, or the screen of a picture display, is called a **raster**.

Interlaced Scanning Fields

Because of the half-line offset for the start of the beam return to the top of the raster and for the start of the second field, the lines of the second field lie in-between the lines of the first field. Thus, the lines of the two are **interlaced**. The two interlaced fields constitute a single television *frame*. Figure 1.24 shows a frame scan with interlacing of the lines of two fields.
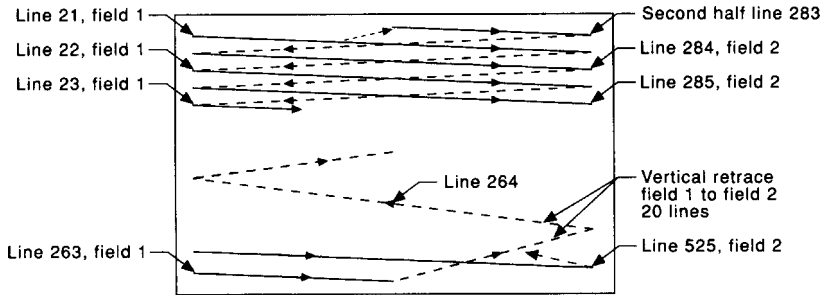
FIGURE 1.24 The interlaced scanning pattern (raster) of the television image. (*Source:* Electronic Industries Association.)

Reproduction of the camera image on a cathode ray tube (CRT) or solid-state display is accomplished by an identical operation, with the scanning beam modulated in density by the video signal applied to an element of the electron gun or control element, in the case of a solid-state display device. This control voltage to the display varies the brightness of each picture element on the screen.

Blanking of the scanning beam during the return trace is provided for in the video signal by a "blacker-than-black" pulse waveform. In addition, in most receivers and monitors another blanking pulse is generated from the horizontal and vertical scanning circuits and applied to the display system to ensure a black screen during scanning retrace. The retrace lines are shown as diagonal dashed lines in Figure 1.24.

The interlaced scanning format, standardized for monochrome and compatible color, was chosen primarily for two partially related and equally important reasons:

- To eliminate viewer perception of the intermittent presentation of images, known as *flicker*
- To reduce video bandwidth requirements for an acceptable flicker threshold level

Perception of flicker is dependent primarily upon two conditions:

- The brightness level of an image
- The relative area of an image in a picture

The 30-Hz transmission rate for a full 525-line television frame is comparable to the highly successful 24-frame-per-second rate of motion-picture film. However, at the higher brightness levels produced on television screens, if all 483 lines (525 less blanking) of a television image were to be presented sequentially as single frames, viewers would observe a disturbing flicker in picture areas of high brightness. For a comparison, motion-picture theaters on average produce a screen brightness of 10 to 25 ft·L (footlambert), whereas a direct-view CRT may have a highlight brightness of 50 to 80 ft·L. It should be noted also that motion-picture projectors flash twice per frame to reduce the flicker effect.

Through the use of interlaced scanning, single field images with one-half the vertical resolution capability of the 525-line system are provided at the high flicker-perception threshold rate of 60 Hz. Higher resolution of the full 483 lines of vertical detail is provided at the lower flicker-perception threshold rate of 30 Hz. The result is a relatively flickerless picture display at a screen brightness of well over 50 to 75 ft·L, more than double that of motion-picture film projection. Both 60-Hz fields and 30-Hz frames have the same horizontal resolution capability.

The second advantage of interlaced scanning, compared to progressive scanning, where the frame is constructed in one pass over the display face (rather than in two through interlace), is a reduction in video bandwidth for an equivalent flicker threshold level. Progressive scanning of 525 lines would have to be completed in 1/60 s to achieve an equivalent level of flicker perception. This would require a line scan to be completed in half the time of an interlaced scan. The bandwidth then would double for an equivalent number of pixels per line.

The standards adopted by the Federal Communications Commission (FCC) for monochrome television in the United States specified a system of 525 lines per frame, transmitted at a frame rate of 30 Hz, with each frame composed of two interlaced fields of horizontal lines. Initially in the development of television transmission standards, the 60-Hz power line waveform was chosen as a convenient reference for vertical scan. Furthermore, in the event of coupling of power line hum into the video signal or scanning/deflection circuits, the visible effects would be stationary and less objectionable than moving **hum bars** or distortion of horizontal-scanning geometry. In the United Kingdom and much of Europe, a 50-Hz interlaced system was chosen for many of the same reasons. With improvements in television receivers, the power line reference was replaced with a stable crystal oscillator, rendering the initial reason for the frame rate a moot point.

The existing 525-line monochrome standards were retained for color in the recommendations of the National Television System Committee (NTSC) for compatible color television in the early 1950s. The NTSC system, adopted in 1953 by the FCC, specifies a scanning system of 525 horizontal lines per frame, with each frame consisting of two interlaced fields of 262.5 lines at a field rate of 59.94 Hz. Forty-two of the 525 lines in each frame are blanked as black picture signals and reserved for transmission of the vertical scanning synchronizing signal. This results in 483 visible lines of picture information. Because the vertical blanking interval represents a significant amount of the total transmitted waveform, the television industry has sought ways to carry additional data during the blanking interval. Such applications include closed captioning and system test signals.

Synchronizing Video Signals

In monochrome television transmission, two basic synchronizing signals are provided to control the timing of picture-scanning deflection:

- Horizontal sync pulses at the line rate.
- Vertical sync pulses at the field rate in the form of an interval of wide horizontal sync pulses at the field rate. Included in the interval are **equalizing pulses** at twice the line rate to preserve interlace in each frame between the even and odd fields (offset by a half line).

In color transmissions, a third synchronizing signal is added during horizontal scan blanking to provide a frequency and phase reference for color signal encoding circuits in cameras and decoding circuits in receivers. These synchronizing and reference signals are combined with the picture video signal to form a **composite video** waveform.

The scanning and color-decoding circuits in receivers must follow the frequency and phase of the synchronizing signals to produce a stable and geometrically accurate image of the proper color **hue** and **saturation**. Any change in timing of successive vertical scans can impair the interlace of the even and odd fields in a frame. Small errors in horizontal scan timing of lines in a field can result in a loss of resolution in vertical line structures. Periodic errors over several lines that may be out of the range of the horizontal scan automatic frequency control circuit in the receiver will be evident as jagged vertical lines.

Television Industry Standards

There are three primary color transmission standards in use today:

- *NTSC* (National Television Systems Committee): Used in the United States, Canada, Central America, most of South America, and Japan. In addition, NTSC is used in various countries or possessions heavily influenced by the United States.
- *PAL* (Phase Alternation each Line): Used in England, most countries and possessions influenced by the British Commonwealth, many western European countries and China. Variation exists in PAL systems.
- *SECAM* (Sequential Color with [Avec] Memory): Used in France, countries and possessions influenced by France, the USSR (generally the former Soviet Bloc nations), and other areas influenced by Russia.

The three standards are incompatible for a variety of reasons (see Benson and Whitaker, 1991).

Television transmitters in the United States operate in three frequency bands:

- Low-band VHF (very high frequency), channels 2 through 6
- High-band VHF, channels 7 through 13
- UHF (ultra-high frequency), channels 14 through 83 (UHF channels 70 through 83 currently are assigned to mobile radio services)

Table 1.6 shows the frequency allocations for channels 2 through 83. Because of the wide variety of operating parameters for television stations outside the United States, this section will focus primarily on TV transmission as it relates to the Unites States.

Maximum power output limits are specified by the FCC for each type of service. The maximum **effective radiated power** (ERP) for low-band VHF is 100 kW; for high-band VHF it is 316 kW; and for UHF it is 5 MW. The ERP of a station is a function of transmitter power output (TPO) and antenna gain. ERP is determined by multiplying these two quantities together and subtracting transmission line loss.

The second major factor that affects the coverage area of a TV station is antenna height, known in the broadcast industry as *height above average terrain* (HAAT). HAAT takes into consideration the effects of the geography in the vicinity of the transmitting tower. The maximum HAAT permitted by the FCC for a low- or high-band VHF station is 1000 ft (305 m) east of the Mississippi River and 2000 ft (610 m) west of the Mississippi. UHF stations are permitted to operate with a maximum HAAT of 2000 ft (610 m) anywhere in the United States (including Alaska and Hawaii).

TABLE 1.6 Frequency Allocations for TV Channels 2 through 83 in the U.S.

Channel Designation	Frequency Band, MHz	Channel Designation	Frequency Band, MHz	Channel Designation	Frequency Band, MHz
2	54–60	30	566–572	58	734–740
3	60–66	31	572–578	59	740–746
4	66–72	32	578–584	60	746–752
5	76–82	33	584–590	61	752–758
6	82–88	34	590–596	62	758–764
7	174–180	35	596–602	63	764–770
8	180–186	36	602–608	64	770–776
9	186–192	37	608–614	65	776–782
10	192–198	38	614–620	66	782–788
11	198–204	39	620–626	67	788–794
12	204–210	40	626–632	68	794–800
13	210–216	41	632–638	69	800–806
14	470–476	42	638–644	70	806–812
15	476–482	43	644–650	71	812–818
16	482–488	44	650–656	72	818–824
17	488–494	45	656–662	73	824–830
18	494–500	46	662–668	74	830–836
19	500–506	47	668–674	75	836–842
20	506–512	48	674–680	76	842–848
21	512–518	49	680–686	77	848–854
22	518–524	50	686–692	78	854–860
23	524–530	51	692–698	79	860–866
24	530–536	52	698–704	80	866–872
25	536–542	53	704–710	81	872–878
26	542–548	54	710–716	82	878–884
27	548–554	55	716–722	83	884–890
28	554–560	56	722–728		
29	560–566	57	728–734		

The ratio of visual output power to **aural** output power can vary from one installation to another; however, the aural is typically operated at between 10 and 20% of the visual power. This difference is the result of the reception characteristics of the two signals. Much greater signal strength is required at the consumer's receiver to recover the visual portion of the transmission than the aural portion. The aural power output is intended to be sufficient for good reception at the fringe of the station's coverage area but not beyond. It is of no use for a consumer to be able to receive a TV station's audio signal but not the video.

In addition to high power stations, two classifications of low-power TV stations have been established by the FCC to meet certain community needs: They are:

- *Translator:* A low-power system that rebroadcasts the signal of another station on a different channel. Translators are designed to provide "fill-in" coverage for a station that cannot reach a particular community because of the local terrain. Translators operating in the VHF band are limited to 100 W power output (ERP), and UHF translators are limited to 1 kW.
- *Low-Power Television (LPTV):* A service established by the FCC designed to meet the special needs of particular communities. LPTV stations operating on VHF frequencies are limited to 100 W ERP, and UHF stations are limited to 1 kW. LPTV stations originate their own programming and can be assigned by the FCC to any channel, as long as sufficient protection against interference to a full-power station is afforded.

Composite Video

The composite video waveform is shown in Figure 1.25. The actual radiated signal is inverted, with modulation extending from the synchronizing pulses at maximum carrier level (100%) to reference picture white at 7.5%. Because an increase in the amplitude of the radiated signal corresponds to a decrease in picture brightness, the polarity of modulation is termed *negative*.
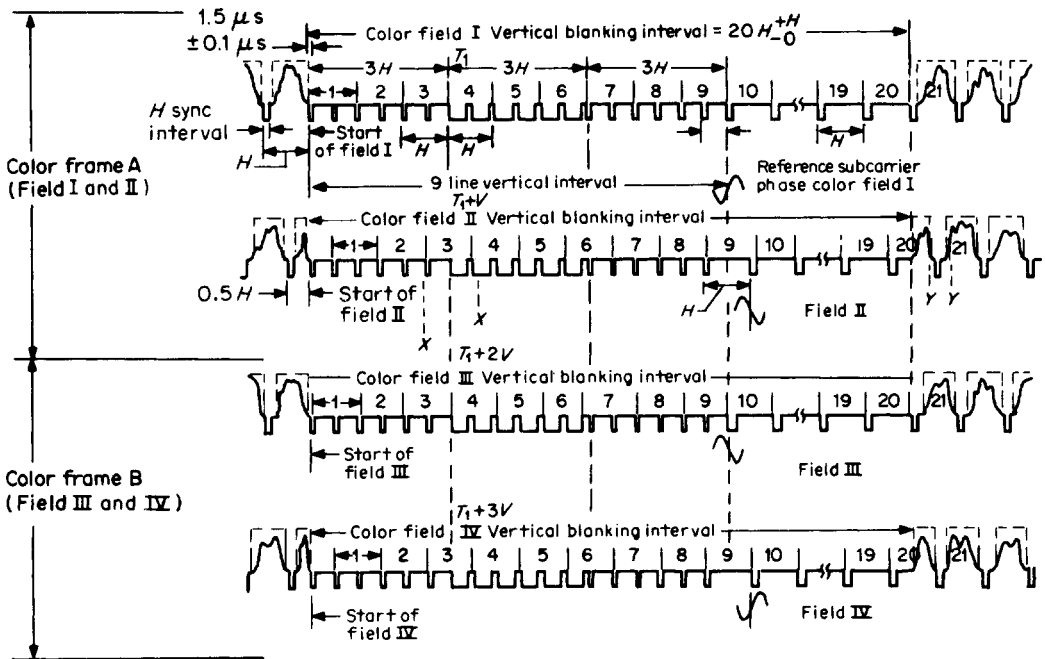
FIGURE 1.25 The principal components of the NTSC color television waveform. (*Source:* Electronic Industries Association.)

The term *composite* is used to denote a video signal that contains:

- Picture luminance and chrominance information
- Timing information for synchronization of scanning and color signal processing circuits

The negative-going portion of the waveform shown in Figure 1.25 is used to transmit information for synchronization of scanning circuits. The positive-going portion of the amplitude range is used to transmit luminance information representing brightness and, for color pictures, chrominance.

At the completion of each line scan in a receiver or monitor, a horizontal synchronizing (**H**-*sync*) pulse in the composite video signal triggers the scanning circuits to return the beam rapidly to the left of the screen for the start of the next line scan. During the return time, a horizontal blanking signal at a level lower than that corresponding to the blackest portion of the scene is added to avoid the visibility of the retrace lines. In a similar manner, after completion of each field, a vertical blanking signal blanks out the retrace portion of the scanning beam as it returns to the top of the picture to start the scan of the next field. The small-level difference between video reference black and blanking level is called **setup**. Setup is used as a guard band to ensure separation of the synchronizing and video-information functions and adequate blanking of the scanning retrace lines on receivers.

The waveforms of Figure 1.26 show the various reference levels of video and sync in the composite signal. The unit of measurement for video level was specified initially by the Institute of Radio Engineers (IRE). These **IRE** *units* are still used to quantify video signal levels. The primary IRE values are given in Table 1.7.
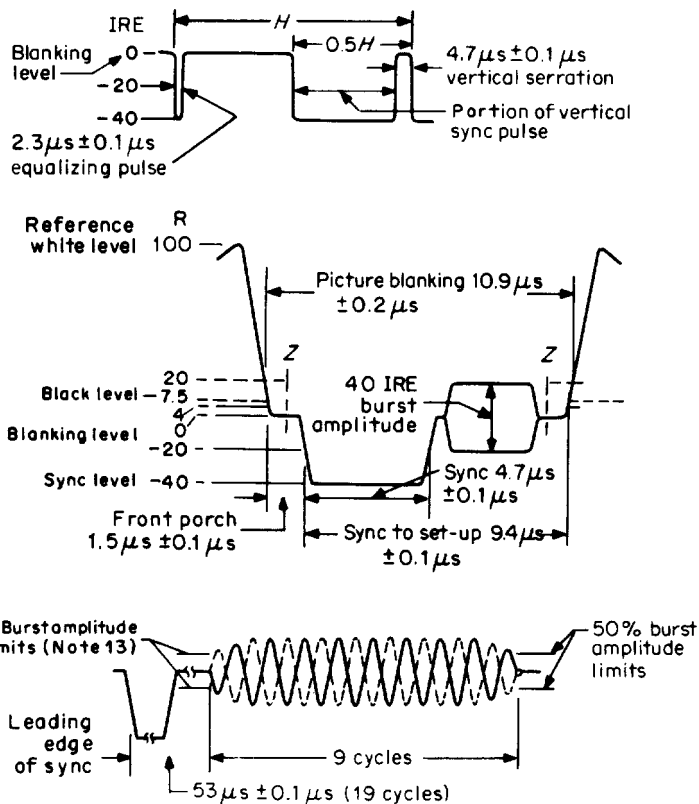
FIGURE 1.26 Sync pulse widths for the NTSC color system. (*Source:* Electronic Industries Association.)

Color Signal Encoding

To facilitate an orderly introduction of color television broadcasting in the United States and other countries with existing monochrome services, it was essential that the new transmissions be compatible. In other words, color pictures would provide acceptable quality on unmodified monochrome receivers. In addition, because of the limited availability of the RF spectrum, another related requirement was the need to fit approximately 2-MHz bandwidth of color information into the 4.2-MHz

TABLE 1.7   Video and Sync Levels in IRE Units

Signal Level	IRE Level
Reference white	100
Blanking level width measurement	20
Color burst sine wave peak	+20 to −20
Reference black	7.5
Blanking	0
Sync pulse width measurement	−20
Sync level	−40

video bandwidth of the existing 6-MHz broadcasting channels with little or no modification of existing transmitters. This is accomplished by using the band-sharing color signal system developed by the NTSC and by taking advantage of the fundamental characteristics of the eye regarding color sensitivity and resolution.

The video-signal spectrum generated by scanning an image consists of energy concentrated near harmonics of the 15,734-Hz line scanning frequency. Additional lower-amplitude sideband components exist at multiples of 60 Hz (the field scan frequency) from each line scan harmonic. Substantially no energy exists halfway between the line scan harmonics, that is, at odd harmonics of one half line frequency. Thus, these blank spaces in the spectrum are available for the transmission of a signal for carrying color information and its sideband. In addition, a signal modulated with color information injected at this frequency is of relatively low visibility in the reproduced image because the odd harmonics are of opposite phase on successive scanning lines and in successive frames, requiring four fields to repeat. Furthermore, the visibility of the color video signal is reduced further by the use of a subcarrier frequency near the cutoff of the video bandpass.

In the NTSC system, color is conveyed using two elements:

- A luminance signal
- A chrominance signal

The luminance signal is derived from components of the three primary colors — red, green, and blue — in the proportions for *reference white*, E_y, as follows:

$$E_y = 0.3E_R + 0.59E_G + 0.11E_B$$

These transmitted values equal unity for white and thus result in the reproduction of colors on monochrome receivers at the proper luminance level. This is known as the *constant-luminance* principle.

The color signal consists of two chrominance components, I and Q, transmitted as amplitude-modulated sidebands of two 3.579545-MHz subcarriers in quadrature. The subcarriers are suppressed, leaving only the sidebands in the color signal. Suppression of the carriers permits demodulation of the color signal as two separate color signals in a receiver by reinsertion of a carrier of the phase corresponding to the desired color signal (**synchronous demodulation**).

I and Q signals are composed of red, green, and blue primary color components produced by color cameras and other signal generators. The phase relationship among the I and Q signals, the derived primary and complementary colors, and the color synchronizing burst can be shown graphically on a **vectorscope** display. The horizontal and vertical sweep signals on a vectorscope are produced from R-Y and B-Y subcarrier sine waves in quadrature, producing a circular display. The chrominance signal controls the intensity of the display. A vectorscope display of an Electronic Industries Association (EIA) standard color bar signal is shown in Figure 1.27.

Color-Signal Decoding

Each of the two chroma signal carriers can be recovered individually by means of synchronous detection. A reference subcarrier of the same phase as the desired chroma signal is applied as a gate to a balanced demodulator. Only the modulation of the signal in the same phase as the reference will be present in

FIGURE 1.27 Vectorscope representation for chroma and vector amplitude relationships in the NTSC system. (*Source: Electronic Industries Association.*)

the output. A low-pass filter may be added to remove second harmonic components of the chroma signal generated in the process.

Transmission Equipment

Television transmitters are classified in terms of their operating band, power level, type of final amplifier stage, and cooling method. The transmitter is divided into two basic subsystems:

- The *visual* section, which accepts the video input, amplitude modulates an RF carrier, and amplifies the signal to feed the antenna system
- The *aural* section, which accepts the audio input, frequency modulates a separate RF carrier and amplifies the signal to feed the antenna system

The visual and aural signals are combined to feed a single radiating system.

Transmitter Design Considerations

Each manufacturer has a particular philosophy with regard to the design and construction of a broadcast TV transmitter. Some generalizations can, however, be made with respect to basic system design.

When the power output of a TV transmitter is discussed, the visual section is the primary consideration. Output power refers to the *peak power* of the visual section of the transmitter (*peak of sync*). The FCC-licensed ERP is equal to the transmitter power output minus feedline losses times the power gain of the antenna.

A low-band VHF station can achieve its maximum 100-kW power output through a wide range of transmitter and antenna combinations. A 35-kW transmitter coupled with a gain-of-4 antenna would work, as would a 10-kW transmitter feeding an antenna with a gain of 12. Reasonable pairings for a high-band VHF station would range from a transmitter with a power output of 50 kW feeding an antenna with a gain of 8, to a 30kW transmitter connected to a gain-of-12 antenna. These combinations assume reasonable feedline losses.

To reach the exact power level, minor adjustments are made to the power output of the transmitter, usually by a front panel power trim control.

UHF stations that want to achieve their maximum licensed power output are faced with installing a very high-power transmitter. Typical pairings include a transmitter rated for 220 kW and an antenna with a gain of 25, or a 110-kW transmitter and a gain-of-50 antenna. In the latter case, the antenna could pose a significant problem. UHF antennas with gains in the region of 50 are possible, but not advisable for most installations because of the coverage problems that can result. High-gain antennas have a narrow vertical radiation pattern that can reduce a station's coverage in areas near the transmitter site.

At first examination, it might seem reasonable and economical to achieve licensed ERP using the lowest transmitter power output possible and highest antenna gain. Other factors, however, come into play that make the most obvious solution not always the best solution. Factors that limit the use of high-gain antennas include:

- The effects of high-gain designs on coverage area and signal penetration
- Limitations on antenna size because of tower restrictions, such as available vertical space, weight, and windloading
- The cost of the antenna

The amount of output power required of a transmitter will have a fundamental effect on system design. Power levels dictate whether the unit will be of solid-state or vacuum-tube design; whether air, water, or vapor cooling must be used; the type of power supply required; the sophistication of the high-voltage control and supervisory circuitry; and many other parameters.

Solid-state devices are generally used for VHF transmitters and for low-power UHF transmitters. Tetrodes may also be used in these ranges. As solid-state technology advances, the power levels possible in a reasonable transmitter design steadily increase. In the realm of high power UHF transmitters, the **klystron** is a common power output device. Klystrons use an *electron bunching* technique to generate high power (55 kW from a single tube is not uncommon) at microwave frequencies. The klystron, however, is relatively inefficient in its basic form. A stock klystron with no efficiency-optimizing circuitry might be only 40 to 50% efficient, depending on the type of device used. Various schemes have been devised to improve klystron efficiency, the best known of which is **beam pulsing**. Two types of pulsing are in common used:

- *Mod-anode pulsing*, a technique designed to reduce power consumption of the klystron during the color burst and video portion of the signal (and thereby improve overall system efficiency)
- *Annular control electrode* (ACE) pulsing, which accomplishes basically the same thing by incorporating the pulsing signal into a low-voltage stage of the transmitter, rather than a high-voltage stage (as with mod-anode pulsing).

Still another approach to improving UHF transmitter efficiency involves entirely new classes of vacuum tubes: the *inductive output tube* (IOT) and the **multistage depressed collector (MSDC) klystron**. The IOT is a device that essentially combines the cathode/grid structure of the tetrode with the drift tube/collector structure of the klystron. The MSDC klystron incorporates a collector assembly that operates at progressively lower voltage levels. The net effect for the MSDC is to recover energy from the electron stream rather than dissipating the energy as heat.

Elements of the Transmitter

A television transmitter can be divided into four major subsystems:

- The exciter
- Intermediate power amplifier (IPA)
- Power amplifier (PA)
- High-voltage power supply

Figure 1.28 shows the audio, video, and RF paths for a typical television transmitter.
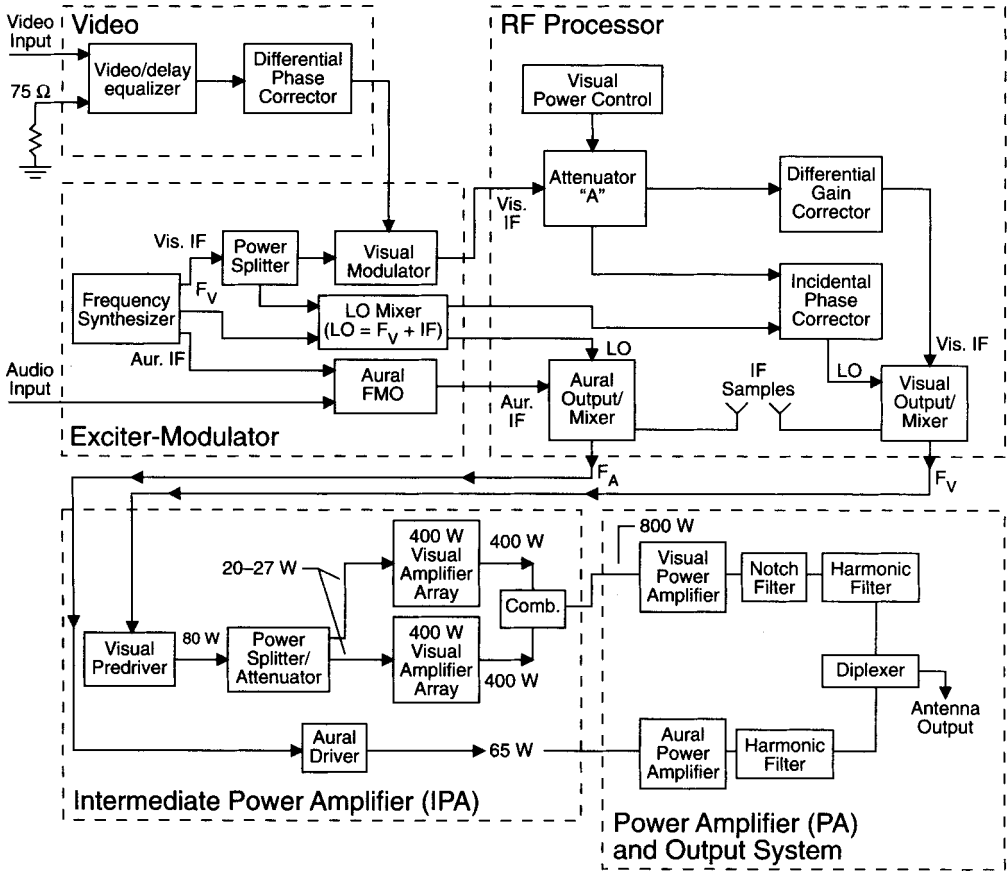
FIGURE 1.28 Simplified block diagram of a VHF television transmitter.

The modulated visual intermediate frequency (IF) signal is band-shaped in a vestigial sideband filter, typically a surface-acoustic-wave (SAW) filter. Envelope-delay correction is not required for the SAW filter because of the uniform delay characteristics of the device. Envelope-delay compensation may, however, be needed for other parts of the transmitter. The SAW filter provides many benefits to transmitter designers and operators. A SAW filter requires no adjustments and is stable with respect to temperature and time. A *color-notch filter* is required at the output of the transmitter because imperfect linearity of the IPA and PA stages introduces unwanted modulation products.

The power amplifier raises the output energy of the transmitter to the desired RF operating level. Tetrodes in television service are operated in the class B mode to obtain reasonable efficiency while maintaining a linear transfer characteristic. Class B amplifiers, when operated in tuned circuits, provide linear performance because of the flywheel effect of the resonance circuit. This allows a single tube to be used instead of two in push-pull fashion. The bias point of the linear amplifier is chosen so that the transfer characteristic at low modulation levels matches that at higher modulation levels. The plate (anode) circuit of a tetrode PA is usually built around a coaxial resonant cavity, which provides a stable and reliable tank circuit.

Solid state transmitters typically incorporate a massively parallel design to achieve the necessary power levels. So-called power blocks of 1 kW or greater are combined as required to meet the target transmitter power output. Most designs use MOSFETs running in a class D (or higher) switching mode. Any one of several combiner schemes may be used to couple the power blocks to the load. Depending on the design, high-reliability features may be incorporated into the transmitter, including automatic disconnection of failed power blocks and hot-changing of defective modules.

UHF transmitters using a klystron in the final output stage must operate class A, the most linear but also most inefficient operating mode for a vacuum tube. Two types of klystrons have traditionally been used: *integral cavity* and *external cavity* devices. The basic theory of operation is identical for each tube, but the mechanical approach is radically different. In the **integral cavity klystron**, the cavities are built into the device to form a single unit. In the **external cavity klystron**, the cavities are outside the vacuum envelope and are bolted around the tube when the klystron is installed in the transmitter. A number of factors come into play in a discussion of the relative merits of integral vs. external cavity designs. Primary considerations include operating efficiency, purchase price, and life expectancy.

Transmitters based on IOT or MSDC klystron final tubes have much in common with traditional klystron-based systems. There are, however, a number of significant differences, including:

- Low-level video waveform precorrection circuitry
- Drive power requirements
- Power supply demands and complexity
- Fault/arc suppression and protection
- Cooling system design and complexity
- Overall system efficiency

The transmitter block diagram of Figure 1.28 shows separate visual and aural PA stages. This configuration is normally used for high-power transmitters. Low-power designs often use a combined mode (*common amplification*) in which the aural and visual signals are added prior to the PA. This approach offers a simplified system but at the cost of additional precorrection of the input video signal.

PA stages often are configured so that the circuitry of the visual and aural amplifiers is identical, providing backup protection in the event of a visual PA failure. The aural PA can then be reconfigured to amplify both the aural and the visual signals at reduced power.

The aural output stage of a television transmitter is similar in basic design to a frequency modulated (FM) broadcast transmitter. Tetrode output devices generally operate class C; solid-state devices operate in one of many possible switching modes for high efficiency. The aural PA for a UHF transmitter may use a klystron, IOT, MSDC, tetrode, or a group of solid-state power blocks.

Harmonic filters are employed to attenuate out-of-band radiation of the aural and visual signals to ensure compliance with FCC requirements. Filter designs vary depending upon the manufacturer; however, most are of coaxial construction utilizing L and C components housed within a prepackaged assembly. Stub filters are also used, typically adjusted to provide maximum attenuation at the second harmonic of the operating frequency of the visual carrier and the aural carrier.

The filtered visual and aural outputs are fed to a hybrid diplexer where the two signals are combined to feed the antenna. For installations that require dual-antenna feedlines, a hybrid combiner with quadrature-phased outputs is used. Depending upon the design and operating power, the color-notch filter, aural and visual harmonic filters, and diplexer may be combined into a single mechanical unit.

Antenna System

Broadcasting is accomplished by the emission of coherent electromagnetic waves in free space from one or more radiating-antenna elements that are excited by modulated RF currents. Although, by definition, the radiated energy is composed of mutually dependent magnetic and electric vector fields, conventional practice in television engineering is to measure and specify radiation characteristics in terms of the electric field only.

The field vectors may be polarized horizontally, vertically, or circularly. Television broadcasting, however, has used horizontal polarization for the majority of installations worldwide. More recently interest in the advantages of circular polarization has resulted in an increase in this form of transmission, particularly for VHF channels. Both horizontal and circular polarization designs are suitable for tower-top or side-mounted installations. The latter option is dictated primarily by the existence of a previously installed tower-top antenna. On the other hand, in metropolitan areas where several antennas must be located on the same structure, either a stacking or candelabra-type arrangement is feasible. Another approach to TV transmission involves combining the RF outputs of two or more stations and feeding a single wideband antenna.

This approach is expensive and requires considerable engineering analysis to produce a combiner system that will not degrade the performance of either transmission system.

Television Reception

The broadcast channels in the United States are 6 MHz wide for transmission on conventional 525-line standards. The minimum signal level at which a television receiver will provide usable pictures and sound is called the *sensitivity level*. The FCC has set up two standard signal level classifications, Grades A and B, for the purpose of licensing television stations and allocating coverage areas. Grade A refers to urban areas relatively near the transmitting tower; Grade B use ranges from suburban to rural and other fringe areas a number of miles from the transmitting antenna.

Many sizes and form factors of receivers are manufactured. Portable personal types include pocket-sized or hand-held models with picture sizes of 2 to 4 in. diagonal for monochrome and 5 to 6 in. for color. Large screen sizes are available in monochrome where low cost and light weight are prime requirements. However, except where portability is important, the majority of television program viewing is in color. The 19- and 27-in. sizes dominate the market.

Television receiver functions may be broken down into several interconnected blocks. With the increasing use of large-scale integrated circuits, the isolation of functions has become less obvious in the design of receivers. The typical functional configuration of a receiver using a trigun picture tube is shown in Figure 1.29.

Display Systems

Color video displays may be classified under the following categories:

- Direct-view CRT
- Large-screen display, optically projected from a CRT
- Large-screen display, projected from a modulated light beam
- Large-area display of individually driven light-emitting CRTs or incandescent picture elements
- Flat-panel matrix of transmissive or reflective picture elements
- Flat-panel matrix of light-emitting picture elements

The CRT remains the dominant type of display for both consumer and professional 525-/625-line television applications. Light-valve systems using a modulated light source have found wide application for
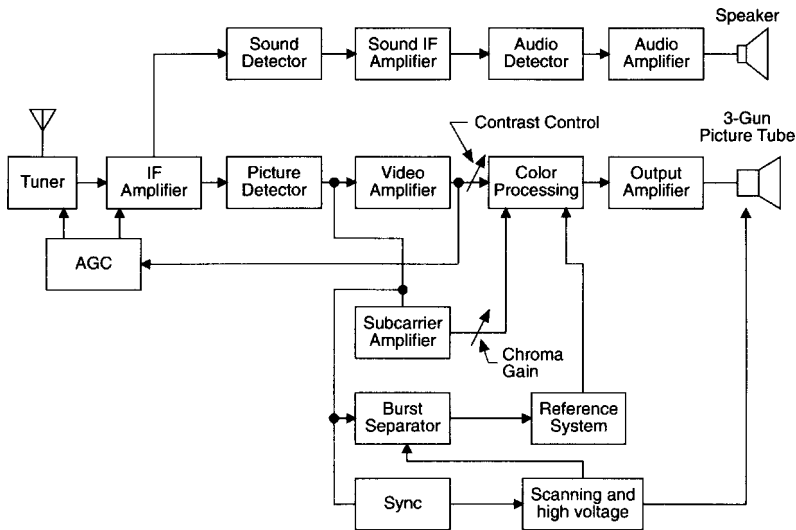
FIGURE 1.29 Simplified schematic block diagram of a color television receiver.

presentations to large audiences in theater environments, particularly where high screen brightness is required. Matrix-driven flat-panel displays are used in increasing numbers for television receivers and for portable projector units. Video and data projectors using LCD technology have gained wide acceptance.

Cathode Ray Tube Display

The direct-view CRT is the dominant display device in television. The attributes offered by CRTs include the following:

- High brightness
- High resolution
- Excellent gray-scale reproduction
- Low cost compared to other types of displays

From the standpoint of television receiver manufacturing simplicity and low cost, packaging of the display device as a single component is attractive. The tube itself is composed of only three basic parts: an electron gun, an envelope, and a shadow-mask phosphor screen. The luminance efficiency of the electron optical system and the phosphor screen is high. A peak beam current of under 1 μA in a 25-in. tube will produce a highlight brightness of up to 100 ft·L. The major drawback is the power required to drive the horizontal sweep circuit and the high accelerating voltage necessary for the electron beam. This requirement is partially offset through generation of the screen potential and other lower voltages by rectification of the scanning flyback voltage.

   As consumer demands drive manufacturers to produce larger picture sizes, the weight and depth of the CRT and the higher power and voltage requirements become serious limitations. These are reflected in sharply increasing receiver costs. To withstand the atmospheric pressures on the evacuated glass envelope, CRT weight increases exponentially with the viewable diagonal. Nevertheless, manufacturers have continued to meet the demand for increased screen sizes with larger direct-view tubes. Improved versions of both tridot delta and in-line guns have been produced. The tridot gun provides small spot size at the expense of critical convergence adjustments for uniform resolution over the full-tube faceplate. In-line guns permit the use of a self-converging deflection yoke that will maintain dynamic horizontal convergence over the full face of the tube without the need for correction waveforms. The downside is slightly reduced resolution.

Defining Terms

Aural: The sound portion of a television signal.

Beam pulsing: A method used to control the power output of a klystron in order to improve the operating efficiency of the device.

Blanking: The portion of a television signal that is used to blank the screen during the horizontal and vertical retrace periods.

Composite video: A single video signal that contains luminance, color, and synchronization information. NTSC, PAL, and SECAM are all examples of composite video formats.

Effective radiated power: The power supplied to an antenna multiplied by the relative gain of the antenna in a given direction.

Equalizing pulses: In an encoded video signal, a series of 2X line frequency pulses occurring during vertical blanking, before and after the vertical synchronizing pulse. Different numbers of equalizing pulses are inserted into different fields to ensure that each field begins and ends at the right time to produce proper interlace. The 2X line rate also serves to maintain horizontal synchronization during vertical blanking.

External cavity klystron: A klystron device in which the resonant cavities are located outside the vacuum envelope of the tube.

Field: One of the two (or more) equal parts of information into which a frame is divided in interlace video scanning. In the NTSC system, the information for one picture is divided into two fields. Each field

contains one-half the lines required to produce the entire picture. Adjacent lines in the picture are contained in alternate fields.

Frame: The information required for one complete picture in an interlaced video system. For the NTSC system, there are two fields per frame.

H (horizontal): In television signals, *H* may refer to any of the following: the horizontal period or rate, horizontal line of video information, or horizontal sync pulse.

Hue: One of the characteristics that distinguishes one color from another. Hue defines color on the basis of its position in the spectrum (red, blue, green, yellow, etc.). Hue is one of the three characteristics of television color. Hue is often referred to as *tint*. In NTSC and PAL video signals, the hue information at any particular point in the picture is conveyed by the corresponding instantaneous phase of the active video subcarrier.

Hum bars: Horizontal black and white bars that extend over the entire TV picture and usually drift slowly through it. Hum bars are caused by an interfering power line frequency or one of its harmonics.

Inductive output tube: An amplifier device for UHF-TV signals that combines aspects of a tetrode (grid modulation) with a klystron (velocity modulation of an electron beam). The result is a more efficient, less expensive device for many applications.

Integral cavity klystron: A klystron device in which the resonant cavities are located inside the vacuum envelope of the tube.

Interlaced: A shortened version of *interlaced scanning* (also called *line interlace*). Interlaced scanning is a system of video scanning whereby the odd- and even-numbered lines of a picture are transmitted consecutively as two separate interleaved fields.

IRE: A unit equal to 1/140 of the peak-to-peak amplitude of a video signal, which is typically 1 V. The 0 IRE point is at blanking level, with the sync tip at –40 IRE and white extending to +100 IRE. IRE stands for *Institute of Radio Engineers*, an organization preceding the IEEE, which defined the unit.

Klystron: An amplifier device for UHF and microwave signals based on velocity modulation of an electron beam. The beam is directed through an input cavity, where the input RF signal polarity initializes a *bunching effect* on electrons in the beam. The bunching effect excites subsequent cavities, which increase the bunching through an energy flywheel concept. Finally, the beam passes an output cavity that couples the amplified signal to the load (antenna system). The beam falls onto a collector element that forms the return path for the current and dissipates the heat resulting from electron beam bombardment.

Low-power TV (LPTV): A television service authorized by the FCC to serve specific confined areas. An LPTV station may typically radiate between 100 and 1000 W of power, covering a geographic radius of 10 to 15 mi.

Multistage depressed collector (MSDC) klystron: A specially designed klystron in which decreasing voltage zones cause the electron beam to be reduced in velocity before striking the collector element. The effect is to reduce the amount of heat that must be dissipated by the device, improving operating efficiency.

Pixel: The smallest distinguishable and resolvable area in a video image. A pixel is a single point on the screen. The word pixel is derived from *picture element*.

Raster: A predetermined pattern of scanning the screen of a CRT. *Raster* may also refer to the illuminated area produced by scanning lines on a CRT when no video is present.

Saturation: The intensity of the colors in the active picture, the voltage levels of the colors. Saturation relates to the degree by which the eye perceives a color as departing from a gray or white scale of the same brightness. A 100% saturated color does not contain any white; adding white reduces saturation. In NTSC and PAL video signals, the color saturation at any particular instant in the picture is conveyed by the corresponding instantaneous amplitude of the active video subcarrier.

Scan: One sweep of the target area in a camera tube or of the screen in a picture tube.

Setup: A video term relating to the specified base of an active picture signal. In NTSC, the active picture signal is placed 7.5 IRE units above blanking (0 IRE). Setup is the separation in level between the *video blanking* and *reference black* levels.

Synchronous detection: A demodulation process in which the original signal is recovered by multiplying the modulated signal by the output of a synchronous oscillator locked to the carrier.

Translator: An unattended television or FM broadcast repeater that receives a distant signal and retransmits the picture and/or audio locally on another channel.

Vectorscope: An oscilloscope-type device used to display the color parameters of a video signal. A vectorscope decodes color information into R-Y and B-Y components, which are then used to drive the X and Y axis of the scope. The total lack of color in a video signal is displayed as a dot in the center of the vectorscope. The angle, distance around the circle, magnitude, and distance away from the center indicate the phase and amplitude of the color signal.

References

K.B. Benson and J. Whitaker (Eds.), *Television Engineering Handbook*, rev. ed., New York: McGraw-Hill, 1991.

K.B. Benson and J. Whitaker, *Television and Audio Handbook for Technicians and Engineers*, New York: McGraw-Hill, 1990.

J. Whitaker, *Radio Frequency Transmission Systems Handbook*, Boca Raton: CRC Press, 1991.

J. Whitaker, *Electronic System Maintenance Handbook*, Boca Raton: CRC Press, 1991.

Further Information

Additional information on the topic of television system technology is available from the following sources:

Broadcast Engineering magazine, a monthly periodical dealing with television technology. The magazine, published by Primedia Business located in Overland Park, Kan., is free to qualified subscribers.

The Society of Motion Picture and Television Engineers, which publishes a monthly journal and holds conferences in the fall and winter. The SMPTE is headquartered in White Plains, N.Y.

The Society of Broadcast Engineers, which holds technical conference is throughout the year. The SBE is located in Indianapolis, Ind.

The National Association of Broadcasters, which holds an annual engineering conference and trade show in the spring. The NAB is headquartered in Washington, D.C.

In addition, the following books are recommended:

J. Whitaker and K.B. Benson (Eds.), *Standard Handbook of Video and Television Engineering*, 4th ed., New York: McGraw-Hill, 2003.

J. Whitaker and K.B. Benson (Eds.), *Standard Handbook of Audio and Radio Engineering*, 2nd ed., New York: McGraw-Hill, 2001.

J. Whitaker and K.B. Benson (Eds.), *Standard Handbook of Broadcast Engineering*, New York: McGraw-Hill, 2005.

National Association of Broadcasters Engineering Handbook, 9th ed., Washington, D.C.: NAB, 1997.

1.4   High-Definition Television

Martin S. Roden

History

When standards were first developed for television, few people dreamed of its evolution into a type of universal communication terminal. While the historical standards are acceptable for entertainment video, they are not adequate for many emerging applications. We had to evolve into a high-resolution standard. High-definition TV (HDTV) is a term applied to a broad class of systems in which developments have received worldwide attention.

HDTV can trace its beginning to Japan during the late 1960s. In 1987, the United States FCC (Federal Communications Commission) was petitioned by broadcasters to set aside frequencies for HDTV

broadcasting. The FCC worked for a number of years to develop standards. In the early 1990s, four competing standards were tested. Instead of choosing one of these, a **Grand Alliance** was formed to develop a single set of standards. The FCC approved the resulting standard on December 24, 1996, and the first receivers were introduced in the United States during 1998.

Traditional TV Standards

We begin with a brief review of the traditional television standards.

Japan and North America use the National Television Systems Committee (NTSC) standard that specifies 525 scanning lines per picture, a field rate of 59.94 per second (nominally 60 Hz), and 2:1 interlaced scanning (although there are about 60 fields per second, there are only 30 new frames per second). The aspect ratio (ratio of width to height) is 4:3. The bandwidth of the television signal, including both video and sound, is 6 MHz. In Europe and some countries outside of Europe, the phase-alternation line (PAL), or the sequential color, and memory (SECAM) standard are used. This specifies 625 scanning lines per picture and a field rate of 50 per second. The bandwidth of this type of television signal is 8 MHz.

HDTV systems significantly increase the number of scan lines in a frame and change the aspect ratio to 16:9. Of course, if we were willing to start from scratch and abandon all existing television systems, we could set the bandwidth of each channel to a number greater than 6 (or 8) MHz, thereby achieving higher resolution. The Japanese Broadcasting Corporation (NHK[NAL2]) has done this in their HDTV system. This system permits 1125 lines per frame with 30 frames per second and 60 fields per second (2:1 interlaced scanning). The **aspect ratio** is 16:9. The system is designed for a bandwidth of 10 MHz per channel. With the 1990 launching of the BS-3 satellite, two channels were devoted to this form of HDTV. To fit the channel within a 10-MHz bandwidth (instead of the approximately 50 MHz that would be needed to transmit the HDTV signal using traditional techniques), bandwidth compression was required. It should be noted that the Japanese system primarily uses analog frequency modulation (FM) (the sound is digital). The approach to decreasing bandwidth is multiple sub-Nyquist encoding (**MUSE**). The sampling below Nyquist lowers the bandwidth requirement, but moving images suffer from less resolution.

Europe began its HDTV project in mid-1986 with a joint initiative involving West Germany (Robert Bosch GmbH), the Netherlands (NV Phillips), France (Thomson SA), and the United Kingdom (Thorn/EMI Plc.). The system, termed **Eureka 95** or D2-MAC, has 1152 lines per frame, 50 fields per second, 2:1 interlaced scanning, and a 16:9 **aspect ratio**. A more recent European standard specifies 1250 scanning lines at 50 fields per second. This is known as the Eureka **EU95**. It is significant to note the number of lines specified by Eureka EU95 is exactly twice that of the PAL and SECAM standard currently in use. The field rate is the same, so it is possible to devise compatible systems that would permit reception of HDTV by current receivers (of course, with adapters and without enhanced definition). The HDTV signal requires nominally 30 MHz of bandwidth.

HDTV Formats

In the United States, the FCC ruled (in March 1990) that any new HDTV system must permit continuation of service to contemporary NTSC receivers. This significant constraint applies to terrestrial broadcasting (as opposed to videodisk, videotape, satellite, and cable television). The HDTV signals can be sent on **taboo channels**, those not used in metropolitan areas, so adequate channel separation can be maintained. Thus, these currently unused channels would be used for simulcast signals. Since the HDTV system for the United States uses digital transmission, transmitter power can be less than that used for conventional television—this reduces interference with adjacent channels. Indeed, in heavily populated urban areas (where many stations are licensed for broadcast), the HDTV signals must be severely limited in power.

There are a total of 18 formats specified for digital television, and six of these fall in the HDTV category. The differences between these formats depend on the resolution and on whether progressive or interlaced scanning is used.

The **interlaced scanning** format evolved with the beginnings of broadcast television. The number of frames per second must be high enough to avoid the perception of flicker. But the higher the number of frames per second, the higher the bandwidth of the resulting video signal. In interlaced scanning, the entire picture area is *painted* twice for every frame. On the first scan, the odd numbered lines are traced, while on the second scan, the even numbered lines are filled in. Thus, the screen shows an image twice for each frame. If, for example, you wanted to transmit 30 frames per second, you could do this with 60 scans per second. The human eye would not detect flicker at the 60 per second rate (i.e., your eye integrates the signal and fill in between the scans), but at a rate of 30 per second most people would detect the stroboscopic effect. Of course, in every engineering decision there is a trade off. While interlaced scanning saves bandwidth, it is less effective for rapid motion (i.e., if you were bionic, you would see a type of comb effect when objects moved, like slicing documents in a one-directional paper shredder). Also, with higher definition formats, the eye becomes more sensitive to certain types of flicker.

The alternative to interlaced scanning is *progressive scanning*. In this format, all lines are scanned in sequence, and the entire frame is scanned.

The three formats most often used in HDTV are 720p, 1080i, and 1080p. The number refers to the number of vertical lines, while the letter indicates whether the scanning is progressive or interlaced.

The 720p format forms the field using 720 vertical lines. Each line contains 1280 pixels horizontally. Progressive scanning is used, and 60 complete pictures are sent every second. Although the resolution quality of 720p is lower than the 1080 formats, it provides smooth motion.

The 1080i format forms the field using 1080 vertical lines. Each line contains 1920 pixels horizontally. Interlaced scanning is used, just as in traditional TV, and 30 complete pictures are sent every second. 1080i represents a compromise between 720p and 1080p. It provides the highest possible resolution, but because it uses interlaced scanning, it does not do as well on motion rendition as the progressive scan formats. While this does not have the motion rendition of 1080p, in certain applications your eye cannot tell the difference. When there is very little motion, and when a signal is derived from a traditional motion picture film are two examples of this. In the latter case, traditional celluloid film uses 24 pictures per second flashed on the screen 48 times per second (twice per picture). This was undoubtedly the original inspiration for interlaced scanning.

Signal Processing

When a color television signal is converted from analog to digital (A/D), the luminance, hue, and saturation signals must each be digitized using eight bits of A/D per sample. Therefore, digital transmission of conventional television requires a nominal bit rate of about 216 megabits/s, while uncompressed HDTV nominally requires about 1200 megabits/s. If we were to use a digital modulation system that transmits one bit per hertz of bandwidth, we see that the HDTV signal requires over 1 GHz of bandwidth, yet only 6 MHz is allocated. Significant data compression is required.

In the early 1990s, four digital HDTV approaches were submitted for FCC testing. The four were proposed by General Instrument Corporation, the Advanced Television Research Consortium (composed of NBC, David Sarnoff Research Center, Philips Consumer Electronics, and Thomson Consumer Electronics, Inc.), Zenith Electronics in cooperation with AT&T Bell Labs and AT&T Microelectronics, and the American Television Alliance (General Instrument Corporation and MIT). There were many common aspects to the four proposals, but major differences existed in the data compression approaches. The data compression techniques can be viewed as two-dimensional extensions of techniques used in voice encoding.

Something unprecedented happened in the spring of 1993. The various competing parties decided, with some encouragement from an FCC advisory committee, to merge to form a Grand Alliance. The Alliance consists of seven members: AT&T, General Instrument Corp., MIT, Philips, Sarnoff, Thomson, and Zenith. This permitted the selection of the *best* features of each of the proposals. The advisory committee was then able to spend the fall of 1995 on completion of the proposed HDTV standard. In the following, we describe a generic system. The reader is referred to the references for details.
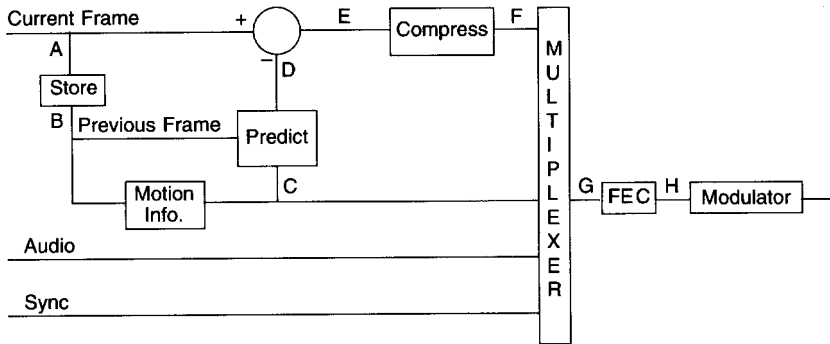
FIGURE 1.30 Block diagram of HDTV transmitter.

Figure 1.30 shows a general block diagram of a digital HDTV transmitter. Each frame from the camera is digitized, and the system has the capability of storing one entire frame. Thus, the processor works with two inputs—the current frame (A) and the previous frame (B). The current frame and the previous frame are compared in a motion detector that generates coded motion information (C). Algorithms used for motion estimation attempt to produce three-dimensional parameters from sequential two-dimensional information. Parameters may include velocity estimates for blocks of the picture.

The parameters from the motion detector are processed along with the previous frame to produce a prediction of the current frame (D). Since the motion detector parameters are transmitted, the receiver can perform a similar prediction of the current frame.

We are describing a compression scheme known as **MPEG-2**. If only a small section of the picture changes, the MPEG-2 encoder only changes that area and leaves the rest of the picture unchanged. On the next frame in the video, only that section of the picture is changed.

MPEG-2 has some problems, but it is a good compression scheme, and it is already an industry standard for digital video for DVD-Videos and some satellite television services. One problem is that MPEG-2 is a *lossy* compression method. That means a higher compression rate gives a poorer picture. There is some loss in picture quality between the digital video camera and the final image produced on the television. However, the quality is still significantly better than an average NTSC image. Using these compression schemes, MPEG-2 can reduce the amount of bits by about 55 to 1. With this compression ratio, much of the information is lost, but the change is not easily noticeable visually. In fact, the compression is carefully designed to match the characteristics of human vision. As one example, the human eye is more sensitive to variation in luminance than to variation in chrominance, and this can be used in the design of the compression algorithms.

The human ear is not as forgiving as the eye. The ear is much more sensitive to subtle changes in sound. Digital TV improves the sound over traditional analog television by using advances in digital sound developed over the last two decades.

Referring back to the block diagram, the predicted current frame is compared to the actual current frame, and a difference signal (E) is generated. This difference signal generally has a smaller dynamic range than the original signal. For example, if the television image is static (i.e., is not changing with time), the difference signal will be zero.

The difference signal is compressed to form the transmitted video signal (F). This compression is performed both in the time and transform domains. Entropy coding of the type used in facsimile can be incorporated to take spatial continuity into account (i.e., a picture usually does not change over the span of a single picture element, so variations of *run length* coding can often compress the data). The compression technique incorporates the MPEG-2 syntax. The actual compression algorithms, based on the **discrete cosine transform**, are adaptive; therefore, a variety of formats can be accommodated (e.g., 1080-line interlaced scanning, 720-line progressive, bi-directional). The main feature is the decreased data rate caused by extracting essential parameters that describe the waveform.
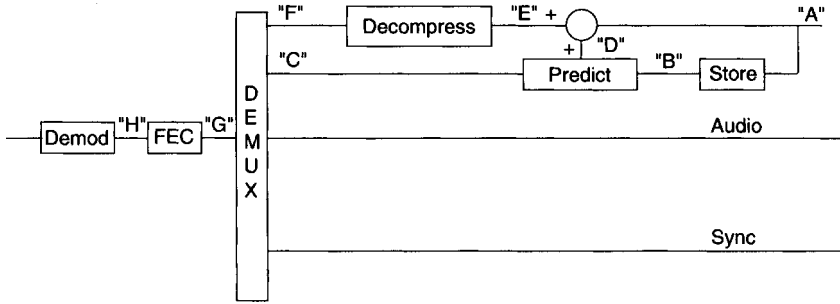
FIGURE 1.31   Block diagram of HDTV receiver.

Four data streams are asynchronously multiplexed to form the information to be transmitted (G). These four signals consist of the coded differential video, the **motion detector** parameters, the digital audio signal (using Dolby Labs' AC-3 digital audio), and the synchronizing signals. Other information can be multiplexed, including various control signals that may be needed by cable operators.

Forward error correction is applied to the multiplexed digital signal to produce an encoded signal (H) that makes the transmission less susceptible to uncorrected bit errors. This is needed because of the anticipated low transmission power rates. Error control is also important because compression can amplify error effects; a single bit error can affect many picture elements.

The encoded data signal forms the input to the modulator. To further conserve bandwidth, a type of quadrature modulation is employed. The actual form is 8-VSB, a variation of **digital vestigial sideband** that includes **trellis coding**. This possesses many of the advantages of quadrature amplitude modulation (QAM).

The corresponding receiver is shown in Figure 1.31. The receiver simply forms the inverse of each transmitter operation. The received signal is first demodulated. The resulting data signal is decoded to remove the redundancy and correct errors. A demultiplexer separates the signal into the original four (or more) data signals. The audio and synchronization signals need no further processing.

The demultiplexed video signal should be the same as the transmitted signal ("F"). We use letters with quotation marks to indicate that the signals are estimates of their transmitted counterpart. This reproduced video signal is decompressed, using the inverse algorithm of that used in the transmitter to yield an estimate of the original differential picture signal ("E"). The predict block in the receiver implements the same algorithm as the transmitter. Its inputs are the reconstructed motion signal ("C") and the previous reconstructed frame ("B"). When the predictor output ("D") is added to the reconstructed differential picture signal ("E"), the result is a reconstructed version of the current frame.

Defining Terms

Aspect ratio: Ratio of frame width to height.
Digital vestigial sideband: A form of digital modulation where a portion of one of the sidebands is partially suppressed.
Discrete cosine transform: A popular format for video compression. The spatial signal is expanded in a cosine series, where the higher frequencies represent increased video resolution.
Entropy coding: A form of data compression that reduces a transmission to a shorter length by reducing signal redundancy.
Eureka 95 and EU95:   European HDTV systems.
Grand Alliance: A consortium formed of seven of the organizations proposing HDTV systems.
Interlaced scanning: A bandwidth reduction technique wherein every other scan line is first transmitted followed by the in-between lines.

Motion detector: A system that compares two adjacent frames to detect differences.
MPEG-2:   Video compression standard devised by the Moving Picture Experts Group.
MUSE: Multiple sub-Nyquist encoding, a technique used in Japanese HDTV system.
Taboo channels: Channels that the FCC does not currently assign in order to avoid interference from adjacent channels.
Trellis coding: A form of digital encoding that provides a constraint (i.e., a structure) to a stream of digital data.

References

S. Banerjee, "Brief overview of the evolution of an technology for high definition television," *Proc. SPIE*, vol. 3582, 1998.

G.W. Beakley, "Channel coding for digital HDTV terrestrial broadcasting," *IEEE Trans. Broadcast.*, vol. 37, no. 4, 1991.

A. Bock and G. Drury, "The introduction of HDTV into digital television networks," *SMPTE J.*, August, 1998.

R. Hopkins, "Digital HDTV broadcasting," *IEEE Trans. Broadcast.*, vol. 37, no. 4, 1991.

P.C. Jain and V. Mitra, "Digital television and video compression," *IETE J. Res.*, vol. 46, no. 5, September-October, 2000.

R.K. Jurgen, Ed., "The challenges of digital HDTV," *IEEE Spectrum*, April, 1991.

Conrad Persson, *Guide to HDTV Systems*, New York: Delmar Learning, 1999.

M.S. Roden, *Analog and Digital Communication Systems*, 5th ed., Los Angeles, CA: Discovery Press, 2003.

W.Y. Zou, "Digital HDTV compression techniques," *IEEE Trans. Broadcast.*, vol. 37, no. 4, 1991.

Further Information

The IEEE Transactions on Broadcasting and Consumer Electronics continue to have periodic articles relating to the HDTV standards and implementation of these standards. Another source of information, though not overly technical, is the periodical Broadcasting and Cable (available by subscription online at www.broadcastingcable.com). Of course, the Web can be used in conjunction with any popular search engine. The vast majority of Web sites have commercial messages, but informational articles and sometimes even university class notes with tutorial information are available.

1.5 Digital Audio Broadcasting

Stanley Salek and Almon H. Clegg

Digital audio broadcasting (DAB) is the term applied to a few independently developed emerging technologies that promise to provide consumers with a new and better aural broadcast system. DAB offers dramatically better reception over existing terrestrial analog AM and FM broadcasts by providing improved audio quality and by integrating technology to provide resistance to interference in stationary and mobile/portable reception environments. Additionally, the availability of a digital data stream direct to consumers opens the prospect of providing extra services to augment basic sound delivery.

At this time, DAB transmission and reception systems are available from at least five system developers. Two of the systems, from Sirius Satellite Radio and XM Satellite Radio, transmit via satellite, using ground-based augmentation transmitters as needed, while the system from the Eureka 147/DAB consortium can be deployed as a satellite, terrestrial, or hybrid satellite/terrestrial system. The remaining systems, from iBiquity Digital Corporation and the Digital Radio Mondiale consortium, are designed for terrestrial signal delivery, operating in established broadcasting frequency bands. This chapter provides a general overview of the technical aspects of DAB systems, as well as a summary of three example transmission systems.

The Need for DAB

Since the early 1980s, the consumer marketplace has undergone a great shift toward digital electronic technology. The explosion of personal computer use has led to greater demands for information, including full multimedia integration. Over the same time period, compact disc (CD) digital audio technology overtook long-playing records and analog audio cassettes as the consumer audio playback medium of choice. Similar digital transcription methods and effects have also been incorporated into commonly available audio and video equipment, including digital videodisc (DVD) and hard disk/flash memory-based recorders and players. Additionally, the ongoing transition to high definition television broadcasting systems has fully incorporated digital methods for video and audio transmission. Because of these market pressures, the radio broadcast industry determined that the existing analog methods of broadcasting needed updating to keep pace with the expectations of the advancing audio marketplace.

Along with providing significantly enhanced audio quality, DAB systems have been developed to overcome the technical deficiencies of existing AM and FM analog broadcasting systems. The foremost problem of current broadcast technology, as perceived by the industry, is its susceptibility to interference. AM medium-wave broadcasts operating in the 530- to 1700-kHz frequency range are prone to disruption by fluorescent lighting and power system distribution networks, as well as by numerous other unintentional radiators, including computer and telephone systems. Additionally, natural effects such as nighttime skywave propagation, interference between stations, and nearby lightning discharges cause irritating service disruption to AM reception. FM broadcast transmissions in the 88- to 108-MHz band are much more resistant to these types of interference. However, multipath propagation and abrupt signal fading, especially found in urban and mountainous areas containing a large number of signal reflectors and shadowers (e.g., buildings and terrain), can seriously degrade FM reception, particularly to mobile receivers.

DAB System Design Goals

DAB systems are designed with several technical goals in mind. The first goal is to create a service that delivers compact disc quality stereo sound for broadcast to consumers. The second goal is to overcome the interference problems of current AM and FM broadcasts, especially under portable and mobile reception conditions. Third, DAB must be spectrally efficient; total bandwidth should be no greater than the bandwidth currently used for corresponding analog broadcasts. Fourth, the DAB system should provide space in its data stream to allow for the addition of ancillary services such as textual program information display, software downloading, or subscription data services. Finally, to foster rapid consumer acceptance DAB receivers must not be overly cumbersome, complex, or expensive.

In addition to these goals, desired features include reduced RF transmission power requirements (when compared to AM and FM broadcast stations having the same signal coverage), a mechanism to seamlessly fill-in coverage areas that are shadowed from the principal transmitted signal, and the ability to easily integrate DAB receivers into personal, home, and automotive sound systems.

Historical Background

DAB development work began in Europe in 1986, with the initial goal to provide high-quality audio services to consumers directly by satellite. Companion terrestrial systems were developed to evaluate the technology, as well as to provide fill-in service in small areas where the satellite signals were shadowed. A consortium of European technical organizations known as Eureka-147/DAB demonstrated the first working terrestrial DAB system in Geneva, Switzerland, in September 1988. Subsequent terrestrial system demonstrations followed in Canada during the summer of 1990 and in the United States in 1991 (http://www.worlddab.org/eureka.aspx).

VHF and UHF transmission frequencies between 200 and 900 MHz were used for the demonstrations with satisfactory results. Because most VHF and UHF frequency bands suitable for DAB are already in use (or in use by/reserved for digital television and other services), an additional Canadian study in 1991 evaluated frequencies near 1500 MHz (L-band) for use as a potential worldwide DAB allocation. This study concluded

that L-band frequencies would support a DAB system such as Eureka-147/DAB, while continuing to meet the overall system design goals.

During the scheduled 1992 World Administrative Radio Conference (WARC-92), frequency allocations for many different radio systems were debated. As a result of WARC-92, a worldwide L-band standard allocation at 1452 to 1492 MHz was designated for both satellite and terrestrial digital radio broadcasting. However, because of existing government and military uses in L-band, the United States was excluded from the standard. Instead, an S-band allocation at 2310 to 2360 MHz was substituted. Additionally, nations including Japan, China, and Russia opted for an extra S-band allocation in the 2535- to 2655-MHz frequency range.

During the same time period, most DAB system development work in the United States shifted from out-band (i.e., UHF, L-band, and S-band) to in-band, because it was uncertain whether S-band frequencies were suitable for terrestrial broadcasting. In-band terrestrial systems merged DAB services with existing medium wave AM and VHF FM broadcasts, using novel adjacent- and co-channel modulating schemes. Between 1992 and 2000, competing system proponents demonstrated proprietary methods of extracting a compatible digital RF signal from co-channel analog AM and FM broadcast transmissions.

Standardization activities for DAB systems were started in the early 1990s by the Consumer Electronics Group of the Electronic Industries Association (EIA), now known as the Consumer Electronics Association (CEA). In 1996, the EIA conducted field-testing of some the systems in the San Francisco Bay area, California. Shortly thereafter, standardization and field-testing activities for IBOC AM and FM systems were started by the National Radio Systems Committee, a joint committee of the National Association of Broadcasters (NAB) and the CEA. In October 2002, the FCC conditionally approved deployment of proprietary analog/digital in-band on-channel (**IBOC**) AM and FM technology developed by system proponent iBiquity Digital Corporation (http://www.ibiquity.com). As of mid-2004, IBOC standards activities continue with a permanent Federal Communications Commission (FCC) authorization of systems usage by all U.S. broadcasters anticipated in the near future.

In 1998, a worldwide consortium known as Digital Radio Mondiale (DRM) was formed, with the goal of producing near FM quality digital broadcasting in AM broadcast bands below 30 MHz (http//www.drm.org). Although there is some technical similarity to the iBiquity AM IBOC system, the non-proprietary DRM system primarily uses an all-digital transmission without a modulated co-channel analog AM component. In 2001, DRM received International Telecommunications Union (ITU) approval, and a technical standard was published. As of mid-2004, a number of international broadcasters have been conducting over-air DRM system testing.

The inception of S-band satellite broadcasting in the United States began in 1997, when the FCC auctioned two 12.5 MHz-wide frequency bands to private operators. Commercial Satellite DAB commenced in late 2001 and 2002, when systems licensees XM Satellite Radio and Sirius Satellite Radio began broadcasting to subscribers. Both systems offer numerous commercial-free music channels and other content (http://www.xmsatelliteradio.com and http://www.sirius.com).

Technical Overview of DAB

Regardless of the actual signal delivery system used, all DAB systems share a common overall topology. Figure 1.32 presents a block diagram of a typical DAB transmission system.

To maintain the highest possible audio quality, program material must originate from digital sources such as CD players and digital audio recorders, or digital audio feeds from network sources. Analog sources such as microphones are converted to a digital audio data stream using an analog-to-digital (A/D) converter, prior to switching or summation with the other digital sources.

The linear digital audio data stream from the studio is then applied to the input of a **source encoder**. The purpose of this device is to reduce the required bandwidth of the audio information, helping to produce a spectrally efficient RF broadcast signal. For example, a 16-bit linear digital audio sampled at 48 kHz (the standard professional rate) requires a data stream of 1.536 megabits/s to transmit a stereo program in a serial format. This output represents a bandwidth of approximately 1.5 MHz, which is much greater than that used by an equivalent analog audio modulating signal [Smyth, 1992]. Source encoders can reduce the data rate by factors of 8:1 or more, yielding a much more efficient modulating signal.
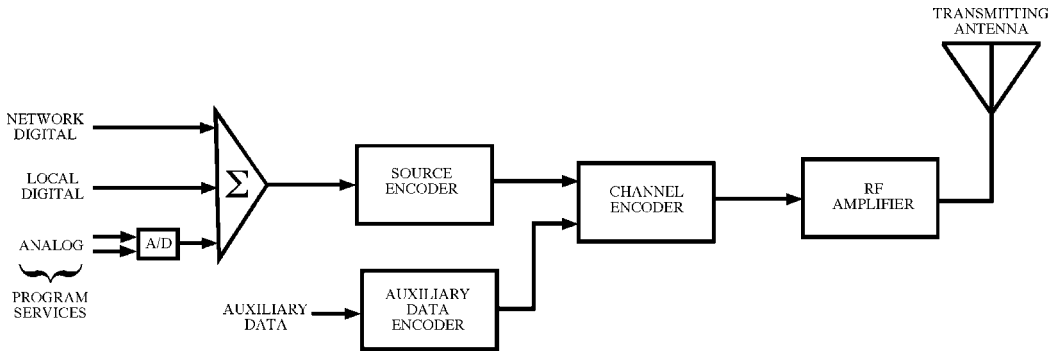
FIGURE 1.32   An example DAB transmission system. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

Following the source encoder, the resulting serial digital signal is applied to the input of the channel encoder, a device that modulates the transmitted RF wave with the reduced-rate audio information. Also auxiliary serial data such as program information and/or receiver control functions can be input to the channel encoder (or combined within the source encoder) for simultaneous transmission.

The **channel encoder** uses sophisticated modulating techniques to accomplish the goals of interference cancellation and high spectral efficiency. Methods of interference cancellation include expansion of time and frequency diversity of the transmitted information, as well as the inclusion of error correction codes in the data stream. Time diversity involves transmitting the same information multiple times by using a pre-determined time interval. Frequency diversity produced by systems such as spread-spectrum, multiple-carrier, or frequency-hopping provides the means to transmit identical data on several different frequencies within the bandwidth of the system. At the receiver, real-time mathematical processes are used to locate the required data on a known frequency at a known time. If the initial information is found to be unusable because of signal interference, the receiver simply uses the same data found on another frequency and/or at another time, producing seamless demodulation.

Spectral efficiency is a function of the modulation system used. Among the modulation formats that have been evaluated for DAB transmission are QPSK, M-ary QAM, and MSK [Springer, 1992]. Using these and other formats, digital transmission systems that use no more spectrum than their analog counterparts have been designed.

The RF output signal of the channel encoder is amplified to the appropriate power level for transmission. Because the carrier-to-noise (C/N) ratio of the modulated waveform is not generally as critical as that required for analog communications systems, relatively low transmission power can often be used. Depending on the sophistication of the data recovery circuits contained in the DAB receiver, the use of C/N ratios as low as 6 dB are possible without causing degradation to the received signal.

DAB reception is largely the inverse of the transmission process, with the inclusion of sophisticated error correction circuits. Figure 1.33 shows a typical DAB receiver.

DAB signal reception is similar to the process used by virtually all radio receivers. A receiving antenna feeds an appropriate stage of RF selectivity and amplification, producing a sample of the coded DAB signal. This signal drives a channel decoder, which reconstructs the audio and auxiliary data streams. To accomplish this task, the channel decoder must demodulate and de-interleave the data contained on the RF carrier and then apply appropriate computational and statistical error correction functions.

The source decoder converts the reduced bit-rate audio stream back to pseudolinear at the original sampling rate. The decoder computationally expands the mathematically reduced data and fills the gaps left from the extraction of irrelevant audio information with averaged code or other masking data. The output of the source decoder feeds audio digital-to-analog (D/A) converters, and the resulting analog stereo audio signal is amplified for the listener.
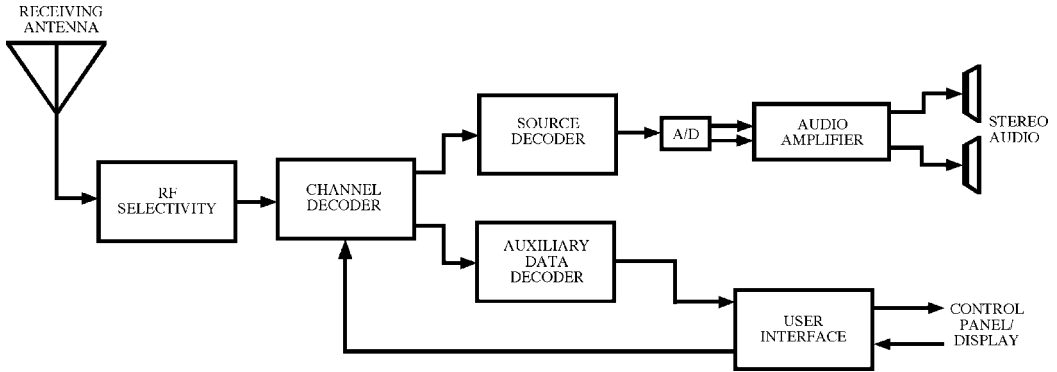
FIGURE 1.33 An example DAB receiver. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

In addition to audio extraction, DAB receivers are generally capable of decoding auxiliary data. This data can be used in conjunction with the user interface to control receiver functions, or for a completely separate purpose. A typical user interface contains a data display screen in addition to the usual receiver tuning and audio controls. This data screen can be used to obtain information about the programming, news reports, sports scores, advertising, or any other useful data sent by the station or an originating network. Also, external interfaces could be used to provide a software link to personal computer systems, or other digital information, and productivity devices.

Audio Compression and Source Encoding

The development of digital audio encoding started with research into pulse-code modulation (PCM) in the late 1930s and then evolved to include work on the principles of digital PCM coding. Linear predictive coding (LPC) and adaptive delta pulse-code modulation (ADPCM) algorithms had evolved in the early 1970s and later were adopted into standards such as C.721 (published by the CCITT) and CD-I (Compact Disc-Interactive). At the same time, algorithms were being invented for use with phoneme-based speech coding. Phonetic coding, a first-generation model-based speech-coding algorithm, was mainly implemented for low bit-rate speech and text-to-speech applications. These classes of algorithms for speech further evolved to include both CELP (Code Excited Linear Predictive) and VSELP (Vector Selectable Excited Linear Predictive) algorithms by the mid-1980s. In the late 1980s, these classes of algorithms were also shown to be useful for high-quality audio music coding. They were used commercially from the late 1970s to the latter part of the 1980s.

Sub-band coders evolved from the early work on quadrature mirror filters in the mid-1970s and continued with polyphase filter-based schemes in the mid-1980s. Hybrid algorithms employing both subband and ADPCM coding were developed in the latter part of the 1970s and standardized (e.g., CCITT G.722) in the mid- to late 1980s. Adaptive transform coders for audio evolved in the mid-1980s from speech coding work done in the late 1970s. By employing psychoacoustic noise-masking properties of the human ear, perceptual encoding evolved from early work of the 1970s and where high-quality speech coders were employed. Music quality bit-rate reduction schemes such as MPEG (Motion Picture Expert Group), PASC (Precision Adaptive Subband Coding), and ATRAC (Adaptive TRansform Acoustic Coding) have been developed. Further refinements to the technology will focus attention on novel approaches such as wavelet-based coding and the use of entropy coding schemes.

Audio coding for digital broadcasting in the early days employed one of the many perceptual encoding schemes previously mentioned or some variation thereof. Fundamentally, they depend on two basic psychoacoustic phenomena: (1) the threshold of human hearing and (2) masking of nearby frequency components. In the early days of hearing research, Harvey Fletcher, a researcher at Bell Laboratories, measured

the hearing of many human beings and published the well-known Fletcher–Munson threshold-of-hearing chart. Basically, it states that depending on the frequency, audio sounds below certain levels cannot be heard by the human ear. Also, the masking effect occurs when two frequencies are very close to each other; when one is higher in acoustic level than the other, the weaker of the two is masked and will not be heard. These two principles allow for as much as 80% of the data representing a musical signal to be discarded.

Figure 1.34 shows how the introduction of frequency components affects the ear's threshold of hearing versus frequency. Figure 1.35 shows how the revised envelope of audibility results in the elimination of components that would not be heard.

The electronic implementation of these algorithms employs a digital filter that breaks the audio spectrum into many subbands, and various coefficient elements are built into the program to decide when it is permissible to remove one or more of the signal components. The details of how the bands are divided and coefficients are determined are usually proprietary to the individual system developers. Standardization groups and system proponents have spent many hours of evaluation, attempting to determine the most accurate coding system for given compression ratios or data transfer rates.

During recent years, the algorithms used for encoding music signals have significantly improved, even though the fundamental methods remain as outlined above. Experience in the art of producing both hardware and software encoders and the frequent use of listening test panels has brought about great improvements. Tweaking of the overall technology has improved sound quality while keeping the compression and transfer rates as needed for the broadcast service to which it is applied. This optimization has benefited the prospects for high quality sound in digital transmission and broadcasting systems.
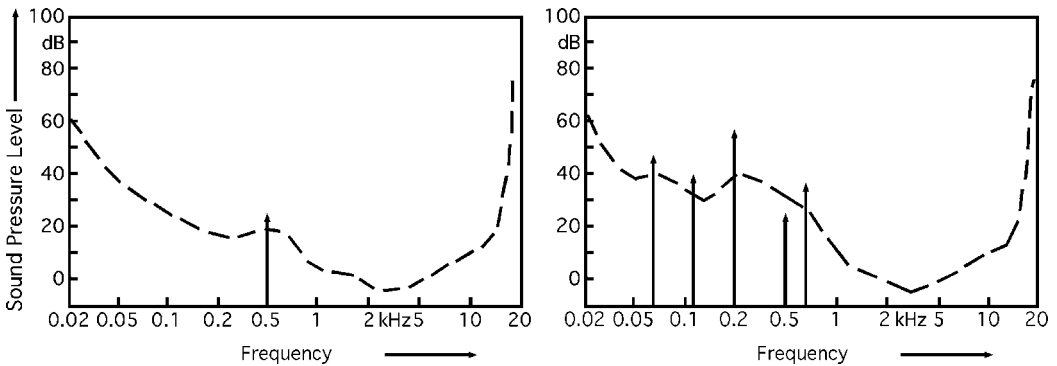
FIGURE 1.34   An example of the masking effect. Based on the hearing threshold of the human ear (dashed line), a 500-Hz sinusoidal acoustic waveform, shown on the left graph, is easily audible at relatively low levels. However, it can be masked by adding nearby higher-amplitude components, as shown on the right. (*Source*: Almon H. Clegg.)
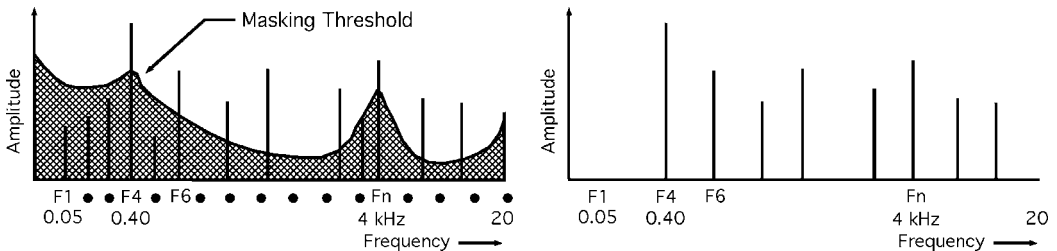
FIGURE 1.35   Source encoders use an empirically derived masking threshold to determine which audio components can be discarded (left). As shown on the right, only the audio components with amplitudes above the masking threshold are retained. (*Source*: Almon H. Clegg.)

Some examples of coding schemes considered for use in broadcasting channels are:

The Perceptual Audio Coder (PAC) extends the idea of perceptual coding to stereo pairs. It uses both L/R (left/right) and M/S (sum/difference matrixing to abate noise generation) coding, switched in both frequency and time in a signal dependent fashion.

The Adaptive Transform Acoustic Coder (ATRAC) has been developed for portable digital audio, but because of its robust coding scheme and availability of commercial integrated circuits, it also lends itself to broadcast use. This coder uses hybrid frequency mapping employing a signal splitting scheme that breaks the band into three sub-bands (0–5.5, 5.5–11, and 11–22 kHz) followed by a suitable dynamically windowed modified discrete cosine transform (MDCT).

Advanced Audio Coding (AAC) is a high-quality perceptual audio coding technology appropriate for many broadcast and electronic music-distribution applications. Coding efficiency is superior to that of MP3, providing higher quality audio at lower transmitted bit rates. Developed and standardized as an ISO/IEC specification by industry leaders, AAC is supported by a growing number of hardware and software manufacturers as the logical successor to MP3 for consumers, computer products, and broadcasting (www.aac-audio.com, http://www.iis.fraunhofer.de/amm/techinf/aac/, and http://www.mp3prozone.com).

Broadcast studios must have the necessary transforms available in their facilities to convert the many digital formats to their broadcast algorithm of choice. Hence, within the broadcast community, expert knowledge of all types and generations of audio compression formats is necessary. If an audio signal passes through several successive compression or bit reduction algorithms, there may be unexpected audible artifacts produced. It is the responsibility of the broadcast engineers to learn the compatibility issues and adjust the signal processing methods to make the necessary conversions with minimum degradation of audio quality. Consideration must be given to the format of the source audio program content and its legacy and compatibility with the coder being used for the final transmission or download. In the case of Internet streaming or downloading, the source coding used may sound best when listened to on a particular software decoder. Hence, there are many choices in the marketplace, each claiming certain improvements over its competitors.

An interesting modern source coding example is CT-aacPlus audio compression. CT-aacPlus is the combination of AAC, a highly efficient global standard based on the collaborative work of several experts on perceptual audio encoding, and additional coding from Coding Technologies' Spectral Band Replication (SBR^{TM}) technology. SBR creates additional bitrate efficiency and utilizes subjective improvements worked out by the German Fraunhofer Institute, the inventor of MP3 (http://www.codingtechnolgies.com and http://www.xmradio.com). With this combination of AAC and SBR, CT-aacPlus has been tested by audio experts and found to provide certain audio improvements. In a double-blind listening test, AAC alone was 33% more efficient when compared to previous generations of competing algorithms. Double-blind listening tests have established that the CT-aacPlus combination is at least 30% more efficient than AAC. As a result of these tests, CT-aacPlus has been adopted by the DRM consortium and accepted by the MPEG standardization group as the reference model for the upcoming versions of MPEG-4 (a standard for a new form of audio and video compression) (http://www.ebu.ch/trev_291-dietz.pdf).

System Example: Eureka-147/DAB

Eureka-147/DAB was the first fully developed DAB system that has demonstrated a capability to meet virtually all the described system goals. Developed by a European consortium, it is an out-band system in that its design is based on the use of a frequency spectrum outside the AM and FM radio broadcast bands. An out-band operation is required because the system packs up to 16 stereophonic broadcast channels (plus auxiliary data) into one contiguous band of frequencies that can occupy a total bandwidth of several megahertz. Thus, overall efficiency is maintained with 16 digital program channels occupying about the same total bandwidth as 16 equivalent analog FM broadcast channels. System developers have promoted Eureka-147/DAB for satellite transmission, as well as for terrestrial applications in locations that have a suitable block of unused spectrum in the L-band frequency range or below.

The ISO/MPEG-2 source encoding/decoding system has been defined for use with the Eureka-147/DAB system. Originally developed by IRT (Institut für Rundfunktechnik) in Germany as MUSICAM (Masking

pattern-adapted Universal Subband Integrated Coding And Multiplexing), the system works by dividing the original digital audio source into 32 subbands. As with the source encoders described earlier, each of the bands is digitally processed to remove redundant information and sounds not perceptible to the human ear. Using this technique, the original audio is sampled at a rate of 768 kilobits/s per channel and is reduced to as little as 96 kilobits/s per channel, representing a compression ratio of 8:1.

The Eureka-147/DAB channel encoder operates by combining the transmitted program channels into a large number of adjacent narrowband RF carriers: each are modulated using QPSK and grouped to maximize the spectrum efficiency known as orthogonal frequency-division multiplex (OFDM). The information to be transmitted is distributed among the RF carriers and is also time-interleaved to reduce the effects of selective fading. A guard interval is inserted between blocks of transmitted data to improve system resistance to intersymbol interference caused by multipath propagation. Convolutional coding is used in conjunction with a Viterbi maximum-likelihood decoding algorithm at the receiver to make constructive use of echoed signals and correct random errors (Alard and Lassalle, 1988).

RF power levels of just a few tens of watts per program channel have been used to provide a relatively wide coverage area, depending on the height of the transmitting antenna above surrounding terrain. This low power level is possible because the system can operate at a C/N ratio of less than 10 dB, in contrast to the more than 30 dB that is required for high-fidelity demodulation of analog FM broadcasts.

Another demonstrated capability of the system is its ability to use gap filler transmitters to augment signal coverage in shadowed areas. A gap filler is simply a system that directly receives the DAB signal at an unobstructed location, provides RF amplification, and retransmits the signal on the same channel into the shadowed area. Because the system can make constructive use of signal reflections (within a time window defined by the guard interval and other factors), the demodulated signal is uninterrupted on a mobile receiver when it travels between an area served by the main signal into the service area of the gap filler.

System Example: iBiquity FM IBOC

As disclosed in recent public FCC proceedings, the iBiquity FM IBOC DAB system provides a novel proprietary method for the transition from analog FM broadcasting to a hybrid digital/analog technology and finally to an all-digital system, using existing broadcast spectrum in the 88–108 MHz FM broadcast band. Source coding uses a proprietary method known as High Definition Coder (HDC), which like CT-aacPlus has its roots in AAC. The present implementation operates at 96 kilobits/s for both stereo channels, representing a compression ratio of 16:1. This ratio is larger if the stream is split to allow both audio and data transmission, or transmission of two or more separate but simultaneously delivered audio programs.

As shown in Figure 1.36, groupings of primary low level OFDM carriers are added to the FM analog signal, both below and above the host signal. Optional extended carriers also may be added, although their presence is not required for basic system operation. In addition to the FM analog host carrier frequency, reference subcarriers (used by the receiver/decoder for synchronization) are included with the IBOC carriers. As illustrated, the injection level of the individual carriers is quite low and is easily contained within the spectral emission mask specified by the U.S. FCC Rules.

Because the ultimate system goal is for a transition to an all-digital broadcast system, the methodology and spectrum loading has been predefined to be compatible with receivers marketed to operate using the hybrid system. Figure 1.37 shows the spectral content of the defined all-digital system. Note,the power of the primary carrier groupings, previously adjacent to the analog FM signal, have been increased, and groupings of secondary carriers have been added in the location of the former FM analog signal spectrum. The higher level OFDM carriers are maintained in an approximate 100- to 200-kHz offset from the carrier frequency so that interference to or from co-channel and adjacent-channel stations is not degraded subsequent to the all-digital broadcasting transition.

For the hybrid (analog compatible) system, the IBOC DAB signal can be combined with the analog FM signal by way of a number of potential methods. Three of the more common combining methods are illustrated in the following block diagrams. In all cases, the FM transmitter analog audio is delayed by several seconds to achieve time alignment with the recovered digital audio signal in the receiver. Thus, the system can
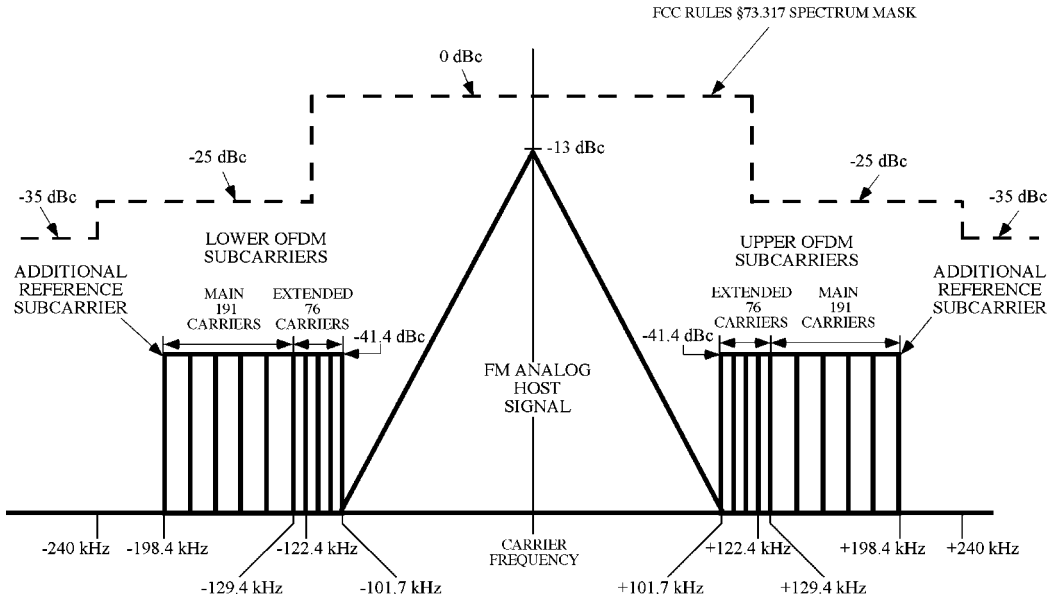
FIGURE 1.36 Spectral content of the iBiquity hybrid FM analog/IBOC DAB system (dBc = decibels relative to carrier level). (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

FIGURE 1.37 Spectral content of the iBiquity all-digital IBOC DAB system intended for use in the present FM broadcast band allocation. The injection level of the secondary carriers is variable, depending on carrier loading. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

automatically and seamlessly return to analog audio demodulation in locations where digital demodulation becomes impaired.

Figure 1.38 depicts the high-level combining method by which full-power analog and digital transmission systems operate in parallel; the two signals are combined prior to the input of the transmitting antenna system. This method has the benefit of maintaining high linearity of the IBOC DAB signal, since it never passes
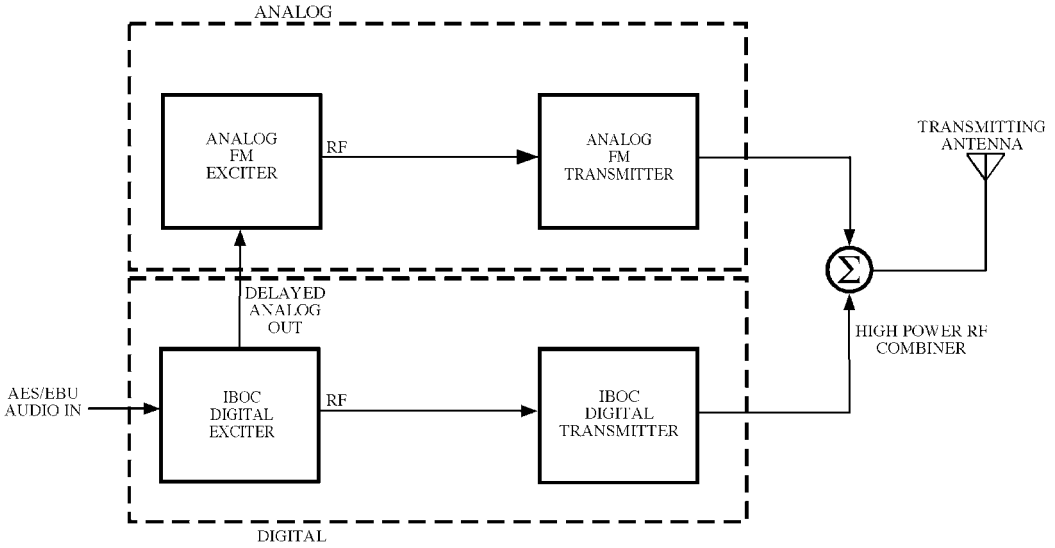
FIGURE 1.38 High-level combining of IBOC DAB and analog FM signals. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)
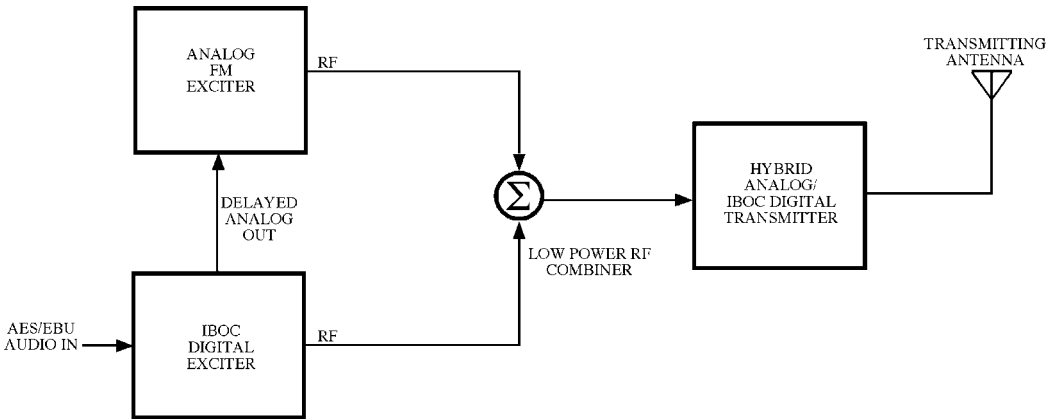
FIGURE 1.39 Low-level combining of IBOC DAB and analog FM signals. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

through the same active stages as the analog FM signal. However, the combiner must be carefully designed to minimize losses, particularly to the analog FM signal, which for many FM stations operates at power levels of 10 kilowatts or more. In order to increase the analog combining efficiency, the digital combining efficiency must be decreased. This requires a significantly higher DAB power than is actually transmitted, often on the order of 10 dB higher. For example, to achieve 100 watts of DAB power at the input to the antenna system, 1,000 watts of DAB power must be generated. This method is most popular with moderate power FM stations, since the amount of DAB power wasted in the combining process remains reasonable.

One alternative to high-level combining is low-level combining, shown in Figure 1.39. This method combines the low power output signals of the analog FM and IBOC exciter units and applies the composite signal to the transmitter. Low level combining requires exceptional linearity in the transmitter RF amplifier stages to prevent distortion of the IBOC carriers. It is generally suitable for transmitters operating at an FM analog output power of 10 kilowatts or less.
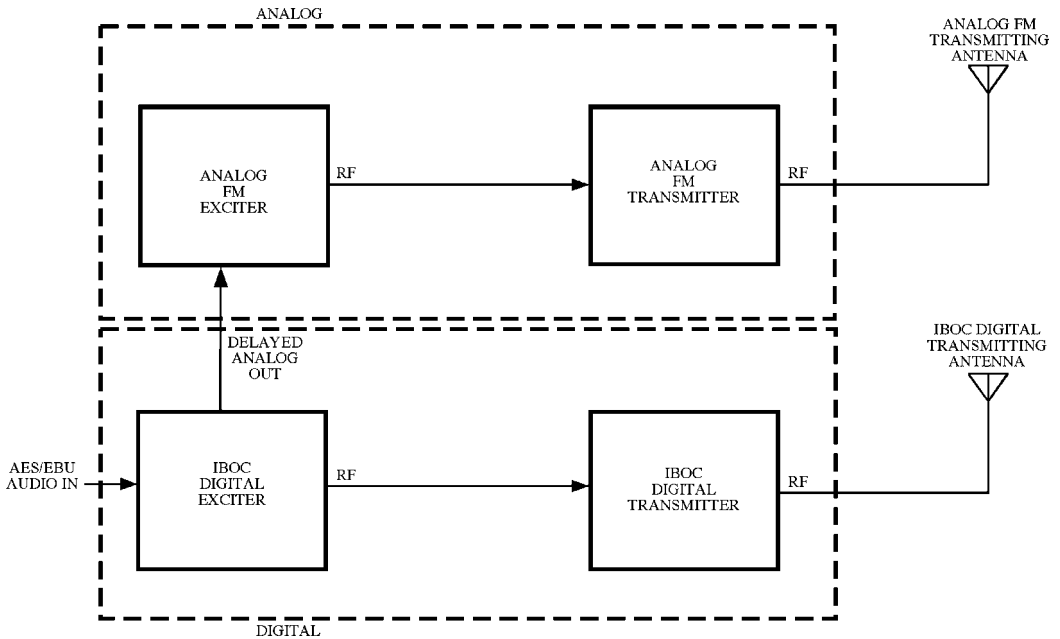
FIGURE 1.40   Combining of IBOC DAB and analog FM signals using separate transmitting antennas. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

A third method of combining analog FM and IBOC carriers is to employ separate transmitting antennas as depicted in Figure 1.40. Field testing has shown this method to be acceptable provided the transmitting antennas have similar radiation patterns, are located close to each other, and are installed at comparable radiation center heights. This method is of significant interest to stations that operate with very high analog FM transmitter power of 30 kilowatts or more, where the high level combining method is impractical due to efficiency issues, and low level combining is not possible due to the prohibitive cost of producing a high power linear transmitter to operate at VHF frequencies. The use of separate antennas is also of interest to many other FM broadcasters, since the method allows the existing analog FM facility to be left in place.

System Example: iBiquity AM IBOC

As with the FM IBOC system, the proprietary iBiquity AM IBOC transmission system is capable of operating in two modes, hybrid and all-digital, with only the hybrid mode exhibiting interoperability with existing analog broadcasting. As with the FM IBOC system, source coding employs HDC and operates at 36 kilobits/s for both stereo channels, representing a compression ratio of almost 43:1. A 400 bit/s data stream is reserved for receiver program identification, but the data path can be expanded further at the expense of using a higher audio compression ratio. According to the developers, HDC is capable of producing acceptable audio quality at bit rates as low as 20 kilobits/s.

Figure 1.41 provides an illustration of the hybrid AM IBOC transmission system spectrum occupancy. Similar to the related FM IBOC system, transmissions are carried by a series of OFDM carriers and separated into primary, secondary, and tertiary groupings. The primary digital carrier groupings are offset ±10 to ±15 kHz from the carrier frequency, while the secondary digital carrier groupings are offset ±5 to ±10 kHz of the carrier frequency. The tertiary carriers are located within ±5 kHz of the carrier frequency, sharing same spectrum as the AM analog host sidebands. As shown by the dashed line, this arrangement of digital carriers fits within the U.S. FCC Rules spectrum mask.
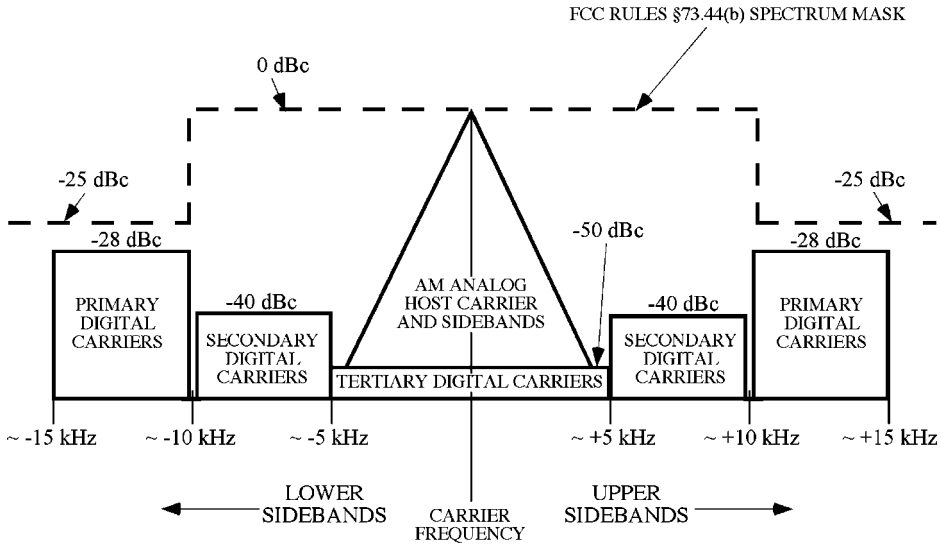
FIGURE 1.41   Spectral content of the iBiquity hybrid AM analog/IBOC DAB system. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

In general, the system requires the analog audio bandwidth be limited to 5 kHz, to protect the secondary digital carriers. An alternate mode of the system allows up to 7.5 kHz analog audio bandwidth, which generates some interference to the secondary digital carriers. Because the digital sidebands carry duplicative information that is inverted about the carrier frequency, recovery of the lower portion of the lower secondary digital subcarriers, along with the upper portion of the upper secondary digital subcarriers, still results in full recovery of the secondary digital information. The drawback is that recovery of both secondary sidebands is required for use of 7.5 kHz analog audio bandwidth, while recovery of just one secondary sideband is required for use of 5 kHz analog audio bandwidth. Thus, IBOC digital service area for 7.5 kHz analog audio bandwidth may be impaired in the presence of adjacent-channel interfering signals.

Figure 1.42 provides a representation of the spectrum occupancy for the defined all-digital IBOC AM system. As with the FM IBOC system, an anticipated transition to this system would take place after most stations have implemented the hybrid IBOC system, and when market indicators have shown IBOC receiver penetration exceeding a defined threshold.

Note, the unmodulated carrier signal is retained in the all-digital system, but the primary digital carrier groupings now immediately surround the carrier, replacing the hybrid system tertiary digital carriers. Defined secondary and tertiary carriers occupy the former positions of the hybrid system secondary digital carrier groupings, and no information is transmitted beyond ±10 kHz of the center carrier frequency. As a result of this design, the anticipated migration to an all-digital transmission should enhance station digital coverage over the hybrid system and show improved performance on transmitting antenna systems exhibiting narrow bandwidth.

Figure 1.43 provides a block diagram of a typical AM IBOC installation. The conversion of an analog station requires the addition of an exciter, an auxiliary service unit (ASU), and appropriate audio processing.

The Auxiliary Service Unit (ASU) accepts AES/EBU digital audio as input, handles sampling rate conversions, and feeds the digital audio inputs of separate AM and IBOC audio processors. The ASU also contains a GPS-synchronized clock to generate the carrier reference and other synchronizing signals used by the exciter. The DAB audio processor controls and levels audio of the IBOC digital transmission, while the analog audio processor controls analog audio. The exciter contains a delay circuit for the analog

FCC RULES §73.44(b) SPECTRUM MASK

0 dBc

-25 dBc

-13 dBc

-13 dBc

-25 dBc

PRIMARY DIGITAL CARRIERS

PRIMARY DIGITAL CARRIERS

-28 dBc

-28 dBc

TERTIARY DIGITAL CARRIERS

SECONDARY DIGITAL CARRIERS

~ -10 kHz　　~ -5 kHz　　　　　　　　~ +5 kHz　　　~ +10 kHz
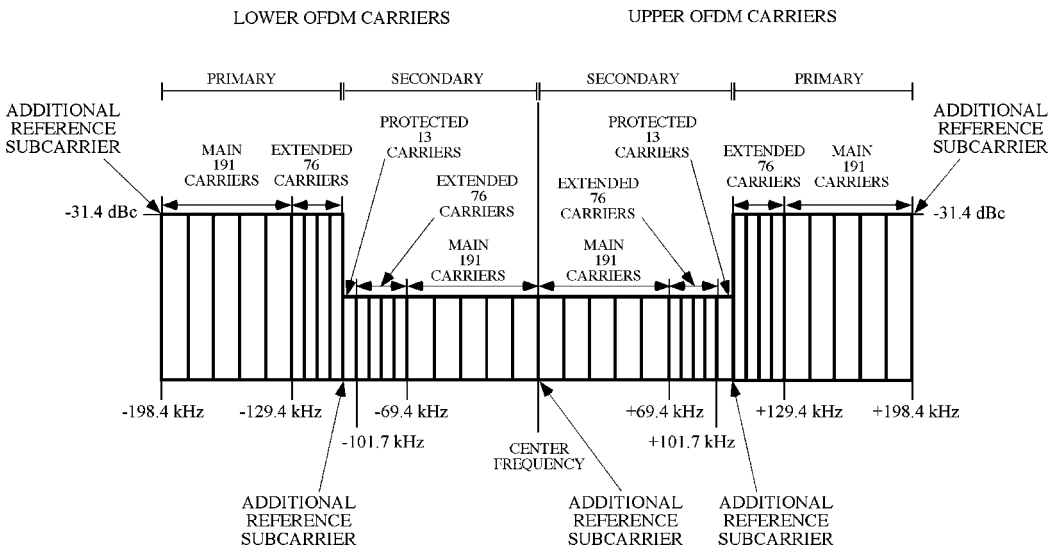
CARRIER FREQUENCY

LOWER SIDEBANDS

UPPER SIDEBANDS
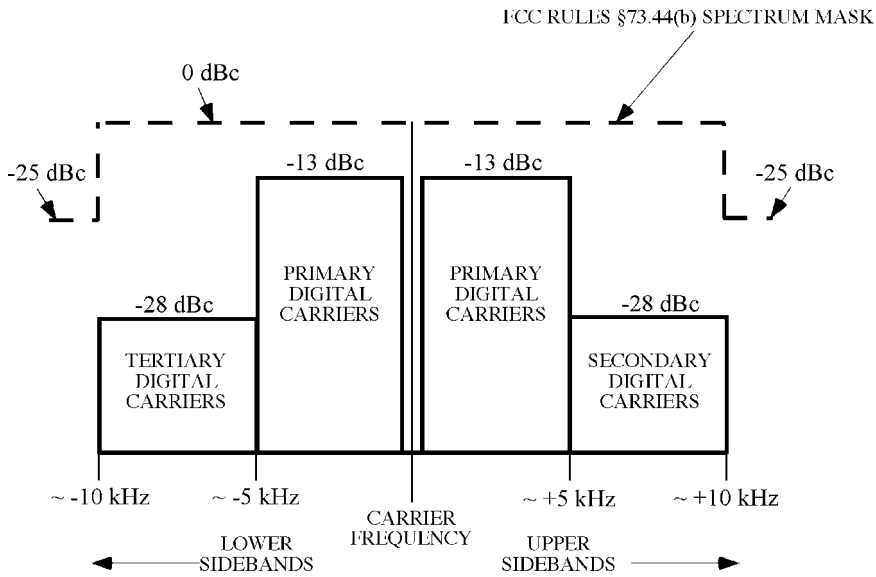
FIGURE 1.42　Spectral content of the iBiquity all-digital IBOC DAB system intended for use in the present medium wave AM broadcast band allocation. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

ANTENNA SYSTEM

GPS RECEIVE ANTENNA

DAB AUDIO PROCESSOR

AUDIO ENCODER

ERROR CORRECTION & INTERLEAVING

AES/EBU AUDIO IN

AUXILIARY SERVICE UNIT

AES/EBU

OFDM MODULATOR

Σ

PHASE / MAGNITUDE CONVERSION

TRANSMITTER

AUDIO

FREQUENCY AND PHASE

ANALOG AUDIO PROCESSOR

AUDIO DELAY

AUDIO LIMITER

10 MHZ, 1PPS REFERENCES AND GPS DATA TO EXCITER
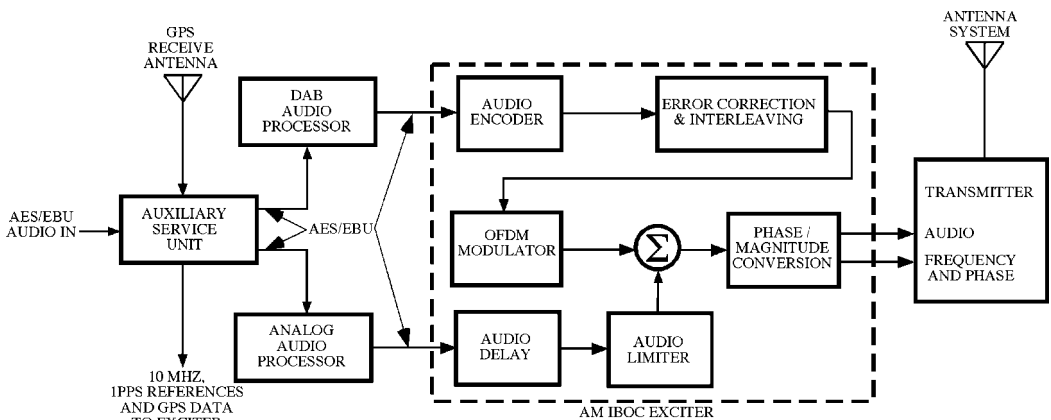
AM IBOC EXCITER

FIGURE 1.43　Combining of IBOC DAB and analog AM signals. (*Source*: Hammett & Edison, Inc., Consulting Engineers.)

audio path so that digital-to-analog transitions in receivers are time-aligned, similar to the FM IBOC system.

The exciter DAB path encodes the source audio, introduces error correction and interleaving overhead, and generates the coded OFDM carriers. The DAB and analog paths are then summed and converted to magnitude/phase signals that feed the broadcast transmitter. The exciter output signals feed the transmitter analog audio input and frequency-determining stages of the transmitter. The hardware topology of an all-digital conversion is the same as that shown, except the analog audio processor and related exciter circuitry are not used.

Defining Terms

Channel encoder: A device that converts source-encoded digital information into an analog RF signal for transmission. The type of modulation used depends on the particular digital audio broadcasting (DAB) system, although most modulation techniques employ methods by which the transmitted signal can be made more resistant to frequency-selective signal fading and multipath distortion effects.

In-Band On-Channel (IBOC): A method of combining digital audio broadcasting signals, intended for reception by radio listeners, with the signals of existing analog broadcast services operating in the medium wave AM and VHF FM broadcast bands.

Gap filler: A low-power transmitter that boosts the strength of transmitted DAB RF signals in areas that normally would be shadowed due to terrain obstruction. Gap fillers can operate on the same frequency as DAB transmissions, or on alternate channels that can be located by DAB receivers using automatic switching.

Source encoder: A device that substantially reduces the data rate of linearly digitized audio signals by taking advantage of the psychoacoustic properties of human hearing, eliminating redundant and subjectively irrelevant information from the output signal. Transform source encoders work entirely within the frequency domain, while time-domain source encoders work primarily in the time domain. Source decoders reverse the process, using various masking techniques to simulate the properties of the original linear data.

References

M. Alard and R. Lassalle, "Principles of modulation and channel coding for digital broadcasting for mobile receivers," *Advanced Digital Techniques for UHF Satellite Sound Broadcasting*, European Broadcasting Union (collected papers), 1988, pp. 47–69.

R. Bruno, "Digital audio and video compression, present and future," presented to the Delphi Club, Tokyo, Japan, July, 1992.

G. Chouinard and F. Conway, "Broadcasting systems concepts for digital sound," *Proc. 45th Annu. Broadcast Eng. Conf.*, National Association of Broadcasters, 1991, pp. 257–266.

F. Conway, R. Voyer, S. Edwards, and D. Tyrie, "Initial experimentation with DAB in Canada," *Proc. 45th Annu. Broadcast Eng. Conf.*, National Association of Broadcasters, 1991, pp. 281–290.

S. Kuh and J. Wang, "Communications systems engineering for digital audio broadcast," *Proc. 45th Annu. Broadcast Eng. Conf.*, National Association of Broadcasters, 1991, pp. 267–272.

P.H. Moose and J.M. Wozencraft, "Modulation and coding for DAB using multi-frequency modulation," *Proc. 45th Annu. Broadcast Eng. Conf.*, National Association of Broadcasters, 1991, pp. 405–410.

S. Smyth, "Digital audio data compression," *Broadcast Eng. Mag.*, 1992, pp. 52–60.

K.D. Springer, Interference Between FM and Digital M-PSK Signals in the FM Band, National Association of Broadcasters, 1992.

First Report and Order to Mass Media Docket No. 99-325, "Digital Audio Broadcasting Systems and their Impact on the Terrestrial Radio Broadcast Service", Federal Communications Commission, Washington, D.C., 2002.

Further Information

The National Association of Broadcasters (http://www.nab.org) publishes periodic reports on the technical, regulatory, and political status of DAB in the United States. Additionally, their Broadcast Engineering Conference proceedings published since 1990 contain a substantial amount of information on emerging DAB technologies.

IEEE Transactions on Broadcasting is typically published quarterly by the Institute of Electrical and Electronics Engineers, Inc. and periodically includes papers on digital broadcasting (http://www.ieee.org).

Additionally, the periodic newspaper publication *Radio World* (http://www.rwonline.com) provides continuous coverage of DAB technology, including proponent announcements, system descriptions, field test reports, and broadcast industry reactions.

2

Equalization

Richard C. Dorf
University of California

Zhen Wan
University of Texas

In bandwidth-efficient digital communication systems the effect of each symbol transmitted over a time dispersive channel extends beyond the time interval used to represent that symbol. The distortion caused by the resulting overlap of received symbols is called **intersymbol interference** (ISI) [Lucky et al., 1968]. ISI arises in all pulse-modulation systems, including frequency-shift keying (FSK), phase-shift keying (PSK), and quadrature amplitude modulation (QAM) [Lucky et al., 1968]. However, its effect can be most easily described for a baseband PAM system.

The purpose of an **equalizer,** placed in the path of the received signal, is to reduce the ISI as much as possible to maximize the probability of correct decisions.

2.1 Linear Transversal Equalizers

Among the many structures used for equalization, the simplest is the transversal (tapped delay line or non-recursive) equalizer shown in Figure 2.1. In such an equalizer the current and past values $r(t - nT)$ of the received signal are linearly weighted by equalizer coefficients (tap gains) c_n and summed to produce the output. In the commonly used digital implementation, samples of the received signal at the symbol rate are stored in a digital shift register (or memory), and the equalizer output samples (sums of products) $z(t_0 + kT)$ or z_k are computed digitally, once per symbol, according to

$$z_k = \sum_{n=0}^{N-1} c_n r(t_0 + kT - nt)$$

where N is the number of equalizer coefficients and t_0 denotes sample timing.

The equalizer coefficients, c_n, $n = 0, 1,\ldots,N - 1$, may be chosen to force the samples of the combined channel and equalizer impulse response to zero at all but one of the NT-spaced instants in the span of the equalizer. Such an equalizer is called a *zero-forcing* (ZF) equalizer [Lucky, 1965].

If we let the number of coefficients of a ZF equalizer increase without bound, we would obtain an infinite-length equalizer with zero ISI at its output. An infinite-length zero-ISI equalizer is simply an inverse filter, which inverts the folded frequency response of the channel. Clearly, the ZF criterion neglects the effect of noise altogether. A finite-length ZF equalizer is approximately inverse to the folded frequency response of

FIGURE 2.1 Linear transversal equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 648. With permission.)

the channel. Also, a finite-length ZF equalizer is guaranteed to minimize the peak distortion or worst-case ISI only if the peak distortion before equalization is less than 100% [Lucky, 1965].

The *least-mean-squared* (LMS) equalizer [Lucky et al., 1968] is more robust. Here the equalizer coefficients are chosen to minimize the mean squared error (MSE)—the sum of squares of all the ISI terms plus the noise power at the output of the equalizer. Therefore, the LMS equalizer maximizes the signal-to-distortion ratio (S/D) at its output within the constraints of the equalizer time span and the delay through the equalizer.

Automatic Synthesis

Before regular data transmission begins, automatic synthesis of the ZF or LMS equalizers for unknown channels may be carried out during a training period. During the training period, a known signal is transmitted and a synchronized version of this signal is generated in the receiver to acquire information about the channel characteristics. The automatic adaptive equalizer is shown in Figure 2.2. A noisy but unbiased estimate:

$$\frac{\delta e^2 k}{\delta c_n(k)} = 2e_k r(t_0 + kT - nT)$$

is used. Thus, the tap gains are updated according to

$$c_n(k+1) = c_n(k) - \Delta e_k r(t_0 + kT - nT), \quad n = 0, 1, \ldots, N-1$$

where $c_n(k)$ is the nth tap gain at time k, e_k is the error signal, and Δ is a positive adaptation constant or step size, error signals $e_k = z_k - q_k$ can be computed at the equalizer output and used to adjust the equalizer coefficients to reduce the sum of the squared errors. Note $q_k = \hat{x}_k$.

FIGURE 2.2 Automatic adaptive equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 651. With permission.)

The most popular equalizer adjustment method involves updates to each tap gain during each symbol interval. The adjustment to each tap gain is in a direction opposite to an estimate of the gradient of the MSE with respect to that tap gain. The idea is to move the set of equalizer coefficients closer to the unique optimum set corresponding to the minimum MSE. This symbol-by-symbol procedure developed by Widrow and Hoff [Feher, 1987] is commonly referred to as the *stochastic gradient* method.

Adaptive Equalization

After the initial training period (if there is one), the coefficients of an adaptive equalizer may be continually adjusted in a *decision-directed* manner. In this mode the error signal $e_k = z_k - q_k$ is derived from the final (not necessarily correct) receiver estimate $\{q_k\}$ of the transmitted sequence $\{x_k\}$ where q_k is the estimate of x_k. In normal operation the receiver decisions are correct with high probability, so that the error estimates are correct often enough to allow the adaptive equalizer to maintain precise equalization. Moreover, a decision-directed adaptive equalizer can track slow variations in the channel characteristics or linear perturbations in the receiver front end, such as slow jitter in the sampler phase.

2.2 Nonlinear Equalizers

Decision-Feedback Equalizers

A decision-feedback equalizer (DFE) is a simple nonlinear equalizer [Monsen, 1971], which is particularly useful for channels with severe amplitude distortion and uses decision feedback to cancel the interference from symbols which have already been detected. Figure 2.3 shows the diagram of the equalizer.

The equalized signal is the sum of the outputs of the forward and feedback parts of the equalizer. The forward part is like the linear transversal equalizer discussed earlier. Decisions made on the equalized signal are fed back via a second transversal filter. The basic idea is that if the values of the symbols already detected are

FIGURE 2.3 Decision-feedback equalizer. (*Source:* K. Feher, *Advanced Digital Communications,* Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 655. With permission.)

known (past decisions are assumed to be correct), then the ISI contributed by these symbols can be canceled exactly, by subtracting past symbol values with appropriate weighting from the equalizer output.

The forward and feedback coefficients may be adjusted simultaneously to minimize the MSE. The update equation for the forward coefficients is the same as for the linear equalizer. The feedback coefficients are adjusted according to

$$b_m(k+1) = b_m(k) + \Delta e_k \hat{x}_{k-m} \quad m = 1, \dots, M$$

where $\hat{x}_k$ is the kth symbol decision, $b_m(k)$ is the mth feedback coefficient at time k, and there are M feedback coefficients in all. The optimum LMS settings of b_m, $m = 1, \dots, M$, are those that reduce the ISI to zero, within the span of the feedback part, in a manner similar to a ZF equalizer.

Fractionally Spaced Equalizers

The optimum receive filter in a linear modulation system is the cascade of a filter matched to the actual channel, with a transversal T-spaced equalizer [Forney, 1972]. The fractionally spaced equalizer (FSE), by virtue of its sampling rate, can synthesize the best combination of the characteristics of an adaptive matched filter and a T-spaced equalizer, within the constraints of its length and delay. A T-spaced equalizer, with symbol-rate sampling at its input, cannot perform matched filtering. A *fractionally spaced equalizer* can effectively compensate for more severe delay distortion and deal with amplitude distortion with less noise enhancement than a T-equalizer.

A fractionally spaced transversal equalizer [Monsen, 1971] is shown in Figure 2.4. The delay-line taps of such an equalizer are spaced at an interval τ, which is less than, or a fraction of, the symbol interval T. The tap spacing τ is typically selected such that the bandwidth occupied by the signal at the equalizer input is $|f| < 1/2\tau$: that is, τ-spaced sampling satisfies the sampling theorem. In an analog implementation, there is no other restriction on τ, and the output of the equalizer can be sampled at the symbol rate. In a digital implementation τ must be KT/M, where K and M are integers and $M > K$. (In practice, it is convenient to choose $\tau = T/M$, where M is a small integer, e.g., 2.) The received signal is sampled and shifted into the equalizer delay line at a rate M/T, and one input is produced each symbol interval (for every M input sample). In general, the equalizer output is given by

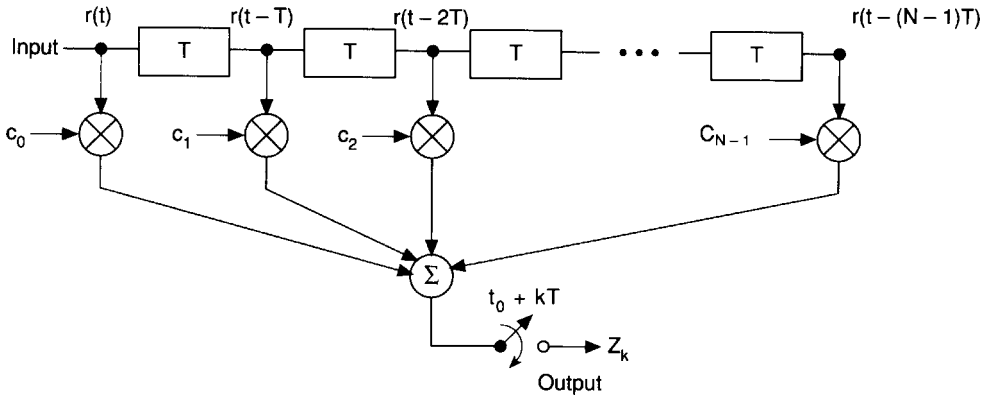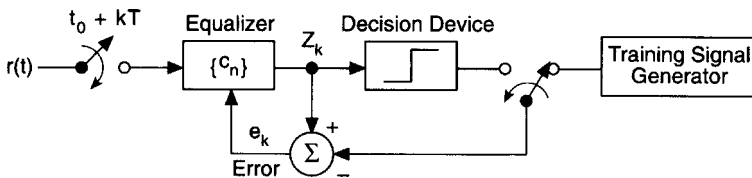$$z_k = \sum_{n=0}^{N-1} c_n r\left(t_0 + kT - \frac{nKT}{M} \right)$$
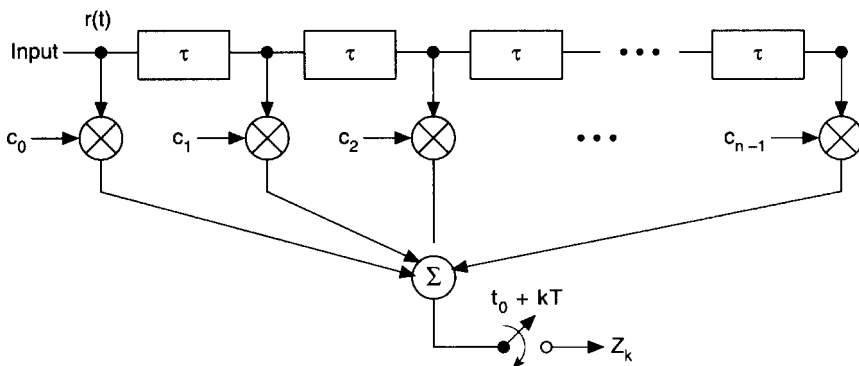
FIGURE 2.4 Fractionally spaced equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, p. 656. With permission.)

The coefficients of a *KT/M* equalizer may be updated once per symbol based on the error computed for that symbol according to

$$c_n(k+1) = c_n(k) - \Delta e_k r\left(t_0 + kT - \frac{nKT}{M}\right), \quad n = 0, 1, \ldots, N-1$$

2.3 Linear Receivers

When the channel does not introduce any amplitude distortion, the linear receiver is optimum with respect to the ultimate criterion of minimum probability of symbol error. The *conventional linear receiver* consists of a matched filter, a symbol-rate sampler, an infinite-length *T*-spaced equalizer, and a memoryless detector. The linear receiver structure is shown in Figure 2.5.

In the conventional linear receiver, a memoryless threshold detector is sufficient to minimize the probability of error; the equalizer response is designed to satisfy the zero-ISI constraint, and the matched filter is designed to minimize the effect of the noise while maximizing the signal.

Matched Filter

The matched filter is the linear filter that maximizes $(S/N)_{\text{out}} = s_0^2(t)/n_0^2(t)$ of Figure 2.6 and has a transfer function given by

$$H(f) = K \frac{S^*(f)}{P_n(f)} e^{-j\omega t_0}$$

where $Sf = F[s(t)]$ is the Fourier transform of the known input signal $s(t)$ of duration T sec. $P_n(f)$ is the PSD of the input noise, t_0 is the sampling time when $(S/N)_{\text{out}}$ is evaluated, and K is an arbitrary real nonzero constant.

A general representation for a matched filter is illustrated in Figure 2.6. The input signal is denoted by $s(t)$ and the output signal by $s_0(t)$. Similar notation is used for the noise.

2.4 Nonlinear Receivers

When amplitude distortion is present in the channel, a memoryless detector operating on the output of this receiver filter no longer minimizes symbol error probability. Recognizing this fact, several authors have
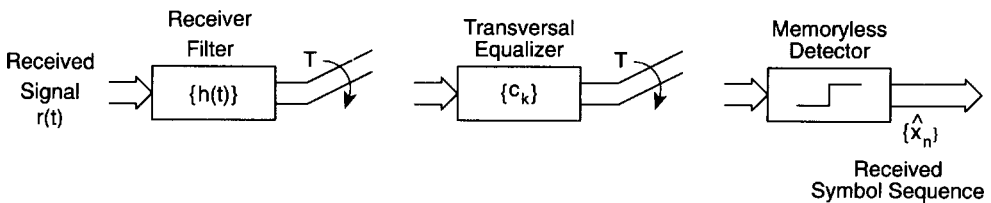
FIGURE 2.5 Conventional linear receiver.
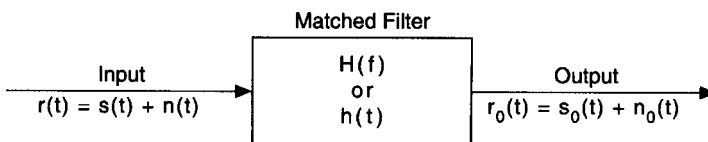
FIGURE 2.6 Matched filter. (*Source:* L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, p. 497, 1990. With permission.)

investigated optimum or approximately optimum nonlinear receiver structures subject to a variety of criteria [Lucky, 1973].

Decision-Feedback Equalizers

A DFE takes advantage of the symbols that have already been detected (correctly with high probability) to cancel the ISI due to these symbols without noise enhancement. A DFE makes memoryless decisions and cancels all trailing ISI terms. Even when the whitened matched filter (WMF) is used as the receive filter for the DFE, the DFE suffers from a reduced effective signal-to-noise ratio, and error propagation, due to its inability to defer decisions.

An infinite-length DFE receiver takes the general form (shown in Figure 2.7) of a forward linear receive filter, symbol-rate sampler, canceler, and memoryless detector. The symbol-rate output of the detector is then used by the feedback filter to generate future outputs for cancellation.

Adaptive Filters for MLSE

For unknown and/or slowly time-varying channels, the receive filter must be adaptive in order to obtain the ultimate performance gain from MLSE (maximum-likelihood sequence estimation). Secondly, the complexity of the MLSE becomes prohibitive for practical channels with a large number of ISI terms. Therefore, in a practical receiver, an adaptive receive filter may be used prior to Viterbi detection to limit the time spread of the channel as well as to track slow time variation in the channel characteristics [Falconer and Magee, 1973].

Several adaptive receive filters are available that minimize the MSE at the input to the Viterbi algorithm. These methods differ in the form of constraint [Falconer and Magee, 1973] on the desired impulse response (DIR) which is necessary in this optimization process to exclude the selection of the null DIR corresponding to no transmission through the channel. The general form of such a receiver is shown in Figure 2.8.

One such constraint is to restrict the DIR to be causal and to restrict the first coefficient of the DIR to be unity. In this case the delay (LT) in Figure 2.8 is equal to the delay through the Viterbi algorithm and the first coefficient of $\{b_k\}$ is constrained to be unity.

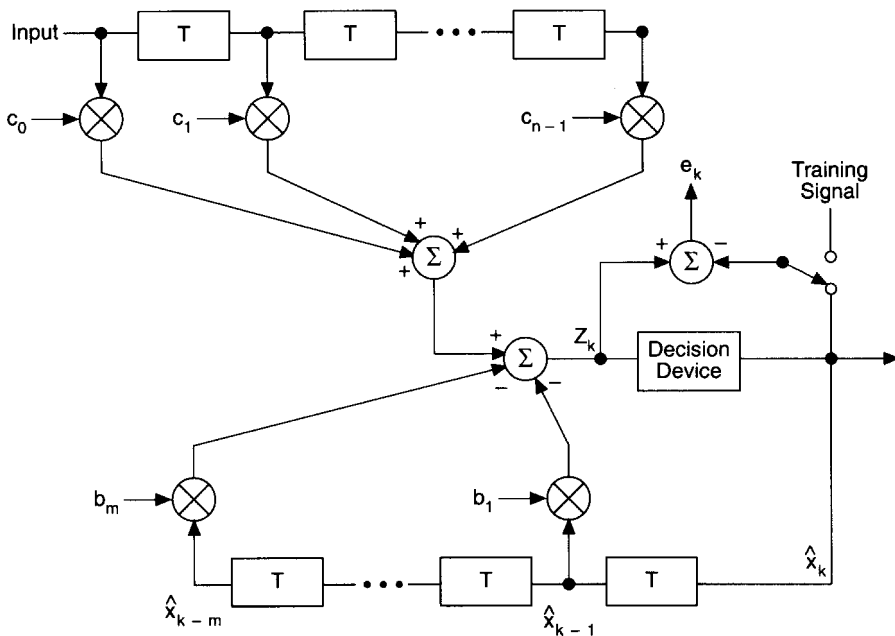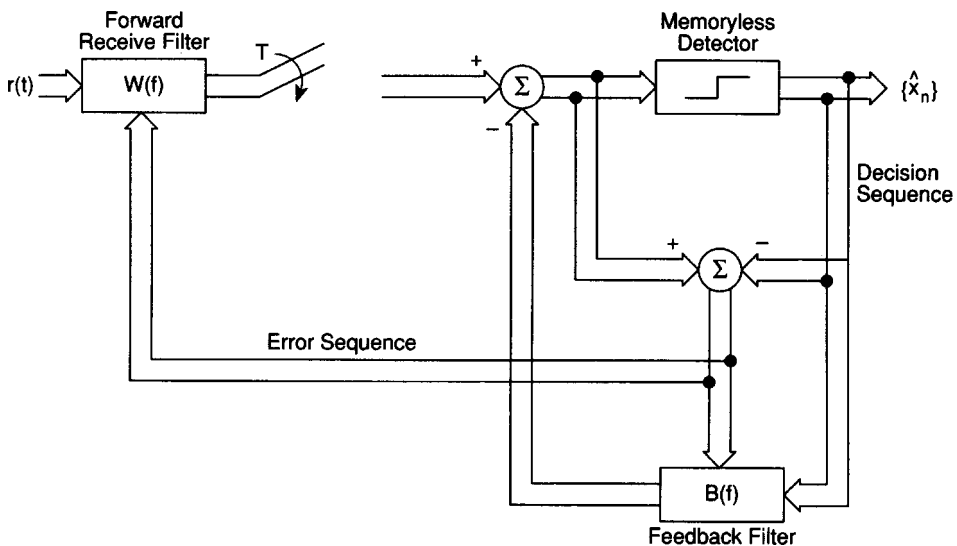**FIGURE 2.7** Conventional decision-feedback receiver. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 675. With permission.)
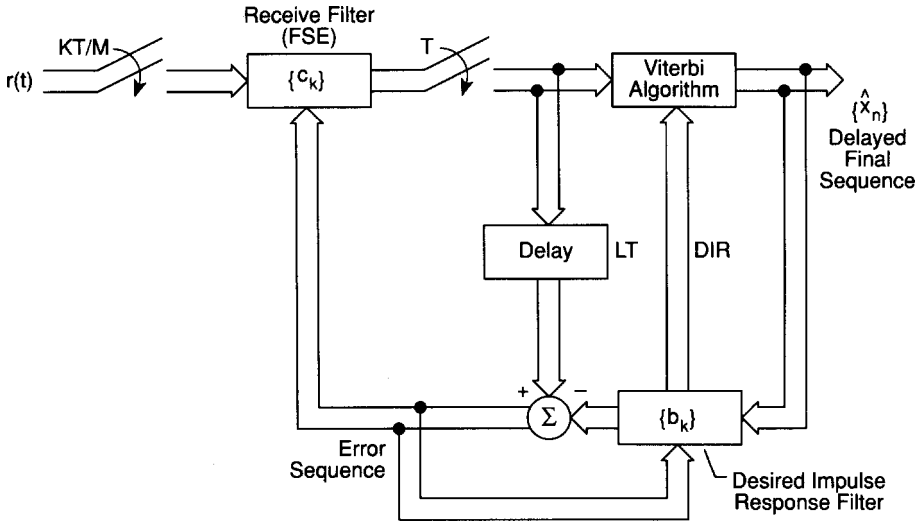
FIGURE 2.8 General form of adaptive MLSE receiver with finite-length DIR. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 684. With permission.)

The least restrictive constraint on the DIR is the unit energy constraint proposed by Falconer and Magee [1973]. This leads to yet another form of the receiver structure as shown in Figure 2.8. However, the adaptation algorithm for updating the DIR coefficients $\{b_k\}$ is considerably more complicated [Falconer and Magee, 1973]. Note that the fixed predetermined WMF and *T*-spaced prefilter combination of Falconer and Magee [1973] has been replaced in Figure 2.8 by a general fractionally spaced adaptive filter.

Defining Terms

Equalizer: A filter used to reduce the effect of intersymbol interference.
Intersymbol interference: The distortion caused by the overlap (in time) of adjacent symbols.

References

L.W. Couch, *Digital and Analog Communication Systems,* New York: Macmillan, 1990.
D.D. Falconer and F.R. Magee, Jr., "Adaptive channel memory truncation for maximum likelihood sequence estimation," *Bell Syst. Technical Journal,* vol. 5, pp. 1541–1562, November 1973.
K. Feher, *Advanced Digital Communications,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.
G.D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Information Theory,* vol. IT-88, pp. 363–378, May 1972.
R.W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. Journal,* vol. 44, pp. 547–588, April 1965.
R.W. Lucky, "A survey of the communication theory literature: 1968–1973," *IEEE Trans. Information Theory,* vol. 52, pp. 1483–1519, November 1973.
R.W. Lucky, J. Salz, and E.J. Weldon, Jr., *Principles of Data Communication,* New York: McGraw-Hill, 1968.
P. Monsen, "Feedback equalization for fading dispersive channels," *IEEE Trans. Information Theory,* vol. IT-17, pp. 56–64, January 1971.

This page intentionally left blank

3

Optical Communication

Thomas E. Darcie
AT&T Bell Laboratories

Joseph C. Palais
Arizona State University

Alan E. Willner
University of Southern California

Reza Khosravani
Sonoma State University

3.1 Lightwave Technology for Video Transmission

Thomas E. Darcie

Lightwave technology has revolutionized the transmission of analog and, in particular, video information. Because the light output intensity from a semiconductor laser is linearly proportional to the injected current, and the current generated in a photodetector is linearly proportional to the incident optical intensity, analog information is transmitted as modulation of the optical intensity. The lightwave system is analogous to a **linear** electrical link, where current or voltage translates linearly into optical intensity. High-speed semiconductor lasers and photodetectors enable intensity-modulation bandwidths greater than 10 GHz. Hence, a wide variety of radio frequency (RF) and microwave applications have been developed [Darcie, 1990].

Converting microwaves into intensity-modulated (IM) light allows the use of optical fiber for transmission in place of bulky inflexible coaxial cable or microwave waveguide. Since the fiber attenuation is 0.2–0.4 dB/km, compared with several decibels per meter for waveguide, entirely new applications and architectures are possible. In addition, the signal is confined tightly to the core of single-mode fiber, where it is immune to electromagnetic interference, cross talk, or spectral regulatory control.

To achieve these advantages, several limitations must be overcome. The conversion of current to light intensity must be linear. Several nonlinear mechanisms must be avoided by proper laser design or by the use of various linearization techniques. Also, because the photon energy is much larger than in microwave systems, the signal fidelity is limited by quantum or **shot noise**.

This section describes the basic technology for the transmission of various video formats. We begin by describing the most common video formats and defining transmission requirements for each. Sources of noise, including shot noise, **relative intensity noise** (RIN), and receiver noise are then quantified. Limitations imposed by source nonlinearity, for both **direct modulation** of the laser bias current and **external modulation** using an interferometric LiNbO$_3$ modulator, are compared. Finally, several other impairments caused by **fiber nonlinearity** or **fiber dispersion** are discussed.

Video Formats and Applications

Each video format represents a compromise between transmission bandwidth and robustness or immunity to impairment. With the exception of emerging digital formats, each is also an entrenched standard that often reflects the inefficiencies of outdated technology.

FM Video

Frequency-modulated (FM) video has served for decades as the basis for satellite video transmission [Pratt and Bostian, 1986], where high signal-to-noise ratios (SNRs) are difficult to achieve. Video information with a bandwidth of $B_v = 4.2$ MHz is used to FM modulate an RF carrier. The resulting channel bandwidth B is given by

$$B \sim \Delta f_{pp} + 2f_m \tag{3.1}$$

where Δf_{pp} is the frequency deviation (22.5 MHz) and f_m is the audio subcarrier frequency (6.8 MHz). As a result of this bandwidth expansion to typically 36 MHz, a high SNR can be obtained for the baseband video bandwidth B_v even if the received carrier-to-noise ratio (CNR) over the FM bandwidth B is small. The SNR is given by

$$SNR = CNR + 10 \, \log\left[\frac{3B}{2B_v}\left(\frac{\Delta f_{pp}}{B_v}\right)\right] + W + PE \tag{3.2}$$

where W is a weighting factor (13.8 dB) that accounts for the way the eye responds to noise in the video bandwidth, and PE is a pre-emphasis factor (0–5 dB) that is gained by emphasizing the high-frequency video components to improve the performance of the FM modulator. High-quality video (SNR = 55 dB) requires a CNR of only 16 dB. This is achieved easily in a lightwave transmission system.

Applications for lightwave FM video transmission include links to satellite transmission facilities, transport of video between cable television company head-ends (super-trunking), and perhaps delivery of video to subscribers over large fiber distribution networks [Way et al., 1988; Olshansky et al., 1988].

AM-VSB Video

The video format of choice, both for broadcast and cable television distribution, is AM-VSB. Each channel consists of an RF carrier that is amplitude modulated (AM) by video information. Single-sideband vestigial (VSB) filtering is used to minimize the bandwidth of the modulated spectrum. The resultant RF spectrum is dominated by the remaining RF carrier, which is reduced by typically 5.6 dB by the AM, and contains relatively low-level signal information, including audio and color subcarriers. An AM-VSB channel requires a bandwidth of only 6 MHz, but CNRs must be at least 50 dB.

For cable distribution, many channels are frequency-division multiplexed (FDM), separated nominally by 6 MHz (8 MHz in Europe), over the bandwidth supported by the coaxial cable. A typical 60-channel cable system operates between 55.25 and 439.25 MHz. Given the large dynamic range required to transmit both the remaining RF carrier and the low-level sidebands, transmission of this multichannel spectrum is a challenge for lightwave technology.

The need for such systems in cable television distribution systems has motivated the development of suitable high-performance lasers. Before the availability of lightwave AM-VSB systems, cable systems used long (up to 20 km) trunks of coaxial cable with dozens of cascaded electronic amplifiers to overcome cable loss. Accumulations of distortion and noise, as well as inherent reliability problems with long cascades, were serious limitations.

Fiber AM-VSB trunk systems can replace the long coaxial trunks so that head-end quality video can be delivered deep within the distribution network [Chiddix et al., 1990]. Inexpensive coaxial cable extends from the optical receivers at the ends of the fiber trunks to each home. Architectures in which the number of electronic amplifiers between each receiver and any home is approximately three or fewer offer a good compromise between cost and performance. The short spans of coaxial cable support bandwidths approaching 1 GHz, two or three times the bandwidth of the outdated long coaxial cable trunks. With fewer active components, reliability is improved. The cost of the lightwave components can be small compared to the overall system cost. These compelling technical and economic advantages resulted in the immediate demand for lightwave AM-VSB systems.

Compressed Digital Video

The next generation of video formats will be the product of compressed digital video (CDV) technology [Netravali and Haskel, 1988]. For years digital "NTSC-like" video required a bit rate of approximately 100 Mbps. CDV technology can reduce the required bit rate to less than 5 Mbps. This compression requires complex digital signal processing and large-scale circuit integration, but advances in chip and microprocessor design have made inexpensive implementation of the compression algorithms feasible.

Various levels of compression complexity can be used, depending on the ultimate bit rate and quality required. Each degree of complexity removes different types of redundancy from the video image. The image is broken into blocks of pixels, typically 8×8. By comparing different blocks and transmitting only the differences (DPCM), factors of 2 reduction in bit rate can be obtained. No degradation of quality need result. Much of the information within each block is imperceptible to the viewer. Vector quantization (VQ) or discrete-cosine transform (DCT) techniques can be used to eliminate bits corresponding to these imperceptible details. This intraframe coding can result in a factor of 20 reduction in the bit rate, although the evaluation of image quality becomes subjective. Finally, stationary images or moving objects need not require constant retransmission of every detail. Motion compression techniques have been developed to eliminate these interframe redundancies. Combinations of these techniques have resulted in coders that convert NTSC-like video (100 Mbps uncompressed) into a few megabits per second and HDTV images (1 Gbps uncompressed) into less than 20 Mbps.

CDV can be transmitted using time-division multiplexing (TDM) and digital lightwave systems or by using each channel to modulate an RF carrier and transmitting using analog lightwave systems. There are numerous applications for both alternatives. TDM systems for CDV are no different from any other digital transmission system and will not be discussed further.

Using RF techniques offers an additional level of RF compression, wherein advanced multilevel modulation formats are used to maximize the number of bits per hertz of bandwidth [Feher, 1987]. Quadrature-amplitude modulation (QAM) is one example of multilevel digital-to-RF conversion. For example, 64-QAM uses 8 amplitude and 8 phase levels and requires only 1 Hz for 5 bits of information. As the number of levels, hence the number of bits per hertz, increases, the CNR of the channel must increase to maintain error-free transmission. A 64-QAM channel requires a CNR of approximately 30 dB.

A synopsis of the bandwidth and CNR requirements for FM, AM-VSB, and CDV is shown in Figure 3.1. AM-VSB requires high CNR but low bandwidth. FM is the opposite. Digital video can occupy a wide area, depending on the degree of digital and RF compression. The combination of CDV and QAM offers the possibility of squeezing a high-quality video channel into 1 MHz of bandwidth, with a required CNR of 30 dB. This drastic improvement over AM-VSB or FM could have tremendous impact on future video transmission systems.

Intensity Modulation

As mentioned in the introduction, the light output from the laser should be linearly proportional to the injected current. The laser is prebiased to an average output intensity L_0. Many video channels are combined electronically, and the total RF signal is added directly to the laser current. The optical modulation depth (m) is defined as the ratio of the peak modulation L_0 for one channel, divided by L_0. For 60-channel AM-VSB systems, m is typically near 4%.

The laser (optical carrier) is modulated by the sum of the video channels that are combined to form the total RF signal spectrum. The resultant optical spectrum contains sidebands from the IM superimposed on unintentional frequency modulation, or **chirp**, that generally accompanies IM. This complex optical spectrum must by understood if certain subtle impairments are to be avoided.

A photodetector converts the incident optical power into current. Broadband InGaAs photodetectors with responsivities (R_0) of nearly 1.0 A/W and bandwidths greater than 10 GHz are available. The detector generates a dc current corresponding to the average received optical power L_r and the complete RF modulation spectrum that was applied at the transmitter. An accoupled electronic preamplifier is used to remove the dc component and boost the signal to usable levels.

FIGURE 3.1 Bandwidth versus carrier-to-noise ratio (CNR) required for AM-VSB, FM, and digital video. Increasingly complex digital compression techniques reduce the bit rate required for NTSC-like video from 100 Mbps to less than 5 Mbps. Bandwidth efficient RF techniques like QAM minimize the bandwidth required for each bit rate but require greater CNRs.

Noise Limitations

The definition of CNR deserves clarification. Depending on the video format and RF modulation technique, the RF power spectrum of the modulated RF carrier varies widely. For AM-VSB video the remaining carrier is the dominant feature in the spectrum. It is thereby convenient to define the CNR as the ratio of the power remaining in the carrier to the integrated noise power in a 4-MHz bandwidth centered on the carrier frequency. For FM or digitally modulated carriers, the original carrier is not generally visible in the RF spectrum. It is then necessary to define the CNR as the ratio of the integrated signal power within the channel bandwidth to the integrated noise power.

Shot Noise

Shot noise is a consequence of the statistical nature of the photodetection process. It results in a noise power spectral density, or electrical noise power per unit bandwidth (dBm/Hz) that is proportional to the received photocurrent I_r ($= R_0 L_r$). The total shot noise power in a bandwidth B is given by

$$N_s = 2eI_rB \tag{3.3}$$

where e is the electronic charge.

With small m, the detected signal current is a small fraction of the total received current. The root mean square (rms) signal power for one channel is

$$P_s = \frac{1}{2}(mI_r)^2 \tag{3.4}$$

The total shot noise power then limits the CNR (P_s/N_s) to a level referred to as the quantum limit. Received powers near 1 mW are required if CNRs greater than 50 dB are to be achieved for 40- to 80-channel AM-VSB systems.

Receiver Noise

Receiver noise is generated by the electronic amplifier used to boost the detected photocurrent to usable levels. The easiest receiver to build consists of a *pin* photodiode connected directly to a low-noise 50- to 75-Ω amplifier, as shown in Figure 3.2(a). The effective input current noise density, (*n*), for this simple receiver is given by

$$n^2 = \frac{4kTF}{R_{\mathrm{L}}} \tag{3.5}$$

where *k* is the Boltzmann constant, *T* is the absolute temperature, *F* is the **noise figure** of the amplifier, and R_{L} is the input impedance. For a 50-Ω input impedance and $F = 2$, $n = 20\,\mathrm{pA}/\sqrt{\mathrm{Hz}}$.

A variety of more complicated receiver designs can reduce the noise current appreciably [Kasper, 1988]. The example shown in Figure 3.2(b) uses a high-speed FET. R_{L} can be increased to maximize the voltage developed by the signal current at the FET input. Input capacitance becomes a limitation by shunting high-frequency components of signal current. High-frequency signals are then reduced with respect to the noise generated in the FET, resulting in poor high-frequency performance. Various impedance matching techniques have been proposed to maximize the CNR for specific frequency ranges.
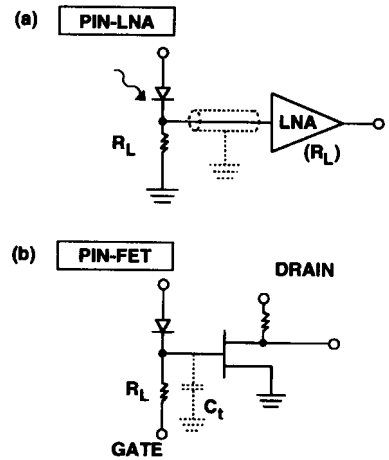
FIGURE 3.2 Receivers for broadband analog lightwave systems. Coupling a *pin* to a low-noise amplifier (a) is simple, but improved performance can be obtained using designs like the *pin* FET (b). C_{t} is the undesirable input capacitance.

Relative Intensity Noise

Relative intensity noise (RIN) can originate from the laser or from reflections and **Rayleigh backscatter** in the fiber. In the laser, RIN is caused by spontaneous emission in the active layer. Spontaneous emission drives random fluctuations in the number of photons in the laser which appear as a random modulation of the output intensity, with frequency components extending to tens of gigahertz. The noise power spectral density from RIN is I_{r}^2 RIN, where RIN is expressed in decibels per hertz.

RIN is also caused by component reflections and double-Rayleigh backscatter in the fiber, by a process called multipath interference. Twice-reflected signals arriving at the detector can interfere coherently with the unreflected signal. Depending on the modulated optical spectrum of the laser, this interference results in noise that can be significant [Darcie et al., 1991].

The CNR, including all noise sources discussed, is given by

$$\mathrm{CNR} = \frac{m^2 I_{\mathrm{r}}^2}{2B[n^2 + 2eI_{\mathrm{r}} + I_{\mathrm{r}}^2 \mathrm{RIN}]} \tag{3.6}$$

All sources of intensity noise are combined into RIN. Increasing *m* improves the CNR but increases the impairment caused by nonlinearity, as discussed in the following subsection. The optimum operating value for *m* is then a balance between noise and distortion.

Figure 3.3 shows the noise contributions from shot noise, receiver noise, and RIN. For FM or digital systems, the low CNR values required allow operation with small received optical powers. Receiver noise is then generally the limiting factor. Much larger received powers are required if AM-VSB noise requirements are to be met. Although detecting more optical power helps to overcome shot and receiver noise, the ratio of signal to RIN remains constant. RIN can be dominant in high-CNR systems, when the received power is large. AM-VSB systems require special care to minimize all sources of RIN. The dominant noise source is then shot noise, with receiver noise and RIN combining to limit CNRs to within a few decibels of the quantum limit.

Linearity Requirements

Source linearity limits the depth of modulation that can be applied. Linearity, in this case, refers to the linearity of the current-to-light-intensity (I–L) conversion in the laser or voltage-to-light (V–L) transmission for an external modulator. Numerous nonlinear mechanisms must be considered for direct modulation, and no existing external modulator has a linear transfer function.

A Taylor-series expansion of the I–L or V–L characteristic, centered at the bias point, results in linear, quadratic, cubic, and higher-order terms. The linear term describes the efficiency with which the applied signal is converted to linear intensity modulation. The quadratic term results in second-order distortion, the cubic produces third-order distortion, and so on.

Requirements on linearity can be derived by considering the number and spectral distribution of the distortion products generated by the nonlinear mixing between carriers in the multichannel signal. Second-order nonlinearity results in sum and difference ($f_i \pm f_j$) mixing products for every combination of the two channels. This results in as many as 50 second-order products within a single channel, in a 60-channel AM-VSB system with the standard U.S. frequency plan. Similarly, for third-order distortion, products result from mixing among all combinations of three channels.

FIGURE 3.3   Current noise densities from receivers, RIN, and shot noise as a function of total received photocurrent. Receiver noise is dominant in FM or some digital systems where the total received power is small. The solid line for receiver noise represents the noise current for a typical 50-Ω low-noise amplifier. More sophisticated receiver designs could reduce the noise to the levels shown approximately by the dotted lines. RIN and shot noise are more important in AM-VSB systems.

However, since the number of combinations of three channels is much larger than for two, up to 1130 third-order products can interfere with one channel. The cable industry defines the **composite second-order** (**CSO**) distortion as the ratio of the carrier to the largest group of second-order products within each channel. For third-order distortion, the **composite triple beat** (**CTB**) is the ratio of the carrier to the total accumulation of third-order distortion at the carrier frequency in each channel.

The actual impairment from these distortion products depends on the spectrum of each RF channel and on the exact frequency plan used. A typical 42-channel AM-VSB frequency plan, with carrier frequencies shown as the vertical bars on Figure 3.4, results in the distributions of second- and third-order products shown in

FIGURE 3.4   Second-order (a) and third-order (b) distortion products for a 42-channel AM-VSB system. The maximum number of second-order products occurs at the lowest frequency channel, where 30 products contribute to the CSO. The maximum number of third-order products occurs near the center channel, where 530 products contribute to the CTB.

Figure 3.4(a) and (b), respectively. Since the remaining carrier is the dominant feature in the spectrum of each channel, the distortion products are dominated by the mixing between these carriers. Because high-quality video requires that the CSO is –60 dBc (dB relative to the carrier), each sum or difference product must be less than –73 dBc. Likewise, for the CTB to be less than 60 dB, each product must be less than approximately –90dB.

FM or CDV systems have much less restrictive linearity requirements, because of the reduced sensitivity to impairment. Distortion products must be counted, as with the AM-VSB example described previously, but each product is no longer dominated by the remaining carrier. Because the carrier is suppressed entirely by the modulation, each product is distributed over more than the bandwidth of each channel. The impairment resulting from the superposition of many uncorrelated distortion products resembles noise. Quantities analogous to the CSO and CTB can be defined for these systems.

Laser Linearity

Several factors limit the light-versus-current (L–I) linearity of directly modulated lasers. Early work on laser dynamics led to a complete understanding of resonance-enhanced distortion (RD). RD arises from the same carrier-photon interaction within the laser that is responsible for the relaxation-oscillation resonance.

The second-harmonic distortion ($2f_i$) and two-tone third-order distortion ($2f_i - f_j$) for a typical 1.3-μm wavelength directly modulated semiconductor laser are shown in Figure 3.5 [Darcie et al., 1986]. Both distortions are small at low frequencies but rise to maxima at half the relaxation resonance frequency. AM-VSB systems are feasible only within the low-frequency window. FM or uncompressed digital systems require enough bandwidth per channel that multichannel systems must operate in the region of large RD. Fortunately, the CNR requirements allow for the increased distortion. The large second-order RD can be avoided entirely by operating within a one-octave

FIGURE 3.5 Resonance distortion for directly modulated laser with resonance frequency of 7 GHz. Both the second-harmonic $2f_i$ and two-tone third-order $2f_i$ f_j distortion peak near half the resonance frequency and are small at low frequency. Also shown is the same third-order distortion for an external modulator biased at the point of zero second-order distortion.

frequency band (e.g., 2–4 GHz), such that all second-order products are out of band.

Within the frequency range between 50 and 500 MHz, nonlinear gain and loss, intervalence-band absorption, and, more importantly, spatial-hole burning (SHB) and carrier leakage can all be significant. Carrier leakage prevents all of the current injected in the laser bond wire from entering the active layer. This leakage must be reduced to immeasurable levels for AM-VSB applications.

SHB results from the nonuniform distribution of optical power along the length of the laser. In DFB lasers, because of the grating feedback, the longitudinal distribution of optical power can be highly nonuniform. This results in distortion [Takemoto et al., 1990] that can add to or cancel other distortion, making it, in some cases, a desirable effect.

Clipping

Even if all nonlinear processes were eliminated, the allowable modulation would be limited by the fact that the minimum output power is zero. Typical operating conditions with, for example, 60 channels, each with an average modulation depth (m) near 4%, result in a peak modulation of 240%. Although improbable, modulations of more than 100% result in clipping.

The effects of clipping were first approximated by Saleh [1989], who calculated the modulation level at which the total power contained in all orders of distortion became appreciable. Even for perfectly linear lasers, the modulation depth is bounded to values beyond which all orders of distortion increase rapidly. Assuming that half the total power in all orders of distortion generated by clipping is distributed evenly over each of N channels, clipping results in a carrier-to-interference ratio (CIR) given by

$$\text{CIR} = \sqrt{2\pi}\,\frac{(1 + 6\mu^2)}{\mu^3}\,e^{1/2\mu^2} \tag{3.7}$$

where the rms modulation index μ is

$$\mu = m\sqrt{N/2} \tag{3.8}$$

External Modulation

Laser-diode-pumped YAG lasers with low RIN and output powers greater than 200 mW have been developed recently. Combined with linearized external LiNbO$_3$ modulators, these lasers have become high-performance competitors to directly modulated lasers. YAG lasers with external modulation offer a considerable increase in launched power, and the low RIN of the YAG laser translates into a slight CNR improvement. The most challenging technical hurdle is to develop a linear low-loss optical intensity modulator.

Low-loss LiNbO$_3$ Mach–Zehnder modulators are available with insertion losses less than 3 dB, modulation bandwidths greater than a few gigahertz, and switching voltages near 5 V. The output intensity of these modulators is a sinusoidal function of the bias voltage. By prebiasing to 50% transmission, modulation applied to the Mach–Zehnder results in the most linear intensity modulation. This bias point, which corresponds to the point of inflection in the sinusoidal transfer function, produces zero second-order distortion. Unfortunately, the corresponding third-order distortion is approximately 30 dB worse than a typical directly modulated DFB laser, at low frequencies. This comparison is shown on Figure 3.5. For high-frequency applications where RD is important, external modulators can offer improved linearity. A means of linearizing the third-order nonlinearity is essential for AM-VSB applications.

Various linearization techniques have been explored. The two most popular approaches are feedforward and predistortion. Feedforward requires that a portion of the modulated output signal be detected and compared to the original applied voltage signal to provide an error signal. This error signal is then used to modulate a second laser, which is combined with the first laser such that the total instantaneous intensity of the two lasers is a replica of the applied voltage. In principle, this technique is capable of linearizing any order of distortion and correcting RIN from the laser.

Predistortion requires less circuit complexity than feedforward. A carefully designed nonlinear circuit is placed before the nonlinear modulator, such that the combined transfer function of the predistorter-modulator is linear. Various nonlinear electronic devices or circuits can act as second- or third-order predistorters. Difficulties include matching the frequency dependence of the predistorter with that of the modulator, hence achieving good linearity over a wide frequency range. Numerous circuit designs can provide reductions in third-order distortion by 15 dB.

Miscellaneous Impairments

Laser chirp can cause problems with direct laser modulation. Chirp is modulation of the laser frequency caused by modulation of the refractive index of the laser cavity in response to current modulation. The interaction of chirp and chromatic dispersion in the fiber can cause unacceptable CSO levels for AM-VSB systems as short as a few kilometers. Dispersion converts the FM into IM, which mixes with the signal IM to produce second-order distortion [Phillips et al., 1991]. These systems must operate at wavelengths corresponding to low fiber dispersion, or corrective measures must be taken.

Chirp also causes problems with any optical component that has a transmission that is a function of optical frequency. This can occur if two optical reflections conspire to form a weak interferometer or in an **erbium-doped fiber amplifier** (EDFA) that has a frequency-dependent gain [Kuo and Bergmann, 1991]. Once again, the chirp is converted to IM, which mixes with the signal IM to form second-order distortion.

Although externally modulated systems are immune to chirp-related problems, fiber nonlinearity, in the form of stimulated Brillouin scattering (SBS), places a limit on the launched power. SBS, in which light is scattered from acoustic phonons in the fiber, causes a rapid decrease in CNR for launched powers greater than approximately 10 mW [Mao et al., 1991]. Since the SBS process requires high optical powers within a narrow optical spectral width (20 MHz), it is a problem only in low-chirp externally modulated systems. Chirp in DFB systems broadens the optical spectrum so that SBS is unimportant.

Summary

A wide range of applications for transmission of video signals over optical fiber has been made possible by refinements in lightwave technology. Numerous technology options are available for each application, each with advantages or disadvantages that must be considered in context with specific system requirements. Evolution of these video systems continues to be driven by development of new and improved photonic devices.

Defining Terms

Chirp: Modulation of the optical frequency that occurs when a laser is intensity modulated.

Composite second order (CSO):   Ratio of the power in the second-order distortion products to power in the carrier in a cable television channel.

Composite triple beat (CTB):   Same as CSO but for third-order distortion.

Direct modulation: Modulation of the optical intensity output from a semiconductor diode laser by direct modulation of the bias current.

Erbium-doped fiber amplifier: Fiber doped with erbium that provides optical gain at wavelengths near 1.55 μm when pumped optically at 0.98 or 1.48 μm.

External modulation: Modulation of the optical intensity using an optical intensity modulator to modulate a constant power (cw) laser.

Fiber dispersion: Characteristic of optical fiber by which the propagation velocity depends on the optical wavelength.

Fiber nonlinearity: Properties of optical fibers by which the propagation velocity, or other characteristic, depends on the optical intensity.

Lightwave technology: Technology based on the use of optical signals and optical fiber for the transmission of information.

Linear: Said of any device for which the output is linearly proportional to the input.

Noise figure: Ratio of the output signal-to-noise ratio (SNR) to the input SNR in an amplifier.

Rayleigh backscatter: Optical power that is scattered in the backwards direction by microscopic inhomogeneities in the composition of optical fibers.

Relative intensity noise: Noise resulting from undesirable fluctuations of the optical power detected in an optical communication system.

Shot noise: Noise generated by the statistical nature of current flowing through a semiconductor *p-n* junction or photodetector.

References

J.A. Chiddix, H. Laor, D.M. Pangrac, L.D. Williamson, and R.W. Wolfe, "AM video on fiber in CATV systems, need and implementation," *IEEE J. Selected Areas in Communications,* vol. 8, p. 1229, 1990.

T.E. Darcie, "Subcarrier multiplexing for lightwave networks and video distribution systems," *IEEE J. Selected Areas in Communications,* vol. 8, p. 1240, 1990.

T.E. Darcie, G.E. Bodeep, and A.A.M. Saleh, "Fiber-reflection-induced impairments in lightwave AM-VSB CATV systems," *IEEE J. Lightwave Technol.,* vol. 9, no. 8, pp. 991–995, Aug. 1991.

T.E. Darcie, R.S. Tucker, and G.J. Sullivan, "Intermodulation and harmonic distortion in IaGaAsP lasers," *Electron. Lett.,* vol. 21, 665–666, erratum; vol 22, p. 619, 1986.

K. Feher, Ed., *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987.

B.L. Kasper, "Receiver design," in *Optical Fiber Telecommunications II,* S.E. Miller and I.P. Kaminow, Eds., San Diego: Academic Press, 1988.

C.Y. Kuo and E.E. Bergmann, "Erbium-doped fiber amplifier second-order distortion in analog links and electronic compensation," *IEEE Photonics Technol. Lett.,* vol. 3, p. 829, 1991.

X.P. Mao, G.E. Bodeep, R.W. Tkach, A.R. Chraplyvy, T.E. Darcie, and R.M. Derosier, "Brillouin scattering in lightwave AM-VSB CATV transmission systems," *IEEE Photonics Technol. Lett.,* vol. 4, no. 3, pp. 287–289, 1991.

A.N. Netravali and B.G. Haskel, *Digital Pictures,* New York: Plenum Press, 1988.

R. Olshansky, V. Lanzisera, and P. Hill, "Design and performance of wideband subcarrier multiplexed lightwave systems," in *Proc. ECOC '88,* Brighton, U.K., Sept. 1988, pp. 143–146.

M.R. Phillips, T.E. Darcie, D. Marcuse, G.E. Bodeep, and N.J. Frigo, "Nonlinear distortion generated by dispersive transmission of chirped intensity-modulated signals," *IEEE Photonics Technol. Lett.,* vol. 3, no. 5, pp. 481–483, 1991.

T. Pratt and C.W. Bostian, *Satellite Communications,* New York: Wiley, 1986.

A.A.M. Saleh, "Fundamental limit on number of channels in subcarrier mulitplexed lightwave CATV systems," *Electron. Lett.,* vol. 25, no. 12, pp. 776–777, 1989.

A. Takemoto, H. Watanabe, Y. Nakajima, Y. Sakakibara, S. Kakimoto, U. Yamashita, T. Hatta, and Y. Miyake, "Distributed feedback laser diode and module for CATV systems," *IEEE J. Selected Areas in Communications,* vol. 8, 1359, 1990.

W. Way, C. Zah. C. Caneau, S. Menmocal, F. Favire, F. Shokoochi, N. Cheung, and T.P. Lee, "Multichannel FM video transmission using traveling wave amplifiers for subscriber distribution," *Electron. Lett.,* vol. 24, p. 1370, 1988.

Further Information

National Cable Television Association (NCTA), Proceedings from Technical Sessions, annual meetings, 1724 Massachusetts Ave. NW, Washington D.C., 20036, 1969.

Society of Cable Television Engineers (SCTE), Proceeding from Technical Sessions, biennial meetings, Exton Commons, Exton, Penn.

T.E. Darcie, "Subcarrier multiplexing for lightwave multiple-access lightwave networks," *J. Lightwave Technol.,* vol. LT-5, pp. 1103–1110, Aug. 1987.

T.E. Darcie and G.E. Bodeep, "Lightwave subcarrier CATV transmission systems," *IEEE Trans. Microwave Theory and Technol.,* vol. 38, no. 5, pp. 534–533, May 1990.

IEEE J. Lightwave Technol., Special Issue on "Broadband Analog Video Transmission Over Fibers," to be published Jan./Feb. 1993.

3.2   Long Distance Fiber Optic Communications

Joseph C. Palais

When the first laser was demonstrated in 1960, numerous applications were anticipated. Some predicted laser beams would transmit messages through the air at high data rates between distant stations. Although laser beams can travel through the atmosphere, too many problems prevent this scheme from becoming practical. Included in the objections are the need for line-of-sight paths and the unpredictability of transmission
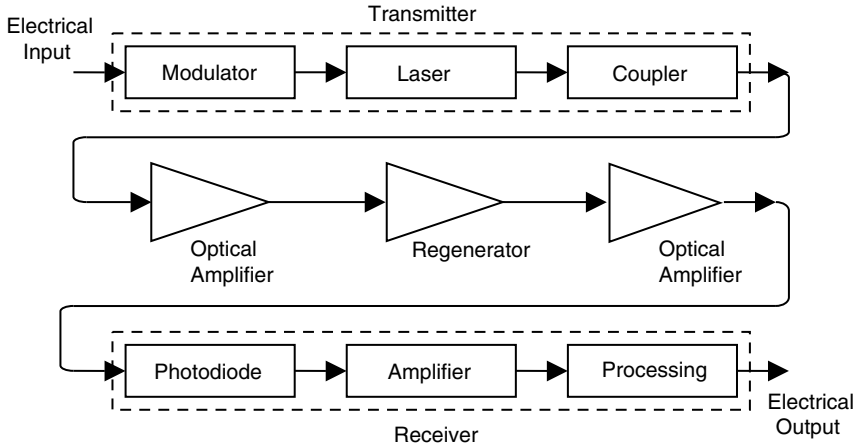
FIGURE 3.6 Long-distance fiber communications system.

through an atmosphere, where weather variations randomly change path losses. Guided paths using optical fibers offer the only practical means of optical transmission over long distances.

Long-distance fiber systems typically have the following operational characteristics: they are more than 10 km long, transmit digital signals (rather than analog), and operate at data rates above a few tens of megabits per second. This section describes systems in this category.

Figure 3.6 illustrates the basic structure of a generalized long-distance fiber optic link. Each of the components are described in the following paragraphs.

A useful figure of merit for these systems is the product of the system data rate and its length. This figure of merit is the well-known rate-length product. The bandwidth of the transmitting and receiving circuits, including the light source and photodetector, limits the achievable system data rate. The bandwidth of the fiber decreases with its length, so that the fiber itself limits the rate-length product. The losses in the system, including those in the fiber, also limit the path length. Systems are bandwidth limited if the rate-length figure is determined by bandwidth restraints and loss limited if determined by attenuation.

The first efficient fiber appeared in 1970, having a loss of 20 dB/km. Seven years later, the first large-scale application was constructed between two telephone exchanges in Chicago. By this time, the loss had been reduced to around 3 dB/km. The digital technology used could accommodate a rate of 45 Mbps over an unrepeatered length of 10 km and a total length of over 60 km with repeaters. The unrepeatered rate-length product for this initial system was a modest 450 Mb/s × km. As fiber technology advanced, this figure steadily increased. Unrepeatered rate-length products have improved to beyond 1000 Gb/s × km (e.g., 10 Gb/s over a path of 100 km). Including regenerators and optical amplifiers in the links increase the net rate-length product considerably. Values beyond 70 Tb/s × km (70,000 Gb/s × km) are achievable with optical amplifiers. This latter figure allows construction of a transmission system operating at 5 Gb/s over a 14,000-km path. The longest terrestrial paths are across the Atlantic and Pacific oceans, distances of about 6,000 and 9,000 km, respectively. Fibers are capable of spanning these distances with high-capacity links.

Fiber

All fibers used for long-distance communications are made of silica glass and only allow a single mode of propagation. The silica is doped with other materials to produce the required refractive index variations for the fiber core and cladding. The important fiber characteristics that limit system performance are its loss and bandwidth. The loss limits the length of the link, and the bandwidth limits the data rate.

Figure 3.7 shows the loss characteristics of single-mode silica fibers. The loss in the 800–900 nm region is dominated by Rayleigh scattering. Total attenuations of approximately 2.5 dB/km make this wavelength region
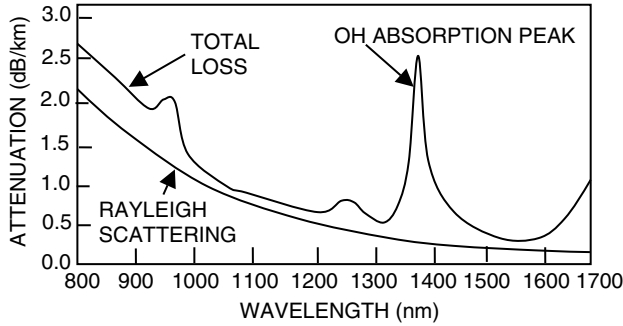
FIGURE 3.7   Spectral attenuation of a silica glass fiber showing contributions from Rayleigh scattering and water vapor absorption.

TABLE 3.1   Spectral Band Classification Scheme

Band	Descriptor	Range (nm)
O-band	Original	1260–1360
E-band	Extended	1360–1460
S-band	Short wavelength	1460–1530
C-band	Conventional	1530–1565
L-band	Long wavelength	1565–1625
U-band	Ultra-long wavelength	1625–1675

unsuitable for long distance links. Losses are approximately 0.5 dB/km around 1300 nm and 0.25 dB/km around 1550 nm. Manufacturing fibers with low OH ion content can reduce the water vapor absorption peak at 1390 nm. As a result, the entire wavelength range from 1260 to 1675 nm can be utilized for fiber systems. The appropriate wavelength designations appear in Table 3.1. The minimum loss near 1550 nm makes this the optimum choice for the very longest links.

Dispersion refers to the spreading of a pulse as it travels along a single-mode fiber as a result of material and waveguide effects. This spreading creates intersymbol interference if allowed to exceed roughly 70% of the original pulse width, causing receiver errors. The dispersion factor M is usually given in units of picoseconds of pulse spread per nanometer of spectral width of the light source and per kilometer of length of fiber.

In the range from 1200 to 1600 nm, the an approximation of the dispersion curve for silica is

$$M = \frac{M_0}{4}\left(\lambda - \frac{\lambda_0^4}{\lambda^3}\right) \tag{3.9}$$

where λ is the operating wavelength, λ_0 is the zero dispersion wavelength, and M_0 is the slope at the zero dispersion wavelength. M_0 is approximately 0.095 ps/(nm$^2 \times$ km). The pulse spread for a path length L, using a light source whose spectral width is $\Delta\lambda$, is then

$$\Delta\tau = ML\Delta\lambda \tag{3.10}$$

The zero dispersion wavelength, which is close to 1300 nm for silica fibers, makes this wavelength attractive for high-capacity links. The dispersion at 1550 nm is typically close to 20 ps/(nm $\times$ km). This is a moderate amount of dispersion. If a proposed 1550-nm system is bandwidth limited because of this spread, several alternatives are available. One solution is to use dispersion-shifted fiber, which is a special fiber with a

refractive index profile designed to shift the zero dispersion wavelength from 1300 nm to 1550 nm. Another solution is to transmit soliton pulses, which use the nonlinearity of the fiber to maintain pulse shape during transmission.

Because of high loss, the 800–900 nm part of the spectrum can only be used for moderate lengths (around 10 km). Because of high dispersion, the data rates are also limited in this region. In the 1300 nm region, the nearly zero dispersion allows high-rate transmission, but the losses limit the distance that can be covered (typically around 50 km). Around 1550 nm, the loss is roughly half the 1300-nm attenuation and twice as much distance can be covered. Dispersion-shifted fiber allows the same high rates as the 1300-nm operation. Regenerators and amplifiers extend the useful distance of fiber links well beyond the distances listed here.

Modulator

A digital electrical signal current modulates the light source. The driver circuit must be fast enough to operate at the system bit rate. As bit rates increase into the multigigabit per second range, this becomes increasingly difficult. Modulation can be completed in the optical domain at very high speeds. In this case, the modulator follows the laser diode rather than preceding it. External modulation is usually accomplished using integrated-optic structures.

Light Source

Laser diodes or light-emitting diodes (LEDs) supply the optical carrier waves for most fiber links. LEDs cannot operate at speeds in the gigabit range, but laser diodes can. For this reason, laser diodes are normally required for high-rate, long-distance links. Laser diodes can be modulated at frequencies beyond 40 GHz.

Laser diodes emitting in the 1200–1700 nm region are semiconductor heterojunctions made of InGaAsP. The exact emission wavelength is primarily determined by the proportions of the constituent atoms. Output powers are commonly in the order of a few milliwatts.

Typical laser diode **spectral widths** are between 1 and 5 nm when operating in more than one longitudinal mode. Single-mode laser diodes can have spectral widths of just a few tenths of a nanometer. As predicted by Equation (3.10), narrow-spectral-width emitters minimize pulse spreading. Minimizing pulse spreading increases the fiber bandwidth and its data capacity.

Solid-state lasers, rather than semiconductor laser diodes, may be useful in specific applications. Examples are the Nd:YAG laser and the erbium-doped fiber laser.

Source Coupler

The light emitted from the diode must be coupled as efficiently as possible to the fiber. Because the beam pattern emitted by a laser diode does not perfectly match the pattern of light propagating in the fiber, there is an inevitable mismatch loss. Good coupler designs, sometimes using miniature lenses, reduce this loss to about 3 dB when feeding a single-mode fiber.

Isolator

An optical isolator is a one-way transmission path. It allows power to flow from the transmitter toward the receiver, but blocks the power flow in the opposite direction. The optical isolator protects the laser diode from back reflections, which tend to increase the laser noise.

Connectors and Splices

Connections between fibers and between the fiber and other components occur at numerous points in a long-distance link. Because there may be many splices in a long system, the loss in each splice must be small. Fusion splices with an average loss of no more than 0.05 dB are often specified. Mechanical splices are also suitable.

They often involve epoxy for fixing the connection. Connectors are used where remateable connections are required. Good fiber connectors introduce losses of only a few tenths of a decibel.

In addition to having low loss, good connectors and splices also minimize back reflections. This is especially important to reduce laser noise in connections near the transmitter. Fusion splices produce little reflection, but mechanical splices and all connectors must be carefully designed to keep reflected power levels low. Reflections occur because of small gaps at the interface between the mated fibers. Successful techniques for reducing reflections include a physical contact connection, where the fiber end faces are polished into hemispheres (rather than flat surfaces), so that the cores of the two mated fibers are in contact with each other. Performance is increased by angling the end faces a few degrees, causing reflected light to be filtered out of the single propagating mode.

Optical Amplifier

Many fiber links are loss limited. One cause is the limited power available from the typical laser diode, which (with the losses in the fiber and the other system components) restricts the length of fiber that can be used. The fiber optic amplifier increases the power level of the signal beam without conversion to the electrical domain. For example, gains of 30 dB are attainable at 1550 nm using the erbium-doped fiber amplifier (EDFA). Note, the EDFA has a bandwidth of over 20 nm, allowing several WDM or numerous OFDM channels (both described later in this section) to be amplified simultaneously.

As indicated in Figure 3.6, there are a number of possible locations for optical amplifiers in a system. An optical amplifier that is only following the transmitter increases the optical power traveling down the fiber. Amplifiers along the fiber path continually keep the power levels above the system noise. An amplifier located at the fiber end acts as a receiver preamplifier, enhancing its sensitivity. Many amplifiers can be placed in a fiber network, extending the total path length to thousands of kilometers.

Regenerator

The regenerator detects the optical signal by converting it into electrical form. It then determines the content of the pulse stream, uses this information to generate a new optical signal, and launches this improved pulse train into the fiber. The new optical pulse stream is identical to the one originally transmitted. The regenerated pulses are restored to their original shape and power level by the repeater.

Many regenerators may be placed in a fiber network, extending the total path length to thousands of kilometers. The advantage of the optical amplifier over the regenerator is its lower cost and improved efficiency. The greater cost of the regenerator arises from the complexity of the conversion between the optical and electrical domains. The regenerator does have the advantage of restoring the signal pulse shape, which increases the system bandwidth. This advantage is negated by a system propagating soliton pulses, which do not degrade with propagation.

Photodetector

The photodetector converts an incoming optical beam into an electrical current. In fiber receivers, the most commonly used photodetectors are semiconductor pin photodiodes and avalanche photodiodes (APD). Important detector characteristics are speed of response, spectral response, internal gain, and noise. Because avalanche photodiodes have internal gain, they are preferred for highly sensitive receivers. Both germanium (Ge) and InGaAs photodiodes respond in the low-loss 1200- to 1700-nm wavelength regions. InGaAs performs better at low signal levels because it has smaller values of dark current (less noisy).

The current produced by a photodetector in response to incident optical power P is

$$i = G\eta eP/hf \tag{3.11}$$

where G is the detector's gain, η is its **quantum efficiency** (close to 0.9 for good photodiodes), h is Planck's constant (6.63×10^{-34} J s), e is the magnitude of the charge on an electron (1.6×10^{-19} coulomb),

and f is the optical frequency. For pin photodiodes (where $G = 1$), typical responsivities are in the order of 0.5 μA/μW.

Receiver

Because of the low power levels expected at the input to the receiver, an electronic amplifier is normally required following the photodetector. The remainder of the receiver includes such electronic elements as band-limiting filters, equalizers, decision-making circuitry, other amplification stages, switching networks, digital-to-analog converters, and output devices (e.g., telephones, video monitors, and computers).

Other Components

There are a number of fiber components, not shown in Figure 3.6, that can be found in some systems. These include passive couplers for tapping off a portion of the beam from the single fiber and wavelength-division multiplexers for coupling different optical carriers onto the transmission fiber.

System Considerations

Long-distance fiber links carry voice, video, and data information. Messages that are not already in a digital format are converted. A single voice channel is normally transmitted at a rate of 64,000 bits per second. Video requires a much higher rate. The rate could be as high as 90 Mbps, but video compression techniques can lower this rate significantly. Fiber systems for the telephone network operate at such high rates that many voice channels can be time-division multiplexed (TDM) onto the same fiber for simultaneous transmission. For example, a fiber operating at a rate of 2.5 Gb/s can carry more than 30,000 digitized voice channels, while one operating at 10 Gb/s can accommodate almost 130,000 voice channels.

Several optical carriers can simultaneously propagate along the same fiber. Such wavelength-division multiplexed (WDM) links further increase the capacity of the system. Systems using up to 32 optical carriers are common. Adding 16 or more channels puts constraints on the multiplexers and light sources. Nonetheless, systems with over 100 channels are feasible. The carrier wavelengths are spaced by a few tenths of a nm or less. In long systems, wideband optical amplifiers are preferred over regenerators for WDM systems because a single amplifier can boost all the individual carriers simultaneously; however, separate regenerators are needed for each carrier wavelength.

Placing numerous fibers inside a single cable increases total cable capacity. This is a cost-effective strategy when installing long fiber cables. The added cost of the extra fibers is small compared to the costs of actually deploying the cable itself. Fiber counts of several hundred are practical. Multifiber cables can have enormous total data capacities.

Still further capacity is possible using optical frequency-division multiplexing (OFDM). In this scheme, many optical carriers very closely spaced in wavelength (maybe a few hundredths of a nanometer) operate as independent channels. Many hundreds of channels can be visualized in 1200–1700 nm wavelength regions. Systems of this type require coherent detection receivers to separate the closely spaced carriers.

Error Rates and Signal-to-Noise Ratio

The signal-to-noise ratio is a measure of signal quality. It determines the error rate in a digital network. At the receiver, it is given by

$$\frac{S}{N} = \frac{(G\rho P)^2 R_{\mathrm{L}}}{G^n 2eR_{\mathrm{L}}B(I_{\mathrm{D}} + \rho P) + 4kTB} \tag{3.12}$$

where P is the received optical power, ρ is the detector's unamplified responsivity, G is the detector gain if an APD is used, n accounts for the excess noise of the APD (usually between 2 and 3), B is the receiver's bandwidth, k is Boltzmann's constant ($k = 1.38 \times 10^{-23}$ J/K), e is the magnitude of the charge on an electron

$(1.6 \times 10^{-19}$ coulomb), T is the receiver's temperature in degrees Kelvin, I_D is the detector's dark current, and R_L is the resistance of the load resistor that follows the photodetector.

The first term in the denominator of Equation (3.12) is caused by shot noise and the second term is attributed to thermal noise in the receiver. If the shot-noise term dominates (and the APD excess loss and dark current are negligible), the system is shot-noise limited. Therefore, the probability of error has an upper bound given by

$$P_e = e^{-n_s} \tag{3.13}$$

where n_s is the average number of photoelectrons generated by the signal during a single bit interval when a binary 1 is received. An error rate of 10^{-9} or better requires about 21 photoelectrons per bit. Shot noise depends on the optical signal level. Because the power level is normally low at the end of a long-distance system, the shot noise is small compared to the thermal noise. Avalanche photodiodes increase the signal level so that shot noise dominates over thermal noise. With APD receivers, ideal shot-noise limited operation can be approached but, not reached because of the APD excess noise and limited gain.

If the thermal noise dominates, the error probability is given by

$$P_e = 0.5 - 0.5 \, \text{erf}(0.354) \tag{3.14}$$

where erf is the error function. An error rate of 10^{-9} requires a signal-to-noise ratio of nearly 22 dB.

System Design

A major part of fiber system design involves the power budget and the bandwidth budget. The next few paragraphs describe these calculations.

In a fiber system, component losses (or gains) are normally given in decibels. The decibel is defined by

$$dB = 10 \log P_2/P_1 \tag{3.15}$$

where P_2 and P_1 are the output and input powers of the component. The decibel describes relative power levels. Similarly, *dBm* and *dBμ* describe absolute power levels. They are given by

$$dBm = 10 \log P \tag{3.16}$$

where P is in milliwatts and

$$dBm = 10 \log P \tag{3.17}$$

where P is in microwatts.

Power budget calculations are illustrated in Table 3.2 for a system that includes an amplifier. A specific numerical example is found in the last two columns. The receiver sensitivity in *dBm* is subtracted from the power available from the light source in *dBm*. This difference is the loss budget (in decibels) for the system. All the system losses and gains are added together (keeping in mind that the losses are negative and the amplifier gains are positive). If the losses are more than the gains (as is usual), the system loss dB_{SL} will be a negative number. The loss margin is the sum of the loss budget and the system loss. It must be positive for the system to meet the receiver sensitivity requirements. The system loss margin must be specified to account for component aging and other possible system degradations. Table 3.2 illustrates a 6-dB margin for the system discussed above. The fiber in the table has a total loss of 24 dB. If its attenuation is 0.25 dB/km, the total length of fiber allowed would be $24/0.25 = 96$ km.

In addition to providing sufficient power to the receiver, the system must also satisfy the bandwidth requirements imposed by the rate at which data are transmitted. A convenient method of accounting for the bandwidth is to combine the rise times of the various system components and compare the result with the rise time needed for the given data rate and pulse coding scheme.

The system rise time is given in terms of the data rate by the expression

$$t = 0.7/R_{\mathrm{NRZ}} \tag{3.18}$$

for non-return-to-zero (NRZ) pulse codes and

$$t = 0.35/R_{\mathrm{RZ}} \tag{3.19}$$

for return-to-zero (RZ) codes.

An example of bandwidth budget calculations appears in Table 3.3. The calculations are based on the accumulated rise times of the various system components.

The system in Table 3.3 runs at 500 Mb/s with NRZ coding for a 100-km length of fiber. Equation (3.18) yields a required system rise time no more than 1.4 ns. The transmitter is assumed to have a rise time of 0.8 ns. The receiver rise time, given as 1 ns in the table, is a combination of the photodetector's rise time and that of the receiver's electronics.

The fiber's rise time was calculated for a single-mode fiber operating at a wavelength of 1550 nm. Equation (3.9) shows that $M = 18$ ps/(nm × km) at 1550 nm. The light source was assumed to have a spectral width of 0.2 nm. Then, the pulse dispersion calculated from Equation (3.10) yields a pulse spread of 0.36 ns. Because the fiber's rise time is close to its pulse spread, this value is placed in the table.

The total system rise time is the square root of the sum of the squares of the transmitter, fiber, and receiver rise times. That is,

$$t_{\mathrm{s}} = \sqrt{t_{\mathrm{t}}^2 + t_{\mathrm{f}}^2 + t_{\mathrm{r}}^2} \tag{3.20}$$

In this example, the system meets the bandwidth requirements by providing a rise time of only 1.33 ns, where as much as 1.4 ns would have been sufficient.

TABLE 3.2 Power Budget Calculations

Source power:	dBm$_{\mathrm{s}}$		3
Receiver sensitivity:	dBm$_{\mathrm{r}}$		−30
Loss budget: dBms-dBmr		dBLB	33
Component efficiencies:			
Connectors	dB$_{\mathrm{c}}$		−5
Splices	dB$_{\mathrm{s}}$		−2
Source coupling loss	dB$_{\mathrm{cl}}$		−5
Fiber loss	dB$_{\mathrm{f}}$		−24
Isolator insertion loss	dB$_{\mathrm{i}}$		−1
Amplifier gain	dB$_{\mathrm{a}}$		10
Total system loss:			
dB$_{\mathrm{c}}$ + dB$_{\mathrm{s}}$ + dB$_{\mathrm{cl}}$ + dB$_{\mathrm{f}}$ + dB$_{\mathrm{i}}$ + dB$_{\mathrm{a}}$		dB$_{\mathrm{SL}}$	−27
Loss Margin: dB$_{\mathrm{LB}}$ + dB$_{\mathrm{SL}}$		dB$_{\mathrm{LM}}$	6

TABLE 3.3 Bandwidth Budget Calculations*

Transmitter:	t_{t}	0.8	
Fiber:	t_{f}	0.36	
Receiver:	t_{r}	1	
System total: $\sqrt{t_{\mathrm{t}}^2 + t_{\mathrm{f}}^2 + t_{\mathrm{r}}^2}$	t_{s}		1.33
System required:	t		1.4

*All quantities in the table are rise time values in nanoseconds.

Defining Terms

Coherent detection: The signal beam is mixed with a locally generated laser beam at the receiver. This results in improved receiver sensitivity and receiver discrimination between closely spaced carriers.

Material dispersion: Wavelength dependence of the pulse velocity. It is caused by the refractive index variation with wavelength of glass.

Quantum efficiency: A photodiode's conversion efficiency from incident photons to generated free charges.

Single-mode fiber (SMF):   A fiber that can support only a single mode of propagation.

Spectral width: The range of wavelengths emitted by a light source.

References

G.P. Agrawal, *Fiber-Optic Communication Systems,* 3rd ed., New York: John Wiley and Sons, 2002.

E.E. Basch, Ed., *Optical-Fiber Transmission*, Indianapolis: Howard W. Sams & Co., 1987.

M. Bass, ed., *Fiber Optics Handbook*, New York: McGraw-Hill, 2002.

C.C. Chaffee, *The Rewiring of America*, San Diego: Academic Press, 1988.

E. Desurvire, *Erbium-Doped Fiber Amplifiers*, New York: John Wiley & Sons, 2002.

M.J.F. Digonnet, *Rare Earth Doped Fiber Lasers and Amplifiers*, 2nd ed., New York: Marcel Dekker, 2001.

R.J. Hoss, *Fiber Optic Communications Design Handbook*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

L.B. Jeunhomme, *Single-Mode Fiber Optics*, 2nd ed., New York: Marcel Dekker, 1990.

N. Kashima, *Passive Optical Components for Optical Fiber Transmission*, Norwood, Mass.: Artech House, 1995.

S.V. Kartalopoulos, *DWDM: Networks, Devices and Technology*, New York: John Wiley & Sons, 2002.

G. Keiser, *Optical Fiber Communications*, 3rd ed., New York: McGraw-Hill, 2000.

R.H. Kingston, *Optical Sources, Detectors and Systems*, New York: Academic Press, 1995.

J.C. Palais, *Fiber Optic Communications*, 5th ed., Upper Saddle River, N.J.: Prentice-Hall, 2005.

S. Shimada, *Coherent Lightwave Communications Technology*, New York: Chapman and Hall, 1994.

A. Yariv, *Optical Electronics in Modern Communications*, 5th ed., New York: Oxford University Press, 1997.

Further Information

Continuing information on the latest advances in long-distance fiber communications is available in several professional society journals and trade magazines including: *IEEE Journal of Lightwave Technology*, *IEEE Photonics Technology Letters, Lightwave, and Laser Focus World*.

3.3   Photonic Networks

Alan E. Willner and Reza Khosravani

Introduction

Due to the high signal attenuation in copper transmission lines for data rates exceeding hundreds of Mbit/s, it has long been recognized that lightwave technology has significant advantages over copper for long-distance data transmission and for cable TV (CATV) distribution networks. Over the past 20 years, the capacity that can be transmitted over an optical fiber has increased and the cost-per-transmitted-bit decreased by several orders of magnitude. Optical transmission is the only viable solution for the ever-increasing bandwidth requirements, even within many local-area networks (LANs). Recently, a single fiber was used for the transmission of 3.7 Tbit/s (3700 Gbit/s) over transoceanic distances [1]. Optical fibers possess an enormous bandwidth that is the base technology for high-capacity networks of present and future.

However, over the past decade, the need for high-capacity multi-user networks has intensified because of significant Internet growth. The World Wide Web, which is founded on wide-area networks (WAN), has become one of the main avenues of demand for information transfer. As demand grows, the copper-based

routing and switching nodes might fail to provide the required throughput. A laudable goal has been to make use of photonic power to bring greater throughput and cost-effectiveness to multi-user networking.

A key technology, which allows straightforward routing and optical data path switching in a photonic network, is wavelength-division-multiplexing (WDM), in which many independent data channels are transmitted in parallel down an optical fiber. By utilizing wavelength-selective component technologies, each data channel's wavelength can be used to determine the routing through the network. Therefore, data is actually transmitted through wavelength-specific "light-paths" on a route arranged by a network controller to optimize throughput.

Significant changes in photonic networks have occurred in the last few years. The widespread deployment of optical fibers and technological breakthroughs in photonic devices (i.e., optical amplifiers, wavelength multiplexers, and optical switches) has significantly changed the achievable architecture of photonic networks. Furthermore, the gradual shift in the nature of the network traffic from voice to data has called for new thinking in protocols, control, and management. Note, the circuit-switch-like SONET ring photonic networks have become a reality, where the wavelength lightpath determines the routing, and the networks are deployed throughout the world in metro-networks and beyond. We hope that some of the more advanced technologies discussed in this chapter will develop during the next decade.

Here, we first review lightwave transmission links that connect network nodes. Then, we discuss architectures and protocols of photonic networks as well as switching techniques. Finally, we discuss some of the enabling technologies and challenges in realizing high throughput photonic networks.

Background: Lightwave Transmission Links Interconnecting Nodes

This section provides the basic background for a simple lightwave transmission link that forms the basis for interconnecting nodes in a photonic network [2].

A lightwave transmission link consists of: a transmitter (Tx) that converts an electrical signal to an optical signal (E/O), an optical fiber as the connecting medium, and a receiver (Rx) that converts the optical signal back to the electrical domain (O/E). The transmitter typically uses a laser diode (LD) as the optical source. The receiver may employ a PIN (positive-intrinsic-negative) or APD (avalanche photodiode) photodetector.

The standard silica fiber has a minimum attenuation of ~0.2 dB/km around 1550 nm. Depending on the core diameter and the optical wavelength, a fiber may be multimode or single mode. Multimode fibers (with a typical core diameter of 62.5 μm) allow several modes to propagate through the fiber, and may be used for short distance communications (e.g., LANs inside a building). The pulse energy is divided among the different modes, and each mode arrives at the receiver with a different delay. This intermodal dispersion effect causes pulse spreading. A single mode fiber has a much smaller core diameter (~10 μm) and allows only one mode to propagate. Although the intermodal dispersion is eliminated in a single mode fiber, the much lower intramodal chromatic dispersion is generated because of the index of refraction's wavelength dependency. This dispersion is significant for long-distance and high-data-rate systems. The chromatic dispersion value for a conventional single mode fiber is approximately 17 ps/nm/km.

As an optical signal propagates through the fiber, the fiber loss causes signal power attenuation. At some point, the optical signal power is too weak for the receiver to detect. Prior to the invention of optical amplifiers, an electronic regenerator was required every several tens of kilometers to detect and re-transmit each data channel. The invention of Erbium-doped fiber amplifiers (EDFAs) [3,4] improved optical communications in two major ways. First, the maximum reach of optical links extended significantly from tens to thousands of kilometers. Second, EDFAs are broadband and can amplify all the multi-wavelength optical signals simultaneously and cost-effectively. For example, a group of four (or 40) optical channels co-propagating on different wavelengths inside the same fiber may be amplified using a single EDFA (see Figure 3.8). As a result, EDFAs enable the practical deployment of wavelength-division-multiplexing that allows the parallel transmission of multiple optical signals on the same fiber. WDM dramatically enhances fiber capacity since "N" wavelength channels will increase the capacity by a factor of N. EDFA amplification is independent of the signal wavelength, bit rate, and modulation format.
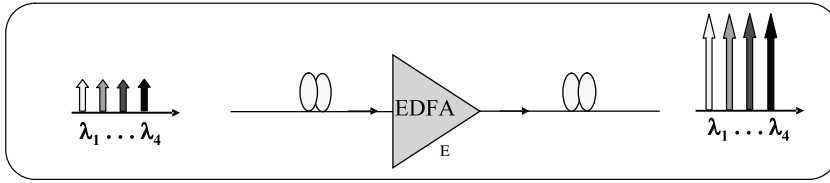
FIGURE 3.8   EDFA amplifies several optical channels at the same time.

Another optical amplification technique, called Raman amplification, has recently gained popularity. Compared to EDFAs, Raman amplifiers [5] can reduce the system degrading effects of fiber nonlinearities. Additionally, the pump wavelength determines the gain wavelength range of Raman amplifiers, allowing other signal wavelengths bands to be amplified. However, the pump efficiency of a Raman amplifier is often significantly lower than an EDFA.

In addition to transmitting on multiple parallel wavelengths, there is also a trend towards higher-bit-rates per channel. Today, commercial optical data channels can operate at 10 Gbit/s, with 40-Gbit/s links about to be deployed. A standard technique to generate a high bit-rate signal is to combine many low speed signals in the time domain. This technique is called time-division multiplexing (TDM). Since chromatic dispersion is a key limitation, high-data-rate long-distance transmission should employ dispersion compensating elements, for which the positive dispersion is compensated periodically along the fiber link with a negative dispersion value. These developments made it possible to build deployed optical networks that operate at 10 or even 40 Gbit/s. For example, single-mode fiber transmission links can support bit-rate-distance product of 500 (Gb/s) $\times$ km and beyond.

Multiple optical signals are transmitted on the same fiber simultaneously to utilize the wide bandwidth of an optical fiber more efficiently. This is accomplished by assigning different wavelengths to the optical channels based on the International Communications Union (ITU) recommendation. In this standard, the available wavelengths are equally spaced (i.e., 50 or 100 GHz frequency spacing) on a grid stabilized to 193.1 THz ($\sim$1552.52 nm) [6]. Each digital data stream is modulated on a different optical carrier wavelength. These optical signals are combined using a wavelength multiplexer and transmitted through the fiber. At the receiving end, the WDM signal is demultiplexed, and each wavelength is detected using a separate receiver. Figure 3.9 shows a typical optical link comprised of multiple transmitters, a wavelength multiplexer, optical fibers and amplifiers, a wavelength demultiplexer, and multiple receivers.

Architecture of Photonic Networks

In the early stages of photonic systems, optics was used only for point-to-point connections. The first generation of photonic networks used the traditional communication networks approach in which data was detected at each node and electronics provided the processing, routing, and switching tasks. With the introduction of WDM and the development of advanced optical components, many functionalities that were not possible in traditional networks are now available. Consequently, the new architectures composed of these optical network elements were introduced [7].
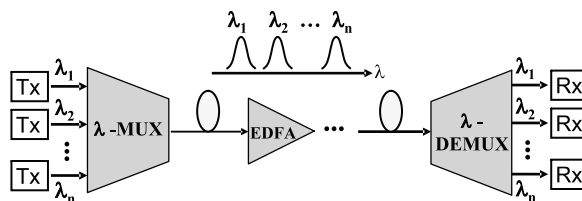
FIGURE 3.9   A typical block diagram of a lightwave transmission link.
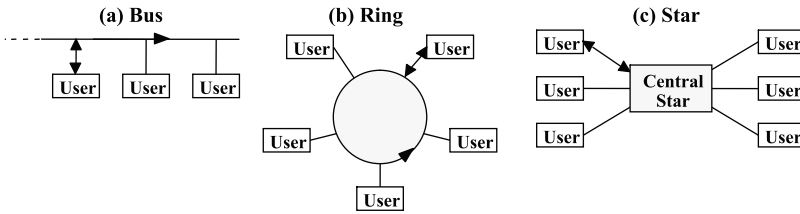
FIGURE 3.10 Different topologies of the networks.

The general topology of a network (the way that end terminals are interconnected) falls into one of the following categories: bus, ring, star, or mesh. In a bus topology, all terminals are connected to the transmission medium though taps (see Figure 3.10a). Data transmitted by each terminal propagates in both directions until it is received by the destination terminal [8]. A ring network is made by connecting several nodes in a closed loop fashion (Figure 3.10b). The real topology may be a good deal more irregular than a circle, depending on the accessibility of stations. Commercial token rings use wire interconnections or optical links to join stations. The transit-time delay increases linearly with the number of stations. In a star topology (Figure 3.10c), all terminals are connected to a central node [8]. Data that is transmitted from a terminal to the central node, will then be retransmitted to all other terminals, including the destination terminal.

Reliability is a problem in both fiber and wire rings. If a station is disabled, or if a fiber breaks, the whole network goes down. To address this problem, a double-ring optical network, also called a "self-healing" ring, is used to bypass the defective stations and loop back around a fiber break (Figure 3.11). Each station has two inputs and two outputs connected to two rings that operate in opposite directions; therefore, cost increases.

As the size of a network increases, a more irregular topology may be needed. Large metropolitan area networks (MANs) and WANs may consist of several ring networks that are interconnected through one or more nodes. In a mesh topology, each node is principally interconnected to its neighboring nodes. Data may be transmitted between any two terminals through intermediate nodes. Since several independent routes may connect the terminals, the mesh networks are less vulnerable to the network failures.

WDM technology has provided a new domain for routing in photonic networks by allowing multiple signals to be transmitted through the same fiber. The new generation of the photonic networks utilizes this extra functionality and makes them superior to the traditional networks. Add/Drop multiplexers and optical crossconnects are the two key network elements for the new generation of the photonic networks. These elements are discussed below.

Add/Drop Multiplexer

In WDM photonic networks, each fiber carries multiple optical signals on different wavelengths. The fiber that connects two nodes carries the data that might be intended for the receiving node as well as the data for subsequent nodes. If special provisions are not made, each node must detect and retransmit all of the wavelengths to receive the data that was intended for the local terminals. Clearly, this scenario decreases
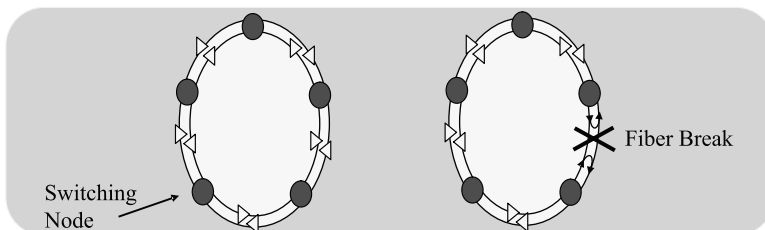
FIGURE 3.11 A self-healing ring network.

network efficiency and increases cost. A more intelligent design would be to demultiplex all the wavelengths, but only detect/transmit the wavelength intended for that node. An optical add/drop multiplexer (OADM) is a network element that provides this functionality [9]. Figure 3.12 shows a simple architecture of an OADM. The wavelengths that are not dropped experience a transparent node and are directed to the output port.

The data on λ_i that is intended for the local terminals is dropped. New data is transmitted on λ_i and is added to the fiber.

The problem with a fixed OADM is that it does not allow reconfiguration after deployment. This imposes a limitation on the network design. It is possible to avoid this limitation by using 2×2 optical switches on each wavelength path between the MUX and DEMUX (reconfigurable OADM). This way, it is possible to choose the desired wavelengths to drop. Tunable transmitters and receivers can be used to reduce the number of transmitters and receivers at each node.

Optical Crossconnects

In large and complex mesh networks, more than one fiber may arrive at each node. Each fiber carries multiple wavelengths that might be intended for different destinations. Optical crossconnects (OXC) are network elements that allow any input fiber wavelength to be switched to any output fiber [10]. Figure 3.13 shows the architecture of an OXC.

The optical switch can connect any wavelength of the input fibers to the desired output.
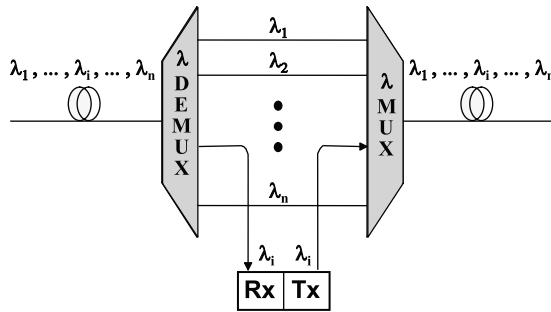
FIGURE 3.12   A fixed optical add/drop multiplexer.
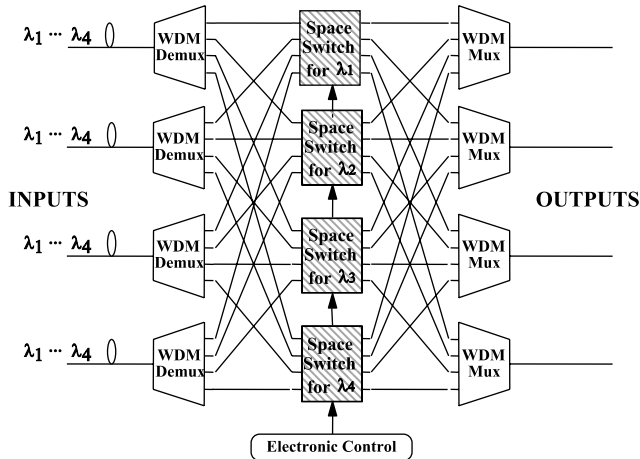
FIGURE 3.13   Optical crossconnect.

The issue of contention is a noticeable problem with the OXC. For example, if two similar wavelengths from two different input fibers contain two packets that are destined to the same output fiber, it is impossible to transmit two optical signals with the same wavelengths on the same fiber. Note, the WDM allows the transmission of multiple optical signals on the same fiber only if they have different wavelengths. This issue causes output-port contention in photonic networks, which will be discussed later.

Passive Optical Networks

The significantly large bandwidth of optical fibers makes them a good transmission media for high capacity MANs and WANs. As the demand for bandwidth increases, it may be necessary to bring the fiber closer to the end users. Fiber-To-The-Curb (FTTC) and Fiber-To-The-Home (FTTH) are network architectures using optical fibers in access networks. The idea is to transmit digital data from the Central Office (CO) to the Optical Network Units (ONUs) using optical fibers and passive optical components such as star couplers and WDM multiplexers. The passive $N \times N$ star coupler has N single-mode fiber inputs and N outputs [2]. In an ideal passive star, a signal incident on any input is divided equally among all the outputs, i.e., the star broadcasts every input to every output. Since only passive optical components are used in these network architectures, they are called Passive Optical Networks (PONs). PONs are generally reliable, simple, and cost-effective. The most common PON architecture broadcasts the data from CO to the ONUs using a passive star coupler. Since the network is transparent to the optical signals, data from all ONUs can be transmitted to the CO using the same network. There are provisions to use WDM technologies in PONs to increase the capacity of the network [7,11]. As an example, WDM multiplexers may be used to divide or combine the wavelengths of different ONUs.

Protocols of Photonic Networks

The synchronous optical network (SONET) and a compatible European version, SDH (Synchronous Digital Hierarchy), are standards that are designed to simplify time division multiplexing in high bit-rate networks and to take advantage of high capacity of optical links [8,12]. The lowest bit-rate for SONET is 51.84 Mbit/s and is called Synchronous Transport Signal level-1 (STS-1) or Optical Carrier level-1 (OC-1). Higher bit-rates at STS-n are generated by interleaving n synchronized STS-1 signals. For example, STS-192 or OC-192 corresponds to 9.95328 Gbit/s.

The gradual shift of the network traffic from voice to data heralded new standards to more efficiently control the "bursty" nature of the traffic. Asynchronous Transfer Mode (ATM) is an ITU-T standard that was developed for packet transfers [8]. The data is broken into fixed size packets (53 bytes) or cells. ATM was designed with the capability to transfer both voice and data traffic. Therefore, the size of the packet is relatively short to reduce the delay in voice networks. ATM also provides different quality-of-service (QoS) for different types of traffic. ATM cells may be transmitted asynchronously or synchronously on top of SONET frames (Figure 3.14).

Internet Protocol (IP) is the most widely used protocol for the Internet [8]. IP converts the data into variable-size packets and adds a header, which includes the source and destination addresses, and error control information. IP packets may be broken into several ATM cells that could be transferred on a SONET network. In general, IP offers no guaranteed QoS. Instead, it uses a best-effort approach to transmit the packets to their destinations. In 1998, Multiprotocol Label Switching (MPLS) technology was introduced to improve the switching efficiency in WDM networks and provide some QoS guarantees [13]. MPLS can provide a label-switched path (LSP) in the network that can be used to establish an end-to-end path between two stations. It also has some measures for protection and failure detection.
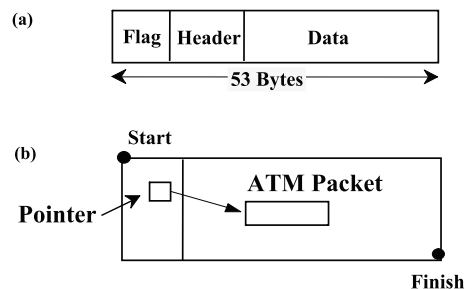
FIGURE 3.14 (a) An ATM cell comprised of flag, header, and data fields; (b) an ATM cell placed on a SONET frame.

Ethernet networks are based on carrier sense multiple access with collision detection (CSMA/CD) and operate at 10 Mbit/s to connect users on a copper bus. Fast Ethernets, operating at 100 Mbit/s and higher, use optical fibers. 100BASE-FX is a fiber-based Fast Ethernet standard that supports a network span of up to 400 meters. Several standards have also been introduced to implement Gigabit Ethernet (e.g., 1000BASE-SX and 1000BASE-LX) and 10-Gbit/s Ethernet (10GBASE-S, 10GBASE-L, 10GBASE-E, and 10GBASE-LX4), depending on the maximum reach and the operation wavelength (850, 1310, or 1550 nm) utilizing the large capacity of optical fiber [8].

Circuit, Packet, and Burst Switching

A communication network provides communication routes between many stations through intermediate nodes. These routes are made by accurately setting the switching devices at each node. There are two general approaches for setting the switches. In a circuit switching network, the status of the intermediate switches remains unchanged during the connection. As a result, a dedicated path with a specific bandwidth is established between the source and destination. Circuit switching networks set up the path before each data transmission and disconnect it afterwards. Once a connection is established, some part of the network resources is allocated to the connection irrespective of the actual data transmission [8]. Public telephone networks use circuit switching to establish a call between two clients.

In a packet switching network, the data transmitted by the source is broken into several packets. Each packet contains a portion of the original data as well as the destination address. A packet switching network transmits the packets through different nodes depending on the availability of resources (bandwidth) at each node. Packet switching networks do not require the establishment and disconnection of circuits, but each packet must carry additional header information that includes the destination address. The header information may be sent as additional bits at the beginning of the packet, on a subcarrier placed outside the spectrum of the transmitted data, or on a different wavelength reserved for this purpose. Figure 3.15 shows the schematic diagram of circuit and packet switching networks.

In the past several years, data traffic has been growing at a much faster rate than voice traffic. Data traffic is bursty in nature, and reserving bandwidth for bursty traffic could be very inefficient. Therefore, circuit switching networks are not optimized for this type of traffic. However, packet switching networks can handle bursty traffic more efficiently through statistical time multiplexing of the packets.

In any packet switching network, there is a possibility that two incoming packets, coming from different sources and intended for the same output port, will arrive at a node at the same time, causing contention. If contention is not resolved, some packets may be dropped and never arrive at their destinations. Packet-loss-probability is the figure of merit in evaluating contention resolution techniques. Conventional routers resolve contention by buffering the incoming packets until the associated output port is available.

To facilitate the increased performance of packet switching networks, O/E and E/O conversions must be avoided and the signals retained in the optical domain at all times. An *all-optical* photonic network avoids the bottleneck of electronic components and the cost of O/E and E/O conversions. In principle, all-optical channels would offer connectivity independent of data-rate and format. Currently, there is significant research
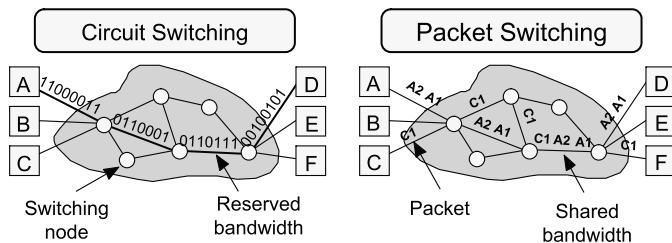
FIGURE 3.15 In a circuit switching network, a path is reserved between two end terminals during the data transmission. In a packet switching network, packets coming from different sources may share the same path.

and effort to design and implement all-optical packet switching (OPS) networks. OPS networks use optical switches that are transparent to the optical signals. One of the most challenging issues in building an OPS network is the issue of contention. Unfortunately, there is no existing optical equivalent to electronic memory that can buffer packets if necessary. In addition, OPS networks require a complex processing unit and fast optical switches [14].

Optical burst switching (OBS) was introduced with a vision to reduce the processing required for switching at each node and to avoid optical buffering [15]. Conceptually, OBS is different from OPS in the following ways:

- Burst size granularity, which is somewhere between packet switching and circuit switching
- Switching period, which is typically on the order of several packets in OBS
- Separation of header information and payload
- Reservation scheme to transmit the bursts
- Variable burst length

Similarly to OPS, OBS takes advantage of time-domain statistical multiplexing to use a single channel's bandwidth to transmit several lower-bandwidth bursty channels. At the network edge, packets are gathered to generate bursts that are sent over the network core. In almost all OBS schemes, the header is sent separately from the payload with an offset time that is dependent on the scheme [16,17]. The header is sent to the switches, and the path is then reserved for the payload that follows. The loose coupling between the control packet and the burst alleviates the need for optical buffering and relaxes the requirement for switching speed.

Enabling Switching Technologies

The conventional networks with optical links replacing copper will have higher throughputs because of the increased bandwidth of the transmission medium. However, an innovative improvement in throughput to terabit-per-second levels with gigabit-per-second access requires a new approach to the physical connectivity, architecture, and access protocols. Most of the photonic technologies can be used in lightwave transmission systems to provide physical connectivity, but devices with new functionalities are needed to implement the proposed architectures. At the same time, with new components functionalities, we can create new architectures. We have already discussed OADM and OXC as network elements that provide new functionalities. In this section, we will discuss other important technologies that enable higher and faster photonic networks.

Optical Switches

One of the most important elements of a network is the switch that controls routing. In the first generation of photonic networks, switching is performed by electronics. The optical signals that are approaching a node first go through O/E conversion. The detected signal is buffered and switched in the electrical domain. After switching, the digital data goes through E/O conversion. In these architectures, high-speed electronics is required at each node, resulting in a bottleneck.

Optical switches would circumvent the electronics bottleneck and allow much higher transmission bit-rates. In addition, optical switches are transparent to the wavelength, modulation format, and bit-rate. Unfortunately, optical switches are more expensive, slower, bulkier, and more lossy than their electrical counterparts, making the transition from electrical to optical switching challenging. Several technologies have been used to produce optical switches. Mechanical switches are usually less expensive with a relatively low insertion loss, but they are slow (milliseconds), bulky, and not scalable. There are also additional issues related to the long term repeatability and reliability. Micro-Electro-Mechanical-System (MEMS) based switches are smaller and more scalable [18]. Thermo-optic switches operate based on changing the index of refraction in a Mach-Zehnder interferometer. The switching speed is approximately a few milliseconds, and the power consumption is comparatively high.

Electro-Optic switches can operate at sub-nanosecond speed [19]. Switching is performed by an electrical control signal that changes the index of refraction of Lithium Niobate (LiNbO3) in a Mach–Zehnder

configuration. Electro-Optic switches usually have a high insertion loss and polarization sensitivity. Semiconductor space switches combine optical amplification in semiconductors and interferometric effects. However, part of the switch insertion loss can be compensated by the optical amplifier gain. These switches can operate at nanosecond speed, but the cost of fabrication is relatively high.

Smart Amplifiers

EDFAs are one of the crucial components of photonic networks. They provide optical amplification to compensate for the fiber loss, components' insertion loss, and splitting loss in couplers. As the input optical power of an EDFA increases, its gain reduces and the output power saturates. In reconfigurable photonic networks, where optical channels are added or dropped dynamically, the amplifier gain changes in the sub-ms range, causing significant channel power fluctuation in the subsequent nodes [20]. The power fluctuation may further degrade the system's performance by increasing fiber nonlinearities [21] or signal-to-noise ratio degradations.

Several approaches for automatic gain control of EDFAs have been demonstrated. In one approach, the EDFA gain is adjusted by using a voltage-controlled attenuator, or by controlling the pump power [22]. A feedback signal from the output of the EDFA controls the pump power or attenuation.

Contention Resolution

One of the biggest challenges facing OPS networks is the lack of random access optical memories required for resolving contention. In the optical regime, contention problems may be addressed in time, wavelength, or space domains. Figure 3.16 summarizes the three contention resolution schemes.

In the time domain, an optical signal is delayed by going through a Fiber Delay Line (FDL) [23]. It is also possible to construct a programmable delay line using a cascade of optical switches and FDLs. There are several drawbacks in using FDLs as optical buffers. First, generating even a small time delay requires a long length of fiber. For example, 200 m of fiber will generate only 1 μs delay. Second, FDLs are first-in first-out (FIFO) memories with a predetermined delay time. Once an optical signal is sent into a fiber delay line, it will not be accessible until it comes out of the other end.

In the space domain or deflection technique, a contended packet is redirected to an empty output port rather than to its destination port [24]. This gives the network a chance to send the packet to its destination at a later time and/or via a different node. The deflected packet usually ends up traveling a longer route to its destination. As a result, packets may arrive out of order at the destination, and may need to be rearranged.
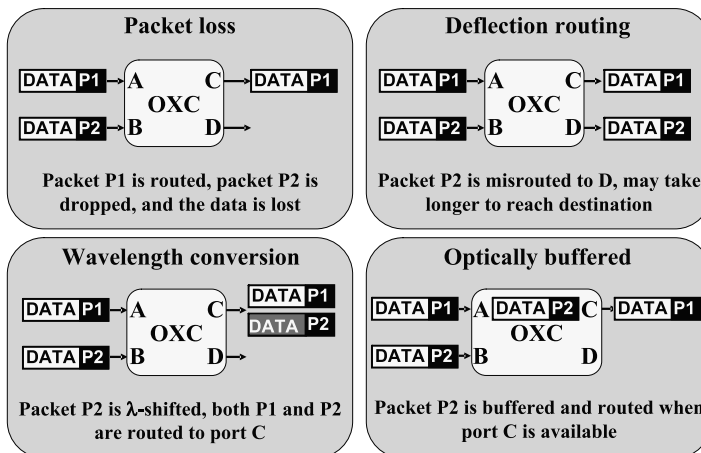
FIGURE 3.16 Contention and contention resolution techniques. Assuming all packets are destined to output port "C", P = packet.

Deflection routing requires a complicated control algorithm to make sure packets are not lost inside the network. Deflection also adds overhead to the network by artificially increasing the network load.

The third technique for contention resolution is based on wavelength conversion [25]. An ideal wavelength converter would change the wavelength of an optical signal independent of its bit-rate, data format, polarization, wavelength, or power as seen in Figure 3.17. However, practical wavelength converters are far from ideal, and depending on the technique, are sensitive to some or all of the input signal parameters. Other important performance metrics for wavelength converters are a low noise figure and a high output extinction ratio.
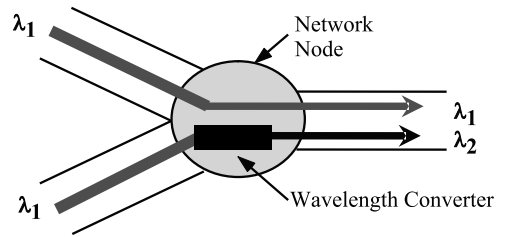
FIGURE 3.17 Wavelength conversion.

Wavelength Conversion

Popular wavelength conversion devices are based on one of three technologies [26]: gain saturation in semiconductor optical amplifiers, interferometric effects, and nonlinear wave-mixing.

Perhaps the simplest technique for wavelength conversion is to use gain saturation in semiconductor optical amplifiers (SOAs). As depicted in Figure 3.18, if the input power of a signal (pump) is large enough, it reduces the gain of the SOA by depleting the carriers. As a result, any other signal (probe) simultaneously propagating through the SOA sees a significantly reduced gain. This effect is called cross-gain-modulation (XGM). As the modulated pump signal drives the SOA into saturation, the inverse data pattern of the pump is imprinted on the continuous wave (CW) probe signal. At the output, a filter is used to select the wavelength-converted, bit-pattern-inverted copy of the original signal (i.e., on the probe wavelength). Since the phase information is lost in this process, this scheme can only be used for on-off-keying (OOK) data formats. The other drawbacks of this technique are signal-to-noise-ratio (SNR) degradation caused by the noise added by the SOA, the low extinction ratio, and the chirp induced on the output signal [27]. Slow gain recovery also limits the maximum bit-rate in an XGM based wavelength converter.

Interferometric wavelength converters take advantage of the index of refraction's dependency on the carrier density in semiconductors. As the index of refraction changes, the phase of the optical signal propagating through the medium changes as well. This phase modulation can easily be converted to amplitude modulation in an optical interferometer.

Nonlinear wave-mixing is the only wavelength conversion technique that makes an exact wavelength-shifted copy of the original signal (amplitude, frequency, conjugated phase). Wave-mixing can be generated in a
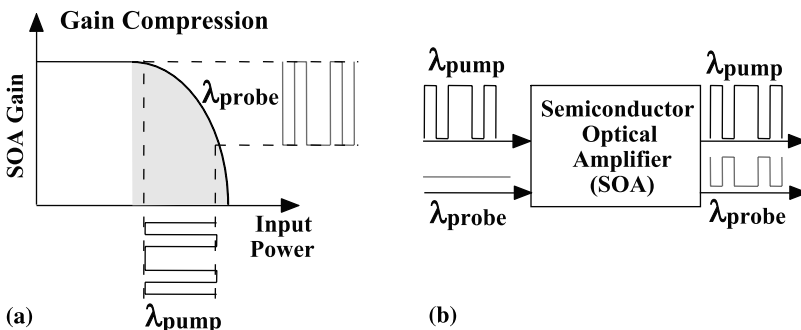
FIGURE 3.18 (a) Gain saturation in SOA caused by high power pump; (b) a continuous wave probe signal in inversely modulated by the pump.

cascaded second-order nonlinear structure, which is referred to as difference frequency generation (DFG). LiNbO3 (Lithium Niobate) is a good example of a material with a high second-order nonlinearity, a high optical bandwidth, and a high dynamic range [28]. When the third-order nonlinearity is used to generate a new wavelength, the process is called four-wave-mixing (FWM).

Cross absorption modulation (XAM) in an electro-absorption modulator has also been used for wavelength conversion [29]. The principle is similar to XGM. However in XAM, the pump signal modulates the absorption instead of the gain of the device, and a copy of the input data is imprinted on the probe wavelength.

Summary

Photonic networks can use light-paths to enable high-throughput and cost-effective data traffic routing. Today's networks are mostly static entities, but more rapid reconfigurability is in the prospective future. The future may even involve full optical packet switching in the optical domain, and the next several years of research will help define its future directions.

References

1. J. Cai, D. Foursa, C. Davidson, Y. Cai, G. Domagala, H. Li, L. Liu, W. Patterson, A. Pilipetskii, M. Nissov, N. Bergano, "A DWDM demonstration of 3.73Tb/s over 11,000 km using 373 RZ-DPSK channels at 10Gb/s," in *Proc. Optical Fiber Communications Conf.*, PD22, 2003.
2. P. Ivan, Kaminow, and Tingye Li, *Optical Fiber Telecommunications IV A and B*, 4th ed., Academic Press, 2002.
3. E. Desurvire, *Erbium-Doped Fiber Amplifiers: Principles and Applications*, Wiley, New York, 1994.
4. E. Desurvire, *Erbium-Doped Fiber Amplifiers, Device and System Developments*, Wiley, New York, 2002.
5. J. Bromage, "Raman amplification for fiber communications systems," *IEEE J. Lightwave Technology*, 22, 79, 2004.
6. ITU-T recommendation G.694.1, "DWDM frequency grid," 2002
7. R. Ramaswami and Kumar K. Sivarajan, *Optical Networks: A Practical Perspective*, 2nd ed., Morgan Kaufmann, San Francisco, 2002
8. W. Stallings, *Data and Computer Communications*, 7th ed., Pearson Prentice Hall, Upper Saddle River, 2004.
9. M. Maier and M. Reisslein, "AWG-based metro WDM networking," *IEEE Communications Magazine*, 42, S19, 2004.
10. J. Lacey, "Tutorial: optical cross-connect and add-drop multiplexers: technologies and applications," in *Proc. Optical Fiber Communication Conference*, 327, 2002.
11. S.-J. Park et al., "Fiber-to-the-home services based on wavelength-division-multiplexing passive optical network," *IEEE J. Lightwave Technology*, 22, 2582, 2004.
12. "SONET Telecommunications Standard Primer," Tektronix, application note, 2001.
13. U. Black, *MPLS and Label Switching Networks*, 2nd ed., Prentice Hall, Upper Saddle River, 2002.
14. M.J. O'Mahony, D. Simeonidou, D. Hunter, A. Tzanakaki, "The application of optical packet switching in future communication networks," *IEEE Communications Magazine*, 128, 2001.
15. C. Qiao, "Labeled optical burst switching for IP-over-WDM integration," *IEEE Communications Magazine*, 104, 2000.
16. M. Jeong, H.C. Cankaya, and C. Qiao, "On a new multicasting approach in optical burst switched networks," *IEEE Communications Magazine*, 96, 2002.
17. I. Baldine et al., "JumpStart: A just-in-time signaling architecture for WDM burst-switched networks," *IEEE Communications Magazine*, 82, 2002.
18. T-W. Yeow, K.L.E. Law, and A. Goldenberg, "MEMS optical switches," *IEEE Communications Magazine*, 39, 158, 2001.

19.   R. Krahenbuhl et al., "Performance and modeling of advanced Ti:LiNbO3 digital optical switches," *IEEE J. Lightwave Technology*, 20, 92, 2002.

20.   A.K. Srivastava et al., "Signal power transients in optically amplified WDM ring networks," in *Proc. Optical Fiber Communication Conf.*, 164, 1998.

21.   M.I. Hayee and A.E. Willner, "Transmission penalties due to EDFA gain transients in add-drop multiplexed WDM networks," *IEEE Photonics Technology Letters*, 11, 889, 1999.

22.   C. Tian and S. Kinoshita, "Analysis and control of transient dynamics of EDFA pumped by 1480- and 980-nm lasers," *J. Lightwave Technology*, 21, 1728, 2003.

23.   I. Chlamtac, A. Fumagalli, and S. Chang-Jin, "Multibuffer delay line architectures for efficient contention resolution in optical switching nodes," *IEEE Transactions on Communications*, 48, 2089, 2000.

24.   M. Baresi et al., "Deflection routing effectiveness in full-optical IP packet switching networks," in *Proc. IEEE International Conference on Communications*, 2, 1360, 2003.

25.   S. Rangarajan et al., "All-optical contention resolution with wavelength conversion for asynchronous variable-length 40 Gb/s optical packets," *IEEE Photonics Technology Letters*, 16, 689, 2004.

26.   J. Elmirghani and H. Mouftah, "All-optical wavelength conversion: Techniques and applications in DWDM networks," *IEEE Communications Magazine*, 86, 2000.

27.   T. Durhuus, B. Mikkelsen, C. Joergensen, S. Lykke Danielsen, and K.E. Stubkjaer, "All-optical wavelength conversion by semiconductor optical amplifiers," *IEEE J. Lightwave Technology*, 14, 942, 1996.

28.   I. Brener, M.H. Chou, and M.M. Fejer, "Efficient wideband wavelength conversion using cascaded second-order nonlinearities in LiNbo3 waveguides," in *Proc. Optical Fiber Communications Conf.*, 39, 1999.

29.   A. Hsu and S.L. Chuang, "Wavelength conversion by cross-absorption modulation using an integrated electroabsorption modulator/laser," in *Proc. the Conference on Lasers and Electro-Optics*, 488, 1999.

This page intentionally left blank

4

Computer Networks

John N. Daigle
University of Mississippi

Sarhan M. Musa
Prairie View A&M University

Matthew N.O. Sadiku
Prairie View A&M University

Richard B. Robrock II
Bell Laboratories Inc.

Apostolis K. Salkintzis
Motorola

Nikos Passas
University of Athens

Stan McClellan
Hewlett Packard Company

Remzi Seker
*University of Arkansas at
Little Rock*

4.1   Computer Communication Networks

John N. Daigle

Introduction

Over the last decade computer communication networks have evolved from the domain of research and business tools for the few, into the mainstream of public life. We have seen continued explosive growth of the Internet, a world-wide interconnection of computer communication networks that allows low-latency person-to-person communications on a global basis. Advancements in **firewall** technology have made it possible to conduct business using the Internet with significantly-reduced fear of compromise of important private information, and broad deployment of the **World Wide Web**, or simply the *Web*. Technology has facilitated the creation of new network-based businesses that span the globe. Significant steps have been taken to extend networking services to the mobile consumer. Perhaps more significantly, the introduction of high speed Internet access to homes and the introduction of the point-to-point protocol (PPP) have made it possible to extend the Internet to every home that has at least a basic twisted-pair telephone line.

   The potential for using networking technology as a vehicle for delivering multimedia voice, data, image, video presentations, multiparty, and multimedia conferencing services has been demonstrated, and the

important problems that must be solved in order to realize this potential are rapidly being defined and focused upon. User-friendly applications that facilitate navigation within the *Web* have been developed and made available to networking users on a non-fee basis via network servers, thus facilitating virtually instantaneous search and retrieval of information on a global basis. For example, it is now common for users to access broadcasts of exciting events, such as *Le Tour de France*, in real time over the Internet using only a dial-up telephone connection. Access to details concerning events of all kinds to all people is unprecedented in history.

By definition, a **computer communication network** is a collection of applications hosted on different machines and interconnected by an infrastructure that provides communications among the communicating entities. While the applications are generally understood to be computer programs, the generic model includes the human being as an application, an example being the people involved in a telephone call.

This article summarizes the major characteristics of computer communication networks. Our objective is to provide a concise introduction that will allow the reader to gain an understanding of the key distinguishing characteristics of the major classes of networks that exist today and some of the issues involved in the introduction of emerging technologies.

There are a significant number of well-recognized books in this area. Among these are the excellent texts by Schwartz [1987] and Spragins [1991], and more recently Kurose and Ross [2001], which have enjoyed wide acceptance both by students and practicing engineers and cover most of the general aspects of computer communication networks. Other books that have been found to be especially useful by many practitioners are those by Rose [1990], Black [1991], and Bertsekas and Gallager [1994.]

The latest developments are, of course, covered in the current literature, conference proceedings, formal standards documents, and the notes of standards meetings. A pedagogically oriented magazine that specializes in computer communications networks is *IEEE Network*, but in addition, *IEEE Communications* and *IEEE Computer* often contain interesting articles in this area. *ACM Communications Review*, in addition to presenting pedagogically oriented articles, often presents very useful summaries of the latest standards activities. Major conferences that specialize in computer communications include the IEEE INFOCOM and ACM SIGCOMM series, which are held annually. It is becoming common at this time to have more and more discussion about personal communication systems, and the mobility issues involved in communication networks are often discussed in *IEEE Network* and *IEEE Personal Communication Systems*. There are also numerous journals, magazines, and conferences specializing in subsets of computer communications technologies.

We begin our discussion with a brief statement of how computer networking came about and a capsule description of the networks that resulted from the early efforts. Networks of this generic class, called **wide area networks** (WANs) are broadly deployed today and there are still a large number of unanswered questions with respect to their design. The issues involved in the design of those networks are basic to the design of most networks, whether wide area or otherwise. In the process of introducing these early systems, we describe and contrast three basic types of communication switching: circuit, message, and packet.

We next turn to a discussion of computer communication **architecture**, which describes the structure of communication oriented processing software within a communication processing system. We introduce the International Standards Organization/Open Systems Interconnection (ISO/OSI) reference model (ISORM). Initially, the ISORM model was intended to guide the development of protocols, but its most important contribution seems to have been the provision of a framework for discussion of issues and developments across the communications field in general and communication networking in particular. We then present the Internet reference model, which parallels the Internet protocol stack that has evolved, and we base our discussion of protocols on protocols selected from the Internet protocol suite. This discussion is necessarily simplified in the extreme, thorough coverage requiring on the order of hundreds of pages, but we hope our brief description will enable the reader to appreciate some of the issues.

Having introduced the basic architectural structure of communication networks, we next turn to a discussion of an important variation on this architectural scheme: the **local area network** (LAN). Discussion of this topic is important because it helps to illustrate what a reference model is and what it is not. Specifically, early network architectures anticipate networks in which individual node pairs are

interconnected via a single link, and connections through the network are formed by concatenating node-to-node connections.

LAN architectures, on the other hand, anticipate all nodes being interconnected in some fashion over the same communication link (or medium). This, then, introduces the concept of adaption layers in a natural way. It also illustrates that if the services provided by an architectural layer are carefully defined, then the services can be used to implement virtually any service desired by the user, possibly at the price of some inefficiency.

Next we discuss ATM and Frame relay, two important link layer technologies. We conclude with a brief discussion of recent developments.

General Networking Concepts

Data communication networks have existed since about 1950. The early networks existed primarily for the purpose of connecting users of a large computer to the computer itself, with the additional capability to provide communications between computers of the same variety and having the same operating software. The lessons learned during the first 20 or so years of operation of these types of networks have been valuable in preparing the way for modern networks. For the purposes of our current discussion, however, we think of communication networks as being networks whose purpose is to interconnect a set of applications that are implemented on hosts manufactured by possibly different vendors and managed by a variety of operating systems. Networking capability is provided by software systems that implement standardized interfaces specifically designed for the exchange of information among heterogeneous computers.

The earliest effort to develop large-scale, general purpose networking capability based on packet switching was lead by the Advanced Research Projects Agency (ARPA) of the Department of the Army in the late 1960s; this effort resulted in the computer communication network called the ARPANET. The end results of the ARPA networking effort, its derivatives, and the early initiatives of many companies such as AT&T, DATAPOINT, DEC, IBM, and NCR have been far-reaching in the extreme. We will concentrate on the most visible product of these efforts, which is a collection of programs that allows applications running in different computers to intercommunicate. Before turning to our discussion of the software, however, we shall provide a brief description of a generic computer communication network.

Figure 4.1 shows a diagram of a generic computer communication network. The most visible components of the network are the **terminals**, the **access lines**, the **trunks**, and the **switching nodes**. Work is accomplished when the users of the network, the terminals, exchange messages over the network.

The terminals, which are usually referred to as end systems, represent the set of communication terminating equipment communicating over the network. Equipment in this class includes, but is not limited to, user terminals, general purpose computers, and database systems. This equipment, either through software or through human interaction, provides the functions required for information exchange between pairs of application programs or between application programs and people. The functions include, but are not limited to, call setup, session management, and message transmission control. Examples of applications include electronic mail transfer, www browsing, and playback of audio streams. An extensive discussion of applications is provided in Kurose and Ross [2000].

Access lines provide for data transmission between the terminals and the network switching nodes. These connections may be set up on a permanent basis or they may be switched connections, and there are numerous transmission schemes and protocols available to manage these connections. The essence of these connections, from our point of view, is a channel that provides data transmission at some number of bits per second (bps), called the channel capacity, (C). The access line capacities may range from a few hundred bps to in excess of millions of bps, and they are usually not the same for all terminating equipment of a given network. The actual information carrying capacity of the link depends upon the protocols employed to effect the transfer; the interested reader is referred to Bertsekas and Gallagher [1987], especially Chapter 2, for a general discussion of the issues involved in transmission of data over communication links.

Trunks, or internodal trunks, are the transmission facilities that provide for transmission of data between pairs of communication switches. These are analogous to access lines, and from our point of view, they simply provide a communication path at some capacity, specified in bps.
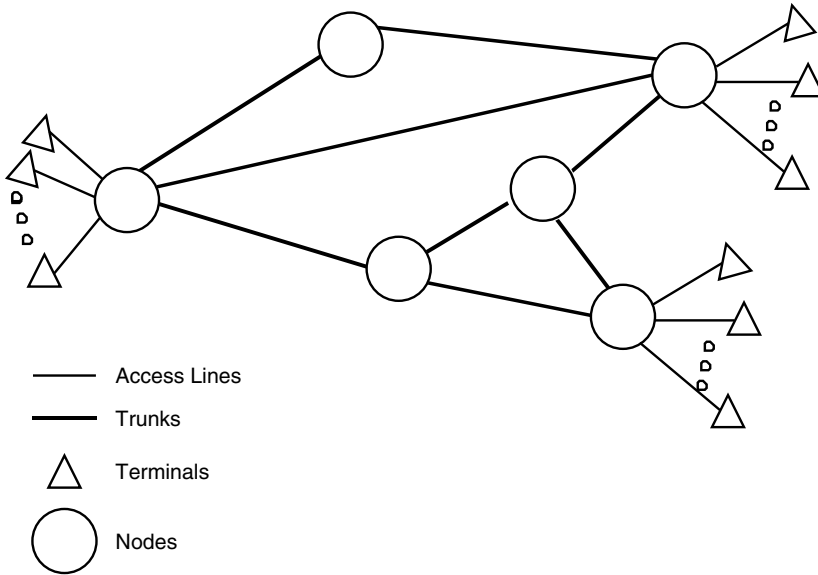
FIGURE 4.1   Generic computer communication network.

There are three basic switching paradigms: circuit, message, and packet switching. **Circuit switching** and **packet switching** are transmission technologies while message switching is a service technology. In circuit switching, a call connection between two terminating pieces of equipment corresponds to the allocation of a prescribed set of physical facilities that provide a transmission path of a certain bandwidth or transmission capacity. These facilities are dedicated to the users for the duration of the call. The primary performance issues, other than those related to quality of transmission, are related to whether or not a transmission path is available at call setup time and how calls are handled if facilities are not available.

Message switching is similar in concept to the postal system. When a user wants to send a message to one or more recipients, the user forms the message and addresses it. The message switching system reads the address and forwards the complete message to the next switch in the path. The message moves asynchronously through the network on a message switch to message switch basis until it reaches its destination. Message switching systems offer services such as mail boxes, multiple destination delivery, automatic verification of message delivery, and bulletin boards. Communication links between the message switches may be established using circuit or packet switching networks as is the case with most other networking applications. An example of a message switching protocol that has been used to build message switching systems is the Simple Mail Transfer Protocol (SMTP).

In the circuit switching case, there is a one-to-one correspondence between the number of trunks between nodes and the number of simultaneous calls that can be carried. That is, a trunk is a facility between two switches that can service exactly one call, and it does not matter how this transmission facility is derived. Major design issues include the specification of the number of trunks between node pairs and the routing strategy used to determine the path through a network in order to achieve a given call blocking probability. When blocked calls are queued, the number of calls that may be queued is also a design question.

A packet switched communication system exchanges messages among users by transmitting sequences of packets comprising the messages. That is, the sending terminal equipment partitions a message into a sequence of packets, the packets are transmitted across the network, and the receiving terminal equipment reassembles the packets into messages. The transmission facility interconnecting a given node pair is viewed as a single trunk, and the transmission capacity of this trunk is shared among all users whose packets traverse

both nodes. While the trunk capacity is specified in bps, the packet handling capacity of a node pair depends both upon the trunk capacity and the nodal processing power.

In some packet switched networks, the path traversed by a packet through the network is established during a call set-up procedure, and the network is referred to as a virtual circuit packet switching network. Other networks provide datagram service, a service that allows users to transmit individually addressed packets without the need for call set-up. Datagram networks have the advantage of not having to establish connections before communications takes place, but have the disadvantage that every packet must contain complete addressing information. Virtual circuit networks have the advantage that addressing information is not required in each packet, but have the disadvantage that a call set-up must take place before communications can occur. Datagram is an example of **connectionless service** while virtual circuit is an example of **connection-oriented service**.

Prior to the late 1970s, signaling for circuit establishment was in-band. That is, in order to set up a call through the network, the call set-up information was sent sequentially from switch to switch using the actual circuit that would eventually become the circuit used to connect the end users. In an extreme case, this amounted to trying to find a path through a maze, sometimes having to retrace one's steps before finally emerging at the destination or just simply giving up when no path could be found. This had two negative characteristics: first, the rate of signaling information transfer was limited to the circuit speed, and second, the circuits that could have been used for accomplishing the end objective were being consumed simply to find a path between the end-points. This resulted in tremendous bottlenecks on major holidays, which were solved by virtually disallowing alternate routes through the toll switching network.

An alternate out-of-band signaling system, usually called **common channel interoffice signaling** (CCIS), was developed primarily to solve this problem. Signaling now takes place over a signaling network that is partitioned from the network that carries the user traffic. This principle is incorporated into the concept of integrated services digital networks (ISDNs), which is described thoroughly in Helgert [1991]. The basic idea of ISDN is to offer to the user some number of 64 Kbps access lines plus a 16 Kbps access line through which the user can describe to an ISDN how the user wishes to use each of the 64 Kbps circuits at any given time. The channels formed by concatenating the access lines with the network inter-switch trunks having the requested characteristics are established using an out-of-band signaling system, the most modern of which is Signaling System #7 (SS#7).

In either virtual circuit or **datagram networks**, packets from a large number of users may simultaneously need transmission services between nodes. Packets arrive to a given node at random times. The switching node determines the next node in the transmission path, and then places the packet in a queue for transmission over a trunk facility to the next node. Packet arrival processes tend to be bursty, that is, the number of packet arrivals over fixed-length intervals of time has a large variance. Because of the "burstiness" of the arrival process, packets may experience significant delays at the trunks. Queues may also build due to the difference in transmission capacities of the various trunks and access lines and delays result. Processing is also a source of delay, and the essence of packet switching technology is to trade delay for efficiency in resource utilization.

Protocol design efforts, which seek to improve network efficiencies and application performance, are frequent topics of discussion at both general conferences in communications and those specialized to networking. The reader is encouraged to consult the proceedings of the conferences mentioned earlier for a better appreciation of the range of issues and the diversity of the proposed solutions to the issues.

Computer Communication Network Architecture

In this section, we begin with a brief, high level, definition of the ISORM, which is discussed in significant detail in Black [1991]. Then we turn to a more detailed discussion of a practical Internet reference model (PIRM). The ISORM has seven layers, none of which can be bypassed conceptually. In general, a layer is defined by the types of services it provides to its users and the quality of those services. For each layer in the ISO/OSI architecture, the user of a layer is the next layer up in the hierarchy, except for the highest layer for which the user is an application. Clearly, when a layered architecture is implemented under this philosophy the
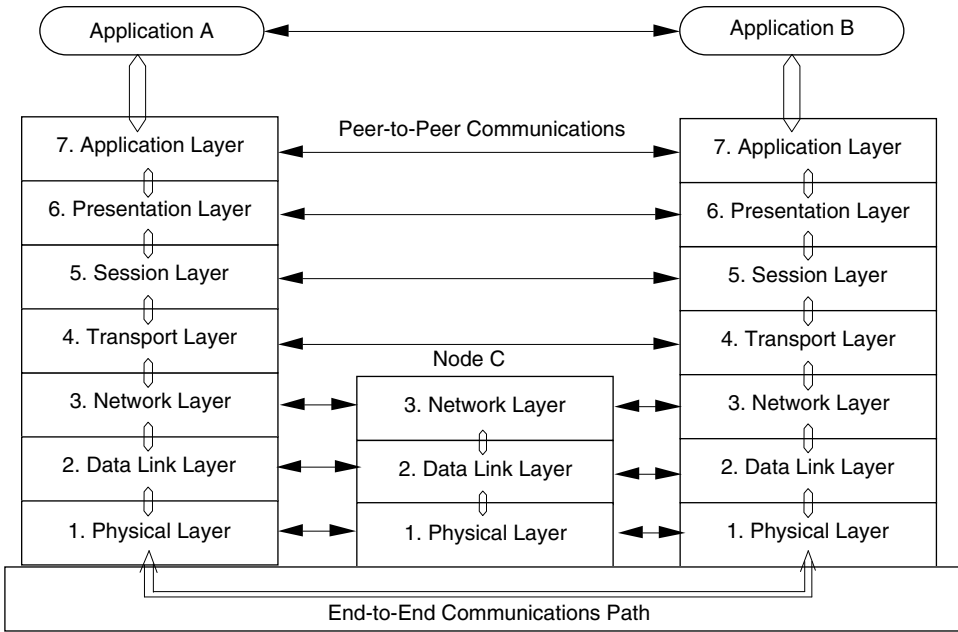
FIGURE 4.2   The ISO reference model.

quality of service obtained by the end user, the application, is a function of the quality of service provided by all of the layers.

Figure 4.2, adapted from Spragins [1991], shows the basic structure of the OSI architecture and how this architecture is envisaged to provide for exchange of information between applications. As shown in the figure, there are seven layers: application, presentation, session, transport, network, data link, and physical. Brief definitions of the layers are now given, but the reader should bear in mind that substantial further study will be required to develop an understanding of the practical implications of the definitions.

Physical layer: Provides electrical, functional, and procedural characteristics to activate, maintain, and deactivate physical data links that transparently pass the bit stream for communication between data link **entities**.

Data link layer: Provides functional and procedural means to transfer data between network entities; provides for activation, maintenance, and deactivation of data link connections, character and frame synchronization, grouping of bits into characters and frames, error control, media access control, and flow control.

Network layer: Provides switching and routing functions to establish, maintain, and terminate network layer connections and transfer data between transport layers.

Transport layer: Provides host-to-host, cost effective, transparent transfer of data, end-to-end flow control, and end-to-end quality of service as required by applications.

Session layer: Provides mechanisms for organizing and structuring dialogues between application processes.

Presentation layer: Provides for independent data representation and syntax selection by each communicating application and conversion between selected contexts and the internal architecture standard.

Application layer: Provides applications with access to the ISO/OSI communication stack and certain distributed information services.

As we have mentioned previously, a layer is defined by the types of services it provides to its users. In the case of a request or a response, these services are provided via invocation of **service primitives** of the layer in question by the layer that wants the service performed. In the case of an indication or a confirm, these services are provided via invocation of service primitives of the layer in question by the same layer that wants the service performed.

The process of invoking a service primitive is not unlike a user of a programming system calling a subroutine from a scientific subroutine package in order to obtain a service, say, matrix inversion or memory allocation. For example, a request is analogous to a CALL statement in a FORTRAN program, and a response is analogous to the RETURN statement in the subroutine that has been called. The requests for services are generated asynchronously by all of the users of all of the services and these join (typically prioritized) queues along with other requests and responses while awaiting servicing by the processor or other resource such as a transmission line.

The service primitives fall into four basic types, which are as follows: request, indication, response, and confirm. These types are defined as follows:

Request: A primitive sent by layer ($N+1$) to layer N to request a service.

Indication: A primitive sent by layer N to layer ($N+1$) to indicate that a service has been requested of layer N by a different layer ($N+1$) entity.

Response: A primitive sent by ($N+1$) to layer N in response to an *indication primitive.*

Confirm: A primitive sent by layer N to layer ($N+1$) to indicate that a response to an earlier *request* primitive has been received.

To understand the basic ideas of a network, it is useful to first ask the question "What functionality would be needed in a point-to-point connection between two computers that are hard-wired together?" We would find that we need everything except the functions provided by layer 3; there is no need for switching and routing functions. Thus, it is natural to think of the networking aspect of a computer communication network as the interconnection of layer 3 entities. In fact, we can think of the network as the collection of layer 3 entities. Everything above layer 3 facilitates activities that are taking place over the network and everything below layer 3 is there simply for the purpose of connecting the layer 3 entities together.

As a practical matter, networking has evolved such that the more realistic reference model has five, rather than seven, layers. In essence, the session and presentation layers are absorbed into the application layer, with the resulting reference model as shown in Figure 4.3.

In order to be more specific about how communications takes place, we now turn to a brief discussion of Internet Protocol (IP), which is the network layer protocol of the Internet. The IP protocol provides a datagram delivery service to the transport layer. In order to do this, IP provides exactly one primitive to its upper layer protocol (ULP): SEND.

As mentioned earlier, an upper layer protocol issuing a primitive is equivalent to a computer program calling a subroutine or procedure. A *primitive* has an associated set of **formal parameters** that are analogous to
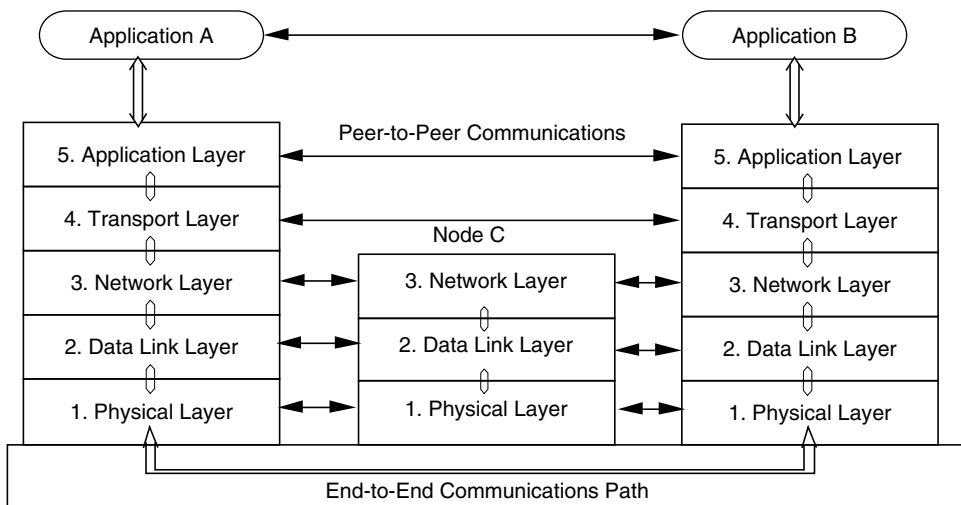

FIGURE 4.3 Practical internet reference model.

the formal parameters of a procedure in a programming language. In this case, the parameters for the SEND.request primitive are the source address, the destination address, the type-of-service parameter set, the data length, the options parameters, and the data. These parameters provide all the information needed by IP to deliver data to the required destination. The data itself is the entity that the ULP wants sent. The source address is the Internet address of the ULP requesting the SEND service and the destination address is the Internet address of the ULP to which the data are to be sent. The type of service and options parameters give IP information about how the delivery is to be handled; for example, what is the priority of the send action? The data length parameter tells IP how much data it has to send.

In general, communications takes place between peer layer protocols by the exchange of **protocol data units** (PDUs), which contain all of the information required for the receiving protocol **entity** to provide the required service. For the PIRM, we say that applications exchange *messages*, layer-4 entities exchange *segments*, layer-3 entities exchange *datagrams*, layer-2 entities exchange *frames*, and layer-1 protocols exchange *1-PDUs*. It should be mentioned that datagrams are packets that traverse networks using connectionless layer-3 protocols. Equivalently, we can say that the packet-switched Internet is a datagram network because IP is the only layer-3 protocol, and it provides connectionless service.

In order to exchange PDUs, entities of a given layer use the services of the next lower layer. To make things concrete, assume the ULP sending the information is a Transmission Control Protocol (TCP) entity in Host A and the ULP to which the information is being sent is the TCP entity in Host B. In the case of the SEND discussed previously, the TCP entity in Host A is using the SEND service to send the information in the data field to its peer destination TCP entity in Host B; the content of this data field is the PDU and it is called a segment since TCP is a layer-4 protocol. At the destination, Host B, IP uses its DELIVER primitive to deliver the packet to the TCP entity in Host B. Thus, the TCP's packet is the data contained in the data field of its SEND request. The destination TCP entity must be able to determine what to do with any packet it receives. Therefore, the PDUs, or packets, must be structured in a standard way so that a computer program can examine the data received and take the appropriate actions. The structure that must be adhered to is specified in the protocol standard for the protocol. For example, TCP is specified in Request for Comments RFC 793.

Similarly, the IP entity in Host A forms a PDU that is sent to its peer IP entity in another mode. The IP-layer PDU, which is called a *datagram* since IP is a layer-3 protocol, is formed by using information provided in the service request issued by the TCP entity and other information available in the IP layer entity itself. Instructive information is packed into the header of the IP datagram while the data part, or sometimes a subset of the data part, of the TCP request forms the data part of the IP datagram.

Some protocols are connection-oriented and some are connectionless, meaning that the service they provide is connection-oriented or connectionless. For example, IP provides a connectionless service while TCP is connection-oriented. For connection mode communications, a connection must be established between two peer entities before they can exchange PDUs. Therefore, connection-oriented protocols must provide primitives that facilitate establishment of a connection. For example, TCP has several forms of OPEN primitives that an application can use to initiate connections and TCP also has CLOSE and ABORT primitives that an application can use to terminate connections.

Once a connection is established, data exchange between the two application layer entities can take place; that is, the entities can exchange PDUs. For example, if an application layer entity in Host A wishes to send a PDU to an application layer entity in Host B, the application layer entity in Host A would issue a T_SEND.request to the appropriate transport link layer entity in Host A. This entity would package the PDU together with appropriate control information into a transport service data unit (TSDU) and send it to its peer transport layer entity in Host B. In order to do this, the transport layer would invoke the services of the network layer. In turn, the network layer would invoke the services of the data link layer and so on. Specifically, the network layer entity in Host A would issue a DL_DATA.request to the appropriate data link layer entity in Host A. This entity would package the PDU together with appropriate control information into a data link service data unit (DLSDU) and send it to its peer at C. The peer data link entity at C would extract the layer-3 PDU, deliver it to the network entity at C, which would forward it to the data link entity in C, providing the connection to Host B. This entity would then send the DLSDU to its peer in Host B,

and this data link entity would extract the layer-3 PDU and pass it to the Host B network entity via a DL_DATA.indication. The receiving network layer entity would extract the SDU and forward it to the transport layer in Host C using the N_Data.indication.

Now, network layer PDUs are called packets and DL layer PDUs are called frames. But the data link layer does not know that the information it is transmitting is a packet; to the DL layer entity, the packet is simply user information. From the perspective of a data link entity, it is not necessary to have a network layer. The network layer exists to add value for the user of the network layer to the services provided by the DL layer. In the example above, value was added by the network layer by providing a relaying capability since Hosts A and C were not directly connected. Similarly, the DL layer functions on a hop-by-hop basis, each hop being completely unaware that there are any other hops involved in the communication. We will see later that the data link need not be limited to a single physical connection.

We now turn to a discussion of LANs, which have inherent properties that make use of sublayers particularly attractive.

Local Area Networks and Internets

In this section, we discuss the organization of communications software for LANs. In addition, we introduce the idea of internets, which were brought about to a large extent by the advent of LANs. We discuss the types of networks only briefly and refer the reader to the many excellent texts on the subject. Layers 4 and 5 for local area communications networks are identical to those of wide area networks. However, because the hosts communicating over a LAN share a single physical transmission facility, the routing functions provided by the network layer, layer 3, are not necessary. Thus, the functionality of a layer 3 within a single LAN can be substantially simplified without loss of utility. On the other hand, a DL layer entity must now manage many simultaneous DL layer connections because all connections entering and leaving a host on a single LAN do so over a single physical link. Thus, in the case of connection oriented communications, the software must manage several virtual connections over a single physical link.

There were several basic types of transmission schemes in use in early local area networks. Three of these received serious consideration for standardization: the **token ring**, **token bus**, and **carrier sense multiple access** (CSMA). All three of those access methods became IEEE standards, the IEEE 802 series, and eventually became ISO standards (ISO 8802 series) because all merited standardization. On the other hand, all existed for the express purposes of exchanging information among peers, and it was recognized at the outset that the upper end of the data link layer could be shared by all three access techniques. The decision to use a common logical link control (LLC) sublayer for all of the LAN protocols and develop a separate sublayer for each of the different media apparently ushered in the idea of adaption sublayers, in this case, the media access control (MAC) sublayer. The most dominant today is the CDMA-based version, which is known as Ethernet or IEEE 802.3.

The idea of sublayering has proven to be valuable as new types of technologies have become available. For example, the new fiber distributed digital interface (FDDI) uses the LLC of all other LAN protocols, but its MAC is completely different from the token ring MAC even though FDDI is a token ring protocol. A thorough discussion of FDDI and related technologies is given in Jain [1994].

Metropolitan area networks (MANs) have been deployed for the interconnection of LANs within a metropolitan area. The primary media configuration for MANs is a dual bus configuration and it is implemented via the distributed queue, dual bus (DQDB) protocol, also known as IEEE 802.6. The net effect of this protocol is to use the dual bus configuration to provide service approaching the FCFS service discipline to the traffic entering the FDDI network, which is remarkable considering that the LANs being interconnected are geographically dispersed. Interestingly, DQDB concepts have recently also been adapted to provide wide area communications. Specifically, structures have been defined for transmitting DQDB frames over standard DS-1 (1.544 megabits per second) and DS-3 (6.312 megabits per second) facilities, and these have been used as the basis for a service offering called Switched Multi-megabit Data Services (SMDS).

One of the more interesting consequences of the advent of local area networking is that many traditional computer communication networks became internets overnight. LAN technology was used to connect stations

to a host computer, and these host computers were already on a WAN. It was then a simple matter to provide a relaying, or bridging, service at the host in order to provide wide area interconnection of stations on LANs to each other. In short, the previously established WANs became networks for interconnection of LANs; that is, they were interconnecting networks rather than stations. Internet performance suddenly became a primary concern in the design of networks. A new business developed to provide specialized equipment, routers, to provide interconnection among networks.

Over the last decade, wireless LANs, which are local area networks in which radio or photonic links serve as cable replacements, have become common. Indeed, wireless LAN technology is now widely deployed in homes as well as in offices. Most of this technology features spread-spectrum wireless transmission with the media access technology being in the IEEE 802.11 family.

We now describe two broadly deployed technologies that provide link-layer connections in communication networks: *frame relay* (FR) technology, which is described in Braun [1994] and **asynchronous transfer mode** (ATM) technology, which is described in McDysan and Spohn [1994] and Pildush [2000].

As we have mentioned previously, there is really no requirement that the physical media between two adjacent data link layers be composed of a single link. In fact, if a path through the network is initially established between two data link entities, there is no reason that DLC protocols need to be executed at intermediate nodes. Through the introduction of adaption layers and an elementary routing layer at the top of the DL layer, DLC frames can be relayed across the physical links of the connection without performing the error checking, flow control, and retransmission functions of the DLC layer on a link-by-link basis. The motivation is that since link transmission is becoming more reliable, extensive error checking and flow control is not needed across individual links; an end-to-end check should be sufficient. Meanwhile, the savings in processing due to not processing at the network layer can be applied to frame processing, which allows interconnection of the switches at higher line speeds. Since bit-per-second costs decrease with increased line speed, service providers can offer savings to their customers through FRNs. Significant issues are frame loss probability and retransmission delay. Such factors will determine the retransmission strategy deployed in the network. The extensive deployment of FR technology at this time suggests that this technology provides improvements over standard packet technology.

Another recent innovation is the ATM. The idea of ATM is to partition a user's data into many small segments, called cells, for transmission over the network. Independent of the data's origin, the cell size is 53 octets, of which 5 octets are for use by the network itself for routing and error control. Users of the ATM are responsible for segmentation and reassembly of their data. Any control information required for this purpose must be included in the 48 octets of user information in each cell. In the usual case, these cells would be transmitted over networks that would provide users with 135 Mb/s and above data transmission capacity (with user overhead included in the capacity).

The segmentation of units of data into cells introduces tremendous flexibility for handling different types of information, such as voice, data, image, and video, over a single transmission facility. As a result, there has been a tremendous investment in developing implementation agreements that will enable a large number of vendors to independently develop interoperable equipment. This effort is focused primarily in the ATM Forum, a private, not-for-profit consortium of over 500 companies of which more than 150 are principal members and active contributors.

LANs, WANs, and MANs based on the ATM paradigm are being deployed as described in Pildush [2000], for example. ATM is an excellent technology for providing worldwide Ethernet LAN interconnection at a full rate of 10 or 100 megabits per second. Originally, numerous vendors were planning to have ATM capabilities at the backplane in much the same way that Ethernet has been provided, but the bulk of deployment of ATM today is at the link layer between IP routers.

ATM and Frame Relay

The asynchronous transfer mode technology was initially developed in hopes of extending ISDN technology towards the concept of a broadband integrated services data network that would provide bandwidth on demand to end users. The ATM architecture consists of three sublayers: the ATM adaption layer (AAL), the

ATM layer, and the Physical Media Dependent (PMD) layer. It is convenient for the purposes of this discussion to think of services as falling into two categories: circuit mode and packet mode, where a circuit mode service, such as voice, is a service that is naturally implemented over a circuit switched facility and a packet-mode service, such as e-mail, is a service that is more naturally implemented over a packet switched connection. From many perspectives, it is natural to implement circuit-mode services directly over ATM while it is more natural to implement packet-mode services at the Internet (or packet) layer.

The implication of this partitioning of service types is that any service that has been developed for deployment over an IP network would naturally be deployed over an ATM network by simply using the ATM network as a packet delivery network. Each packet would traverse the network as a sequence of cells over an end-to-end virtual connection. If the flow control and resource management procedures can be worked out, the net effect of this deployment strategy would be, for example, that an application designed to be deployed over an Ethernet segment could be deployed on a nationwide (or even global) network without a noticeable performance degradation. The implications of this type of capability are obvious as well as mind-boggling.

Recent Developments

Broad deployment of networking technologies and wireless communications devices are the big story in recent times.

A typical home deployment of networking technology today includes a router, which is at the interface of the home network and the outside world. Typically, the router has an IEEE 802.3 (Ethernet) port of the 100 Mb/s variety as well as an IEEE 802.11 wireless LAN port facing the in-home side of the network. The connection from the home may be one of many technologies, ranging from a traditional relatively low-speed dialed connection over a phone line to relatively high speed connection based on some form of digital subscriber loop (DSL) technology, such as asymmetric DSL (ADSL), which delivers service to the home in the hundreds of kb/s to low Mb/s range. Connected together to form a network within a home may be a number of computers and a print server having an attached printer. A number of additional computers may be attached to the same network via wireless IEEE 802.11 connections. The router itself is usually connected to an Internet service provider (ISP) over the point-to-point protocol (PPP). All Internet applications running on all endsystems (computers) in the home share the same PPP link-layer connection to the ISP port to the Internet.

Many of the services that network users enjoy today are the result of enhancements to the development of multimedia applications and protocols that are integrated into Web browsers. The Web browsers allow users to access search engines through which they can find interesting Web content. Once the content is found, the user can generally access the content quickly by simply clicking on a URL. When the object arrives at the user's system, the Web browser automatically invokes the services of a helper application, which renders the object to the user. For example, in case the object is a JPEG file, the Web browser will invoke an application that can display a picture that is stored in JPEG format.

At the time of this writing, networking technologies are being extended to the wireless domain as discussed in Lin and Chlamtac [2001] and Garg [2002]. The intention is to make all networking services available to mobile end systems using essentially the same infrastructure as the cellular telephone system. Alternatively stated, the intention is to evolve the wireless cellular phone system to a point where it is just part of the Internet. Meanwhile the intention is to evolve the Internet to the point where it handles all types of communications services. Thus, over the long term it is expected that communications types be available over the Internet and that access to the Internet will be available to users at any time and place, whether the user is mobile or stationary. The end objective is Universal Personal Communications Services (UPCS) to which we all look forward.

Defining Terms

Introduction

Architecture: The set of protocols defining a computer communication network.

Computer communication network: Collection of applications hosted on different machines and interconnected by an infrastructure that provides intercommunications.

Firewall: Computer communication network hardware and software introduced into an Internet at the boundary of public network and a private network for the purpose of protecting the confidential information and network reliability of the private network.

Local area network: A computer communication network spanning a limited geographic area, such as a building or college campus.

Metropolitan area network: A computer communication network spanning a limited geographic area, such as a city; sometimes features interconnection of LANs.

Wide area network: A computer communication network spanning a broad geographic area, such as a state or country.

World Wide Web: A collection of hypertext-style servers interconnected via Internet services.

General Networking Concepts

Access line: A communication line that connect a user's terminal equipment to a switching node.

Circuit switching: A method of communication in which a physical circuit is established between two terminating equipments before communication begins to take place. This is analogous to an ordinary phone call.

Connectionless service: A mode of packet switching in which packets are exchanged without first establishing a connection. Conceptually, this is very close to message switching, except that if the destination node is not active, then the packet is lost.

Connection-oriented service: A mode of packet switching in which a call is established prior to any information exchange taking place. This is analogous to an ordinary phone call, except that no physical resources need be allocated.

Common channel interoffice signalling: Use of a special network, dedicated to signalling, to establish a path through a communication network, which is dedicated to the transfer of user information.

Message switching: A service-oriented class of communication in which messages are exchanged among terminating equipments by traversing a set of switching nodes in a store and forward manner. This is analogous to an ordinary postal system. The destination terminal need not be active at the same time as the originator in order that the message exchange take place.

Packet switching: A method of communication in which messages are exchanged between terminating equipments via the exchange of a sequence of fragments of the message called packets.

Switching node: A computer or computing equipment that provides access to networking services.

Trunk: A communication line between two switching nodes.

Computer Communication Network Architecture

Entity: A software process that implements a part of a protocol in a computer communication.

Formal parameters: The parameters passed during the invocation of a service primitive; similar to the arguments passed in a subroutine call in a computer program network.

International Standards Organization Reference Model: A model, established by ISO, that organizes the functions required by a complete communication network into seven layers.

Protocol data unit: The unit of exchange of protocol information between entities. Typically, a protocol data unit (PDU) is analogous to a structure in C or a record in Pascal; the protocol is executed by processing a sequence of PDUs.

Service primitive: The name of a procedure that provides a service; similar to the name of a subroutine or procedure in a scientific subroutine library.

Local Area Networks and Internets

Adaption sublayer: Software that is added between two protocol layers to allow the upper layer to take advantage of the services offered by the lower layer in situations where the upper layer is not specifically designed to interface directly to the lower layer.

Token bus: A method of sharing a bus-type communications medium that uses a token to schedule access to the medium. When a particular station has completed its use of the token, it broadcasts the token on the bus, and the station to which it is addressed takes control of the medium.

Token ring: A method of sharing a ring-type communications medium that uses a token to schedule access to the medium. When a particular station has completed its use of the token, it transmits the token on the bus, and the station that is physically next on the ring takes control.

Carrier sense multiple access: A random access method of sharing a Bus-type communication medium in which a potential user of the medium listens before beginning to transmit.

Media access control: A sublayer of the link layer protocol whose implementation is specific to the type of physical medium over which communication takes place and controls access to that medium.

Internet: A network formed by the interconnection of networks.

ATM and Frame Relay

Asynchronous transfer mode (ATM):   A mode of communication in which communication takes place through the exchange of tiny units of information called cells.

Broadband Integrated Services Digital Network (B-ISDN):   A generic term that generally refers to the future network infrastructure that will provide ubiquitous availability of integrated voice, data, imagery, and video services.

Fast packet network: Networks in which packets are transferred by switching at the frame layer rather than the packet layer. Such networks are sometimes called frame relay networks. At this time, it is becoming vogue to think of frame relay as a service, rather than transmission, technology.

References

D. Bertsekas and R. Gallagher, *Data Networks*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1987.

U.D. Black, *OSI: A Model for Computer Communication Standards*, Englewood Cliffs, NJ: Prentice Hall, 1991.

E. Braun, *The Internet Directory*, New York: Fawcett Columbine, 1994.

V.K. Garg, *Wireless Network Evolution: 2 G to 3G*, Upper Saddle River, NJ: Prentice Hall, 2002.

J.L. Hammond and P.J.P. O'Reilly. *Performance Analysis of Local Computer Networks*, Reading, MA: Addison-Wesley, 1986.

H.J. Helgert, *Integrated Services Digital Networks*, Reading, MA: Addison-Wesley, 1991.

R. Jain, *Handbook: High-Speed Networking Using Fiber and Other Media*, Reading, MA: Addison-Wesley, 1994.

J.F. Kurose and K.W. Ross, *Computer Networking: A Top–Down Approach Featuring the Internet*, Reading, MA: Addison-Wesley, 2000.

Y.-B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*, New York: Wiley.

D.E. McDysan and D.E. Spohn, *ATM: Theory and Application*, New York: McGraw-Hill, 1994.

G.D. Pildush, *Cisco ATM Solutions: Master ATM Implementation of Cisco Networks*, Indianapolis, IN: Cisco Systems Press, 2000.

M. Rose, *The Open Book*, Englewood Cliffs, NJ: Prentice Hall, 1990.

M. Schwartz, *Telecommunications Networks: Protocols, Modeling and Analysis*, Reading, MA: Addison-Wesley, 1987.

J.D. Spragins, *Telecommunications: Protocols and Design*, Reading, MA: Addison-Wesley, 1991.

W. Stallings, *Handbook of Computer-Communications Standards: The Open Systems Interconnection (OSI) Model and OSI-Related Standards*, New York: Macmillan Publishing Company, 1990.

Further Information

There are many conferences and workshops that provide up-to-date coverage in the computer communications area. Among these are the IEEE INFOCOM and ACM SIGCOMM conferences and the IEEE Computer Communications Workshop, which are specialized to computer communications and are held annually. In addition, IEEE GLOBCOM (annual), IEEE ICC (annual), IFIPS ICCC (bi-annual), and the

International Telecommunications Congress (bi-annual) regularly feature a substantial number of paper and panel sessions in networking.

The ACM *Communications Review*, a quarterly, specializes in computer communications and often presents summaries of the latest standards activities. *IEEE Network*, a bi-monthly, specializes in tutorially oriented articles across the entire breadth of computer communications and includes a regular column on books related to the discipline. Additionally, *IEEE Communications* and *IEEE Computer*, monthly magazines, frequently have articles on specific aspects of networking. Also, see *IEEE Personal Communication Systems*, a quarterly magazine, for information on wireless networking technology.

For those who wish to be involved in the most up-to-date activities, there are many interest groups on the Internet that specialize in some aspect of networking. Searching for information on the Internet has become greatly simplified with the advent of the World Wide Web and the deployment of a number of search engines on the Web. It is simply a matter of accessing one of these search engines, entering the topic of interest, and waiting for results. For example, one can go to www.google.com and then enter mobile IP as the search keyword, and Google will return ten pages of references on the topic, including the set of standards relating to the topic.

4.2   Local Area Networks

Sarhan M. Musa and Matthew N.O. Sadiku

Introduction

A local area network (LAN) is a high-speed data network that interconnects workstations, personal computers, mainframe computers, printers and other peripheral devices within a small geographical area. A LAN can be as small as one room, or can extend over multiple rooms, multiple floors within a building, and even multiple buildings within a campus or organization. LANs offer computer users many advantages, including shared access to devices (such as printers), file exchange between connected users, and communication between users via electronic mail.

There are different kinds of LANs depending on topology: Ethernet (or CSMA/CD), token ring, token bus, and star. Here, we will consider each of these as well as wireless LAN.

Ethernet

The Ethernet, also known as the carrier sense multiple access with collision detection (CSMA/CD) system, is the most widely used form of LAN. Ethernet was so-named to describe the way that cabling could carry data everywhere throughout the network. Ethernet or CSMA/CD is the most widely installed local area network LAN technology because of its simplicity. It is specified and used as the basis of LAN standards by the IEEE 802.3 standards committee. Since then, it has evolved from 10 Mbps, as traditional Ethernet, to fast Ethernet that operates at 100 Mbps, and then to Gigabit Ethernet at 1 or 10 Gbps.

With CSMA/CD, a station wishing to transmit first listens to the medium to see whether another transmission is in progress (carrier sense). If the medium is idle, the station may transmit. If two or more stations attempt to transmit at the same time, there will be a collision; the data from both transmissions will be garbled and not receive successfully. The station will normally wait $9.6\mu s$ or 96 bit periods before starting the transmission. The CSMA/CD Ethernet technology is collision based; that is to say, collision is normal in an Ethernet network. Collisions can happen early or late during transmission, and are therefore termed as early collision and late collision. Figure 4.4 explains the scenario for an early collision.

The physical layer characteristics include:

- Data rate: 10 Mbps to 10 Gbps
- Maximum station separation: 2.8 km

- Maximum number of stations: 1024
- Medium: twisted pair, coaxial cable, and optical fiber
- Logical topology: bus
- Physical topology: bus, star, hierarchical star
- Maximum frame size: 1518 bytes
- Frames on LAN: single
- Message protocols: variable frame size, "best effort delivery"

Topology

The topology of a LAN usually refers to the structure or geometric layout of the cable used to interconnect stations on the network. Unlike conventional data communications networks that can be configured in a variety of ways by the addition of hardware and software, most LANs are designed to operate based upon the interconnection of stations that follow a specific topology. The most common used in LANs include bus, ring, and star as shown in Figure 4.5.

Bus

In a bus topology structure, a cable is laid out as one long branch on which each station is connected. The bus topology was the first topology used when local area networks became commercially available in the late
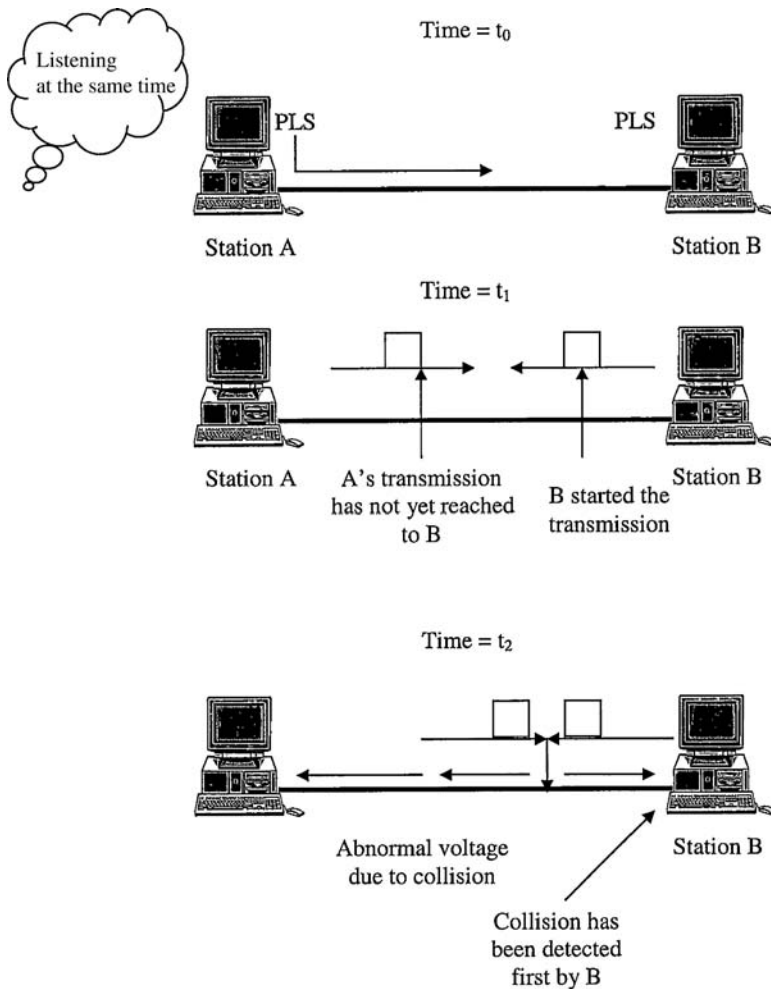
FIGURE 4.4 Explanation of early collision. (Adopted from Chowdhury [2000].)

Ring

Bus

Star
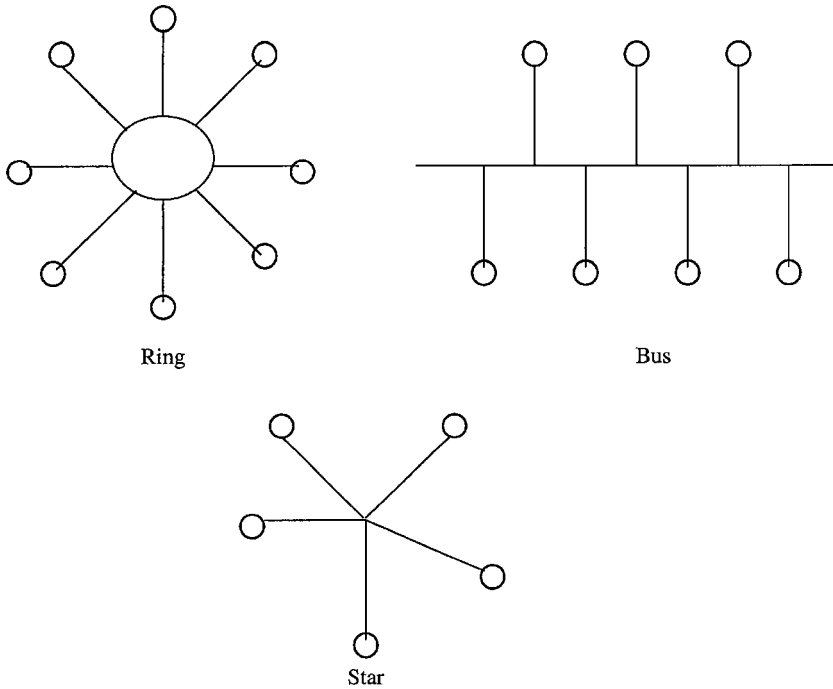
FIGURE 4.5   Local network topologies.

1970s. Since then, its use has diminished significantly. Connecting the cable requires a simple device called a tap. The tap is a passive device, because it does not alter the signal and does not require electricity to operate. On the workstation end of the cable is a network interface card (NIC), which is an electronic device that performs the necessary signal conversions and protocol operations.

LAN Addressing Schemes

A unique 48-bit identifier or address in each unit is installed by the manufacturers of network interface hardware. This is often called the MAC (Media Access Control) number. Only Ethernet chip or board manufacturers are usually involved in obtaining new numbers. Although the Ethernet address is often fixed at the factory, alternative schemes are employed on some systems, and the workstation IP number is used to generate part of the Ethernet address. The Ethernet packet header has three fields, with a trailer field containing a CRC error check number. Ethernet receivers read the 48-bit MAC destination address fields as the packets pass by. Any packet holding the reader's address will then continue to be read into the input buffer. A scheme of this type avoids having to read all the passing packets and also does not involve the main CPU in any activity. If a packet is seen to have a "foreign" destination address it would either be ignored or redirected to the gateway machine that would pass it onto the wider network. This solution becomes impossible to maintain due to daily changes and hardware failures.

As shown in Figure 4.6, IP version 4 numbers have four parts—the top few bits specify the Class of IP number, then the site ID, then the subnet identifier, and finally, the lower part identifies the actual machine. This manner of separating the IP number into fields helps in understanding when networks are being configured and ranges of addresses need to be allocated to each LAN segment. Only the lower subnet + host values can be revised by the local administrator; the upper network part has been assigned by the Internet Assigned Number Authority (NIC).
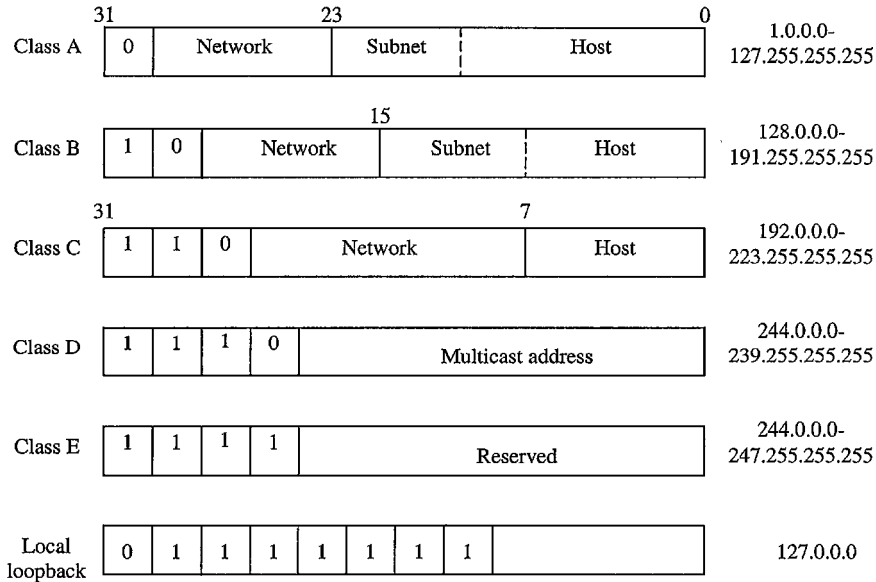
FIGURE 4.6 Five forms of IPv4 numbers and their ranges.

Ring

In a ring topology, a single cable that forms the data highway is shaped into a ring, as shown in Figure 4.7. Such topology allows unicast, multicast, and broadcast addressing. Each station checks the destination address on the received frame.

When the destination address is unicast, it must match the physical address of the station. If it does, the packet is copied and regenerated and passed to the next station in the ring. Otherwise, the packet is only regenerated and passed to the next station without being copied. When the destination address is multicast, the transmission is from one source to many stations that register to receive the traffic. For broadcast addressing, transmission is from one source to every station on the network. A station and the medium are interfaced through a medium interface, which can be an external piece of hardware installed on the medium or as part of the network interface card installed inside the station. In the case of ring topology, both internal and external interfaces are active devices, which means that they act as repeaters.

For a ring topology, signals are propagated in only one direction from one interface to another. This means that each station has a predecessor and successor. Any two medium interfaces are connected point-to-point.

Token Ring

The token ring (or token-passing ring) is a protocol defined in IEEE Project 802.5 standard. It uses a token-passing access method, which implied that stations take turns in sending data. Each station is allowed to transmit only during its turn and may send only one frame during each turn. The mechanism responsible for this rotation is called *token passing*. A token is a special placeholder frame that goes from station to station around the ring. Every station is allowed to send data only when it has possession of the token. The token is passed from station to station until it encounters a station with data to send. The station keeps the token and sends a data frame.

The data frame is passed around the ring, being regenerated by each station. Each intermediate station examines the destination address, finds that the frame is addressed to another station, and relays it to its neighbor. The destination station recognizes it own address, copies the message, and changes four bits in the last byte of the frame to indicate the address was recognized and the frame was copied. The full packet then
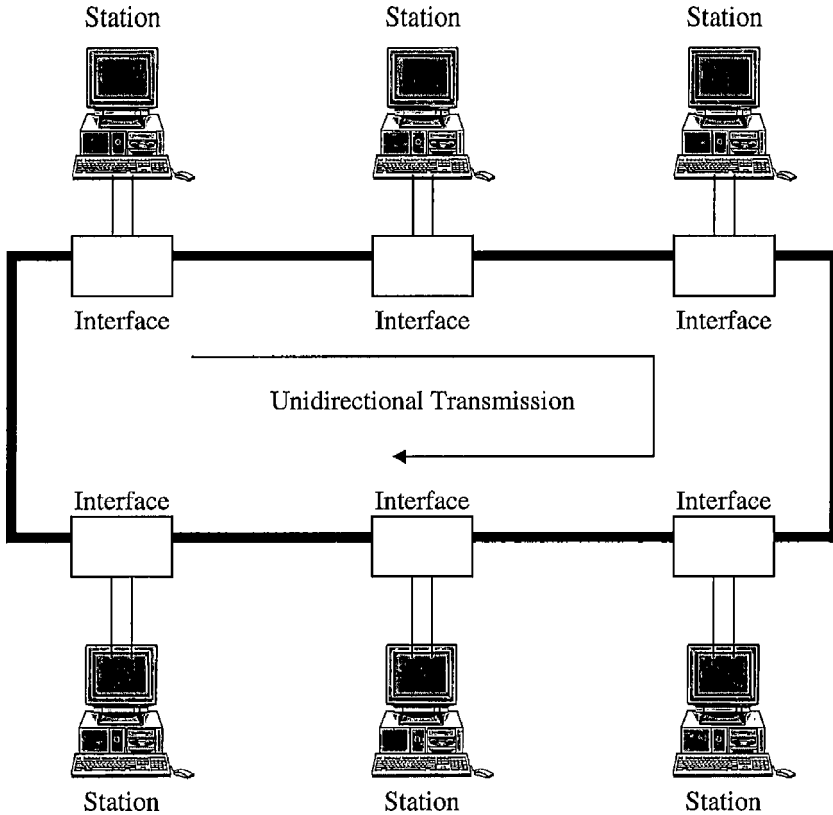
FIGURE 4.7   A typical ring LAN. (Adopted from Forouzan [2003].)

continues around the ring until it returns to the source station that sent it. The source station receives the frame. It releases the token back to the ring. The sender then discards the used data frame.

Star

In star LAN topology, each station is directly connected to a common central node, referred to as the star coupler, via two point-to-point links, one for transmission and one for reception. Endpoints on a network are connected to a common central hub, or switch, by dedicated links. Logical bus and ring topologies are often implemented physically in a star topology. A typical star LAN is shown in Figure 4.8.

VLANs

A virtual local area network (VLAN) is a subnetwork or a segment of a local area network configured by software, not by physical wiring. Segmentation makes broadcasting possible at the data link layer. A LAN can be divided into several logical LANs called *VLANS* with each VLAN acting as a workgroup in the organization. If a person moves from one group to another, the physical configuration does not need to change since the group membership is defined by software, not hardware. Some VLAN vendors use the 32-bit IP address as a membership characteristic. For example, the administrator can define that stations having IP addresses 181.34.23.67, 181.34.23.72 belong to VLAN.

Wireless LAN

The wireless local area network (WLAN) is a new form of communication system. It is basically a local area network, confined to a geographically small area such as a single building, office, store, or campus, that
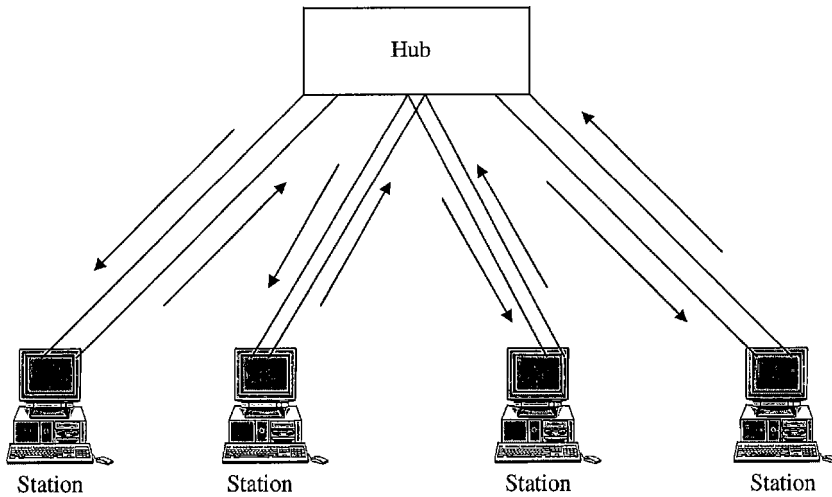
FIGURE 4.8 A typical star LAN. (Adopted from Forouzan [2003].)

provides high data connectivity to mobile stations. Using electromagnetic airwaves (radio frequency or infrared), WLANs transmit and receive data over the air. A WLAN suggests less expensive, fast, and simple network installation and reconfiguration.

The proliferation of portable computers coupled with the mobile user's need for communication is the major driving force behind WLAN technology. WLAN creates a mobile environment for the PC and LAN user. It may lower LAN maintenance and expansion costs since there are no wires that require reconfiguration. Thus, WLANs offer the following advantages over the conventional wired LANs:

- Installation flexibility: allows the network to go where wire cannot go
- Mobility: can provide LAN users with access anywhere
- Scalability: can be configured in a variety of topologies to meet specific needs

However, WLAN does not perform as well as wired LAN because of the bandwidth limitations and may be susceptible to electromagnetic interference and distance. While the initial investment on WLAN hardware can be higher than the cost of wired LAN hardware, overall installation expenses and life-cycle costs can be significantly lower.

Physical Layer and Topology

WLAN does not compete with wired LAN. Rather, WLANs are used to extend wired LANs for convenience and mobility. Wireless links essentially fill in for wired links using electromagnetic radiation at radio or light frequencies between transceivers. A typical WLAN consists of an access point and the WLAN adapter installed on the portable notebook. The access point is a transmitter/receiver (transceiver) device; it is essentially the wireless equivalent of a regular LAN hub. An access point is typically connected with the wired backbone network at a fixed location through a standard Ethernet cable and communicates with wireless devices by means of an antenna. WLANs operate within the prescribed 900 MHz, 2.4 GHz, and 5.8 GHz frequency bands. Most LANs use 2.4 GHz frequency bands because it is most widely accepted.

A wireless link can provide services in several ways, the following among them:

- Replace a point-to-point connection between two nodes or segments on a LAN. A point-to-point link is a connection between two devices for transferring data. A wireless link can be used to bridge two LAN segments, as shown in Figure 4.9. Like a point-to-point link, the link connects two wireless bridges
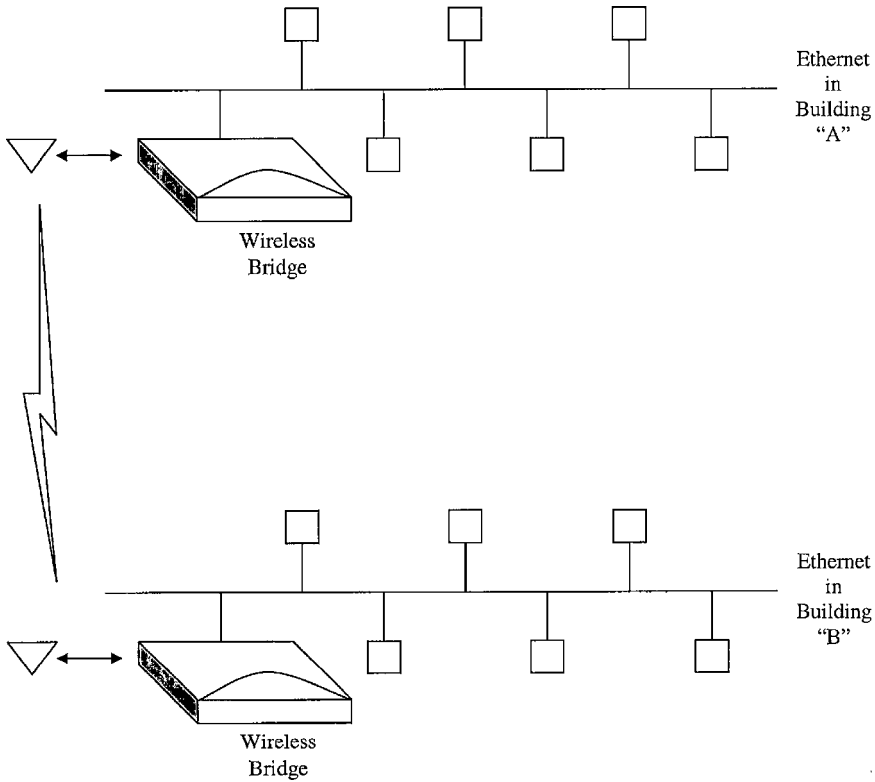
FIGURE 4.9   A wireless link replacing a point-to-point connection.

attached to the two LANs. Such an arrangement is useful for linking LANs in two buildings where a highway or river makes direct connection difficult.

- Provide a connection between a wired LAN and one or more WLAN nodes. In this case, a device is attached to the wired LAN to act as a point of contact (called access point) between the wired LAN and the wireless nodes. The device can be a repeater, bridge, or router.
- Act as a stand-alone WLAN for a group of wireless nodes. This can be achieved using topologies similar to a wired LAN, namely, a star topology can be formed with central hub controlling the wireless nodes, a ring topology with each wireless node receiving or passing information sent to it or a bus topology with each wireless capable of hearing everything said by all the other nodes. The three popular WLAN topologies are star, ring, and bus.

Technologies

When designing WLANs, manufacturers have to choose from two main technologies that are used for wireless communications today: radio frequency (RF) and infrared (IR). Each technology has its own merits and demerits.

RF is used for applications where communications are over long distances and are not line-of-sight. In order to operate in the license-free portion of the frequency spectrum known as the ISM band (industrial, scientific, and medical), the RF system must use a modulation technique called *spread spectrum* (SS). Spread spectrum is wideband radio frequency technology developed by the military during World War II for use in reliable, secure, mission-critical communications systems. The SS system is one in which the transmitted signal is spread over a frequency much wider than the minimum bandwidth required to send the signal. Using spread spectrum, a radio is supposed to distribute the signal across the entire spectrum. This way, no single user can

dominate the band and collectively all users look like noise. The fact that such signals appear like noise in the band makes them difficult to find and jam, thereby increasing security against unauthorized listeners. There are two types of spread spectrum technology: frequency hopping and direct sequence.

Frequency hopping spread spectrum (FHSS) offers a current maximum data rate of 3 Mbps. It uses a narrowband carrier that changes frequency in a pattern known to both transmitter and receiver. It is based on the use of a signal at a given frequency that is constant for a small amount of time and then moves to a new frequency. The sequence of different channels for the hopping pattern is determined in pseudorandom fashion. This means that a very long sequence code is used before it is repeated, over 65,000 hops, making it appear random. Thus it is very difficult to predict the next frequency at which such a system will stop and transmit/receive data as the system appears to be a noise source to an unauthorized listener. This makes the FHSS system very secure against interference and interception. FHSS is characterized by low-cost, low-power consumption, and less range than DSSS but greater range than infrared. Most WLAN systems use FHSS.

Direct sequence spread spectrum takes a signal at a given frequency and spreads it across a band of frequencies where the center frequency is the original signal. The spreading algorithm, which is the key to the relationship of the spread range of frequencies, changes with time in a pseudorandom sequence. When the ratio between the original signal bandwidth and the spread signal bandwidth is very large, the system offers great immunity to interference. For example, if a 10 kbps signal is spread across 1 GHz of spectrum, the spreading ratio is 100,000 times or 50 dB. However, in the ISM band used in WLAN, the available bandwidth critically limits the ratio of spreading and so the advantages of a DSSS scheme against interference is greatly limited. It has been shown that for the WLAN system using DSSS, the spreading ratio is at best 10 times. DSSS is characterized by high-cost, high-power consumption, and more range than FHSS and infrared physical layers.

The second technology used in WLAN is infrared (IR), where the communication is carried by light in the invisible part of the spectrum. It is primarily used for very short distance communications (less than 1 m), where there is a line-of-sight connection. Since IR light does not penetrate solid materials (it is even attenuated greatly by window glass), it is not really useful in comparison to RF in WLAN system. However, IR is used in applications where the power is extremely limited, such as a pager.

Standards

Although a number of proprietary, nonstandard wireless LANs exist, standards have now been developed. Two international organizations have contributed to the development of standards for WLANs: the Institute of Electronics and Electrical Engineers (IEEE) and the European Telecommunications Standards Institute (ETSI).

In 1997, the IEEE 802.11 committee (http://ieee802.org/11) issued a standard for wireless LANs. The standard addresses the physical and MAC layers of the OSI model and includes the following:

- A transmission rate of up to 2 Mbps
- Two different media for transmission over wireless LAN: infrared (IR) and radio frequency (RF)
- The media access control (MAC) protocol as carrier sense multiple access with collision avoidance (CSMA/CA), i.e. devices can interoperate with wired LANs via a bridge
- MAC protocol provides two service types: asynchronous and synchronous (or contention-free). The asynchronous type of service is mandatory while the synchronous type is optional
- MAC layer protocol is tied to the IEEE 802.2 logical link control (LLC) layer making it easier to integrate with other LANs
- Three different physical layers: an optical-based physical-layer implementation that uses IR light to transmit and two RF-based physical-layer choices—direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS) both operating at 2.4 GHz industrial, scientific, and medical (ISM) frequency bands. (The ISM bands 902–928 MHz, 2400–2483.5 MHz, and 5725–5850 MHz do not require a license to operate.) Added features to the MAC that can maximize battery life in portable clients via power-management schemes
- Data security through which the wireless LANs can achieve wired equivalent privacy

The standard basically defines the media and configuration issues, transmission procedures, throughput requirements, and range characteristics for WLAN technology. It avoids rigid requirements and gives room for vendors in the following areas: multiple physical media, common MAC layer irrespective of the physical layer, common frame format, power limit, and multiple on-air data rates.

There are three major problems encountered by an RF LAN. First, frequency allocation is limited for LANs, but since LANs operate with low power, frequency reuse is possible. Second, interference from other wireless LANs controlled by different organizations and other wireless sources is a problem. This problem can be controlled by using spread spectrum techniques. Third, security is at stake because an RF signal can penetrate through walls and hostile operators can intercept RF LAN communications. Encryption can be used to lessen this problem. IR LAN uses both laser diodes and light-emitting diodes as emitters. It is useful in high electromagnetic interference (EMI) environments. It is also secure since IR signals cannot penetrate walls.

CSMA/CA is slightly different from carrier sense multiple access with collision detection (CSMA/CD), which is the MAC protocol used in an Ethernet-wired LAN. In CSMA/CA, when a node has something to transmit, it waits for silence on the network. When no other nodes are heard, it transmits and waits to receive an acknowledgment from the recipient node. If it fails to receive an acknowledgment within a time period, it assumes that collision has occurred and follows a process similar to CSMA/CD. Each node then waits for silence and only transmits after a random waiting time. While CSMA/CA protocol is slower than CSMA/CD due to the need for waiting for acknowledgment, it works well for wireless LANs. Also, WLANs operate in strong multipath fading channels, where channel characteristics can change resulting in unreliable communication.

The ETSI devoted its attention to RF wireless LANs. The ETSI is close to finalizing its standard, which is based on the 2.4 GHz range used for spread-spectrum LANs in several European countries. The European standard WLAN, called HiperLAN, will allow speeds of 24 Mbps [5].

Besides IEEE and ETSI, there are organizations that are more interested in the implementation and interoperability of WLAN products. Such organizations include the Wireless LAN Alliance (WLANA at www.wlana.com) and Wireless Ethernet Compatibility Alliance (WECA at www.wi-fi.org or www.wireless-ethernet.com). WLANA was formed in 1996 with 12 members as a trade association for wireless LAN vendors. WECA is a nonprofit manufacturing consortium with over 60 companies as members; it was formed in 1999 to certify interoperability of IEEE 802.11 products.

Applications

Offering the obvious advantage of no wire installation costs, wireless LANs can be deployed in a dynamic environment where there is frequent reconfiguration of computer networks. Also, without the cables, excavation and long installation waiting times, it is simpler to connect difficult-to-reach customers.

Although several products for RF and IR LANs are already available in the marketplace, the introduction of their applications is just beginning. Typical mobile users are assumed to be laptop or notebook computers and portable base stations. Services provided by WLANs include data applications such as over TCP/IP and multimedia applications.

The most prominent users of WLANs are those whose projects promise quick payoffs for adding mobility. Industries such as security services, banks, retail, manufacturing, and healthcare are more notable for deploying wireless LANs that allow workers to roam while gathering information.

Mobile terminals—personal digital assistants (PDAs), specialized handheld terminals, and barcode scanners—connected to WLANs are increasingly being used to enhance business operations. It has become commonplace for WLANs to be used in applications such as:

- Printer sharing: linking to a distant printer within a department
- Electronic mail: sending and receiving e-mails from anywhere
- Healthcare: access to patient records from practically anywhere and location-independent claims processing

- Financial services such Stock or Community Exchange: implementing hand-held communicators in the trading room to increase the speed, accuracy, and reliability of its price reporting system
- Factory control: data acquisition, inventory control, scoreboards, and robotics

Other applications include trading, banking, restaurants, retail industry, warehousing, manufacturing, education, office environments, petroleum industry, agriculture, and food services. Today, WLAN technology is becoming fairly mature. WLANs are becoming more widely recognized as a general-purpose connectivity alternative for a broad range of customers.

Still, the WLAN market remains small because the technology is new and so components are expensive and the data rates are low. On the one hand, it costs less than $100 to buy the network card needed to connect a PC to a wired Ethernet LAN with the data rate of 10 Mbps. On the other hand, the card needed to interface the same PC to wireless radio LAN costs $500, while the wireless hubs (access points) that connect the portable units to the wired network cost as much as $3000 each for a data rate of 1 to 2 Mbps. However, research groups are working hard to shrink radios into a chip that can be mass produced cheaply. If they succeed, the demand for radio LANs may follow the same trend as cellular phones in recent years.

References

D.D. Chowdhury, *High Speed LAN Technology Handbook*, Berlin: Springer Verlag, 2000.

B.A. Forouzan, *Data Communications and Networking*, 3rd ed., New York: McGraw Hill, 2004.

B.A. Forouzan, *Local Area Networks*, New York: McGraw Hill, 2003.

M. Sadiku and M. Ilyas, *Simulation of Local Area Networks*, Boca Raton, FL: CRC Press, 1995.

W. Stallings, *High-Speed Networks and Internets: Performance and Quality of Service*, Upper Saddle River, NJ: Prentice Hall, 2002.

C.M. White, *Data Communications and Computer Networks, A Business User's Approach, Course Technology*, Boston, MA: Thomson Learning, Inc., 2002.

R. Williams, *Computer Systems Architecture, A Networking Approach*, Upper Saddle River, NJ: Pearson Education, 2001.

G.M. Zobrist, "Local area networks," *IEEE Potentials*, December/January, pp. 6–10, 1995.

4.3 The Intelligent Network

Richard B. Robrock II

The term *intelligent network* refers to the concept of deploying centralized databases in the telecommunications network and querying those databases to provide a wide variety of network services such as 800 service (toll-free service) and credit card calling. The first use of these centralized databases was in AT&T's network in 1981 where they were used to facilitate the setup of telephone calls charged to a calling card. Today such databases are widely deployed throughout North America and support the handling of well over 100 billion telephone calls per year.

The words *intelligent network,* when first used, had a relatively narrow definition, but that definition has broadened considerably with the introduction of the advanced intelligent network, the wireless intelligent network, and soon, the broadband intelligent network. The advanced intelligent network has introduced powerful service creation tools that have empowered network providers to create their own network services. The network providers, in turn, are beginning to broaden the participation in service creation by allowing their customers or third parties to use these tools to create services. The result has been a rapid growth in new network services.

A History of Intelligence in the Network

The first "intelligence" in the telephone network took the form of rows of human telephone operators, sitting side by side, plugging cords into jacks to facilitate the handling of calls. These operators established calls to far-away points, selected the best routes and provided billing information. They were also an information source—providing time or weather or perhaps disseminating the local news. Moreover, they had the opportunity to demonstrate a kind of heroism—gathering volunteers to save a house from fire, helping to catch a prowler, locating a lost child, and on and on. In the early years of telephony, the feats of the telephone operator were indeed legendary.

In the 1920s, however, technology became available that allowed automatic switching of telephone calls through the use of sophisticated electromechanical switching systems. Initially, these switches served as an aid to operators; ultimately, they led to the replacement of operators. The combination of the rotary telephone dial and the electromechanical switch allowed customers to directly dial calls without the assistance of operators. This led to a reduction of human intelligence in the network.

Another dramatic change took place in the telephone network in 1965; it was called software. It came with the marriage of the computer and the telephone switching system in the first stored-program control switch. With the introduction of switching software came a family of custom calling services (speed calling, call waiting, call forwarding, and three-way calling) for residential customers, and a robust set of Centrex features (station attendant, call transfer, abbreviated dialing, etc.) for business customers. The first software programs for these stored-program control switches contained approximately 100,000 lines of code; by 1990 some of these switching systems became enormously complex, containing 10 million lines of code and offering hundreds of different services to telephone users.

During the 1980s, a new architectural concept was introduced; it came to be called the intelligent network. It allowed new telecommunications services to be introduced rapidly and in a ubiquitous and uniform fashion. Feature and service availability in the network ceased to be solely dependent upon the hardware and software in stored-program control switches. Rather some new intelligence was centralized in databases that were accessed using packet switching techniques. Most significantly, the intelligent network started to provide some of the capabilities that operators had made available in the early years of telephony. The remaining sections describe the intelligent network, its characteristics, and its services. They also provide a description of the advanced intelligent network, which dramatically broadens the participation in the creation of new services.

The Intelligent Network

The intelligent network architecture is illustrated in Figure 4.10; its primary elements are a switching system, a signaling network, a centralized database, and an operations support system that supports the database. The architectural concept is a simple one. When a customer places a telephone call that requires special
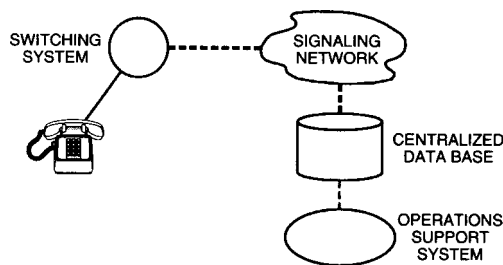

FIGURE 4.10 Intelligent network architecture—telephone calls that require special handling are intercepted in a switching system that launches queries through a signaling network to a centralized database. (*Source:* R.B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1, pp. 7–20, January 1991. © 1991 IEEE.)

handling, such as a toll-free call (800 service) or credit card call, that call is intercepted by the switching system that suspends call processing while it launches a query through a signaling network to a centralized database. The database, in turn, retrieves the necessary information to handle the call and returns that information through the signaling network to the switch so that the call can be completed. The role of the operations support system is to administer the appropriate network and customer information that resides in the database.

It is conceivable that the database in this architecture could reside in the switching system, and the signaling network in this instance would not be required. However, that would magnify the task of administering the customer information, since that information would be contained in thousands of switches instead of dozens of centralized databases. In addition, even more importantly, there are two shortcomings associated with basing many of the potential new services in switches, rather than utilizing centralized databases to provide information for the switches. The first is a deployment problem. As of 1990 there were more than 15,000 switches in the United States, and a single switch can cost millions of dollars. To introduce a new service in local switches and to make it widely available generally requires some not-so-simple changes in those switches or, in some cases, replacement of certain switch types altogether. These switch modifications typically take years to implement and require a tremendous capital investment. As a result, ten years after introduction, custom calling services were available to fewer than 1% of the residential customers in the United States.

A second problem with switch-based services has been that a single service sometimes behaves differently in different switch types. For example, the speed calling access patterns are different in various stored-program control switches. The public is not particularly sensitive to this fact, because speed calling is not associated with an individual but rather an individual's station set. People live in a mobile society, however, and they want to have their services available from any station set and have them behave the same from any station set.

The intelligent network architecture has been the key to solving both the deployment problem and service uniformity problem associated with switch-based services. Services deployed using an intelligent network centralized database are immediately ubiquitous and uniform throughout a company's serving area.

Intelligent Network Systems

In 1981, AT&T introduced into the Bell System a set of centralized databases called network control points; they supported two applications—the billing validation application for calling card service (credit card calling) and the INWATS database used to support 800 service. Queries were launched to these databases through AT&T's common-channel interoffice signaling (CCIS) network.

In 1984, following the divestiture of the Regional Bell Operating Companies from AT&T, the regional companies began planning to deploy their own **common-channel signaling (CCS)** networks and their own centralized databases. They selected the **signaling system 7** protocol for use in their signaling networks, called CCS7 networks, and they named their databases **service control points (SCPs).**

The CCS7 Network

A general architecture for a regional signaling network is shown in Figure 4.11. The network is made up of **signal transfer points (STPs),** which are very reliable, high-capacity packet switches that route signaling messages between network access nodes such as switches and SCPs. To perform these routing functions, the STPs each possess a large routing database containing translation data.

The CCS7 network in Figure 4.11 contains both local STPs and regional STPs. The STPs are typically deployed in geographically separated pairs so that in the event of a natural disaster at one site, such as an earthquake, flood, or fire, the total traffic volume can be handled by the second site. Indeed, redundancy is provided at all key points so that no single failure can isolate a node.

As illustrated in Figure 4.11, the following link types have been designated:

- A-links connect an access node, such as a switching system or SCP, to both members of an STP pair.
- B-links interconnect two STP pairs forming a "quad" of four signaling links where each STP independently connects to each member of the other pair.
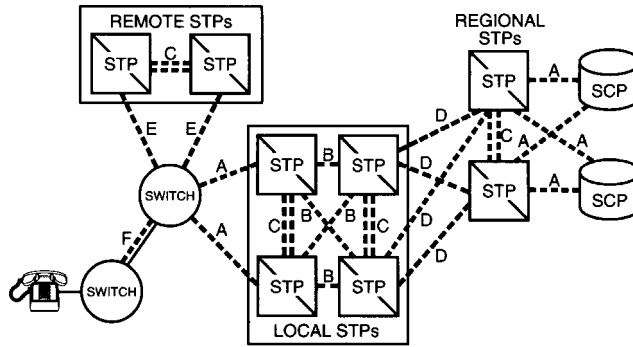
FIGURE 4.11   Link arrangements in a CCS7 signaling network. (*Source:* R.B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1, pp. 7–20, January 1991. © 1991 IEEE.)

- C-links are the high-capacity connections between the geographically separated members of an STP pair.
- D-links connect one STP pair to a second STP pair at another level in the signaling hierarchy or to another carrier.
- E-links connect an access node to a remote STP pair in the signaling network and are rarely used.
- F-links directly interconnect two access nodes without the use of an STP; they are nonredundant.

The CCS7 links normally function at 56 kb/s in North America, while links operating at 64 kb/s are common in Europe.

The CCS7 signaling network provides the underlying foundation for the intelligent network, and the regional telephone companies in the United States began wide-scale deployment of these networks in 1986; several large independent telephone companies and interexchange carriers (ICs) soon followed. They used these networks for both trunk signaling between switches as well as for direct signaling from a switch to a database.

The Service Control Point

The "brains" of the intelligent network is the SCP. It is an on-line, fault-tolerant, transaction-processing database that provides call handling information in response to network queries. The SCP deployed for 800 service is a high-capacity system capable of handling more than 900 queries per second or 3 million per hour. It is a real-time system with a response time of less than one half second, and it is a high-availability system with a downtime of less than 1 minute per year for a mated SCP pair. The SCP is also designed to accommodate growth, which means that processing power or memory can be added to an in-service system without interrupting service. In addition, it is designed to accommodate graceful retrofit, which means that a new software program can be loaded into an in-service SCP without disrupting service.

Data Base 800 Service

SCPs have been deployed throughout the United States in support of the Data Base 800 Service mandated by the Federal Communications Commission. This service provides its subscribers with number portability so that a single 800 number can be used with different carriers. The Data Base 800 Service architecture is shown in Figure 4.12. With this architecture, 800-number calls are routed from an end office to a service switching point (SSP) that launches queries through a CCS7 signaling network to the SCP. The SCP identifies the appropriate carrier, as specified by the 800 service subscriber, and then, if appropriate, translates the 800 number to a plain old telephone (POTS) number. This information is subsequently returned to the SSP so that the call can be routed through the network by handing the call off to the appropriate carrier. This technology allows subscribers to select the carrier and the POTS number as a function of criteria such as time of day, day
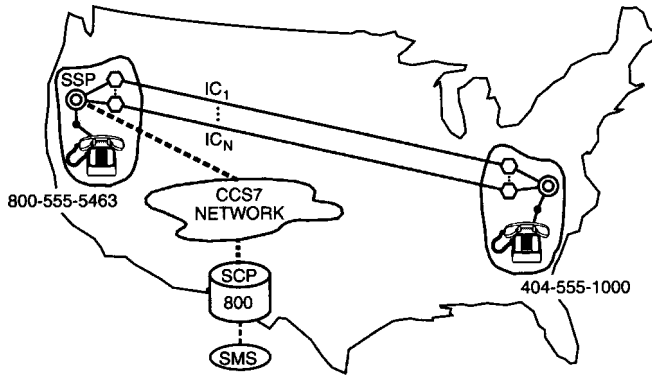
FIGURE 4.12 Data Base 800 Service—800-number calls are routed to an SSP that launches queries through a CCS7 network to an SCP containing the 800 database. In this example, the SCP translates the 800 service number of 800–555–5463 into the POTS number of 404–555–1000. (*Source:* R.B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991. © 1991 IEEE.)

of week, day of year, percent allocation, and the location of the calling station. Thus the SCP provides two customer-specified routing information functions: a carrier identification function and an address translation function.

The SCP 800 Service database is administered by a single national **service management system (SMS).** The SMS is an interactive operations support system that is used to process and update customer records. It is the interface between the customer and the SCP. It translates a language that is friendly to a customer into a language that is friendly to online, real-time databases. Along the way, it validates the customer input.

Demand for toll-free services in the United States became so great that, in 1996, the industry began to introduce additional numbering plan areas (NPAs) to support the service. Indeed, the Intelligent Network systems were enhanced to support a range of toll-free NPAs: 800, 888, 877, 866, 855, 844, 833, and 822.

Alternate Billing Services

Alternate billing services (ABS) have also been implemented using the intelligent network architecture. Alternate billing is an umbrella title that includes calling card service, collect calling, and bill-to-third-number calling. The network configuration supporting ABS is shown in Figure 4.13.
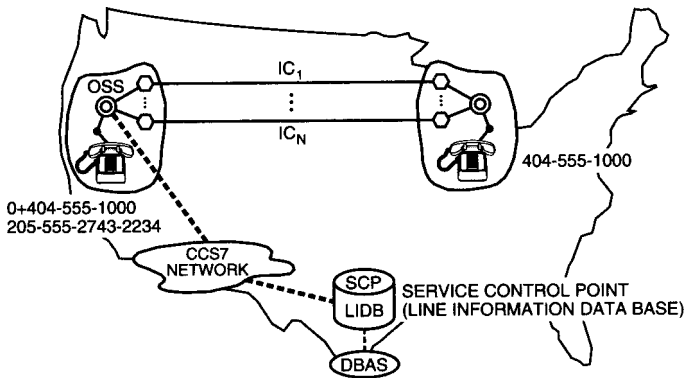

FIGURE 4.13 Alternate billing services—calls are routed to an OSS that launches queries through the CCS7 network to SCPs containing the LIDB application. (*Source:* R.B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991. © 1991 IEEE.)

With this architecture, when a customer places a calling card call, the call is routed to an operator services system (OSS) that suspends call processing and launches a query through a CCS7 signaling network. The query is delivered to an SCP that contains the **line information database (LIDB)** application software. The LIDB application can provide routing information, such as identifying the customer-specified carrier that is to handle the call, as well as provide screening functions, such as the calling card validation used to authorize a call. The LIDB then returns the appropriate information to the OSS so that the call can be completed. The LIDBs are supported by the **database administration system (DBAS),** which is an operations support system that processes updates for calling card service as well as other services. Multiple DBAS systems typically support each LIDB.

During 1991, each of the Regional Bell Operating Companies and a number of large independent telephone companies interconnected their CCS7 networks, mostly through STP hubs, to create a national signaling network; it was a process called LIDB interconnect. When it was finished, it meant that a person carrying a particular company's calling card could, from anywhere in the United States, query the LIDB containing the associated calling card number.

Although the LIDB was originally developed to support calling card service, it has since found wide application in the telecommunications industry. For example, the LIDB is used to translate the telephone number of a calling party to a name as part of calling name delivery service, or to convert that number to a nine-digit ZIP code as part of single number service. The LIDB databases now contain more than a quarter of a billion customer records that are updated at a rate of more than a million changes per day. Although physically distributed, the LIDBs appear logically as a single database. They represent a national resource.

Other Services

For alternate billing services, the SCP is essentially designed to perform two functions: carrier identification and billing authorization. For 800 service, the SCP provides carrier identification and address translation. These basic functions of authorization, address translation, and carrier identification can be used again and again in many different ways. For example, the intelligent network has been used to support private virtual networks (PVNs). PVNs make use of the public telephone network but, by means of software control, appear to have the characteristics of private networks. A PVN serves a closed-user group, and a caller requires authorization to gain access to the network. This screening function on originating calls uses an authorization function. Second, a PVN may offer an abbreviated dialing plan, for example, four-digit dialing. In this instance, the SCP performs an address translation function, converting a four-digit number to a ten-digit POTS number. There may also be a customer-specified routing information function that involves selecting from a hierarchy of facilities; this can be accomplished through use of the SCP carrier identification function.

The SCP in the intelligent network can support a vast number of services ranging from calling name delivery service to messaging service. With calling name delivery, a switch sends a query to the SCP with the ten-digit calling party number; the response is the calling party name that is then forwarded by the switch to a display unit attached to the called party station set. In support of messaging services, the address translation capability of the SCP can be used to translate a person's telephone number to an electronic-mail address. As a result, the sender of electronic mail need only know a person's telephone number.

The Advanced Intelligent Network

The intelligent network architecture discussed thus far is often referred to in the literature as Intelligent Network/1; this architecture has addressed the deployment problem and the service uniformity problem. The next phase in the evolution of this network has come to be called the advanced intelligent network (AIN), with the AIN standards defined by Telcordia Technologies (formerly known as Bellcore).

The concept of AIN is that new services can be developed and introduced into the network without requiring carriers to wait for switch generics to be upgraded. Some AIN applications introduce powerful service-creation capabilities that allow nonprogrammers to invoke basic functions offered in the network and

stitch together those functions, as illustrated in Figure 4.14, to constitute a new service. As a result, AIN promises to dramatically shorten the interval required to develop new services. Perhaps of greater significance, it promises to broaden the participation in service creation. In addition, it offers the opportunity to personalize or customize services. The silicon revolution has driven the cost of memory down to the point where it is economically viable to have enough memory in the network to store the service scripts or call processing scenarios that are unique to individuals.

Many people think of the AIN as a collection of network elements, network systems, and operations systems; this view might be called a technologist's view. Perhaps a better representation is shown in Figure 4.15; it shows a collection of people—people empowered to create services.

Historically, the creation of new services provided by the telephone network has been the sole domain of the network element and network system suppliers. There is perhaps a good analogy with the automobile industry. A market study in the early 1900s predicted that 200,000 was the maximum number of cars that
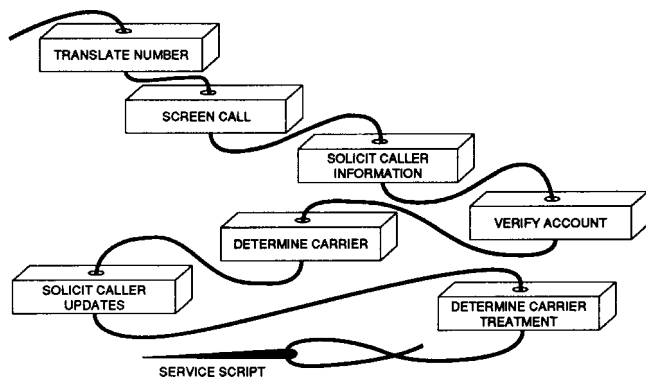
FIGURE 4.14 Creating the service script or scenario for a call by stitching together functional blocks. (*Source:* R.B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991. © 1991 IEEE.)
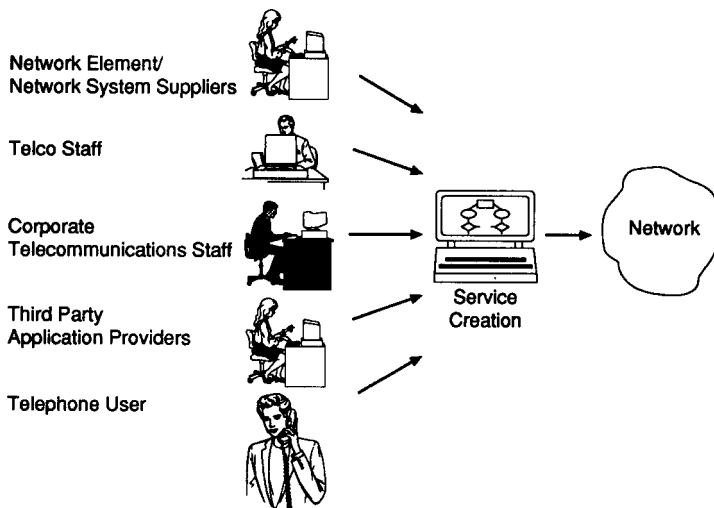
FIGURE 4.15 The advanced intelligent network—a business perspective. (*Source:* R.B. Robrock II, "Putting the Telephone User in the Driver's Seat," International Council for Computer Communication Conference on Intelligent Networks, pp. 144–150, May 1992.)

could ever be sold in a single year in the United States; the reasoning was that 200,000 was the maximum number of chauffeurs that could enter the workforce in a single year. In the telecommunications business, the network element and network system suppliers have been the chauffeurs of the network services business.

The service-creation tools offered by the AIN, however, empower telephone company staff to create new services. Moreover, similar tools may well be used by the telecommunications staff of large corporations, or by third-party application providers or even by some segment of the telephone user population. As a result, we may see an explosion in the number of network services.

The AIN introduces very powerful service-creation tools that are used to produce service-logic scripts (programs). In one arrangement, the service creation is done by assembling service-logic graphs from graphical icons that represent functional components of services. The completed graph is then validated with an expert system and tested off-line by executing every leg of the service-logic graph. At this point the service-logic program can be downloaded into the service control point so that it is ready for execution.

To make use of the new service, it is then necessary to set "triggers" in the appropriate service switching point. These triggers can be set for both originating and terminating calls, and they represent events that, should they occur, indicate the need for the switch to launch a query to the SCP for information the switch needs to process the call.

The first phase of the AIN, called AIN 0, became reality in late 1991 when friendly user trials began in two of the Regional Bell Operating Companies. The AIN 0 call model introduced three trigger check points: the off-hook immediate trigger, the off-hook delayed trigger, and the digit collection and analysis trigger. The release was based on the American National Standards Industry (ANSI) transaction capability application part (TCAP) issue 1, where TCAP is at layer 7 of the SS7 protocol stack.

AIN 0 evolved to AIN 0.1 and then AIN 0.2, with each new version of AIN containing additional triggers. AIN 0.1, based on ANSI TCAP issue 2, separated the formal call model into an originating call model and a terminating call model and introduced three new triggers: the N11 trigger, the 3-6-10-digit trigger, and the termination attempt trigger. AIN 0.2 introduced a new network element, the intelligent peripheral (IP), and supported personal communication service and voice-activated dialing. In the process it introduced busy and no answer triggers.

Today over 100 AIN services—such as single number service, flexible hot line, inbound call restriction, 500 access service, etc.—are deployed in North America, and the number is growing rapidly. Moreover, the U.S. Telecommunications Act of 1996 required local number portability, that is, customer were allowed to keep their telephone numbers if they switched local carriers. The concept soon spread from the wireline network to the wireless or cellular network. The technologies of AIN were the answer.

The concept of AIN is not limited to the United States. The European Telecommunications Standards Institute (ETSI) has defined a European AIN standard referred to as Core INAP, and deployment of Core INAP systems in Europe began in 1996.

The architectural concepts of AIN are now beginning to carry over into wireless networks as well as broadband networks. Although the standards in these domains are just being developed, the value added by the wireless intelligent network (WIN) and the future broadband intelligent network (BIN) promises to surpass the value seen in the narrowband wireline world.

Back to the Future

The intelligent network, with its centralized databases, has offered a means to rapidly introduce new services in a ubiquitous fashion and with operational uniformity as seen by the end user. The advanced intelligent network has gone on to provide a service-independent architecture, and, with its powerful service-creation capabilities, has empowered nonprogrammers to participate in the development of new services. In many ways, as we go into the future, we are going back to a time when operators were the "human intelligence" in the network. The human intelligence was all but eliminated with the introduction of switching systems, but now the intelligent network is working to put the intelligence of the human operator back into the network.

Defining Terms

Common-channel signaling (CCS):   A technique for routing signaling information through a packet-switched network.

Database administration systems (DBAS):   An operations support system that administers updates for the line information database.

Line information database (LIDB):   An application running on the service control point that contains information on telephone lines and calling cards.

Service control point (SCP):   An on-line, real-time, fault-tolerant, transaction-processing database that provides call-handling information in response to network queries.

Service management system (SMS):   An operations support system that administers customer records for the service control point.

Signal transfer point (STP):   A packet switch found in the common-channel signaling network; it is used to route signaling messages between network access nodes such as switches and SCPs.

Signaling system 7 (SS7):   A communications protocol used in common-channel signaling networks.

References

AT&T Bell Laboratories, "Common channel signaling," *The Bell System Tech. J.,* vol. 57, no. 2, pp. 221–477, February 1978.

AT&T Bell Laboratories, "Stored program controlled network," *The Bell System Tech. J.,* vol. 61, no. 7, part 3, pp. 1573–1815, September 1982.

Bell Communications Research, "Advanced intelligent network (AIN) 0.1 switch-service control point (SCP) application protocol interface generic requirements," *Bell Commun. Res. Technical Ref.,* TR-NWT-001285, Issue 1, August 1992.

European Telecommunications Standards Institute, "Intelligent network (IN): Intelligent network capability set 1 (CS1) core intelligent network applications protocol (INAP) part 1: Protocol specification," *Eur. Telecom. Stds. Inst.,* ETS 300 374–1, draft, May 1994.

Globecom '86: The Global Telecommunications Conference, *Conference Record,* vol. 3, pp. 1311– 1335, December 1986.

R.J. Hass and R.W. Humes, "Intelligent network/2: A network architecture concept for the 1990s," International Switching Symposium, *Conference Record,* vol. 4, pp. 944–951, March 1987.

R.B. Robrock, II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991.

R.B. Robrock, II, "Putting the telephone user in the driver's seat," International Council for Computer Communication Intelligent Networks Conference, pp. 144–150, May 1992.

R.B. Robrock, II, "The many faces of the LIDB data base," International Conference on Communications, *Conference Record,* June 1992.

Telcordia Technologies, "Advanced intelligent network (AIN) switch-service control point (SCP)/Adjunct interface generic requirements," *Telcordia Technologies, Generic Requirements,* GR-1299-CORE, Issue 9, November 2003.

Further Information

The magazine *Bellcore Exchange* contained numerous articles on the intelligent network, particularly in the following issues: July/August 1986, November/December 1987, July/August 1988, and March/April 1989. Articles on AIN service creation appeared in the January/February 1992 issue.

   The monthly publication *IEEE Communications Magazine* contains numerous articles on the intelligent network. A special issue on the subject was published in January 1992. Copies are available from the IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854–4150.

   The monthly publication *The Telcordia Technologies Digest* lists recent publications available from Telcordia Technologies. There are a series of technical advisories, technical requirements, generic requirements, and

special reports that have been issued on the intelligent network. Copies are available by contacting Telcordia Customer Service toll-free 1–800–521-CORE (2673).

The bimonthly publication *The AT&T Technical Journal* contains numerous articles on the intelligent network. The advanced intelligent network is the subject of a special issue: Summer 1991, vol. 70, nos. 3–4. Current or recent issues may be obtained from the AT&T Customer Information Center, P.O. Box 19901, Indianapolis, IN 46219.

4.4   Mobile Internet

Apostolis K. Salkintzis and Nikos Passas

Introduction

The mobile Internet can be considered as the migration of the standard Internet applications and services to the mobile environment. The introduction of mobility itself raises a number of concerns and challenges. For instance, what wireless technology is most appropriate for the provision of Internet services? Is this technology equally appropriate for applications with dissimilar requirements such as e-mail and video broadcasting? How do we provide *ubiquitous* Internet services while users move across different locations in which the same wireless service may not be available? Will it be more appropriate to consider several wireless technologies such as cellular data networks, wireless local area networks (WLANs), wireless personal area networks (WPANs)? If yes, then how do we combine them into a seamless wireless service? How do we optimize the utilization of wireless resources in order to accommodate as many mobile Internet users as possible? How do we handle the security issues raised by wireless transmission and possibly by the use of different wireless services provided by different operators?

These are only a small number of the questions we need to address in our attempt to make the mobile Internet a reality. One important clarification is in order here: The fact that we can now use our laptop or personal digital assistant (PDA) along with a wireless device (e.g., cellular phone, WLAN adapter, etc.) to have access to our Internet/intranet, e-mail, or other Internet Protocol (IP) service does not really mean the mobile Internet is already available. In reality, what is defined as mobile Internet is far more complex than that. By definition, the Internet is a network of millions of users who communicate by means of standard IP protocols. Therefore, in the mobile Internet we need also to support millions of *always-connected* mobile/ wireless users. The keywords we need to bear in mind are "always-connected" and "millions of users." To be always-connected means that we do not have to make a connection before every transaction or after we change wireless service. Our assumption is that we always have connectivity to public Internet services and we are always reachable through a public IP address, no matter where we are. This calls for extensive mobility management (discussed in section "Mobility Management Schemes for Mobile Internet") and seamless handover across different wireless networks. We do need several wireless networks (and several wireless technologies, as we will see later on) because no sole wireless network can provide ubiquitous wireless services and hence no sole wireless network can meet the always-connected requirement. A typical mobile Internet user is assumed to move *seamlessly* between different wireless networks (or even between fixed and wireless networks), which may or may not employ the same radio-access technology. This is schematically illustrated in Figure 4.16. The seamless fashion of movement suggests that our (virtual) connection to the public Internet is transferred from one access network to the other without any actions on the user's side. In effect, this creates a virtual wireless network from the user's perspective that provides ubiquitous Internet connectivity.

We need also to support "millions of users" wirelessly connected to the Internet in order to be compliant with the large scale of users supported by the Internet. This creates capacity concerns and further justifies why no single wireless network is sufficient. The capacity concerns have direct implications in several design aspects. For instance, the wireless technology needs to be as spectrum-efficient as possible; it has to implement a random access scheme that can accommodate a large number of users and degrade gracefully in overload
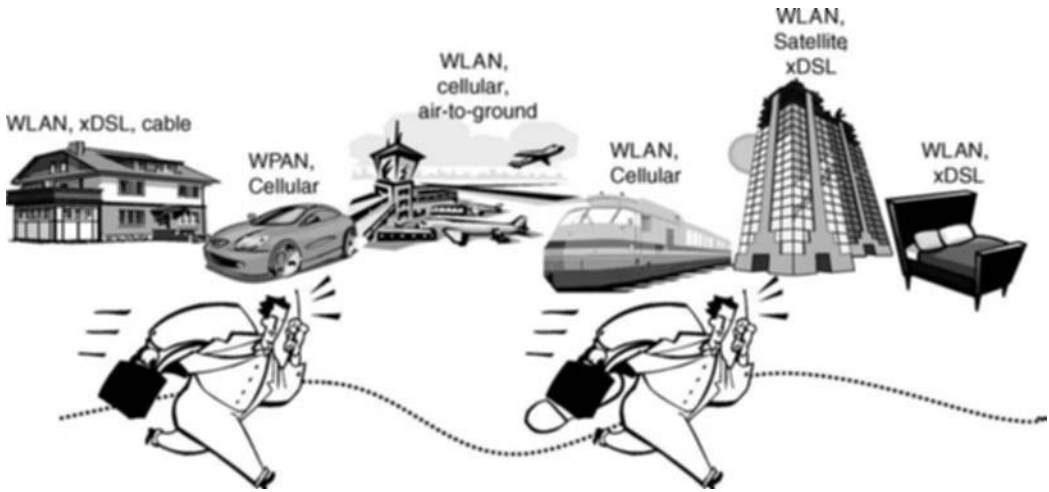
FIGURE 4.16   The concept of always-connected.

conditions. The IP addressing scheme should be able to support the required user capacity, the end-to-end transport schemes, and applications should probably take into account mobility.

It is interesting to note that most recent research projects, as well as standardization activities, are moving around such goals. In this context, the correlation between the mobile Internet and the so-called "beyond 3G" technologies is evident. In reality, most of the beyond 3G technologies are tailored to support the requirements for mobile Internet: increased capacity, QoS, mobility, security, TCP/IP enhanced performance, and integration of diverse technologies into one ubiquitous virtual network.

Despite the fact that herein we focus entirely on wireless networks as potential access means for the mobile Internet, it is important to note that fixed access networks and their associated technologies (e.g., xDSL, cable modems, etc.) do play an important role. In fact, the vision towards the mobile Internet entails both wireless and fixed access technologies, as well as methods to seamlessly integrate them into an IP-based core network (see Figure 4.17). A typical user scenario, which shows the fundamental interworking requirements between fixed and wireless access networks, occurs when a user downloads a file over the Internet using a cable modem, and in the middle of the file transfer takes his laptop to the car and drives to the airport (see Figure 4.16). In a mobile Internet environment, the file download would be seamlessly switched from the cable connection to a wireless connection, e.g., a cellular data connection and would be carried on while on the go.

In our mobile Internet environment, as defined above, it is evident that enhanced mobility is a key requirement, thus we discuss it in more detail in section "Mobility Management Schemes for Mobile Internet."

Key Aspects of the Evolution toward the Mobile Internet

Bearing in mind the above discussion, we list below the key aspects of the evolution towards the mobile Internet. Several organizations worldwide such as 3GPP, 3GPP2, IETF, and IEEE develop activities relevant to these aspects [1].

- Mobile networks will evolve to an architecture encompassing an **IP-based core network** and many wireless access networks. The key aspect in this architecture is that signaling with the core network is based on IP protocols (more correctly, on protocols developed by IETF) and it is independent of the access network (be it UMTS, cdma2000, WLAN, etc.). Therefore, the same IP-based services could be accessed over any access network. An IP-based core network uses IP-based protocols for *all* purposes including data transport, networking, application-level signaling, mobility, etc. The first commercial
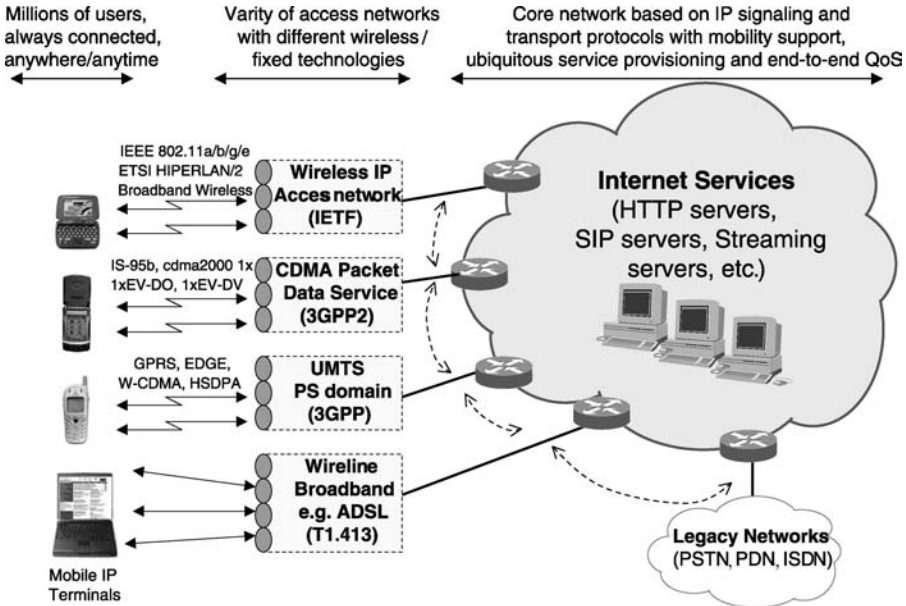
FIGURE 4.17 The architecture of mobile Internet is tightly coupled with the "beyond-3G" network architecture.

approach towards this IP-based core network is the well-known *IP Multimedia Core Network Subsystem* (IMS) standardized by 3GPP and 3GPP2. IMS is further discussed in Ref. [1] and its relevant specifications can be found at www.3gpp.org/ftp/specs/.

- The long-term trend is towards **all-IP mobile networks** where not only the core network, but also the radio access network is solely based on IP technology. In this approach, the base stations in a cellular system are IP access routers and mobility/session management is carried out with IP-based protocols (possibly substituting the cellular-specific mobility/session management protocols).
- Enhanced **IP Multimedia/Internet applications** will be enabled by means of application-level signaling protocols standardized by IETF (e.g., SIP, HTTP, etc.). This again is addressed by the 3GPP/3GPP2 IP Multimedia Core Network Subsystem.
- **End-to-end QoS** provisioning will be important for supporting the demanding multimedia/Internet applications. In this context extended interworking between, e.g., UMTS, QoS, and IP QoS schemes is needed, or more generally, interworking between layer-2 QoS schemes and IP QoS is required for end-to-end QoS provision.
- **Voice over IP** (VoIP) will be a key technology. Several standards organizations are developing activities to enable VoIP, e.g., the ETSI TISPAN project (http://portal.etsi.org/tispan), the IETF SIP Working Group (http://www.softarmor.com/sipwg/), etc.
- The mobile terminals will be based on **software-configurable radios** with capability to support many radio access technologies across many frequency bands.
- The ability to move across hybrid access technologies will be an important requirement, which calls for efficient and fast vertical handovers and **seamless session mobility**. The IETF SEAMOBY (http://www.ietf.org/html.charters/seamoby-charter.html) and the legacy MOBILE-IP (http://www.ietf.org/html.charters/OLD/ mobileip-charter.html) working groups have addressed some of the issues related to seamless mobility. Fast Mobile IP and Micro-mobility schemes (see section "Mobility Management Schemes for Mobile Internet") are key technologies in this area.
- In the highly hybrid access environment of the mobile Internet **security** will also play a key role. IEEE 802.11 task group I (TGi) is standardizing new mechanisms for enhanced security in WLANs. The IETF SEAMOBY WG also addresses the protocols that deal with (security) context transfer during handovers.
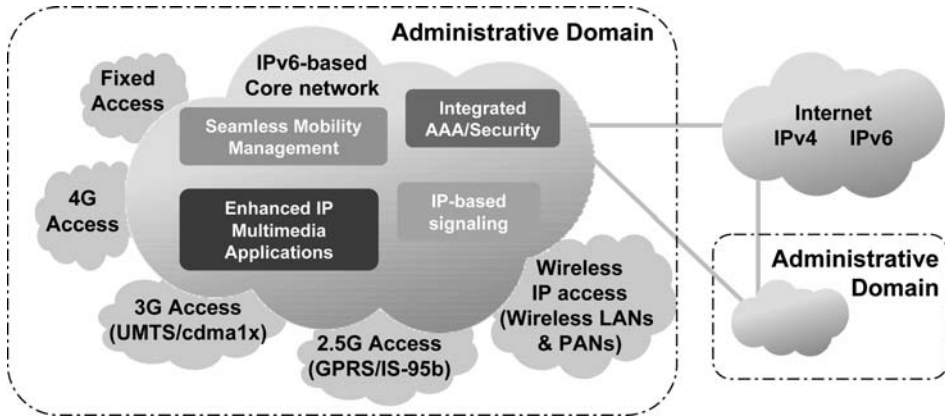
FIGURE 4.18 Simplified architecture of mobile Internet.

- For extended roaming between different administrative domains and/or different access technologies, **advanced AAA protocols** and **AAA interworking mechanisms** will be required. AAA interworking between WLANs and 3GPP networks is being studied by 3GPP and 3GPP2 [1–5].
- Enhanced **networking APIs** for QoS-, multicast-, and location-aware applications will be needed.
- **Wireless Personal Area Networks** (WPANs) will start spreading, initially based on Bluetooth technology (see www.bluetooth.com) and later on IEEE 802.15.3 high-speed wireless PAN technology, which satisfies the requirement of the digital consumer electronic market (e.g., wireless video communications between a PC and a video camera).
- Millions of users are envisioned being "always-connected" to the IP core infrastructure. Market forecasts suggest that about 1 billion addresses will be needed by 2005 for the integration of Internet-based systems into transportation means (cars, aircraft, trains, ships, and freight transport) and associated infrastructures for mobile e-commerce. This indicates that there is a strong **need for IPv6**—IPv6 is being adapted in 3GPP. The European Union, in particular, is pushing hard for the fast adoption of IPv6.
- Wireless communication technology will evolve further and will support higher bit rates. For instance, WLANs will soon support **bit rates of more than 100 Mbps**. This is being addressed by the IEEE Wireless Next Generation Standing Committee (see www.ieee802.org/11). Higher bit rates are typically accompanied by smaller coverage areas, and therefore efficient, fast, and secure horizontal handovers will be required.

The above evolutionary aspects lead to a high-level network architecture as the one depicted in Figure 4.18. Note that each administrative domain represents a network part that is typically operated by a single mobile network operator.

Evolution to IP-based Core Networks

Without doubt the most widely supported evolution towards the mobile Internet is the evolution to IP-based core networks, also referred to as *all-IP* core networks. The term "all-IP" emphasizes the fact that IP-based protocols are used for *all* purposes including transport, application-level signaling, mobility, security, QoS, etc. Typically, several wireless and fixed access networks are connected to an all-IP core network, as illustrated in Figure 4.17 and Figure 4.18. Users will be able to use multimedia applications over terminals with software-configurable radios capable of supporting a vast range of radio access technologies such as WLANs, WPANs, UMTS, and cdma2000. In this environment seamless mobility across the different access technologies is considered as a key issue by many vendors and operators.
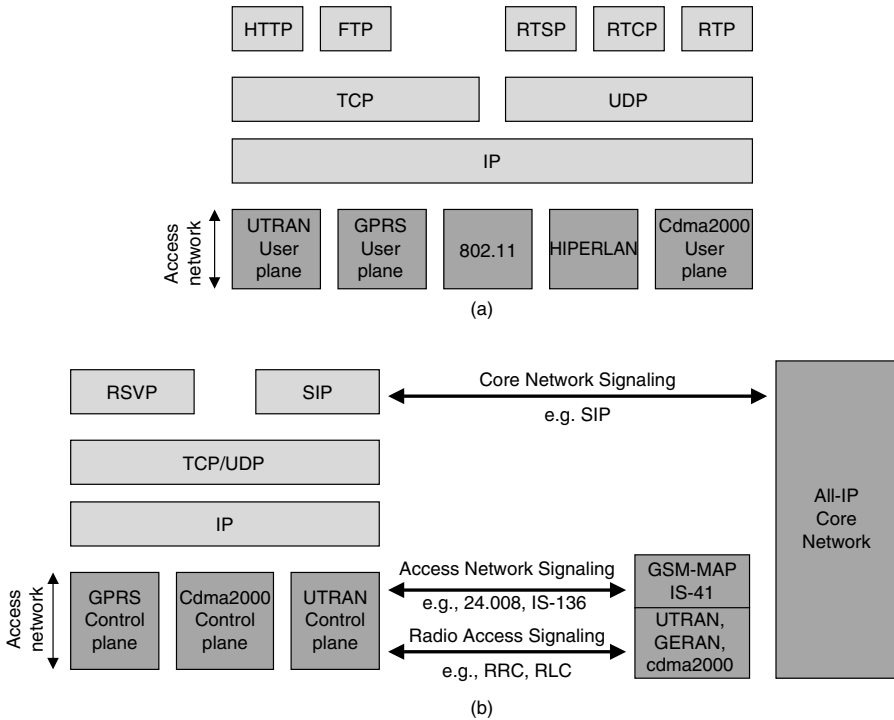
FIGURE 4.19   Simplified protocol architecture in an all-IP network architecture; (a) Control Plane, (b) User Plane.

In the all-IP network architecture the mobile terminals use the IP-based protocols defined by IETF to communicate with the core network and perform, e.g., session/call control and traffic routing. All services in this architecture are provided on top of the IP protocol. As shown in the protocol architecture of Figure 4.19, the mobile networks (such as UMTS, cdma2000, etc.) turn into access networks that provide only mobile bearer services. The teleservices in these networks (e.g., cellular voice) are only used to support the legacy 2G and 3G terminals, which do not support IP-based applications (e.g. IP telephony). For the provision of mobile bearer services, the access networks mainly implement micromobility management, radio resource management, and traffic management for provisioning of quality of service. Micromobility management in 3GPP access networks is based on GPRS Tunneling Protocol (GTP), and uses a hierarchical tunneling scheme for data forwarding. On the other hand, micromobility management in 3GPP2 access networks is typically based on IP micromobility protocols. Macromobility (inter-domain mobility) is typically based on Mobile-IP, as specified in RFC 3220 (see http://www.ietf.org/rfc/rfc3220.txt). All these mobility schemes are discussed in more detail in the section on "Mobility Management Schemes for Mobile Internet."

In the short term the all-IP core network architecture would be based on the IMS architecture specified by 3GPP/3GPP2, which in turn is based on the IP multimedia architecture and protocols specified by IETF. This IMS architecture would provide a new communications paradigm based on integrated voice, video, and data. You could call a user's IMS number, for instance, and be redirected to his Web page where you could have several options, e.g., write a message for him, record a voice message, click on an alternative number to call if he is on vacation, etc. You could place a Session Initiation Protocol (SIP) call to a server and update your communication preferences, e.g., "only my manager can call me, all others are redirected to my Web page" (or vice versa!). At the same time, you could be on a conference call.

IP-based Core Networks in the Enterprise

Figure 4.20 shows how an enterprise could take advantage of an all-IP core network (which is part of a mobile Internet environment) in order to minimize its communication costs and increase communications efficiency.
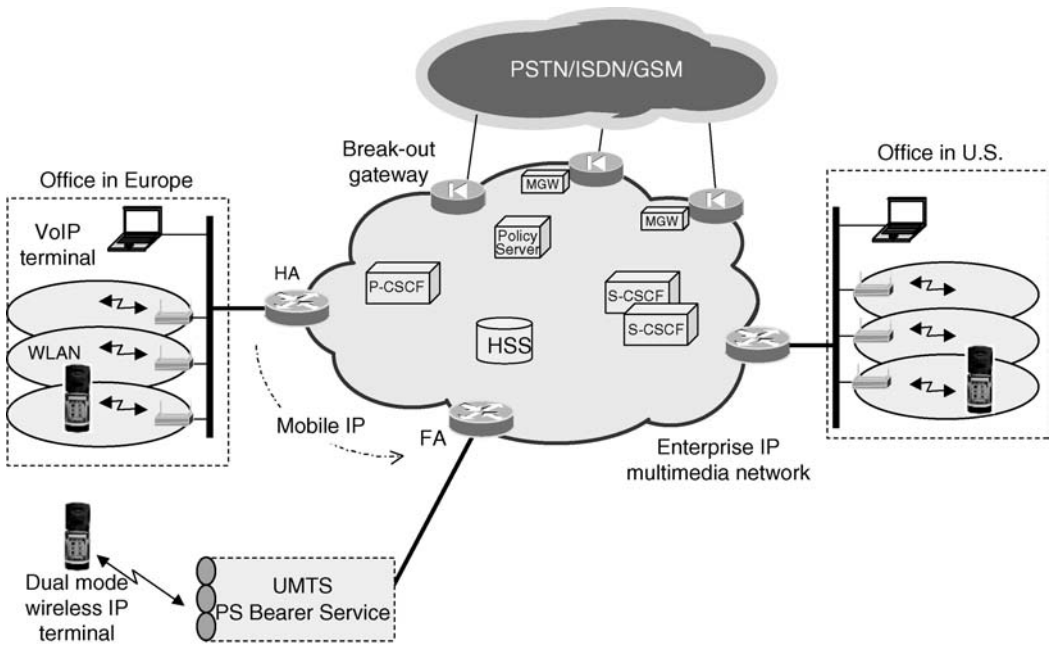
FIGURE 4.20   Deployment of all-IP networks in the enterprise.

The typical IP network of the enterprise could be evolved to an IP multimedia network, which would support (i) IP signaling with the end terminals for establishing and controlling multimedia sessions, (ii) provisioning of QoS, (iii) policy-based admission control, and (iv) authentication, authorization, and possibly accounting. The all-IP network provides an integrated infrastructure for efficiently supporting a vast range of applications with diverse QoS requirements, and, in addition, provides robust security mechanisms. The architecture of this all-IP network could be based on the IMS architecture specified by 3GPP/3GPP2 (see an up-to-date 3GPP TS 23.228 specification at www.3gpp.org/ftp/specs/ for a detailed description).

In the example shown in Figure 4.20, an employee in the European office could request a voice call to another employee, e.g., in the U.S. office. This request would be routed to the default Proxy-Call Session Control Function (P-CSCF) that serves the European office. This P-CSCF relays the request to the Serving CSCF (S-CSCF) of the calling employee, i.e., to the CSCF that this employee has previously registered. This S-CSCF holds subscription information of the calling employee and can verify whether he/she is allowed to place the requested call. In turn, the S-CSCF finds another S-CSCF with which the called subscriber has registered and relays the request to this S-CSCF. Note that if the calling employee in the European office were calling a normal telephone (PSTN) number in the United States, the call would be routed through the IP network to a break-out gateway closest to the called PSTN number. This way, the long-distance charges are saved.

The S-CSCF of the called U.S. employee holds information on the employee's whereabouts and can route the call to the correct location. In case the called U.S. employee happens to be roaming in Europe, the call would be routed to the appropriate P-CSCF, which currently serves this employee. It is important to note that, although signaling can travel a long path (e.g., from Europe to the United States and then back to Europe), the user-plane path would be the shortest possible.

The support of roaming is another important advantage of the above architecture. For instance, a European employee could take his dual-mode mobile device to the U.S. office, and after powering-on his device and registering his current IP address (with his S-CSCF), he would be able to receive multimedia calls at his standard number.

Even when the employee is away from his office and cannot directly attach to the enterprise network (e.g., he is driving on the highway), he can still be reached on his standard number. In this case the employee uses, for example, the UMTS network to establish a mobile signaling channel to his all-IP enterprise network, which assigns him an IPv6 address. The employee registers this IP address with an S-CSCF (via the appropriate P-CSCF), and thereafter he can receive calls at his standard number. The signaling mobile channel remains activated for as long as the employee uses UMTS to access the enterprise network. In such a scenario, the UMTS network is used only to provide access to the enterprise network and to support the mobile bearers required for IP multimedia services. To establish IP multimedia calls the employee would need to request the appropriate UMTS bearers, each one with the appropriate QoS properties. For instance, to receive an audio-video call, two additional UMTS bearers would be required, one for each media component. The mapping between the application-level QoS and the UMTS QoS, as well as the procedures required to establish the appropriate UMTS bearers are specified in 3GPP Rel-5 specifications, specifically in 3GPP TS 24.008 and 3GPP TS 24.229 (available at www.3gpp.org/ftp/specs/).

If the enterprise network supports a macromobility protocol, e.g., Mobile-IP (see section "Mobility Management Schemes for Mobile Internet"), it could be possible to provide session mobility across the enterprise WLAN and the UMTS network. In this case when the employee moves from the WLAN to the UMTS he uses Mobile-IP to register his new IPv6 address with his home agent. After that any subsequent terminating traffic would be tunneled from the employee's home agent to the foreign agent that serves the employee over the UMTS network.

Another roaming scenario is illustrated in Figure 4.21, which involves interworking between the enterprise all-IP network and the mobile operator's all-IP network. In this scenario, the key aspect is that the employee uses a P-CSCF in the mobile operator's domain, i.e., it uses the operator's IMS system. It is noted that this scenario corresponds to the roaming scenario considered in 3GPP Rel-5 specifications, and therefore it might be the most frequently used in practice.
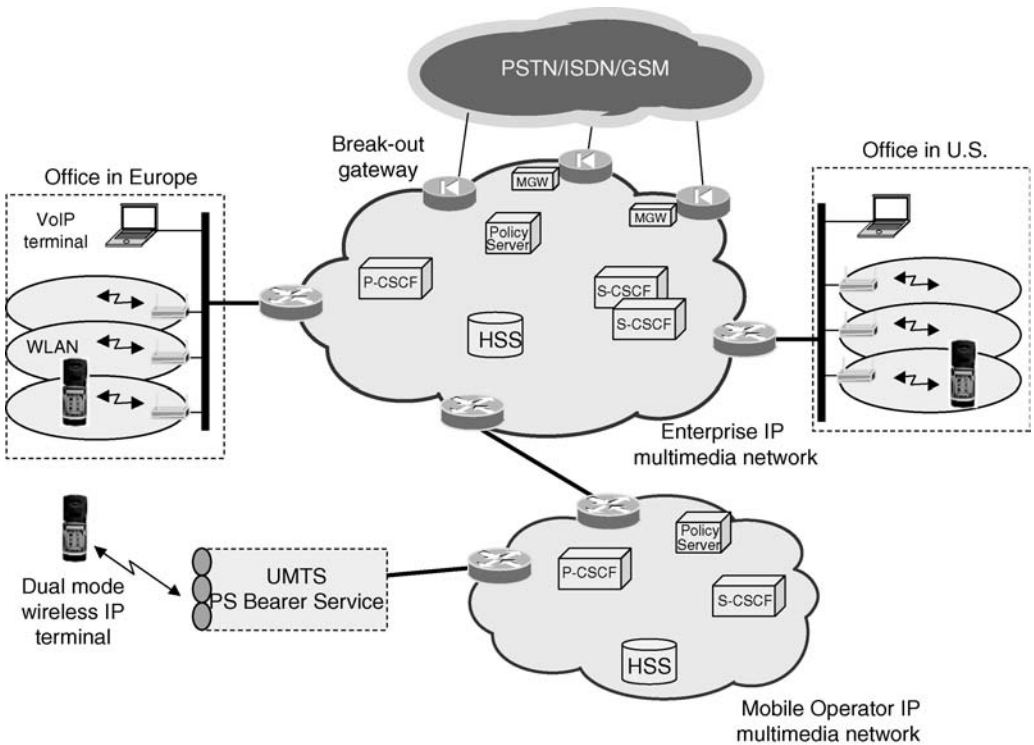

FIGURE 4.21   Deployment of all-IP networks in the enterprise (mobile operator uses IMS).

Ubiquitous Mobile Internet: Combining Cellular and Wireless LAN Networks

The interworking between 3G cellular and Wireless Local Area Networks (WLANs) has been considered as a suitable and viable evolution path towards the next generation of wireless networks, which would be capable of providing a nearly ubiquitous mobile Internet service. A 3G cellular network can provide relatively low-speed (up to 384 kbps per user) Internet services, but over a large coverage area. On the other hand, a WLAN can typically provide high-speed Internet services (with an effective throughput of tens of Mbps), but over a geographically small area. An integrated 3G/WLAN network combines the strengths of each, resulting in a wide-area wireless system capable of providing users with ubiquitous Internet services, which range from low-speed to high-speed in strategic locations.

However, the 3G/WLAN interworking raises considerable challenges, especially when we demand seamless continuity of Internet sessions across the two networks. To deal with these challenges, several 3G/WLAN interworking requirements need to be identified and fulfilled.

Typically the 3G/WLAN interworking requirements are specified and categorized in terms of several usage scenarios [3,4]. For example, a common usage scenario is when a 3G subscriber is admitted to a WLAN environment by reusing his regular 3G credentials, and then obtains an IP connectivity service (e.g., access to the Internet). In this case, the interworking requirements include support of 3G-based access control, signaling between the WLAN and the 3G network for authentication, authorization, and accounting (AAA) purposes, etc. Other scenarios can call for more demanding interworking requirements. We may envision, for instance, a scenario in which a 3G subscriber initiates a video session in his home 3G network and subsequently transits to a WLAN environment, wherein the video session is continued *seamlessly*, i.e., without any noticeable change to the quality of service (QoS). In this case, not only 3G-based access control is required, but also access to 3G-based services is needed over the WLAN network, which in turn calls for appropriate routing enforcement mechanisms. More importantly, however, there is need for QoS consistency across 3G and WLAN, which appears to be not very straightforward given the different QoS features offered by these networks. Indeed, WLANs have initially been specified without paying much attention to QoS aspects and aimed primarily at simple and cost-effective designs. Even with the recent IEEE 802.11e developments [10,11], WLAN QoS still exhibits several deficiencies with respect to the 3G QoS. In contrast, 3G cellular networks were built with the multimedia/Internet applications, trade simplicity, and cost in mind for inherently providing enhanced QoS in wide-area environments.

In the rest of this section we examine some aspects of the seamless continuity of Internet services across UMTS and WLAN. In this context we address several issues such as routing enforcement, access control, differentiation between the traffic of regular WLAN data users and UMTS roamers, etc. The framework for this discussion is the consideration of a practical UMTS/WLAN interworking architecture, which conforms to the 3GPP specifications [4,5] and other interworking proposals found in the technical literature [3].

UMTS/WLAN Interworking Architecture

The end-to-end interworking architecture we are considering is illustrated in Figure 4.22 and it is compliant with the proposals [3,4]. Below we briefly discuss the main characteristics of this architecture. Note that our goal is not to provide a comprehensive description, but rather to define the key aspects of a practical environment that can enable ubiquitous mobile Internet services. For more detailed information on the considered architecture the interested reader is referred to Ref. [3] and Ref. [4].

As shown in Figure 4.22, the 3G network supports access (via the Internet) to a variety of IP multimedia services over two radio access technologies: UMTS Terrestrial Radio Access (UTRA) and WLAN access. Access control and traffic routing for 3G subscribers in UTRA is entirely handled by the UMTS Packet-Switched (PS) network elements, which encompass the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN) [6]. On the other hand, access control and traffic routing for 3G subscribers in WLAN (*UMTS roamers*) is shared among the WLAN and the UMTS network elements as discussed below. The important assumption we make, as shown in Figure 4.22, is that 3G subscribers can change radio access technology and keep using their ongoing multimedia/Internet sessions in a seamless fashion. Thus, we assume that *seamless Internet service continuity* is provided.
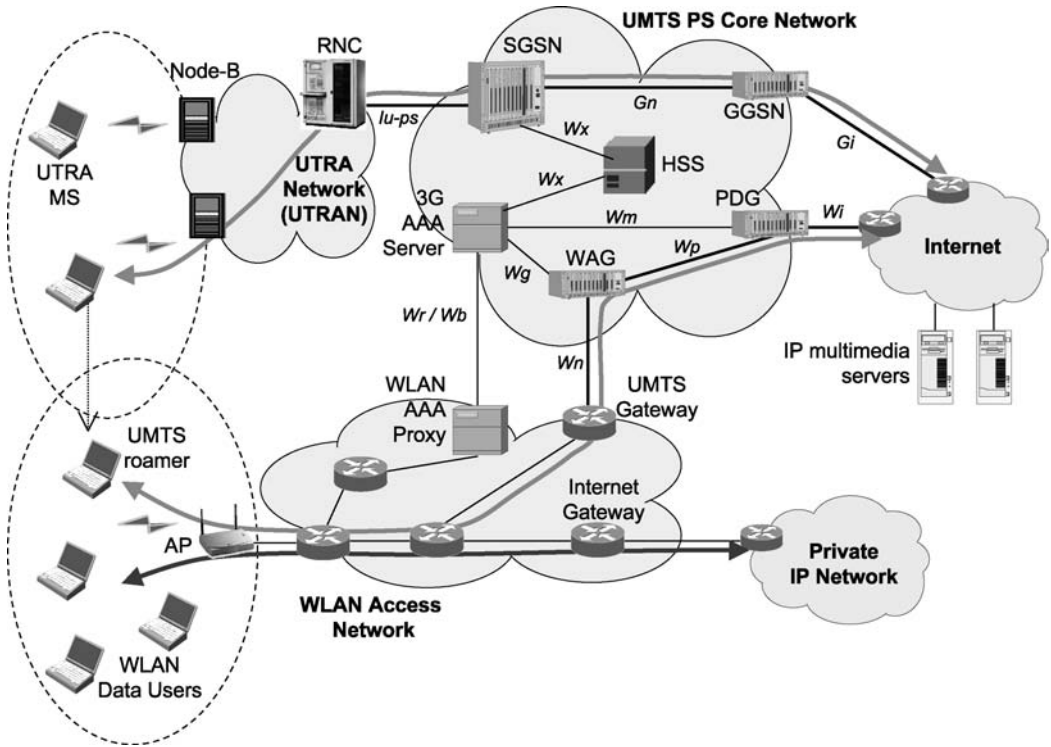
FIGURE 4.22   The considered end-to-end interworking architecture for seamless multimedia session continuity.

The WLAN access network may be owned either by the UMTS operator or by any other party (e.g., a public WLAN operator or an airport authority), in which case the interworking is enabled and governed by appropriate business and roaming agreements. As shown in Figure 4.22 in a typical deployment scenario the WLAN network supports various user classes, e.g., UMTS roamers and regular WLAN data users (i.e., no 3G subscribers). Differentiation between these user classes and enforcement of corresponding policies is typically enabled by employing several *Service Set Identifiers* (SSIDs) [7]. For example, the regular WLAN data users may associate with the SSID that is periodically broadcast by the Access Point (AP) (denoted as SSID(b)), whereas the UMTS roamers may associate with another SSID that is also configured in the AP, but not broadcast (denoted as SSID[g]). In this case, the WLAN can apply distinct access control and routing policies for the two user classes, and can forward the traffic of WLAN data users, e.g., to the Internet and the traffic of UMTS roamers to the UMTS PS core network (as shown in Figure 4.22). Such routing enforcement is vital for supporting seamless service continuity and can be implemented as discussed [3]. Moreover, different AAA mechanisms could be used for the different user classes.

For enabling interworking with WLANs, the UMTS PS core network incorporates three new functional elements: the *3G AAA Server*, the *WLAN Access Gateway* (WAG), and the *Packet Data Gateway* (PDG). The WLAN needs also to support similar interworking functionality to meet access control and routing enforcement requirements. The 3G AAA Server in the UMTS domain terminates all AAA signaling originated in the WLAN that pertains to UMTS roamers. This signaling is securely transferred across the *Wr/Wb* interface, which is typically based on *Radius* [8] or *Diameter* [9] protocols. The 3G AAA Server interfaces with other 3G components such as the WAG, the PDG, and Home Subscriber Server (HSS), which stores information defining the subscription profiles of 3G subscribers. The 3G AAA Server can also route AAA signaling to/from another 3G networks, in which case it serves as a proxy, and it is referred to as *3G AAA Proxy* [3].

As shown in Figure 4.22, traffic from UMTS roamers is routed to the WAG across the *Wn* interface and finally to the PDG across the *Wp* interface. This routing is enforced by establishing appropriate traffic tunnels after a successful access control procedure. The PDG functions much like a GGSN in a UMTS PS core network. It routes the user data traffic between the mobile station (MS) and an external *Packet Data Network* (PDN) (in our case, the Internet), and serves as an anchor point that hides the mobility of the MS within the WLAN domain. The WAG functions mainly as a route policy element, i.e., ensures that user data traffic from authorized MSs is routed to the appropriate PDGs, located either in the same, or in a foreign UMTS network.

Seamless Internet Session Handover

Although Figure 4.22 shows the architecture that can support seamless Internet session continuity, it does not address the dynamics of handover procedure, which is especially important for the provision of seamless continuity. To further elaborate on this key procedure, we depict in Figure 4.23 a typical signaling diagram that pertains to a situation where an MS hands over from UMTS to WLAN in the middle of an ongoing Internet session (e.g., a VoIP session). The establishment of the Internet session is triggered at instant *A* and in response the MS starts the Packet Data Protocol (PDP) context establishment procedure for requesting the appropriate QoS resources (described by the "Req. QoS" Information Element [IE], [12]). The UMTS network acknowledges the request and indicates the negotiated QoS resources (specified by the "Neg. QoS" IE) that could be provided. After that, IP traffic on the user plane commences and the Internet session gets in progress. At some point the MS enters a WLAN coverage area and it starts receiving
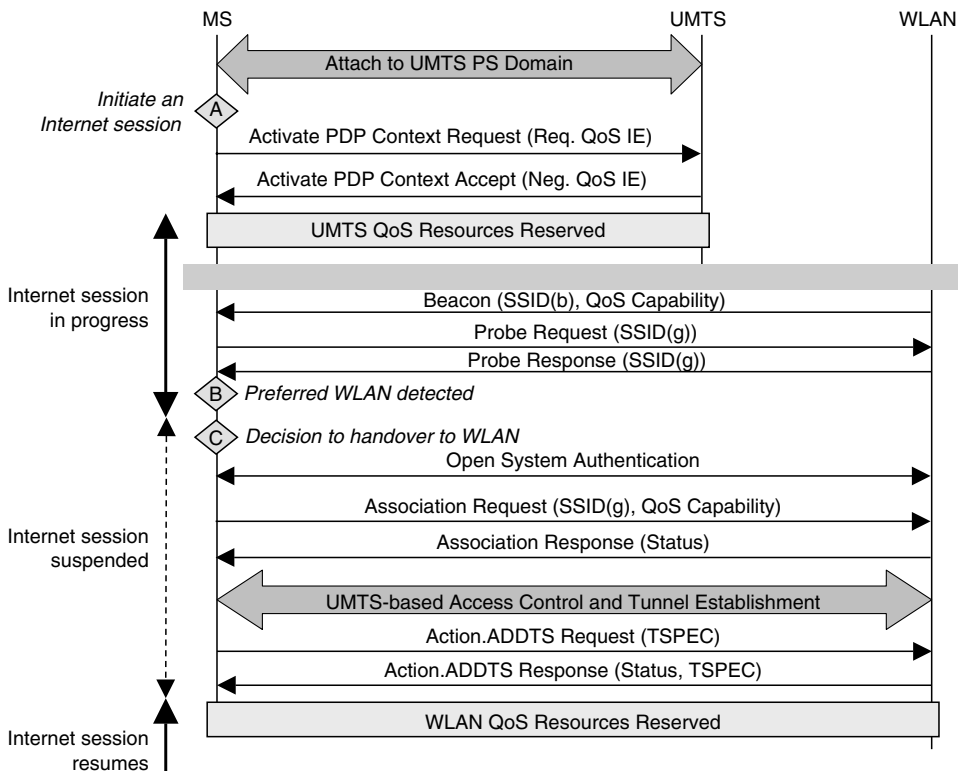

FIGURE 4.23 Typical signaling during handover of an Internet session from UMTS to WLAN (HCCA availability is assumed).

Beacons[1] from the nearby Access Points (APs). We assume that this can happen concurrently with the ongoing Internet session because, although the MS has one transceiver available, it can periodically decode signals on other frequency channels for inter-system handover purposes. The MS may need to check if the detected WLAN supports one of its preferred SSIDs before considering it valid for inter-system change. For this purpose, the MS *probes* for a preferred SSID, denoted as SSID(g), according to the applicable procedures in Ref. [7].

At instant *C* the MS takes the decision to handover to the detected WLAN and thus suspends the ongoing Internet session. This may demand further signaling with the UMTS, but we omit this for simplicity. After switching to the WLAN channel, the normal 802.11 authentication and association procedures [7] are carried out. Subsequently, the UMTS-based access control procedure is executed in which the MS is authenticated and authorized by means of its regular 3G credentials [3,4]. At this stage, a tunnel will also be established for routing further IP traffic from the MS to a UMTS entry point (the WAG according to Figure 4.22). Next, the MS uses 802.11e [10] QoS signaling (assuming it is supported by the WLAN) to reserve the appropriate resources for its suspended Internet session. The *Traffic Specification* (TSPEC) element carries a specification of the requested QoS resources. For the objectives of seamless continuity, it is apparent that TSPEC needs to be set consistently with the QoS negotiated in the UMTS system. After this point, the Internet session is finally resumed in the WLAN, possibly after some high-layer mobility management procedures (e.g., Mobile IP or SIP).

From the above discussion it becomes evident that vertical handovers from UMTS to WLAN (and vice versa) present several challenges, especially for minimizing the associated latencies and the interruption of ongoing Internet sessions. Apart from that, however, the maintenance of consistent QoS across the UMTS and WLAN networks is equally challenging.

Mobility Management Schemes for Mobile Internet

The main aim of mobility management across different access networks is to ensure that the user is "always-connected," or better yet "always-best-connected" (ABC). ABC means that the network offers a set of access technologies and mobility management mechanisms that allow the users to be connected with the most appropriate available technology at all times in order to enjoy the best possible service. "Best" is usually defined separately for each user as part of his/her profile, and it can be a function of service quality, cost, terminal capabilities, personal preferences, etc. In any case, the network should have the flexibility to adjust the access technology, and activate the appropriate mobility management mechanisms in order to be consistent with the user's profile. This should be performed with no or minimum intervention of the user, leading to what is called "invisible network." Consequently, a set of available access technologies and mechanisms should be integrated in a single architecture supporting multiple services, adjustments at all layers, and vertical handover capabilities between different technologies [13].

The ABC concept contains the idea of ubiquitous connectivity at any time and any place. To achieve this goal the underlying assumption is that the user that is always-connected is not hindered in geographic or motion restrictions. The user can move either by foot or through other means (car, train, ship, etc.) and still maintain the best level of connectivity possible. Mobility support is inherent for any new ABC architecture. The first level of mobility support focuses on the infrastructure design. The mobile access networks deployed consist of geographically dispersed base stations connected in a hierarchical fashion that allows the mobile device to connect successively to neighboring base stations as it moves. This is the model that all current cellular networks employ (GSM, GPRS, UMTS), and also the model upon which the Internet community has built mobility support for IP enabled devices through Mobile IP [14].

Macromobility

Mobile IP [14] allows an MS to maintain connectivity while changing its point of attachment. This is not possible with standard IP where routing assumes a permanent position for each terminal. According to

---

[1]From the Beacons the MS discovers what particular QoS features the WLAN supports, if any.
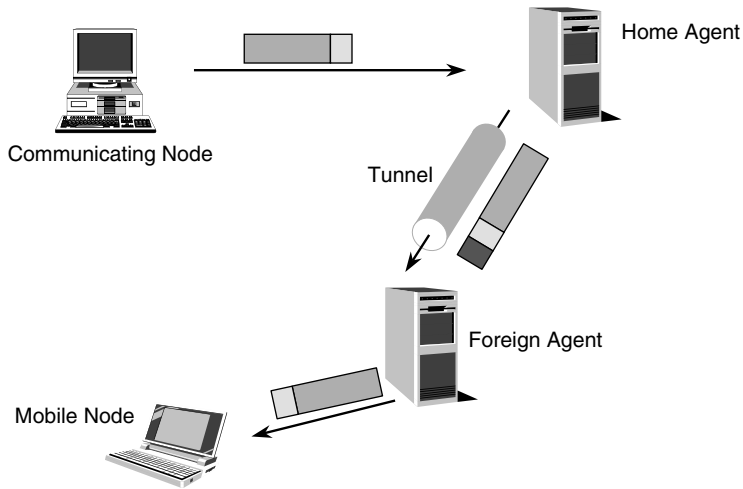
FIGURE 4.24   Basic Mobile IP functionality.

Mobile IP, an MS is given a long-term IP address on the network that it is registered (i.e., the home network), referred to as the "home address." This home address is administered in the same way as a permanent IP address provided to a stationary host. When away from its home network, a temporary "care-of address" is associated with the MS and reflects its current point of attachment. The MS uses its home address as the source address of all IP packets that it sends. Mobile IP defines two basic functional entities to implement its operation, the *Home Agent* (HA) and the *Foreign Agent* (FA). The HA is a router on an MS's home network, which accepts all packets destined to the MS and redirects them, through standard IP tunnelling, to the current position of the MS (i.e., the care-of address). To maintain updated location information, the HA is informed by the MS for every change of the care-of address through a special location update message. The FA is a router on an MS's visited network, which provides routing services. The FA detunnels and delivers to the MS packets that were tunnelled by the HA. For packets sent by the MS, the FA may serve as a default router.

  Although Mobile IP provides a solution to the problem of terminal movement, its performance depends on the distance between the home network and the current point of attachment. More specifically, when traffic is sent to the MS, packets are first routed to the HA, which encapsulates them and tunnels them to the FA for delivery. As is evident from Figure 4.24, the route taken by these packets is triangular. The most extreme case of routing can be observed when the communicating node and the MS are closely located to each other and far away from the HA. Additionally, in cases when the communicating node is behind a firewall, MS's outgoing traffic transmitted from a visiting network may be rejected if the firewall does not allow incoming traffic from the specific direction. Both these problems are addressed by Mobile IPv6, the evolution of Mobile IP that uses capabilities introduced by IPv6 to improve protocol's performance.

  According to Mobile IPv6 [P2], the functionality of the FA is included in every MS, which has the ability to de-tunnel incoming packets without the need of an external agent, making the whole scheme much simpler. Additionally, upon receipt of the first packet from a communicating node, the MS can send a location update message directly to this node with the information about its current care-of address. In this way, the communicating node can change the destination address of its outgoing packets from the home address to the care-of address of the MS and eliminate the triangular routing problem of legacy Mobile IP. Finally, the problem of firewalls is solved by setting up a reverse tunnel, from the MS to the HA, and sending all outgoing packets through this tunnel. Although this results in triangular routing in the reverse direction, packets are routed to the communicating node through the MS's home network, and are accepted by the firewall.

Micromobility

As the Internet technology penetrates more and more into every connectivity aspect in the research community, there has been much work done in optimizing mobility support for IP devices. One of the first observations was that the Mobile IP standard was not suitable for high-mobility, small geographic areas. Mobile IP is optimized for *macromobility* of relatively slow moving hosts, as it requires that, after each migration, a location update message is sent to a possibly distant HA, potentially increasing handoff latency and load on the network. To handle fast moving hosts in relatively small areas, the so-called *micromobility* protocols evolved. These protocols operate within an administrative domain to achieve optimum mobility support for fast moving users within the domain's boundaries. Most micromobility protocols establish and maintain soft-state host-specific routes in the micromobility enhanced routers. The inter-domain mobility support is left to standard Mobile IP, however. Three main representatives of this category are briefly presented: the Cellular IP, HAWAII, and Hierarchical Mobile IP.

According to Cellular IP, none of the nodes in the access network knows the exact location of a mobile host. Packets are routed to the mobile host on a hop-by-hop basis, where each intermediate node only needs to know on which of its outgoing ports to forward packets [16]. To minimize control messaging, regular data packets transmitted by mobile hosts on the uplink direction are used to establish host location information. The path taken by these packets is cached in the intermediate nodes in order to locate the mobile node's current position. To route downlink packets addressed to a mobile host, the path used by recent uplink packets transmitted by the mobile host is reversed. When the mobile host has no data to transmit, it periodically sends a route update packet to the gateway to maintain its downlink routing state. Following the principle of passive connectivity, idle mobile hosts allow their respective soft-state routing cache mappings to time out. These hosts transmit paging update packets at regular intervals defined by a paging update time. The paging update packet is an empty IP packet addressed to the gateway that is distinguished from a route update packet by its IP type parameter. Paging update packets are sent to the base station that offers the best signal quality. Similar to data and route update packets, paging update packets are routed on a hop-by-hop basis to the gateway. Intermediate nodes may optionally maintain paging caches that have the same format and operation as a routing cache except for two differences. First, paging cache mappings have a longer timeout period called paging-timeout. Second, paging cache mappings are updated by any packet sent by mobile hosts including paging-update packets. Paging cache is used to avoid broadcast search procedures found in cellular systems. Intermediate nodes that have paging cache will only forward the paging packet if the destination has a valid paging cache mapping and only to the mapped interface(s). Without any paging cache the first packet addressed to an idle mobile host is broadcast in the access network. While the packet does not experience extra delay it does, however, load the access network. Using paging caches, the network operator can restrict the paging load in exchange for memory and processing cost [17].

HAWAII segregates the network into a hierarchy of domains loosely modeled on the autonomous system hierarchy used in the Internet [18]. The gateway into each domain is called the *domain root route*r. Each host is assumed to have an IP address and a home domain. While moving in its home domain, the mobile host retains its IP address. Packets destined to the mobile host reach the domain root router based on the subnet address of the domain, and are then forwarded over special dynamically established paths to the mobile host. When the mobile host moves into a foreign domain, we revert to traditional Mobile-IP mechanisms. If the foreign domain is also based on HAWAII, then the mobile host is assigned a colocated care-of address from its foreign domain. Packets are tunneled by the home agent to the care-of address, according to Mobile IP. When moving within the foreign domain, the mobile host retains its care-of address unchanged, and connectivity is maintained using dynamically established paths. The protocol contains three types of messages for path setup: power-up, update, and refresh. A mobile host that first powers up and attaches to a domain sends a *path setup power-up messag*e. This has the effect of establishing host specific routes for that mobile host in the domain root router and any intermediate routers on the path towards the mobile host. Thus, the connectivity from that domain root router to the mobile hosts connected through it forms a virtual tree overlay. Note that other routers in the domain have no specific knowledge of this mobile host's IP address. While the mobile host moves within a domain, maintaining end-to-end connectivity to the mobile host requires special techniques

for managing user mobility. HAWAII uses *path setup update messages* to establish and update host-based routing entries for the mobile hosts in selective routers in the domain so that packets arriving at the domain root router can reach the mobile host with limited disruption. The choice of when, how, and which routers are updated constitutes a particular path setup scheme. The HAWAII path state maintained in the routers is characterized as "soft-state." This increases the robustness of the protocol to router and link failures. The mobile host infrequently sends periodic *path refresh messages* to the base station to which it is attached to maintain the host-based entries, failing which they will be removed by the base station. The base station and the intermediate routers, in turn, sends periodic *aggregate hop-by-hop* refresh messages towards the domain root router. Path setup messages are sent to only selected routers in the domain, resulting in very little overhead associated with maintaining soft state.

Hierarchical Mobile IP (HMIP) [19] is an extension of the traditional Mobile IP protocol used to cover micromobility scenarios. It introduces a new function, the Mobility Anchor Point (MAP), and minor extensions to the mobile host operation. The correspondent node and Home Agent operation are not affected. A MAP is a router located in a network visited by the mobile host, and is used as a local home agent. Just like Mobile IP, HMIP is independent of the underlying access technology, allowing mobility within or between different types of access networks. The operation of the protocol can be briefly described as follows. A mobile host entering a foreign network will receive Router Advertisements containing information on one or more local MAPs. The mobile host can bind its current location with a temporary address on the foreign subnet. Acting as a local home agent, the MAP will receive all packets on behalf of the mobile node it is serving and will encapsulate and forward them directly to the mobile node's current address. If the mobile node changes its current address within the foreign network, it only needs to register the new address with the MAP. Hence, only the regional address needs to be registered with correspondent nodes and the home agent, which does not have to change as long as the MS moves within the same network. This makes the mobile node's mobility transparent to the correspondent nodes with which it is communicating and its home network. An HMIP-aware mobile host with an implementation of Mobile IP should choose to use the MAP when discovering such capability in a visited network. However, in some cases the mobile node may prefer to simply use the standard Mobile IP implementation. For instance, the mobile host may be located in a visited network within its home site. In this case, the home agent is located near the visited network and could be used instead of a MAP. In this scenario, the mobile host would have to update the home agent whenever it moves.

QoS and Mobility Support

In order to improve the QoS provided during handovers, the mobility support techniques described above should be combined with IP QoS mechanisms such as the Resource reSerVation Protocol (RSVP) [20]. Designed for fixed networks, RSVP assumes fixed end-points, and for that reason its performance is problematic in mobile networks. When an active MS changes its point of attachment with the network (e.g., in handover), it has to reestablish reservations with all its communicating nodes along the new paths. For an outgoing flow the MS has to issue a PATH message immediately after the routing change, and wait for the corresponding RESV message before starting data transmission through the new attachment point. Depending on the hops between the sender and the receiver, this can cause considerable delays resulting in temporary service disruption. The effects of handover are even more annoying in an incoming flow because the MS has no power to immediately invoke the path reestablishment procedure. Instead, it has to wait for a new PATH message, issued by the sender, before responding with a RESV message in order to complete the path reestablishment. Simply decreasing the period of the soft state timers is not an efficient solution, because this could increase signaling overhead significantly.

A number of proposals can be found in the literature extending RSVP for either inter-subnet or intra-subnet scenarios. For intra-subnet scenarios, proposals that combine RSVP with micromobility solutions, such as Cellular IP, can reduce the effects of handover on RSVP since only the last part of the virtual circuit has to be reestablished. For inter-subnet scenarios, the existing proposals include advance reservations, multicasting, RSVP tunneling, etc. Below we focus in intra-subnet solutions, which can better fit in WLANs. Talukdar et al. [21] proposed Mobile RSVP (MRSVP), an extension of RSVP that allows the MS to preestablish paths to all the neighboring cells. All reservations to these cells are referred to as *passive*

reservations, in contrast to the *active* reservations in the cell that the MS actually is. When the MS moves from an old cell to a new one, the reservations in the new cell become active, while the reservations in the old cell change to passive. Although this proposal reduces the handover delays for path reestablishment, it requires RSVP to be enhanced to support a possible large number of passive reservations, while each AP has to maintain a lot of state information regarding active and passive reservations. Additionally, new real-time flows have to wait for all the necessary (passive and active) reservations before starting transmission, resulting in a possible high blocking rate. Tseng et al. [22] proposed the Hierarchical MRSVP (HMRSVP) in an attempt to reduce the number of required passive reservations. According to HMRSVP, passive reservations are performed only when an MS is moving in the overlapping area of two or more cells.

According to Kuo et al. [23], RSVP is extended with two additional processes, a resource clear and a resource rereservation in order not to release and reallocate reservations in the common routers of the old and the new path. This solution performs well in reducing the path reestablishment time, but modifies the RSVP protocol significantly. Chen et al. [24] proposed an RSVP extension based on IP multicast to support MSs. RSVP messages and actual IP datagrams are delivered to an MS using IP multicast routing. The multicast tree, rooted at each MS, is modified dynamically every time an MS roams to a neighboring cell. Hence, the mobility of an MS is modeled as a transition in multicast group membership. In this way, when the MS moves to a neighboring cell that is covered by the multicast tree, the flow of data packets can be delivered to it immediately. This method can minimize service disruption due to rerouting of the data path during handovers, but it introduces extra overhead for the dynamic multicast tree management and requires for multiple reservations in every multicast tree.

All these approaches, while trying to improve the performance of RSVP in mobile networks, either result in low resource utilization due to advance reservations, or require considerable modifications of protocols and network components operation. In micromobility environments only a small part of the path is changed, while the remaining circuit can be reused. Accordingly, a scheme for partial path reestablishment can be considered, which handles discovery and setup of the new part between the crossover router and the MS. The number of hops required to set up the partial path during a handover depends on the position of the crossover router. Such a scheme can reduce resource reservation delays and provide better performance for real-time services, without affecting the operation of RSVP considerably. Paskalis et al. [25] have proposed a scheme that reduces the delay in data path reestablishment without reserving extra resources, while it requires modifications only in the crossover router between the old and the new path. According to this scheme, an MS may acquire different "local" care-of addresses while moving inside an access network, but is always reachable by a "global" care-of address through tunneling, address translation, host routing, or any other routing variety, as suggested in various hierarchical mobility management schemes. The crossover router, referred to as *RSVP Mobility Proxy*, handles resource reservations in the last part of the path and performs appropriate mappings of the global care-of address to the appropriate local care-of address. A similar approach for partial path reestablishment has also been proposed by Moon et al. [26]. According to this scheme, if an MS is a sender, it sends an RSVP PATH message after the route update is completed. When the RSVP daemon on the crossover router, which is determined by the route update message, receives an RSVP PATH message after a mobility event, it immediately sends an RSVP RESV message to the MS without delivering it to the original receiver. If an MS is a receiver, the RSVP daemon on the crossover router can trigger an RSVP PATH message immediately after detecting any changes to the stored PATH state or receiving a notification from the underlying routing daemon. This PATH message can be generated based on the PATH state stored for the flow during previous RSVP message exchanges.

Mobility Management Architectures

Although the mechanisms described above provide the means of *how* to perform handover in mobile IP networks, they do not include the intelligence to decide *when* to handover, and *where* to handover. In the case of UMTS/WLAN interworking (as described above), the decision for the MS is easy, assuming that there are simple rules based on aspects such as the services used and the terminal capabilities (for example, the WLAN could simply be the preferred access system for all services and under all conditions). But what happens when

the terminal has to choose from a wide set of available access technologies (e.g., UMTS, GSM, WLAN, Satellite) or even multiple access systems of the same technology (e.g., multiple WLANs)? Nowadays, it is relatively standardized for an MS to move seamlessly based on the infrastructure of a single provider, i.e., using its base stations, accounting services, and other facilities. The notion of ABC, however, supports the use of the fittest existing infrastructure at any time. According to this idea, the user is aware of the multiple surrounding mobility support infrastructures and can choose to connect to the fittest at any time (possibly judging from multiple parameters such as cost, bandwidth, technology capabilities, available services, etc.). Multiple connections may be initiated and deployed for the optimum result. The connections can stem from the same mobile device to multiple access networks, or even from cooperating mobile devices to different access network technologies. Such multi-homing capabilities must be built into the operating system of the terminal handling the devices, and are still in research stages. Examples of available infrastructure alternatives include the base station of another network provider (of the same technology), the base station in an access network of another technology, or even connectivity shared in an ad-hoc manner from a peer MS. To provide such an increased flexibility to the user, an advanced system architecture is required consisting of functional entities in both the MS and the network, able to built a distributed decision system that guides the MS to decide when to handover and to which of the available access systems.

A number of research projects have been initiated in the last few years towards designing an integrated architecture that can provide ABC provision to the user [27–29]. To give a hint on the design principles of such an architecture, the basic points of the CREDO system [27] are briefly described. CREDO's approach is sketched in Figure 4.25. A salient feature of this approach is that both the network and the terminals contribute towards optimal system operation. At the network's side, an advanced Network and Service Management System (NSMS) coordinates the various access systems towards achieving joint management of their resources, in particular through traffic load balancing among these segments. NSMS is also capable for QoS provisioning to users and for resource provisioning in reply to requests from Service Providers. The multimode MSs, besides being capable of operating over diverse access systems, incorporate functionality provided through a management module called Terminal Station Management System (TSMS) for the exploitation of this capability. The TSMS on a terminal interacts with NSMS (through message exchanges, according to an appropriate protocol) towards network-driven selection of the access system to which the terminal is assigned. The interaction ensures that the "local view" of the terminal (radio conditions in the area,
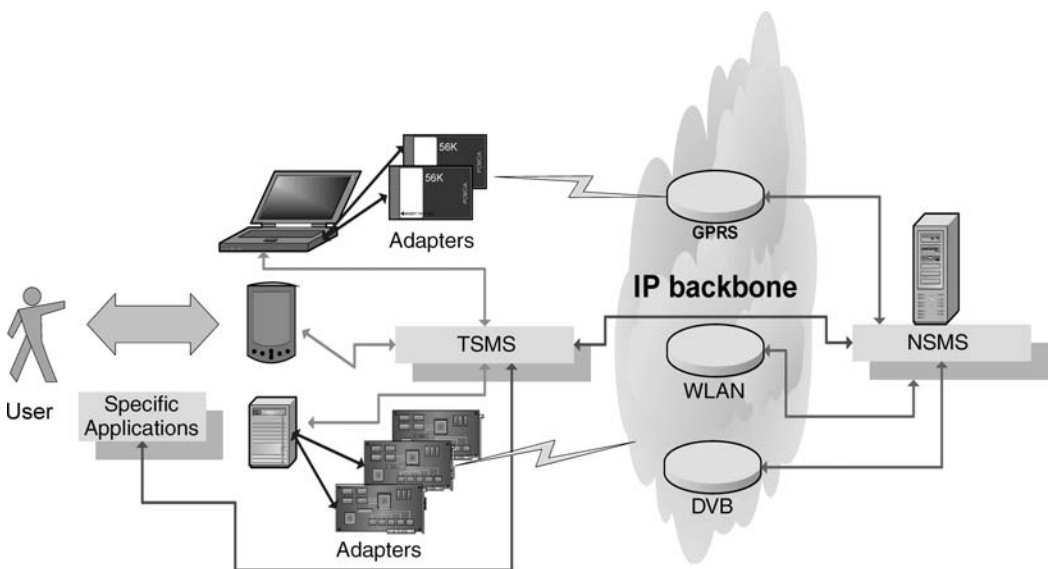

FIGURE 4.25 The CREDO architecture.

specific content services received over that terminal, the QoS levels associated with these services, etc.) and the global view of the network (traffic load over the various segments, the need to avoid congestion towards preserving QoS, etc.), are beneficially combined.

Two primary objectives are addressed by the NSMS in CREDO:

1. To guide individual users (terminals) towards the selection of the appropriate access network and the appropriate QoS level per service, based on the requested services, user preferences, terminal profiles, and network availability. This is achieved through a short-term optimization process, operating in near real-time.
2. To monitor the network infrastructure, assess the network and service-level performance, and employ this assessment for an optimal accommodation of the aggregate demand volume. The latter is achieved through a mid-term optimization process that takes into account system-wide information about the status and policies of the networks in the composite environment and the user preferences and terminal profiles. The results of the mid-term optimization (which also serve as a preconfiguration for subsequent short-term optimizations) are applied through a redistribution of terminals over the available radio networks.

The TSMS, on the other hand, is the core of the CREDO terminal and provides the following functions:

1. It receives service start and stop requests from the applications. This way it can keep track of all the currently running applications.
2. It monitors the terminal status: TCP/IP status, network interface status, application status, etc.
3. It reports all the gathered information to the NSMS.
4. Together with the NSMS, it selects the best access network to use at each moment.
5. It manages the terminal network configuration and provides the Mobile IP implementation. It configures the network drivers and the TCP/IP stack according to the decisions it takes.

As clearly shown from this brief overview, increased functionality is required in both the network and the terminal, in order to implement a decision-making mechanism that allows smooth ABC provision. Research in this area is ongoing, but the complexity of the problem and the large number of parameters ask for more detailed, complete, and yet feasible solutions.

Concluding Remarks

From the above discussion, it is clear that the mobile Internet is both an opportunity and a big challenge. The opportunity is to provide a fully integrated system, offering always-best-connected services without any user intervention. A system that will give the user the confidence that, by simply opening his mobile device, he will get the best possible service in terms of quality, cost, availability, device capabilities, etc. The challenge is to fulfill the large number of requirements imposed towards this provision.

In short, the key aspects of the evolution toward a fully integrated mobile Internet include the following:

- Evolution towards highly heterogeneous networks with several access technologies (both wired and wireless) and intelligent mobility management that supports fast vertical handovers and seamless session mobility. In this context, the interworking between 3G cellular and WLANs provides a key evolutionary path that can pave the way to other types of integration, e.g., with fixed and wireless access systems such as DVB-T and ADSL.
- Evolution towards all-IP-based networks, i.e., networks that support IP-based application signaling, mobility management, quality of service, as well as IP-based transport in the core and network access.
- Advanced IP Multimedia applications, which are tightly integrated with the equivalent applications in the Internet. Voice over IP, video/audio streaming, instant messaging, presence, press to talk, and other services will be enabled through the use of IP-based signaling protocols such as SIP.
- Provisioning of end-to-end QoS in order to support the demanding multimedia applications over diverse access media.

- Evolution of mobile terminals towards software-configurable radios with capabilities to be reprogrammed over the air and support many radio access technologies across many frequency bands.
- Robust and highly sophisticated security and AAA mechanisms and protocols. See Ref. [1] for further discussion.
- Adaptation of IPv6.
- Integration with Wireless Personal Area Network (WPAN) technologies.

References

1. A.K. Salkintzis, *Mobile Internet: Enabling Technologies and Services*, in Electrical Engineering and Applied Signal Processing Series, Boca Raton, FL: CRC Press, 2004, ISBN 0-8493-1631-6.
2. A.K. Salkintzis, C. Fors, and R.S. Pazhyannur, "WLAN-GPRS integration for next generation mobile data networks," *IEEE Wireless Commun.*, vol. 9, no. 5, pp. 112–124, 2002.
3. A.K. Salkintzis, "Interworking techniques and architectures for WLAN/3G integration towards 4G mobile data networks," *IEEE Wireless Commun.*, vol. 11, no. 3, pp. 50–61, 2004.
4. 3GPP TS 23.234 v6.0.0, "3GPP system to WLAN Interworking; System Description (Release 6)," March 2004.
5. 3GPP TR 22.934 v6.2.0, *Feasibility Study on 3GPP System to WLAN Interworking (Release 6)*, September, 2003.
6. 3GPP TS 23.060 v5.6.0, *General Packet Radio Service (GPRS); Service description; Stage 2 (Release 5)*, June 2003.
7. IEEE standard 802.11, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, 1999.
8. C. Rigney et al., *Remote Authentication Dial in User Services (RADIUS)*, IETF RFC 2138, April 1997.
9. P. Calhoun et al., *Diameter Base Protocol*, IETF RFC 3588, September, 2003.
10. IEEE draft standard 802.11e/D8.0, *Medium Access Control (MAC) Quality of Service (QoS) Enhancements*, February 2004.
11. Stefan Mangold et al., "Analysis of IEEE 802.11e for QoS support in wireless LANs," *IEEE Wireless Commun.*, 6, 40–50, 2003.
12. 3GPP TS 24.008 v5.12.0, *Mobile Radio Interface Layer 3 Specification; Core Network Protocols; Stage 3 (Release 5)*, June 2004.
13. E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Commun.*, February 2003.
14. C. Perkins, Ed., *IP Mobility Support for IPv4*, IETF RFC 3344, August 2002.
15. D. Johnson, C. Perkins, and J. Arkko, *Mobility Support in IPv6*, IETF RFC3775, http://www.ietf.org/rfc/rfc3775.txt, June 2004.
16. A.G. Valko, "Cellular IP — a new approach to internet host mobility," *ACM Comput. Commun. Rev.*, January 1999.
17. A.T. Campbell, J.Gomez, and A.G. Valko, *An Overview of Cellular IP*, IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, September 1999.
18. R. Ramjee, K. Varadhan, L. Salgarelli, S. Thuel, S.Y. Wang, and T. La Porta, "HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Trans. Network.*, June 2002.
19. H. Soliman, C. Castelluccia, K. El-Malki, and L. Bellier, *Hierarchical MIPv6 Mobility Management (HMIPv6)*, Internet Draft, http://www.ietf.org/internet-drafts/draft-ietf-mipshop-hmipv6-02.txt, June 2004.
20. R. Braden et al., *Resource ReSerVation Protocol (RSVP) — Version 1 Functional Specification*, IETF RFC2205, September 1997.
21. A. Talukdar et al., "MRSVP: a resource reservation protocol for an integrated services network with mobile hosts," *J. Wireless Networks*, vol. 7, no. 1, 2001.
22. C.-C. Tseng et al., "HMRSVP: a hierarchical mobile RSVP protocol," in *Proc. Int. Workshop Wireless Networks Mobile Comput. (WNMC)*, Valencia, Spain, April 2001.

23. G.-S. Kuo and P.-C. Ko, "Dynamic RSVP for mobile IPv6 in wireless networks," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Tokyo, Japan, May 2000.

24. W.-T. Chen and L.-C. Huang, "RSVP mobility support: a signalling protocol for integrated services internet with mobile hosts," in *Proc. INFOCOM*, Tel Aviv, Israel, March 2000.

25. S. Paskalis et al., "An efficient RSVP/mobile IP interworking scheme," *ACM Mobile Networks J.*, vol. 8, no. 3, 2003.

26. B. Moon and A.H. Aghvami, "Reliable RSVP path reservation for multimedia communications under IP micro mobility scenario," *IEEE Wireless Commun.*, October 2002.

27. H.Y. Lach and M. Catalyna, *Network access co-ordination to complement IP mobility protocols*, http://www.watersprings.org/pub/id/draft-lach-nac-00.txt, June 2003.

28. W. Zhang, J. Jaehnert, and K. Dolzer, "Design and evaluation of a handover decision strategy for fourth generation mobile networks," *Proc. VTC 2003-Spring*, Jeju, Korea, April 2003.

29. M.A. Rónai, R. Tönjes, M. Wolf, and A. Petrescu, "Mobility issues in OverDRiVE mobile networks," *Proc. IST Mobile Wireless Commun. Summit*, Aveiro, Portugal, pp. 287–291, June 2003.

4.5   Quality of Service in Packet-Switched Networks

Stan McClellan and Remzi Seker

This chapter discusses several technologies and concepts that are important in the provisioning of "Quality of Service" (QoS) in packet-switched networks. Unfortunately, "QoS" is an overloaded term often used to describe many things unrelated to data transmission. In packet-based networks, the techniques that provide performance guarantees for data streams are somewhat limited. Here, we use "QoS" to describe a network's ability to provide performance guarantees for the preferential delivery of packet streams, and we discuss the technologies and architectures that are generally incorporated into a QoS framework for IP-based networks. This framework lets network elements discriminate between particular traffic streams and then treat those streams in a particular manner, subject to broad constraints on forwarding performance.

Introduction and Background

Much like beauty, the notion of Quality of Service (QoS) for network communications is in the eye of the beholder. For example, under similar transmission conditions, the "quality" of voice, video, or other data may be interpreted very differently at the receiver. As a result, it is almost impossible to separate the practical deployment of QoS capabilities in a network from a thorough evaluation of those capabilities in the context of an application (or class of applications). Unfortunately, since application-level requirements are many and highly variable, a thorough evaluation of QoS technologies in the context of all applications is impossible. Additionally, due to the real-world constraints of existing network resources, a compromise between "theoretically desirable" and "practically implementable" is an overriding requirement.

Although much progress has been made towards an acceptable balance between the complexity and performance of packet-based QoS technologies, several significant deterrents remain, particularly for networks based on the Internet Protocol suite (IP). Some of the most significant issues relate to the heterogeneous, multi-domain, production nature of the distributed Internet and the lack of a robust approach to ensuring QoS requests across domain boundaries. Additionally, the lack of a unified interface between *user* and *network* for communicating and negotiating QoS requirements prevents active participation by applications in the end-to-end process. Even if these issues had tractable solutions, the difficult issue of performance optimization for complex, IP-based distributed systems remains. End-to-end network performance depends on many factors, from the digital signal processing techniques used in media compression to the architecture, management, and policy enforcement technologies for the overall network. These considerations are

particularly important for multimedia streams, which may demand higher bandwidth or have lower tolerance for delays than other data. Fortunately, recent developments in network optimization may overcome existing IP QoS issues for certain cases.

Because of some fundamental architectural characteristics of IP networks, the techniques available for providing performance guarantees are somewhat limited. These techniques tend to shape the aggregate traffic stream with simplistic regard for the needs of individual streams, or to ensure specific service characteristics for individual streams at the expense of impractical complexity. The shaping and performance guarantees for aggregated traffic are often called "Class of Service" (CoS) approaches. CoS technologies define a small set of generic behaviors that can approximate the actual service requirements of data streams if the streams are mapped into suitable, properly defined classes. In contrast, approaches capable of "true QoS" often require a homogeneous underlying network structure, and exchange a large amount of complexity and effective throughput for highly specific constraints on multiple statistical parameters of the data stream.

Network Architecture

In general, communication networks can be decomposed into functional elements in "planes" or categories of operation. For a discussion of QoS, a useful decomposition is in terms of *Access*, *Transport*, *Management*, and *Application* subnetworks as shown in Figure 4.26. In the figure, *Users* interface with an *Access* network that is responsible for performing or enabling the authentication of subscribers or devices as well as authorizing those entities to use network resources. As part of the authorization process, the *Access* network performs admission control and traffic grooming for the collection of subscriber traffic, subject to some constraints imposed by the *Management* network. The *Access* network interfaces with the *Core* or *Transport* network to exchange traffic that is within the terms established by a *Service Level Agreement* (SLA), or desired boundaries for aggregated data streams. The *Transport* network ensures that conformant data inserted at an ingress point is delivered to its appropriate egress point intact and without disruption or violation of the SLA. As shown in the figure, the *User* may be interacting with an *Application*
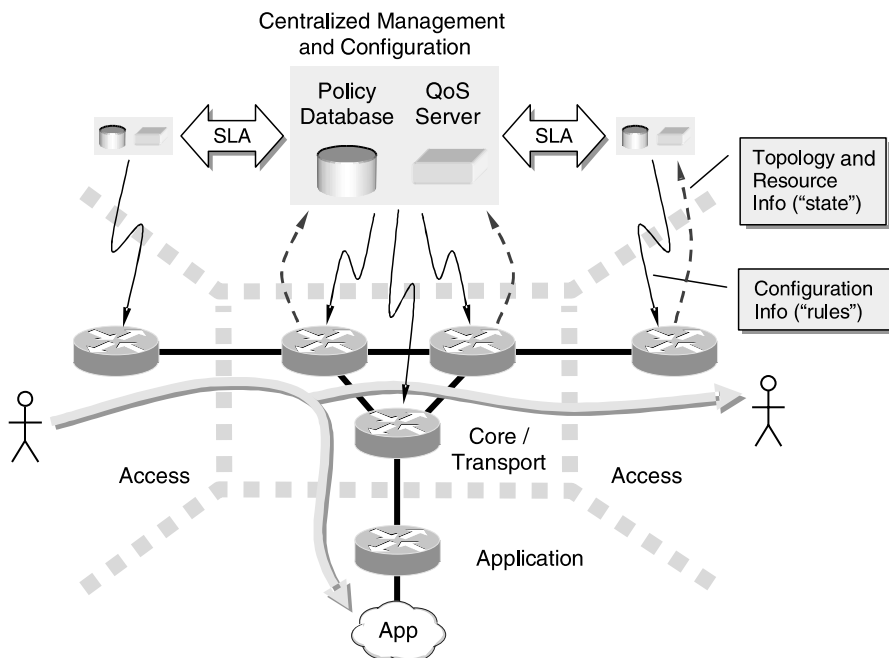

FIGURE. 4.26 Network architecture decomposed into Access, Transport, Management, and Application subnetworks.

resident in a special type of access network adjunct to the transport network, or with another *User* in a different *Access* network.

Regardless of the specifics of the interaction, and assuming that *Application* resources are adequately provisioned, the characteristics and capabilities of the *Access* and *Transport* networks are largely responsible for the Quality of Service (or lack thereof) seen by the *User*; these capabilities are intimately tied to the *Management* architecture of the aggregate network.

Configuration and States.   The network of Figure 4.26 cycles through functional states generally aligned with *Initialization* (or *Configuration*) and *Operation* modes or states.[1] In each mode, the elements in different "functional planes" of the network have key responsibilities, which involve the QoS technologies discussed in later sections.

In the *Initialization* or *Configuration* state, the primary activity is one of configuration or reconfiguration. This activity may be driven by administrative inputs or data related to an autonomous process, and may be initiated via scheduled activities or asynchronous events. For example, the negotiation of SLA parameters between subnetworks or administrative domains may depend on economic factors or business relationships, and typically lead to bounding conditions on network configuration. These items amount to a sort of "initial conditions" for subsequent network operations or configuration inputs. Additionally, alignment between services specified by the SLA and capabilities offered by network elements must logically follow the creation of the SLA in a mapping of "theory" to "reality." The partitioning of resources for end-to-end service fidelity is a significant consideration, particularly given the fundamental impact of QoS technologies on the stability of network elements. Of course, in a dynamically responsive network, the activities involved in the *Initialization* state can lead to reconfiguration of subnetwork or element functions to address time-dependent conditions or priorities. The SLA and element configuration are indicated in Figure 4.26 where topology or resource information is used to drive configuration activity. Note that the reconfiguration or distribution of "rules" can be initiated by user interactions, network state or conditions, or centralized, policy-based administrative actions.

In the *Operation* state, the primary network activities are very different for the *Access* and *Transport* subnetworks, and are carried out according to the configuration or resource partitioning established during *Initialization*. For instance, the operational *Access* network is responsible for complex ingress functions, which include shaping, policing, and logical differentiation of user data streams. In this respect, *shaping* data streams involves the enforcement of stability for pre-established traffic classes, *policing* involves the enforcement of policy or access rules, and *differentiation* involves the parsing and mapping of user or flow data to the available transport facilities. In similar fashion, the operational *Transport* network is responsible for ensuring the fidelity of end-to-end stream differentiation according to the *Access* indication and the boundaries established by the SLA(s). In modern packet-switched networks, this fidelity is generally accomplished by examining a label or tag attached to each packet during ingress differentiation, and placing the packet in a suitable, pre-configured transport queue. In addition to these functions, overall management activities take place within and between subnetworks to effect coarse adjustment between configurable traffic classes and to provide feedback to ingress admission control that is based on the current network state. This feedback component is indicated in Figure 4.26 as topology or resource information flowing from *Transport* entities towards *Management* entities.

As an example of the preceding discussion, the interactions of various subnetworks can be viewed in the context of a single, simplified traffic flow, as shown in Figure 4.27. In the figure, a QoS-bearing request from the user invokes access-control procedures in the *Access* network and interaction with the configuration management and policy entities in the *Management* domain. As part of the authentication, authorization, and access-control process, the *Management* entities must verify resource availability against dynamic and static

---

[1]Of course, since the network may be a very large distributed system, the notion of a homogeneous "state" for all parts of the system is somewhat unrealistic. However, these "states" or "modes" may be identifiable for isolated subnetworks or collections of elements, so the classification is reasonable for the purposes of discussion.
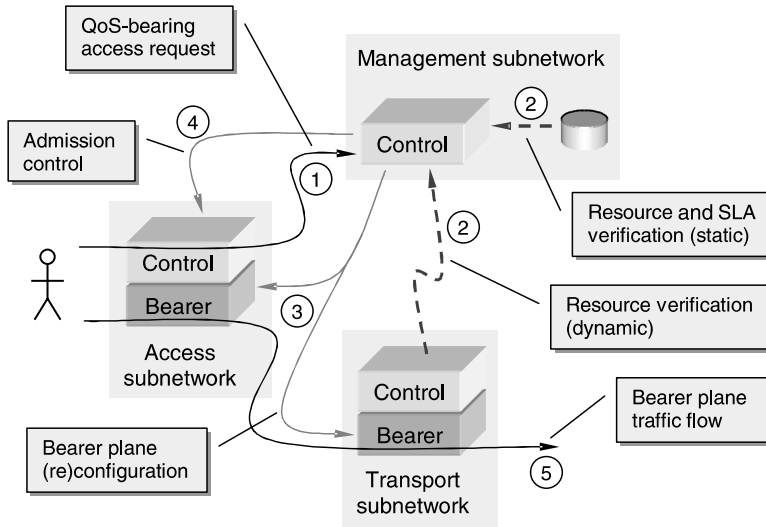
FIGURE 4.27   Subnetwork interaction based on QoS-sensitive traffic request.

network parameters and SLA boundaries. Based on the result of these admission processes, resources in the *Access* and *Transport* network may be adjusted to accommodate the bearer traffic flow, or the flow may be mapped onto an existing transport facility with pre-established service characteristics.

Industry Specifications.   Based in part on the functionally segregated network architecture of Figure 4.26, several Internet Engineering Task Force groups (IETF) [1] have been working on standardized approaches for IP-based QoS technologies. The IETF approaches of interest here fall primarily into the following categories:

- prioritization using differentiated services (RFC 2475),
- reservation using integrated services (RFC 2210), and
- multiprotocol label switching (RFC 3031).

The QoS mechanisms recommended by the IETF are important, because in addition to their implicit functional decomposition (compatible with Figure 4.26), they are designed to enable some forms of QoS without disrupting the basic transport architecture of the Internet. This architecture relies heavily on conventional packet forwarding devices (routers) that must efficiently and accurately process large numbers of statistically multiplexed packets. Unfortunately, IP network design is based on best-effort packet forwarding, an approach that does not distinguish explicitly between the needs of particular streams. In fact, some of the efficiency and scalability of IP networking is based on an architecture where intermediate forwarding devices *do not need to consider* explicit per-packet or per-stream requirements. So, although the IETF QoS mechanisms really only provide performance guarantees for aggregated packet streams (i.e., they are CoS approaches), they may be the only viable approach to service differentiation in an IP network.

Structure of Discussion

QoS technologies that are important in the *Core* or *Transport* sections of packet-switched networks are discussed in Section "Transport Mechanisms." This discussion centers on the definition of "Quality of Service" and technologies enabling packet-switched networks to adapt to the requirements of multiservice dataflows. In concert with the burgeoning popularity of broadband Internet access, several important access network technologies are discussed in Section "Access Mechanisms." The composite effects of

transport, access control, and other architectural factors are particularly important for end-to-end QoS guarantees. Finally, the "Conclusion" section concludes that the deployment of end-to-end QoS mechanisms for application data- flows is a difficult but crucial requirement for next-generation IP networks.

Transport Mechanisms

The current structure of the Internet relies heavily on conventional routers to store, examine, and transmit packets. These routers, with their routing protocols and forwarding mechanisms, are fundamental components of the *Transport* section of Figure 4.26. The reliable, flexible structure of the commodity Internet is largely based on the routing protocols, which determine packet forwarding, and on the deterministic behavior of those forwarding mechanisms. For reference, Figure 4.28 contains a functional diagram of a generic router showing basic components such as the route database and route selection, packet forwarding process, and topology updates.

Routing

The actions of routers are logically similar to the actions of a shipping clerk or production-line worker. Each item from a stream of items (the assembly line) is inspected, wrapped, boxed, and loaded for transport. Broken items must be tossed aside, and maximum efficiency is achieved when all items are similar. The introduction of "special" or "nonstandard" items causes conflicts in resource allocation and efficiency. The introduction of QoS-sensitive data streams into IP-based networks is equivalent to interlacing items needing special handling into a highly optimized assembly line. Depending on the nature of the data (size, shape, characteristics of the item), the ramifications of differentiated transport can affect all subsequent operations [2].

For example, current-generation IP routers use routing protocols to distribute topology and reachability information. This information is used to compute optimal forwarding or next-hop paths. Even in best-effort (conventional) networks, the core forwarding function can be a significant bottleneck because of complex,
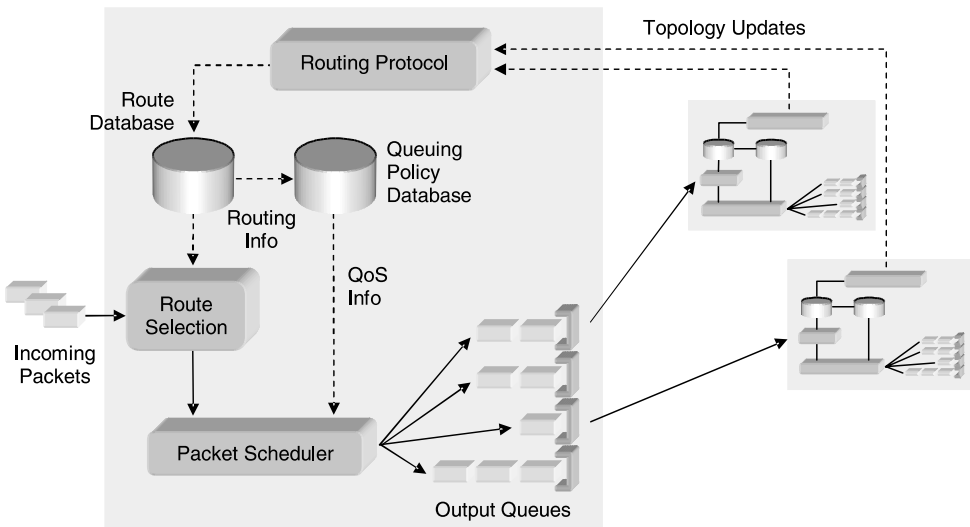

FIGURE 4.28   Generic router architecture, showing routing decisions, packet scheduler, and output queuing.

per-packet calculations.[1] The complexity of these computations does not scale gracefully, even for the one-dimensional service characteristics of current IP traffic. Unfortunately, queuing and forwarding mechanisms in QoS-capable networks may be significantly more complex. In addition, the introduction of higher-order, application-specific information may require significant changes to the "reconnaissance" that supplies data for these computations.

To illustrate these factors, the router in Figure 4.28 is augmented with "QoS-aware" functions such as a queuing policy database, sophisticated packet scheduler, and several customizable, individually prioritized output queues. Additionally, the route updates (reconnaissance) are assumed to contain higher-order information to feed the QoS-aware route computations. It is helpful to view the architecture of IP packet forwarding elements in this rather simplistic fashion to highlight concepts of shortcut routing and switching, integrated and layered routing protocols, and per-stream differentiation via class-of-service mechanisms.

Shortcut Routing and Switching. To keep pace with rapidly increasing data rates, forwarding algorithms are often implemented in hardware-based, connection-oriented "shortcut routing," which combines considerations from the transport, network, and link layers of the canonical Open Systems Interconnection (OSI) model. This integration of routing and switching in IP networks can address some of the special requirements of QoS-sensitive streams. However, although efficient implementation may reduce the computations and memory accesses required for simple forwarding operations, additional poorly defined burdens such as policy-based filtering, QoS assurance metrics, and sophisticated scheduling algorithms will impose additional complexity that may not be amenable to implementation in silicon [4,5].

Switching architectures tend to be implemented in special-purpose hardware, such as "network processors," which can efficiently process structured data streams. The "layer 3 switch" is an evolutionary step towards a two-stage process where *forward tactical information* gathered from the data stream or network topology is used to rapidly and temporarily reconfigure devices for subsequent packet flows. In the context of data networking, *flow detection*, *shortcut routing*, and *IP switching* are terms that are commonly used to describe this process. Typically, the efficient, synchronous data handling capability of a switch architecture is slaved to the slower, higher-layer processes that distribute network topology [2].

This process is part of the *Initialization* or *Reconfiguration* state in the network of Figure 4.26, and is shown in the figure as the distribution of "rules" to network elements. However, a key assumption for this technology is the deployment of a homogeneous transport infrastructure, such as in a local-area or single-domain network. In contrast, layered and integrated approaches to route determination tend to be more flexible and amenable to deployment for larger-scale, multi-domain networks, at the expense of "precision" in the match between QoS requirements and service availability.

Layered Approaches—OSPF. A layered approach to integrating IP with various transport technologies maintains a level of independence between the routing protocols of the separate OSI layers. The independence of these protocols in layered solutions leads to a scattering of layer-2 and layer-3 connectivity information between separately maintained databases.

The predominant IP routing protocol is *Open Shortest Path First* (OSPF, RFC 2328), which was developed for conventional routed IP networks. Large routed networks are typically broken into *areas* that are interconnected by area border routers and backbone networks. Each router has a complete view of the topology for its area, but an incomplete view of the topology for neighboring areas. The primary function of OSPF is to exchange network topology information between network areas, which reduces the amount of routing traffic. When link states change, routing information is flooded between nodes, and topology calculations are performed in parallel. In this fashion, large networks of routers are able to converge on a common topology quickly, and each router is able to use the link state and bandwidth information to calculate the "best" path across the network for subsequent packets. In this case, best generally means fewest router hops. This mechanism is shown in Figure 4.26 and Figure 4.28 as the feedback of state and resource information ("topology updates") from routers to each other or to a

---

[1]For example, OSPF (discussed later) uses Dijkstra's algorithm on distributed topology and reachability information to produce weighted per-hop forwarding rules [3].
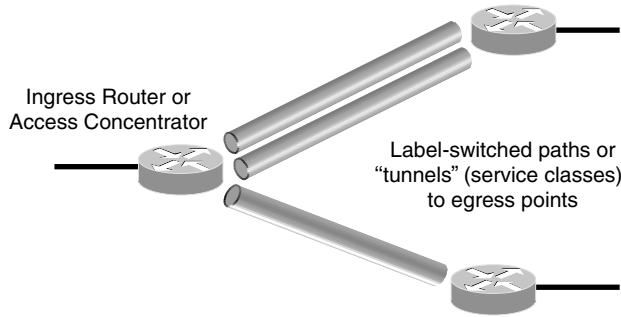
FIGURE 4.29   Label-switched "tunnels" in an IP transport network are independent of network topology.

configuration entity, as well as the computation and use of routing information. However, even though OSPF topology calculations are *efficient*[1] each router must still examine each packet and make a forwarding decision that may not include the multi-dimensional QoS requirements, policy rules, or accurate forwarding configuration for all routers in the data path. This requirement for "high touch" packet handling is cumbersome and not optimal for high-rate or high-complexity forwarding. As a result, other technologies are likely to be prevalent in QoS-enabled networks.

Integrated Approaches—MPLS.   As the leading candidate for next-generation IP routing technology, Multi-Protocol Label Switching (MPLS, RFCs 3031 and 3270) addresses the requirements for QoS and availability in backbone networks by simplifying the packet forwarding function. This simplification is accomplished, in the context of Figure 4.26, by pushing complex analysis and processing of per-packet routing information to the ingress point (*Access* network), instead of repeating the process at intermediate or interior nodes (*Transport* network). This function can be viewed as a compromise between the *signaling* function of true connection-oriented networks and the best-effort *packet forwarding* of IP networks [2,4,6]. In effect, MPLS establishes some number of presignaled paths with particular service characteristics in the *Transport* network, and then maps incoming flows into those paths based on a metric applied in the *Access* network.

The general concept behind label switching refers to a type of traffic engineering which allows a router to efficiently determine a packet's next hop without looking at the header of the packet or referring to routing lookup tables. Instead, an MPLS-capable network architecture, such as shown in Figure 4.26, attaches fixed-length labels to packets at their ingress point in the *Access* network so that the *Transport* network can use the labels to make forwarding decisions. The MPLS concept relies on the definition of preestablished or "overlay" label-switched paths (LSP) or tunnels through the network that, when combined with appropriate resource partitioning, can be used as "service classes" for a Class of Service (CoS) implementation. A simplified view of this configuration is shown in Figure 4.29. In the figure, the structure of the overlay LSP's are independent of the actual network topology or physical connections. However, the level of service differentiation depends directly on the characteristics of forwarding nodes along the LSP. During operation, ingress routers in the *Access* network classify packets into particular service classes (tunnels) based on the best match between data requirements and path availability. The data requirements used in the classification process may be inferred from a flow's source or destination, user profile, or other data characteristics, or it may be dictated by administrative means. The mapping function between the data requirements and the available service classes (tunnels) is a function of network implementation. This mapping is also an important aspect of the SLA because subsequent packet forwarding in the *Transport* network automatically moves packets into paths indicated by their ingress label.

The limitation of available service classes (tunnels) can be viewed as a finite number of lanes on a highway, or as a computing process with independent threads. While bottlenecks may occur in a particular traffic lane or individual thread, other options are available and are relatively independent. A result of this compromise between best-effort forwarding and application-specific QoS is a sort of quantization of the possible QoS

---

[1]OSPF uses less than 1 percent of link bandwidth and less than 2 percent of CPU capacity for routing calculations [3].

mechanisms. As a result, some distortion must be introduced in the servicing of individual data streams [2]. The resulting distortion is an abstract combination of the *Access* network's classification and the *Transport* network's service classes and queuing/forwarding functions.

Traffic differentiation methods in the *Access* section, combined with MPLS in the *Transport* section and some form of dynamic resource partitioning, may be a natural and effective hybrid approach to QoS in IP networks. However, the issues of end-to-end coordination between the specific meaning of traffic classifications, the accurate mapping of these classifications to labels, and the deployment of consistent queuing and forwarding behaviors at intermediate nodes is an extremely complex undertaking.

These issues are far more complex if fault management is required for the LSPs as a condition of the SLA. In this case, the failure of an LSP may require the use of a different tunnel with equivalent service characteristics. This problem may require dual provisioning of paths, with accompanying difficulties in traffic engineering. Maintaining network stability during the dynamic allocation of resources for these tunnels is an open research area due to the tight coupling between routing architecture and forwarding behavior [7].

Per-Stream Differentiation

A key factor to enabling effective service differentiation (CoS or QoS) in data communication lies in the *Transport* network's ability to handle multiple classes of traffic. On one extreme (exact QoS), the network must dynamically allocate resources based on per-stream requirements. On another extreme (CoS), the network is preconfigured for a subset of services, and the streams passing through the network must be mapped onto the "closest match" for their requirements. The current Internet is an example of the latter architecture, where a single service class is available (best effort forwarding) that has arbitrary bandwidth but nondeterministic latency and jitter characteristics. This service class is very good for applications such as file transfers, which are insensitive to latency. The existing public telephone network is another example of the latter architecture, where the single available service class (a dedicated circuit) has excellent latency and jitter characteristics, but limited bandwidth. This service class is optimized for voice transport, but can perform well for small volumes of data.

In IP networks, two general QoS frameworks have been explored: *Differentiated Services* or *"DiffServ"* (RFC 2475) and *Integrated Services* or *"IntServ"* (RFC 2210). Both of these approaches rely on some form of advance configuration or resource partitioning at intermediate forwarding nodes, including deterministic buffering, queuing, and packet scheduling algorithms. However, *DiffServ* is specifically designed to achieve low complexity and stability in existing IP network deployments, whereas an *IntServ* architecture would result in higher implementation complexity and cost of deployment.

The Intserv model refers to an approach that emulates circuit switching to reserve resources according to requests made by individual data streams. In the IntServ architecture, a protocol such as the Resource Reservation Protocol (RSVP, RFC 2205) is used to specify requirements for QoS parameters such as bandwidth, packet loss, delay, and jitter, in hopes that the intervening network will be able to satisfy this request on a per-flow basis. Note that the IntServ model infers a requirement for core transport nodes to maintain some amount of per-flow state information, which is not particularly compatible with typical IP architecture. IntServ also depends on the forwarding behaviors of the core nodes, and has been extended to interface with some of these mechanisms (for instance, RFC 2750 and RFC 2872).

In contrast, the DiffServ model aggregates similar flows into *equivalence classes*, which eliminates the need for transport nodes to maintain per-flow state or signaling information. With DiffServ, examination of per-flow characteristics is done once at ingress, with simple per-hop behaviors and aggregate policing performed by task-optimized core nodes. Clearly, these mechanisms are compatible with the decomposed network of Figure 4.26. Unfortunately, as is the case with IntServ, the transport effectiveness of DiffServ-marked packets depends heavily on the per-hop forwarding behaviors implemented in the intervening nodes.

Diffserv. The DiffServ model refers to an approach for implementing service differentiation that has been defined to be broadly compatible with the existing Internet architecture. In the DiffServ architecture, service differentiation is not provided on a per-flow or per-request basis. Instead, packets are classified into predefined per-hop behaviors (PHBs) that *approximate* the QoS requirements of the data stream. The DiffServ
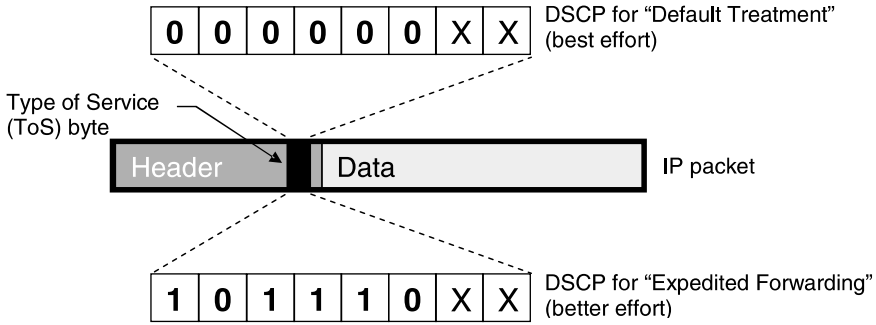
FIGURE 4.30   Examples of DiffServ codepoints (DSCP). The DSCP bits occupy the six most significant bits of the type-of-service byte in the IP header. The two least significant ToS bits are used for congestion notification, and are unrelated to QoS.

classification mechanism uses a tag called the *DiffServ code point* (DSCP) that is attached to the six most-significant bits of the type-of-service byte (ToS, RFC 791) in the IP header. DSCP values are defined by RFC 2474 and illustrated in Figure 4.30 in relation to the structure of an IP packet. The idea behind DSCP classification is to enable transport routers to easily and quickly classify packets into different types of output queues or PHB's based on the DSCP. Square tags go in square queues. Round tags go in round queues. Packets in the square queue get treated differently than packets in the round queue [8].

To enable distributed configuration, independently administrated networks implement an all-knowing, overlord node, sometimes called a *bandwidth broker*. This node acts as a proxy on behalf of a user or application to request traffic flow through transit networks. This overarching function is shown in Figure 4.26 as the "Policy Database" and "QoS Server" of the *Management* network. An issue with this architecture is that separately administered networks may be implemented with differing technologies, policies, and network management strategies, making end-to-end coordination difficult. Additionally, since DiffServ leverages the architecture where per-stream complexity is pushed to the edge of the network, traffic policing functions may be performed only in the *Access* network. This configuration is tremendously sensitive to misconfiguration, since *implied internal trust domains* mean that internal forwarders do not perform shaping, etc. As a result, SLA verification and enforcement are *critical* to prevent the effects of "relative ethics" between the definitions of particular traffic classes in different network domains.

Although the DiffServ model defines a collection of PHB's and traffic classifications, it doesn't specify explicit mechanisms for achieving the service classes indicated by the PHB's. The DiffServ *expedited forwarding class* (EF) is intended to provide the highest level of aggregate QoS for streams requiring minimal delay and jitter. Other classes, such as *assured forwarding* (AF) and *best effort* (BE) classes detail coarse loss bounds and relative prioritization between streams.

Expedited Forwarding

The highest priority DiffServ classification is called *expedited forwarding* (EF, defined in RFC 3246). The idea behind EF is to simulate a "virtual leased line" by ensuring minimal queuing of packets within each router along the transport path. In this fashion, the EF class hopes to provide guarantees on delay and jitter, which are important for isochronous data streams (i.e., video and audio). Unfortunately, due to an inability to distinguish between individual traffic streams, only the *aggregate* EF flow receives the desired treatment. This produces jitter in the delivery of *individual* EF streams. The only way to minimize these effects is to practice "gross overprovisioning," where only a small percentage of the available bandwidth is made available to the EF class, and only a few EF streams are allowed [8].

Assured Forwarding

The most complex Diffserv classification is called *assured forwarding* (AF, defined in RFC 2597). The AF designation creates four independent classes of AF, each with three subtypes. This classification requires a total of 12 DSCP values, as shown in Table 4.1.

TABLE 4.1   Assured Forwarding (AF) DSCP values in binary and decimal. The DSCP values occupy the six most significant bits of the type-of-service byte in the IP header

CoS	Drop Precedence (likelihood)	Subtype (x)	AF1x	AF2x	AF3x	AF4x
"Gold"	Low (unlikely)	$x = 1$	001010 (10)	010010 (18)	011010 (26)	100010 (34)
"Silver"	Medium (moderate)	$x = 2$	001100 (12)	010100 (20)	011100 (28)	100100 (36)
"Bronze"	High (likely)	$x = 3$	001110 (14)	010110 (22)	011110 (30)	100110 (38)

The difference between AF classes is related to different levels of forwarding assurance. The subtypes in each AF class indicate a "drop precedence," or relative importance within the class, as shown in Table 4.1. Each forwarding node (router) allocates resources such as buffer space, bandwidth, etc., for each AF class. Based on these resources, traffic streams have some level of assurance that packets from each class will be forwarded as desired. Transmissions can exceed these resources at their own peril, described by the "drop precedence." A congested DiffServ node is more likely to drop AF packets that have a higher drop precedence. So, within the AF designation, forwarding depends on the relationships between the instantaneous traffic load at a router, the available resources compared to the desired resources, and the drop precedence of each packet [8].

Best-Effort Forwarding
The lowest Diffserv classification is the well-known *best effort* (BE) behavior of the current Internet. So, coarse differentiation between service levels is made by classifying packets as BE (poor), AF (better, with conditions), or EF (best).

Queuing and Forwarding Mechanisms.   Of course, the Diffserv classifications or DSCP tags are merely *suggestions* by the *Access* network as to how the packets should be handled in transport. These suggestions must be acted upon by individual forwarding nodes in the *Transport* network. Some popular queuing and forwarding mechanisms (PHB's) are listed in Table 4.2 and described below. Under the proper circumstances, these mechanisms can appropriately handle a variety of service classes, including isochronous (or time-sensitive) traffic.

Weighted Fair Queueing (WFQ)
WFQ is a congestion management algorithm that enables routers to provide consistent, fair response time for packet forwarding when traffic load is high. A general diagram of a WFQ mechanism is shown in Figure 4.31, where incoming packets are classified into output queues based on some innate characteristic of the traffic. Common classification schemes for IP-based networks include transport address or protocol, application protocol, or user information.

After classification, packets are scheduled for output (forwarding) based on priorities or weights derived from some external inputs. The WFQ weights can be derived from administrative sources, feedback from network conditions, traffic handling policies, or other inputs. The weight that WFQ assigns to a flow

TABLE 4.2.   Some queuing mechanisms for IP service differentiation

| Queuing Mechanism	Comment
First In First Out (FIFO) or Best-Effort (BE)	Packets are forwarded in the order of their arrival, and no service differentiation is performed or guaranteed.
Weighted Fair Queuing (WFQ)	Interactive traffic is inserted into the front of the queue to reduce response time; then, remaining bandwidth is shared between additional flows.
Class-Based Weighted Fair Queuing (CBWFQ)	Packets are parsed into multiple categories ("classes") of queues for finer-grained differentiation. WFQ scheduling is applied separately for each category.
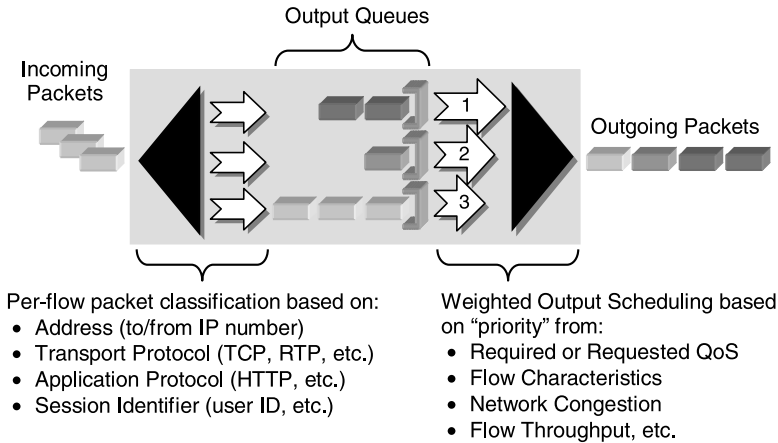
FIGURE 4.31 Weighted Fair Queuing (WFQ) packet flow diagram showing classification, queuing, and priority-based output scheduling.

determines the transmit order for packets queued in that flow. With proper configuration, WFQ can be used to schedule interactive or isochronous traffic to the front of the output queue to reduce response time. Remaining traffic then fairly shares the remaining bandwidth according to the assigned weights.

Class Based Weighted Fair Queueing (CBWFQ)

Class Based Weighted Fair Queueing (CBWFQ) is a specific type of WFQ that uses aspects of the incoming traffic stream to classify or forward packets. CBWFQ extends the basic WFQ to include support for user-defined traffic classes with particular characteristics. In this case, traffic classes can be designed with target parameters or behaviors such as minimum bandwidth or maximum queue depth. These characteristics provide boundaries for packet forwarding during network congestion.

The behavior of CBWFQ when the forwarding node encounters these boundary conditions can include class-specific policy-based actions for queue management, such as "tail drop" or "time in queue" metrics. Generally, traffic not matching any of the configured classes is given best-effort treatment. Once classified, the traffic classes are prioritized and scheduled for output based on a policy map that may be attached to particular physical interfaces.

Various queuing disciplines can be implemented within a traffic class. For example, *strict priority queuing* in a traffic class would allow delay-sensitive data such as voice to be scheduled for transmission before non-voice packets in other queues. This configuration effectively gives isochronous traffic preferential treatment over other traffic.

Other Approaches

As mentioned previously, a fundamental issue in conventional IP networks is the workload inferred on core forwarders (routers). This workload might include complex per-packet calculations or use of unavailable or voluminous per-packet data. Alternative approaches to the fundamental tradeoff between "per flow state" (QoS) and "scalability" (CoS) may be related to preprocessing packet flows, adapting network-based congestion control mechanisms, or augmenting packet headers with dynamic per-flow, per-packet state.

Dynamic Packet Scheduling (DPS). One of the significant problems with IntServ-type mechanisms such as RSVP is that per-flow state can grow exponentially, which is not a scalable requirement for core forwarding devices. Dynamic Packet State (DPS) is an alternative approach to per-flow state that augments the packet header to include flow-specific information that is updated by forwarding nodes [9]. The DPS architecture uses ingress classification and efficient core forwarding as in the DiffServ model, but generalizes the concept of

a "tag" to be an executable object or indicator with variable per-hop content. With DPS, ingress routers in the *Access* network precompute necessary parameters and classify packets, while *Transport* nodes dynamically combine per-packet information ("flow state") with instantaneous network conditions ("network state") and update the packet headers for processing at subsequent nodes. This architecture is a compromise between the relatively static DiffServ architecture and the explicit per-flow state requirements of RSVP, where core routers perform simplified scheduling based on packet classification and network conditions. The structure is logically very similar to modern, efficient medical provider networks where "ingress" general practice physicians preclassify patients before referring them to upstream specialists, along with partial diagnosis and treatment. However, DPS technology could be very difficult to implement with legacy router capabilities.

Congestion Control in Traffic Classes. The joint optimization between QoS and congestion control mechanisms is another approach to IP QoS that could have significant merit, particularly since it is compatible with the decomposed network architecture of Figure 4.26 and DiffServ/MPLS [9]. In one sense, the legacy requirement for new IP technologies to exhibit "TCP-friendly" behavior fundamentally limits the potential for QoS alternatives. However, flow aggregation schemes such as DiffServ/MPLS require some form of congestion control for each aggregate. In the CoS paradigm, the isolation of traffic classes may alter the strict requirement for TCP-friendliness across all traffic classes, which could lead to better QoS mechanisms in certain traffic classes. In particular, per-class congestion control mechanisms could bridge the gap between CoS and QoS by segregating isochronous traffic from other packet streams and applying fundamentally different, more appropriate congestion control.

Access Mechanisms

In addition to the dependence on core technologies, the thorough evaluation of access technologies is crucial in developing and validating a functional, broadly deployable QoS framework. Here, the terminology "access technologies" is used to include the interaction between application-level requirements, the network protocols that are present at the boundary between application and network, the logical topology of the local network, and the unique physical layer capabilities that are provided by various network architectures. In discussing QoS capabilities of various access technologies, it is clear that QoS requirements are expressed at the edge of the network, or at the point of ingress, prior to (or instead of) propagation through/between subdomains. Additionally, there are generally very distinct physical contexts through which ingress is achieved. These QoS requirements must be coordinated with overall network policies, as well as access control and resource authorization mechanisms. With this understanding, the DiffServ model can be an important aspect of *Access* network architecture [10].

Policies and Access Control

Aside from the transport-related issues of network and traffic engineering, service provider handoff agreements, and architectures that satisfy latency, jitter, and forwarding constraints, an important QoS consideration is the interplay between subscriber profile, dynamic service provisioning, and service-level validation in the access network. Although transport networks can be overprovisioned to effectively handle traffic in certain "behavior aggregates," bandwidth in access networks is often strictly limited, or shared in a nonhierarchical manner. For instance, the use of IEEE 802.11 wireless LAN (WLAN) "hotspots" for broadband Internet access has extensive commercial appeal [11]. However, the integration of network usage policies, access control and authorization mechanisms, and OSI layer 1, 2, and 3 QoS technologies of WLAN "hotspots" with an upstream DiffServ/MPLS network is neither trivial nor well-addressed by standards.

Policy-Based Networking. Technologies that enable, manage, and validate QoS based on network-prescribed policies will be critical in modern IP-based access networks. In particular, QoS provisioning and SLA management in the interworking of "edge" or access capabilities (Layer 2) and "upstream" or transport capabilities (Layer 3) is not well defined. Although *Access* networks may support QoS mechanisms, the disconnect between layer-2 and layer-3 QoS or access control results in a lack of integrated QoS provisioning,
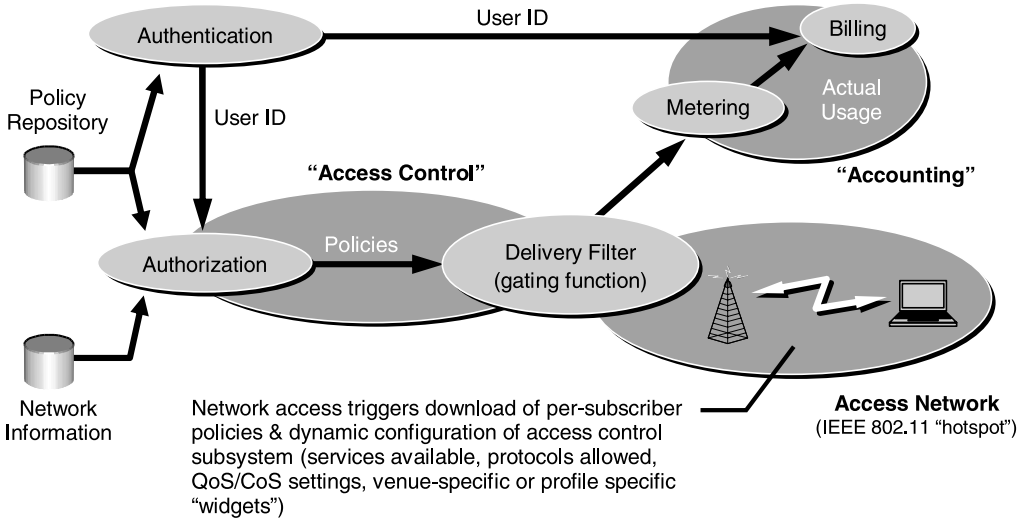
FIGURE 4.32 General architecture for policy-based networking via subscriber profile.

policy distribution/enforcement, and SLA verification for end-to-end flows. For instance, although the IEEE has proposed standards to address some aspects of QoS in the WLAN arena, these approaches are distinct from IP-based mechanisms, and may not be directly deployable without some interworking function.

An area of commonality between most technology propositions is the implicit reliance on a central infrastructure to define, propagate, and manage QoS features. Of course, these centralized facilities depend on a tightly controlled, distributed infrastructure to enforce, evaluate, and report such policies. This architecture is compatible with the network decomposition of Figure 4.26, and important elements of QoS-aware, policy-based access control are indicated in Figure 4.32.

In Figure 4.32, the tasks of authentication, authorization, and accounting are shown as logically different functions hosted in separate parts of the network. Additionally, the *Access* subnetwork (in this case, a WLAN hotspot) is shown as distinct from the access control function, which is comprised of an authorization capability, network policies, and a "delivery filter" or gating function. As shown, the attachment of a subscriber to the network first triggers the authentication process, which depends both on network policies ("static") and network information ("dynamic"). After authentication, the authorization process also uses these network parameters to compile and distribute specific, per-subscriber policies ("access control") to elements in the *Access* network. These elements control access to network resources and communication of traffic or usage statistics to the accounting function. The technologies of interest in this architecture include *Transport* considerations such as Diffserv/IntServ, MPLS, SLA verification and provisioning, and *Access* considerations such as Digital Subscriber Loop (DSL), wireless LAN technologies (IEEE 802.11), and their relationships with telephony networks. Additionally, multi-network access control architectures such as IEEE 802.1x may play an important role in enabling QoS in a variety of situations.

Of course, in commercial enterprises, the most important functional elements are contained in the "accounting" area. Although these functions are gated and fed by authentication and access control functions, specific discussions of rating and billing mechanisms are out of scope here.

Access Control. Access control and policy distribution in a managed network is a difficult proposition. An existing technology which may be very useful in this respect for IP networks is the IEEE 802.1x framework for port-based authentication [12]. In the 802.1x architecture, network access is limited to rudimentary functions until the authentication process is completed. Based on the result of authentication, network access can be gated and services can be authorized. The 802.1x architecture is also interesting

because it explicitly prescribes centralized control of user credentials and coordination with network-wide QoS or CoS policies, as described in Figure 4.32 and Figure 4.26.

In 802.1x, the participants in an authentication conversation are the *Supplicant*, *Authenticator*, and *Authentication Server*. Typically, the *Authentication Server* is part of a *Management* subnetwork, and the *Authenticator* is part of the *Access* network that the *Supplicant* is trying to use. Before using network services, the *Supplicant* initiates an authentication request to the *Authentication Server*. This authentication conversation is brokered by the *Authenticator* and can be based on standardized protocols such as the Extensible Authentication Protocol (EAP, defined in RFC 2284). For mobile subscribers, the authentication conversation is typically relayed between *Authentication Server* entities in different network domains. Until the authentication process completes successfully, the *Authenticator* prevents the *Supplicant* from gaining network access while translating between downstream, point-to-point network transports and the upstream authentication infrastructure.

Typically, the authentication conversation is transported by a well-known Authentication, Authorization, and Accounting (AAA) protocol such as RADIUS (RFC 2865). The interworking of EAP with RADIUS is well defined (RFC 3579), and several authentication protocols can be mediated via the EAP/RADIUS combination, including public keys and digital certificates. The RADIUS authentication architecture is a very flexible request/response conversation, where information is exchanged in "Attribute/Value" pairs. In this architecture, the *Authentication Server* can consult a local (internal) or remote (external) subscriber database, as well as proxy authentication requests to RADIUS servers in other administrative domains.

Because 802.1x does not directly depend on the AAA transport protocol, it is also compatible with enhanced authentication infrastructures such as DIAMETER (RFC 3588), which may become more important as IP QoS is deployed. In comparison to RADIUS, the DIAMETER transport has improved scalability, session control, security, and other refinements that can be leveraged in QoS-sensitive environments. Table 4.3 summarizes some of the important differences between RADIUS and DIAMETER. Aspects of DIAMETER that may be particularly useful in administration of QoS-based requests are related to session control (command extensions and resolution of accounting events) and security (flexible encryption).

In addition to simple RADIUS authorization, and in combination with sophisticated authentication infrastructures, the 802.1x framework can be used in a QoS/CoS environment to dynamically provide privileges, policies, configurations, and accounting requirements to network components in both the *Transport* and *Access* subnetworks. Gated by the authentication process, 802.1x enables network elements to modify access privileges according to individual entitlement. Although RADIUS-authenticated sessions can already provide limited service authorization, 802.1x creates a robust framework for service configuration, resulting in the potential for fine-grained, dynamic privilege authorization. This dynamic user/network configuration control is particularly important in popular broadband access networks.

Broadband Access

With the accelerating price erosion of broadband access, QoS is becoming an important differentiator for network operators. With reference to Figure 4.26 and Figure 4.32, the *Access* network, such as a DSL interface or WLAN hotspot, may be operated by an independent "network access provider" (NAP) that is operationally distinct from the "network service provider" (NSP) operating the *Transport* network. Either of these networks may contain *Application* subnetworks, and both likely contain a *Management* subnetwork.

For application services in the *Application* network, the NSP and/or NAP may also have a business relationship with an external "application service provider" (ASP). The boundaries between these separate enterprises are complicated by the technologies, functional responsibilities, and network policies that affect the realization of end-to-end QoS. In the following section, we focus on some particular aspects of broadband access networks that are compatible with the prevailing QoS/CoS architecture of Figure 4.26 and Figure 4.32, and the scenario of Figure 4.27.

Wireless LAN and Wireless Telephony. The Universal Mobile Telecommunications System (UMTS) has been extensively promoted as the future high-speed, high-bandwidth mobile telecommunications system to support multimedia services. However, due in part to the economic and logistical issues of "upgrading"

TABLE 4.3   Diameter vs. Radius

Category	Item	DIAMETER	RADIUS	Comment
Transport	IP Transport Protocols	TCP, SCTP	UDP	TCP and SCTP are connection-oriented transports, with flow-control and congestion avoidance mechanisms, whereas RADIUS uses connectionless UDP. DIAMETER has reliable message delivery (with per-hop retransmission and heartbeat), which enhances failover and makes reachability status for peer servers more precise.
Scalability	Attribute Length	24-bit	8-bit	DIAMETER allows larger Attribute values per transmission ("payload").
	Identifier	32-bit	8-bit	DIAMETER increases the number of concurrent pending messages ("window").
	Alignment	32-bit	8-bit	DIAMETER Header entries and Attribute data must align on 32-bit boundaries for processing efficiency.
Indirection	Proxy support	Distributed	Centralized	Localized (per-hop) failure detection in DIAMETER allows proxy servers to initiate failover and retransmission. RADIUS uses end-to-end retransmission by peer servers.
Session Control	Vendor-specific extensions	Attributes and Commands	Attributes only	DIAMETER has vendor-specific Commands as well as vendor-specific Attributes to enable customization and preserve interoperability.
	Server capability increased	Peer-to-peer	Client-Server	The peer-to-peer relationship of DIAMETER servers enable session termination and user reauthentication or reauthorization, whereas the client/server RADIUS architecture often requires proprietary protocol extensions.
	Accounting decoupled from Authentication and Authorization	Independent STOP	Joint STOP	DIAMETER Authentication/Authorization messages can be routed and handled differently than Accounting messages to improve session accounting fidelity
	Accounting resolution	Various	Session only	DIAMETER allows User, Session, Subsession, and Multisession IDs to correlate accounting.
Security	Encryption	IPsec, TLS	Shared secret	Individual DIAMETER Attribute/Value pairs can be digitally signed/encrypted to prevent observation/tampering by intervening proxy/relay servers.

existing networks to UMTS standards, wireless LAN technologies have been used as an *Access* network for some UMTS services [11,13].

The use of WLAN-based technology requires specific QoS levels that are acceptable by the standards of UMTS networks. Unfortunately, coordination between the OSI layer-1/layer-2 WLAN QoS technologies and IP-based layer-3 technologies is not addressed specifically by disparate standards bodies. Additionally, provisioning and billing of commercial UMTS data services can be complex when alternative *Access* network technologies such as WLAN allow for intrinsically shared network attachment. To bridge this gap, the Third Generation Partnership Program (3GPP) has described various scenarios for interworking UMTS and WLAN technologies, which include the billable use of UMTS core network services via a WLAN *Access* network [14].

In addition to issues of separate QoS technologies for various network layers, and in a twist on issues of separately owned and managed *Access*, *Transport*, and *Management* networks, the use of WLAN "hotspots" introduces another permutation: an *Access* network shared by competing network operators who use QoS as a leverage point for subscriber affinity. In this scenario, multiple UMTS operators share a WLAN-based *Access* network, which may be independently owned and operated. In such an interconnection, the architecture of Figure 4.28 is essentially replicated in each participating subnetwork, and the policy management framework

between subnetworks becomes an explicit, critical part of the overall SLA between subscribers, operators, and subnetworks.

In general, a "master" policy broker is established in each of the independent network areas. These master brokers then control QoS policies related to the flow of subscriber data *into* their respective networks. In this case, the QoS enforcement mechanisms of the *Access* network, which may include network-specific CoS mapping and classification functions, must be applied independently for data streams subject to the policy structures of different *Transport* networks. Clearly, this scenario presents tremendous logistical problems, particularly in cases where conflicts, omissions, or inaccuracies are present in the QoS policies of the constituent networks. Although the UMTS network operator would like to control the delivery of services to the user as if their terminal were directly connected to the UMTS network, the intervening non-UMTS *Access* network creates a level of "indirection" in the datapath. In particular, dynamic service requirements may be dependent on variable resources or capabilities in both networks, and the determination of enforceable network-level policies may be subject to some negotiation. In these scenarios, the relationship between QoS and policy entities in different subnetworks may be hierarchical, peered, or a hybrid architecture. The structure and administration of this "QoS/policy architecture" is particularly important to facilitate dynamic traffic streams [13].

Digital Subscriber Loop.   Another example of access technologies that require QoS/CoS mechanisms is fixed broadband networks such as Digital Subscriber Loop (DSL). Various mechanisms exist that are compatible with the architecture described by Figure 4.26 and Figure 4.32. In fact, it is not uncommon for DSL subscribers to have privately owned WLAN access points interfacing to a provider-offered DSL-based *Access* network. This setup is only a slight perturbation of Figure 4.32, and closely related to the UMTS/WLAN scenarios discussed previously. For DSL-based *Access* networks, the entities controlling each subnetwork may be combined, or the Internet Service Provider (ISP) or NAP may be distinct from the NSP, which is also distinct from the ASP or collection of ASPs. Regardless of the business arrangements and relationships, if the ASP offers applications that require service guarantees such as streaming video, VoIP, etc., the NSP must reserve *Transport* resources or provide a suitable mapping to existing CoS tunnels, and the NAP must reserve or enable *Access* resources that map appropriately to the requested QoS. In such scenarios, the technological barriers tend to be more tractable than the "business to business" policy coordination issues between independent entities, and simpler than in the case of a single WLAN *Access* network being shared amongst multiple UMTS service providers.

In the DSL scenario, resource enforcement in the *Access* network (NAP) based on configuration parameters indicated by the *Transport* network (NSP) is a preferable hierarchy. Also, due to bearer network security and stability issues, a "transparent" proxy-based request/configuration architecture is necessary, as indicated in the "QoS server" of Figure 4.26. With this centralized architecture, the potential for an explicit relationship between a subscriber profile and a dynamic network configuration, particularly in the context of RADIUS/ DIAMETER AAA protocols, is very reasonable—as long as the *Transport* network is either overprovisioned or configured with sufficient CoS resolution.

To realize such an architecture, an entity in either the *Access* or *Transport* networks must perform "route alignment" between the physical location of ingress/egress for a subscriber ("network service access point" or NSAP) and the available overlay tunnels or QoS facilities. For instance, for an ethernet-based access network, frame-based CoS could be used to map onto preconfigured circuits at an access multiplexer. After the authentication process is complete, the subscriber and NSAP information must be forwarded to a centralized broker to cross-reference network policies with subscriber privileges, and then downloaded to reconfigure ingress and core devices. In any case, admission control feedback from the NSP's *Transport* network may result in different queuing/forwarding strategies in border or ingress routers, as well as traffic control (shaping, policing) in the upstream sections of the DSL-based *Access* network [15].

Conclusion

True "QoS" guarantees are a necessary characteristic of next generation packet-based networks that will have a profound impact on the deployment of advanced, network-sensitive applications. In particular, the interaction

between application-specific traffic requirements and the real-world constraints of existing technologies must be carefully considered. Unfortunately, the effective delivery of QoS for data *and* multimedia streams presupposes sufficient bandwidth, latency, jitter, and higher-order guarantees than most networks can generally accommodate. Adequate QoS for packet-based multimedia is difficult for telephone networks, which are structured for voice transport. Adequate QoS for isochronous streams is difficult for IP networks, which are structured for best-effort data. In IP-based networks, the result of these constraints is an overriding requirement for practical implementation. This situation has led to the definition of CoS mechanisms where *Transport* networks are provisioned with some collection of service aggregates, and ingress *Access* networks are tasked with mapping the requirements of incoming streams onto these available transports—subject to some network-imposed metrics.

Since end-to-end performance optimization on large-scale distributed systems is difficult at best, the notion of *strict* guarantees on QoS/CoS will be dependent on the Service Level Agreements (SLAs) defined between neighboring network domains. The negotiation and enforcement mechanisms that comprise these SLAs at the boundaries between network operators and subnetwork domains must be coupled with, and measured against, the *actual performance* of *real applications,* in order to establish their validity. In addition, each user's expectation for QoS will depend on the configuration and capabilities of the *Access*, *Transport*, and *Application* sections of each network.

An area of particular interest in the quest for end-to-end QoS/CoS guarantees is the *Access* subnetwork and its relation to the *Transport* and *Management* subnetworks. In an IP-based paradigm, the bulk of per-stream complexity is explicitly transferred to the *Access* network, where it is dealt with *once* in a fashion that is assumed to be acceptable and easily aggregated. In a dynamic environment, the mapping between per-stream needs and special *Access* capabilities, such as those available in WLAN or DSL implementations, is more effectively addressed by architectures where *Access* network capabilities are slaved to the administration of CoS capabilities of upstream subnetworks. In essence, this implementation requires specific *Access* network capabilities to be integrated with the *Transport* network's configuration (available service classes) and per-hop-behaviors (available queuing mechanisms).

In this context, a viable next step in the evolution of network-based QoS guarantees may be the enablement of "feedback" between the various subnetworks of Figure 4.26. With the proper combination of network-based policies, programmable network elements, and dynamic feedback between subnetworks and applications, even limited service classes may be adjusted to provide the responsiveness necessary for QoS-sensitive applications.

References

1. The Internet Engineering Task Force (IETF), The complete text of IETF Requests for Comments (RFCs), http: //www.ietf.org/rfc/rfcNNNN.txt, "NNNN" is the RFC number.
2. S. McClellan, K. Burst, and G. Grimes, "Issues and techniques in network-based distributed healthcare: advanced networking techniques," in *Proc. Fourth World Conf. Integ. Des. Process Technol.*, Kusadasi, Turkey, 1999.
3. J.T. Moy, *OSPF: Anatomy of an Internet Routing Protocol*, Reading, MA: Addison-Wesley, 1998.
4. C. Metz, "Ingredients for better routing? Read the label," *IEEE Internet Comput.*, vol. 2, no. 5, pp. 10–15, 1998.
5. P. Dumortier, "Toward a new IP over ATM routing paradigm," *IEEE Commun. Mag.*, vol. 36, no. 1, pp. 82–86, 1988.
6. G. Hagard and M. Wolf, "Multiprotocol label switching in ATM networks," *Ericsson Rev.*, no. 3, 1–12, 1998.
7. J.L. Marzo, E. Calle, C. Scoglio, and T. Anjali, "QoS online routing and MPLS multilevel protection: a survey," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 126–131, 2003.
8. M. Stricklen, B. Cummings, and S. McClellan, "Linux and the next generation Internet," *Linux J.*, 73, 90–98, 2000.
9. M. Welzl and M. Muhlhauser, "Scalability and quality of service: a trade-off?," *IEEE Commun. Mag.*, vol. 7, no. 6, pp. 32–36, 2003.

10. J. Evans and C. Filsfils, "Deploying diffserv at the network edge for tight SLAs, part 1," *IEEE Internet Comput.*, vol. 8, no. 1, pp. 61–65, 2004.

11. S. McClellan, S. Low, and W.-T. Tan, "Disruptive technologies and their affect on global telecommunications," in *Advances in Computers*, M. Zelkowitz Ed., vol. 61, Academic Press–Elsevier: San Diego, CA, 2004.

12. IEEE standard for local and metropolitan area networks — port-based network access control, IEEE Std 802.1X-2001, 2001.

13. W. Zhuang, Y.-S. Gan, K.-J. Loh, and K.-C. Chua, "Policy-based QoS management architecture in an integrated UMTS and WLAN environment," *IEEE Commun. Mag.*, vol. 41, no. 11, pp. 118–125, 2003.

14. 3GPP TR 22.934, ver. 6.2, *Feasibility Study on 3GPP System to Wireless Local Area Network (WLAN) Interworking (Release 6)*, September 2003.

15. C. Bouchat, S. van den Bosch, and T. Pollet, "QoS in DSL access," *IEEE Commun. Mag.*, vol. 7, no. 9, pp. 108–114, 2003.

This page intentionally left blank

5

Ad Hoc Wireless Networks

Mohammad Ilyas
Florida Atlantic University

Ad hoc wireless networks are communication networks without a well-defined infrastructure. These networks are established on demand to satisfy a specific need for communication and exist for a shorter time period. The communication devices in an ad hoc network may be mobile and without ties to a fixed topological infrastructure. As the devices move, the network topology changes. The devices in ad hoc networks are the source and destination of information being exchanged, and they also act as intermediate devices to relay information between two devices located outside the communication range of each other. These networks have dynamic topology, bandwidth-constrained variable capacity wireless links, energy-constrained operation, and limited physical security. These networks have tremendous potential for commercial and military applications. They are useful for providing support where no communication infrastructure exists, or its deployment is economically not feasible. Their potential applications include emergency situations, healthcare, home networking, and disaster recovery operations. In this chapter, we discuss opportunities and challenges posed by ad hoc wireless communication networks.

5.1   Introduction

To meet the growing need for fast and reliable information exchange, communications networks have become an integral part of our society. Recent advancements in information technology and miniaturization of mobile communication devices have increased the use and efficiency of the wireless communication environment manifold. Because of this significant growth, wireless networks have witnessed rapid changes and development of new applications. One such change is the development of mobile ad hoc wireless networks [4,8].

Wireless networks have been in existence for centuries, but the medium of communication and its usage mechanisms have changed throughout history. Earlier wireless networks could only use audible and visual means of communication such as human voice, smoke signals, and/or reflective surfaces for conveying the information. With the advancement of technology, new and efficient means (electromagnetic and optical wireless) of communication and efficient mechanisms for their use are now available.

Ad hoc wireless networks have communication devices without a fixed topological infrastructure. They are self-organizing and adaptive [5]. As the devices move the network topology changes. The devices in these networks are not only source and destination of information exchanged, but they act as intermediate devices

to relay information between devices located outside the communication range. These networks are characterized by dynamic topologies, bandwidth-constrained variable capacity wireless links, energy-constrained operation, and limited physical security [4,11].

Ad hoc networks have tremendous potential in commercial and military applications [8,9,11,15]. These networks are particularly useful for providing communication support where no fixed infrastructure exists, or where the deployment of a fixed infrastructure is economically not feasible. They can be used in the area of military operations, emergency situations, healthcare, academic settings, and home networking. Examples of disaster situations are: earthquake and flooding, when the rescue teams need to coordinate themselves without the availability of fixed networks; military operations, when communication is in a hostile environment; businesses, where employees share information in a conference; students using laptop computers to participate in an interactive lecture; and many other similar situations. These networks meet a temporary networking need for a specific duration of time; when the need disappears, so do the networks.

Ensuring effective communication among the devices is one of the major challenges in wireless ad hoc networks. Because of the limited range of each device's wireless transmission, a device needs to act as a relay station to communicate with devices outside its transmission range by forwarding packets to the destination. Therefore, a routing protocol that addresses a diverse range of issues such as low bandwidth, mobility, and low power consumption is necessary in ad hoc wireless networks. The mobile devices work together in a dynamic but cooperative environment to maintain the communication channels.

It is possible for the mobile devices in an ad hoc wireless network to drift and lose communication with other mobile devices. In this situation, an ad hoc network may be divided into two or more independent ad hoc networks. Moreover, it is also possible when mobile devices in two or more ad hoc networks are in close proximity to each other, they can fuse together into one larger ad hoc network. One can imagine the challenges of managing such a dynamic communication environment [2,12,13].

Sensor networks are another form of ad hoc wireless networks. In sensor networks, the mobile devices could be as small as a grain of rice and their numbers in the tens of thousands. These devices are self-sufficient in transmitting, receiving, processing, and power. The sensors are programmable for any given application [8].

This chapter discusses opportunities and challenges faced in development of ad hoc wireless networks. The next section discusses application and opportunities that ad hoc wireless networks provide. Section "Challenges" discusses the challenges of using ad hoc wireless networks. Finally, Section "Summary and Conclusions" provides a summary and conclusions for this chapter.

5.2   Applications and Opportunities

Ad hoc wireless networks are used when a networking environment is needed for a limited duration of time. These networks provide significant opportunities and are used in numerous situations where a communication infrastructure is nonexistent or difficult to establish within timing constraints. Typically, these applications include:

- Search and rescue applications in disaster situations
- Defense (army, navy, air force) applications
- Healthcare applications
- Academic environment applications
- Industrial/corporate environment applications

Many other applications can utilize ad hoc wireless networks. However, in this chapter, we will only focus on five listed above [6].

Search and Rescue Applications

When we face an unfortunate situation such as an earthquake, hurricane, or similar disasters, ad hoc wireless networks can be very useful in search and rescue operations. In populated areas, disasters deplete power and communication capabilities because they destroy infrastructures. Ad hoc wireless networks can be established

without infrastructures and provide communication among various relief organizations for coordination of rescue operations. Wireless sensor networks (another form of ad hoc network) can be used to conduct search for survivors and provide care in a timely manner.

Rescue operations also use robots to search for survivors. These robots communicate with each other using wireless ad hoc networks to coordinate their activities. Based on the size of area affected by a disaster, robots can expedite the search by forming an ad hoc network, searching the area, and gathering information. The information gathered can be analyzed and processed, and appropriate relief/help can be readily directed where needed.

Defense Applications

Secure communications is one of the key aspects of any successful defense operation. Many defense operations take place in locations where communication infrastructure is not available. The use of wireless ad hoc and sensor networks in these situations are very useful and expedient. Different units (army, navy, and air force) involved in defense operations also need to maintain communication with each other. Air force planes flying information may establish an ad hoc wireless network for communicating, sharing images, and data among themselves. Army and navy groups on the move can also establish use ad hoc wireless networks. The advantage of this type of communication is that the ad hoc network moves with the individuals and equipment.

Information gathering is one of the many ad hoc wireless (particularly sensor) network applications. Intelligence gathering for defense purposes can make use of sensor networks very effectively. The sensors used for these applications are essentially disposable and are used only once. The sensors are deployed in large quantities to gather intelligence over a selected area by air or other appropriate means. Because of their tiny size, these sensors will remain suspended in the air for some time. While suspended, they can gather the information they have been programmed to collect, process the information, share among nearby sensors, reach a consensus, and transmit information to a central location. This information can then be analyzed at a central processing facility and a decision concerning the next step can be made. Sensor networks can also be used for tracking objects or targets, which is one of the critical applications in defense settings.

With rapid advancements in semiconductor technologies, the size of electronic devices is decreasing. At the same time, these devices are able to muster higher and higher processing power on tiny chips. These advancements led to the development of wearable computers. The idea of wearable computers is not new, but the idea of a smart dress, consisting of many tiny computers (or sensors), is relatively recent. A smart dress is essentially an ad hoc network of tiny computers that goes with you wherever you go because your wear it. Tiny computers connected by tiny wires or by wireless means, can exchange information with each other, process information, and take an action that they are programmed to do if all of the prerequisite conditions are satisfied. A smart dress may be programmed to monitor certain conditions and vital signs of an individual on a regular basis. This could become very useful for defense personnel in combat situation. The monitored information can be processed and appropriate action by your own dress, if needed. A smart dress may even be able to indicate the exact location of the problem, or call for help if needed.

Healthcare Applications

Exchanging multimedia (audio, video, and data) information between the patient and healthcare facilities is very helpful in critical and emergency situations. An individual in transit to a hospital by an ambulance may exchange information using ad hoc communication networks. In many situations, a healthcare professional can better diagnose and prepare a treatment plan for an individual if he has video information instead of audio or data information alone. For instance, the video information may be helpful in assessing the patient's reflexes and coordination. In addition, determining the degree of injury by visual information rather than by audio or other descriptive information alone can expedite treatment when the patient reaches the hospital.

Real-time ultrasound scans of a patient's kidneys, heart, or other organs may be very helpful in preparing a treatment plan for a patient who is in transit to a hospital, prior to his/her arrival. This information can be

transmitted through wireless communication networks, from an ambulance to a hospital, or to healthcare professionals who are converging toward the hospital from multiple locations to treat the patient.

An ad hoc wireless network established within a home (smart homes) can also be very useful for monitoring homebound patients. Such homes may be able to make some basic decisions (based on information exchanged between various sensors participating in an ad hoc network) that are beneficial to an elderly population. Some of the actions that smart homes can take include monitoring the movement patterns inside a home, recognizing when a person has fallen, recognizing an unusual situation, and informing a relevant agency so that appropriate help can be provided.

The concept of the smart dress discussed in the subsection on defense applications can also be used for monitors health conditions of patients. Such dresses may become very useful for providing healthcare for our elderly population.

Academic Environment Applications

Most academic institutions already have, or are in process of establishing, wireless communication networks. These networks provide students and faculty with a convenient environment in which to interact and accomplish their studies or research. In this setting, ad hoc wireless networks can enhance this type of environment and add many attractive features. For instance, an ad hoc wireless communication network between an instructor and the students enrolled in his/her class can provide an easy and convenient mechanism for the instructor to distribute handouts to all the students in the class, and for students to submit their assignments. Sharing information among the class participants can be as easy as click of a key on the keyboard. Because of the mobility of ad hoc wireless networks, they can be established while on field trips and industrial visits. Staying in touch cannot be any easier than this.

Industrial/Corporate Environment Applications

Most industrial/corporate sites have wireless communication networks in place, particularly in manufacturing environments. Manufacturing facilities, in general, have numerous electronic devices that are interconnected. Having wired connectivity leads to cluttering and crowding of space and poses not only safety hazards, but also adversely affects reliability. Using wireless communication networks eliminates many safety and spatial concerns. Connectivity in the form of ad hoc wireless communication networks adds many attractive aspects, including mobility. These devices can be relocated, and the networks reconfigured based on their requirements. Additionally, communication among various entities can be maintained, and corporate meetings can take place without employees gathering in the same room.

5.3   Challenges

Although ad hoc wireless networking represents technological advancements, there are many challenges associated with fully using its benefits. As with all mobile communication environments, ad hoc wireless communications operate with the following constraints:

- Limited communication bandwidth and capacity
- Limited battery power and life
- Size of the mobile devices
- Information security
- Communication overhead

Wireless communication operates with limited bandwidth, which implies only a limited amount of information can be transmitted over a period of time. Efficient transmission techniques will pave the way for increased capacity. However, there is a need for innovative approaches to optimal use of available bandwidth and capacity. The concept of cellular communication structures and the use of transmission techniques such as

CDMA are very helpful. Additional research is needed to provide more efficient mechanisms for using the available communication bandwidth in a wireless communication environment [8].

Mobile communication devices do not have access to unlimited power. They use batteries and have a limited supply of power. Higher power usage shortens the battery life. Efforts are being made to design a device that consumes less power and adjusts the strength of communication signals based on the distance between communicating points. In addition, efficient signal processing techniques and algorithms requiring less power usage are being developed. We have made significant progress in these areas, but more research is needed to make it even better [4,7,10,16].

With the advancements in semiconductor technologies, more electronic components can be placed on smaller chips. That has led to the development of mobile devices that are more powerful and less power hungry. As the size of these mobile devices decreases, more features can be added to these devices without increased power usage. The challenge is to maintain that trend.

Wireless communication environments are more prone to security risks than other communication networks, and ad hoc wireless networks are no exception. The level of desired information security can be achieved, but it adds processing overhead and requires additional bandwidth for transmission. Researchers are working to discover mechanisms that will provide secure information transfer and at the same time will not add prohibitive overhead [9,10,18].

Reducing the communication overhead for transferring information in ad hoc wireless communication networks is one of the biggest and most formidable challenges. When information needs to be transmitted from one node to another, a route or path needs to be established. In addition, some procedure for sharing the common pool of resources such as bandwidth has to be established. Bluetooth technology and IEEE 802.11 protocols provide mechanisms for sharing the resources [1,2,14,15]. In addition, there were several routing mechanisms proposed to establish a route between two communicating devices. The challenge that ad hoc wireless networks pose is that they have dynamic topology. In order to establish a route between two communicating devices, the network components need to be aware of their location. To make things complicated, mobile devices may keep changing their locations. The procedures for establishing routes need to be dynamic and adaptive. The route established at the start of information transfer between two devices may change by the time the information reaches its destination. Therefore, routing information needs to be as current as possible all the time [7,12,13,16,17].

There are several possibilities for establishing and maintaining routes. Routes can be established proactively or on-demand. The procedures that establish routes proactively incur more overhead because the establishment of all routes may not be necessary. If routes are established proactively and frequently, they will be current and immediately available to any communication device that needs to send information to another device. If routes are established on-demand, the overhead incurred will be less because the routes will be established as needed. However, the on-demand routing mechanism will introduce delay for devices because they will have to wait for an established route before initiating communication. Many hybrid routing mechanisms have been proposed. The challenge remains establishing the best possible routes with the least possible overhead [8,9].

5.4   Summary and Conclusions

This chapter has discussed opportunities and challenges related to ad hoc wireless communication networks. Development of ad hoc wireless networks and sensor networks are useful in many areas including disaster recovery, defense, healthcare, academic, and industrial environments. However, there are many other challenges. These include the development of mechanisms that efficiently use limited bandwidth and communication capacity, mechanisms for reducing power consumption and extending the battery life, developing smaller and more powerful mobile devices, developing algorithms for enhancing information security, and developing efficient routing procedures. These are major challenges to overcome, but steady progress is being made to address them.

References

1. "Bluetooth Wireless Technology," Official Bluetooth Web site: http://www.bluetooth.com.
2. S. Chakrabarti and A. Mishra, "QoS issues in ad hoc wireless networks," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 142–148, 2001.
3. I. Chlamtac, M. Conti, and J. Liu, "Mobile ad hoc networking: imperatives and challenges," *Ad Hoc Networks*, vol. 1, no. 1, pp. 13–64, 2003.
4. A. Ephermides, "Energy concerns in wireless networks," *IEEE Mag. Wireless Commun.*, vol. 9, no. 4, pp. 48–59, 2002.
5. M. Frodigh, "Wireless ad hoc networking — the art of networking without a network," *Ericsson Rev.*, 2001.
6. B. Furht and M. Ilyas, Eds., *Wireless Internet Handbook: Technologies, Standards, and Applications*, Boca Raton, FL: CRC Press, 2003.
7. A.J. Goldsmith and S.B. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Commun.*, vol. 9, no. 4, pp. 8–27, 2002.
8. M. Ilyas, Ed., *The Handbook of Ad Hoc Wireless Networks*, Boca Raton, FL: CRC Press, 2003.
9. M. Ilyas and I. Mahgoub, Eds., *Mobile Computing Handbook*, Boca Raton, FL: Auerbach Publications, 2005.
10. P. Papadimitratos and Z.J. Haas, "Securing mobile ad hoc networks," in *Mobile Computing Handbook*, M. Ilyas and I. Mahgoub, Eds., Boca Raton, FL: Auerbach Publications, 2005.
11. C. Perkins, *Ad Hoc Networking*, Reading, MA: Addison-Wesley, 2001.
12. R. Sankar, "Routing and mobility management in wireless ad hoc networks," in *Mobile Computing Handbook*, M. Ilyas and I. Mahgoub, Eds., Boca Raton, FL: Auerbach Publications, 2005.
13. Y. Shu, O. Yang, and L. Wang, "Adaptive routing in ad hoc networks," in *The Handbook of Ad Hoc Wireless Networks*, M. Ilyas, Ed., Boca Raton, FL: CRC Press, 2003.
14. C-K. Toh, *Ad Hoc Mobile Wireless Networks: Protocols and Systems*, Englewood Cliffs, NJ: Prentice Hall, 2002.
15. C-K. Toh, M. Delawar, and D. Allen, "Evaluating the communication performance of an ad hoc wireless networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 3, pp. 402–414, 2002.
16. C-K. Toh, "Maximum battery life routing to support ubiquitous mobile computing in wireless ad hoc networks," *IEEE Commun. Mag.*, 39, 138–147, 2001.
17. J. Wu and F. Dai, "Broadcasting in ad hoc networks based on self-pruning," in *Proc. IEEE INFOCOM*, 2003, pp. 2240–2250.
18. C-H. Yeh, "Protection and restoration in ad hoc wireless networks," *IEEE 10th Int. Conf. Networks*, August, 266–273, 2002.

6
Information Theory

H. Vincent Poor
Princeton University

Carl G. Looney
University of Nevada

Robert J. Marks II
University of Washington

Sergio Verdú
Princeton University

Joy A. Thomas
Stratify

Thomas M. Cover
Stanford University

6.1 Signal Detection

H. Vincent Poor

The field of signal detection and estimation is concerned with the processing of information-bearing signals for the purpose of extracting the information they contain. The applications of this methodology are quite broad, ranging from areas of electrical engineering such as automatic control, digital communications, image processing, and remote sensing, into other engineering disciplines and the physical, biological, and social sciences.

There are two basic types of problems of interest in this context. *Signal detection* problems are concerned primarily with situations in which the information to be extracted from a signal is discrete in nature. That is, signal detection procedures are techniques for deciding among a discrete (usually finite) number of possible alternatives. An example of such a problem is the demodulation of a digital communication signal, in which the task of interest is to decide which of several possible transmitted symbols has elicited a given received signal. *Estimation* problems, on the other hand, deal with the determination of some numerical quantity taking values in a continuum. An example of an estimation problem is that of determining the phase or frequency of the carrier underlying a communication signal.

Although signal detection and estimation is an area of considerable current research activity, the fundamental principles are quite well developed. These principles, which are based on the theory of statistical inference, explain and motivate most of the basic signal detection and estimation procedures used in practice. In this section, we will give a brief overview of the basic principles underlying the field of signal detection. A more complete introduction to these subjects is found in Poor [1994].

General Considerations

The basic principles of signal detection can be conveniently discussed in the context of decision-making between two possible statistical models for a set of real-valued measurements, $Y_1, Y_2, \ldots, Y_n$. In particular, on observing $Y_1, Y_2, \ldots, Y_n$, we wish to decide whether these measurements are most consistent with the model

$$Y_k = N_k, \qquad k = 1, 2, \ldots, n \tag{6.1}$$

or with the model

$$Y_k = N_k + S_k, \qquad k = 1, 2, \ldots, n \tag{6.2}$$

where $N_1, N_2, \ldots, N_n$ is a random sequence representing noise, and where $S_1, S_2, \ldots, S_n$ is a sequence representing a (possibly random) signal.

In deciding between Equation (6.1) and Equation (6.2), there are two types of errors possible: a *false alarm,* in which Equation (6.2) is falsely chosen, and a *miss,* in which Equation (6.1) is falsely chosen. The probabilities of these two types of errors can be used as performance indices in the optimization of rules for deciding between Equation (6.1) and Equation (6.2). Obviously, it is desirable to minimize both of these probabilities to the extent possible. However, the minimization of the **false-alarm probability** and the minimization of the **miss probability** are opposing criteria. So, it is necessary to effect a trade-off between them in order to design a signal detection procedure. There are several ways of trading off the probabilities of miss and false alarm: the **Bayesian detector** minimizes an average of the two probabilities taken with respect to prior probabilities of the two conditions Equation (6.1) and Equation (6.2), the *minimax* detector minimizes the maximum of the two error probabilities, and the **Neyman-Pearson detector** minimizes the miss probability under an upper-bound constraint on the false-alarm probability.

If the statistics of noise and signal are known, the Bayesian, minimax, and Neyman-Pearson detectors are all of the same form. Namely, they reduce the measurements to a single number by computing the **likelihood ratio**

$$L(Y_1, Y_2, \ldots, Y_n) \triangleq \frac{p_{S+N}(Y_1, Y_2, \ldots, Y_n)}{p_N(Y_1, Y_2, \ldots, Y_n)} \tag{6.3}$$

where p_{S+N} and p_N denote the probability density functions of the measurements under signal-plus-noise (Equation (6.2)) and noise-only (Equation (6.1)) conditions, respectively. The likelihood ratio is then compared to a *decision threshold*, with the signal-present model (Equation (6.2)) being chosen if the threshold is exceeded, and the signal-absent model (Equation (6.1)) being chosen otherwise. Choice of the decision threshold determines a trade-off of the two error probabilities, and the optimum procedures for the three criteria mentioned above differ only in this choice.

There are several basic signal detection structures that can be derived from Equation (6.1) to Equation (6.3) under the assumption that the noise sequence consists of a set of independent and identically distributed (i.i.d.) Gaussian random variables with zero means. Such a sequence is known as **discrete-time white Gaussian noise**. Thus, until further notice, we will make this assumption about the noise. It should be noted that this assumption is physically justifiable in many applications.

Detection of Known Signals

If the signal sequence S_1, S_2,..., S_n is known to be given by a specific sequence, say s_1, s_2,..., s_n (a situation known as *coherent detection*), then the likelihood ratio (Equation (6.3)) is given in the white Gaussian noise case by

$$\exp\left\{\left(\sum_{k=1}^{n} s_k Y_k - \frac{1}{2}\sum_{k=1}^{n} s_k^2\right)\Big/\sigma^2\right\} \tag{6.4}$$

where σ^2 is the variance of the noise samples. The only part of Equation (6.4) that depends on the measurements is the term $\sum_{k=1}^{n} s_k Y_k$ and the likelihood ratio is a monotonically increasing function of this quantity. Thus, optimum detection of a coherent signal can be accomplished via a correlation detector, which operates by comparing the quantity

$$\sum_{k=1}^{n} s_k Y_k \tag{6.5}$$

to a threshold, announcing signal presence when the threshold is exceeded.

Note that this detector works on the principle that the signal will correlate well with itself, yielding a large value of Equation (6.5) when present, whereas the random noise will tend to average out in the sum Equation (6.5), yielding a relatively small value when the signal is absent. This detector is illustrated in Figure 6.1.

Detection of Parametrized Signals

The correlation detector cannot usually be used directly unless the signal is known exactly. If, alternatively, the signal is known up to a short vector θ of random parameters (such as frequencies or phases) that are independent of the noise, then an optimum test can be implemented by threshold comparison of the quantity

$$\int_{\Lambda} \exp\left\{\left(\sum_{k=1}^{n} s_k(\boldsymbol{\theta})Y_k - \frac{1}{2}\sum_{k=1}^{n} [s_k(\boldsymbol{\theta})]^2\right)\Big/\sigma^2\right\} p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{6.6}$$

where we have written $S_k = s_k(\boldsymbol{\theta})$ to indicate the functional dependence of the signal on the parameters, and where Λ and $p(\boldsymbol{\theta})$ denote the range and probability density function, respectively, of the parameters.
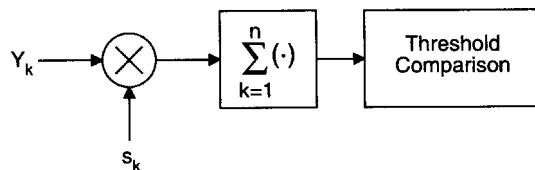

FIGURE 6.1 Correlation detector for a coherent signal in additive white Gaussian noise.

The most important example of such a parametrized signal is that in which the signal is a modulated sinusoid with random phase; i.e.,

$$S_k = a_k \cos(\omega_c k + \theta), \qquad k = 1, 2, \ldots, n \tag{6.7}$$

where $a_1, a_2, \ldots, a_n$ is a known amplitude modulation sequence, ω_c is a known (discrete-time) carrier frequency, and the random phase θ is uniformly distributed in the interval $[-\pi, \pi]$. In this case, the likelihood ratio is a monotonically increasing function of the quantity

$$\left[\sum_{k=1}^{n} a_k \cos(\omega_c k) Y_k \right]^2 + \left[\sum_{k=1}^{n} a_k \sin(\omega_c k) Y_k \right]^2 \tag{6.8}$$

Thus, optimum detection can be implemented via comparison of Equation (6.8) with a threshold, a structure known as an **envelope detector**. Note that this detector correlates the measurements with two orthogonal components of the signal, $a_k \cos(\omega_c k)$ and $a_k \sin(\omega_c k)$. These two correlations, known as the in-phase and quadrature components of the measurements, respectively, capture all of the energy in the signal, regardless of the value of θ. Since θ is unknown, however, these two correlations cannot be combined coherently, and thus they are combined noncoherently via Equation (6.8) before the result is compared with a threshold. This detector is illustrated in Figure 6.2.

Parametrized signals also arise in situations in which it is not appropriate to model the unknown parameters as random variables with a known distribution. In such cases, it is not possible to compute the likelihood ratio (Equation (6.6)) so an alternative to the likelihood ratio detector must then be used. (An exception is that in which the likelihood ratio detector is invariant to the unknown parameters—a case known as *uniformly most powerful detection*.) Several alternatives to the likelihood ratio detector exist for these cases.

One useful such procedure is to test for the signal's presence by threshold comparison of the *generalized likelihood ratio*, given by

$$\max_{\theta \in \Lambda} L_\theta(Y_1, Y_2, \ldots, Y_n) \tag{6.9}$$

where L_θ denotes the likelihood ratio for Equation (6.1) and Equation (6.2) for the known-signal problem with the parameter vector fixed at θ. In the case of white Gaussian noise, we have

$$L_\theta(Y_1, Y_2, \ldots, Y_n) = \exp\left\{ \left(\sum_{k=1}^{n} s_k(\boldsymbol{\theta}) Y_k - \frac{1}{2} \sum_{k=1}^{n} [s_k(\boldsymbol{\theta})]^2 \right) \Big/ \sigma^2 \right\} \tag{6.10}$$

It should be noted that this formulation is also valid if the statistics of the noise have unknown parameters, e.g., the noise variance in the white Gaussian case.

One common application in which the generalized likelihood ratio detector is useful is that of detecting a signal that is known except for its time of arrival. That is, we are often interested in signals parametrized as

$$s_k(\theta) = a_{k-\theta} \tag{6.11}$$

where $\{a_k\}$ is a known finite-duration signal sequence and where θ ranges over the integers. Assuming white Gaussian noise and an observation interval much longer than the duration of $\{a_k\}$, the generalized likelihood ratio detector in this case announces the presence of the signal if the quantity

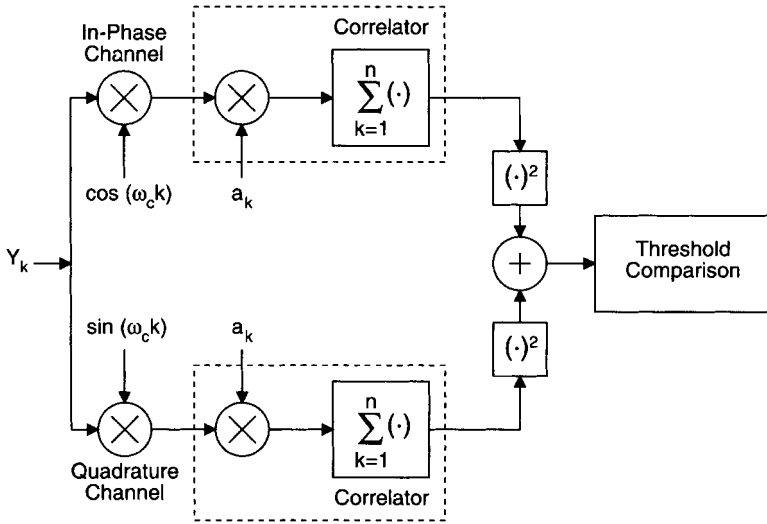$$\max_{\theta} \sum_{k} a_{k-\theta} Y_k \tag{6.12}$$

FIGURE 6.2 Envelope detector for a noncoherent signal in additive white Gaussian noise.

exceeds a fixed threshold. This type of detector is known as a *matched filter*, since it can be implemented by filtering the measurements with a digital filter whose pulse response is a time-reversed version of the known signal $\{a_k\}$ (hence it is "matched" to the signal), and announcing the signal's presence if the filter output exceeds the decision threshold at any time.

Detection of Random Signals

In some applications, particularly in remote sensing applications such as sonar and radio astronomy, it is appropriate to consider the signal sequence $S_1, S_2, \ldots, S_n$ itself to be a random sequence, statistically independent of the noise. In such cases, the likelihood ratio formula of Equation (6.6) is still valid with the parameter vector θ simply taken to be the signal itself. However, for long measurement records (i.e., large n), Equation (6.6) is not a very practical formula except in some specific cases, the most important of which is the case in which the signal is Gaussian.

In particular, if the signal is Gaussian with zero-mean and autocorrelation sequence $r_{k,l} \overset{\Delta}{=} E\{S_k S_l\}$, then the likelihood ratio is a monotonically increasing function of the quantity

$$\sum_{k=1}^{n} \sum_{l=1}^{n} q_{k,l} Y_k Y_l \tag{6.13}$$

with $q_{k,l}$ the element in the kth row and lth column of the positive-definite matrix

$$\mathbf{Q} \overset{\Delta}{=} \mathbf{I} - (\mathbf{I} + \mathbf{R}/\sigma^2)^{-1} \tag{6.14}$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix, and $\mathbf{R}$ is the covariance matrix of the signal, i.e., it is the $n \times n$ matrix with elements $r_{k,l}$.

Note that Equation (6.13) is a quadratic function of the measurements; thus, a detector based on the comparison of this quantity to a threshold is known as a **quadratic detector.** The simplest form of this detector results from the situation in which the signal samples are, like the noise samples, i.i.d. In this case, the

quadratic function (Equation (6.13)) reduces to a positive constant multiple of the quantity

$$\sum_{k=1}^{n} Y_k^2 \tag{6.15}$$

A detector based on Equation (6.15) simply measures the energy in the measurements and then announces the presence of the signal if this energy is large enough. This type of detector is known as a *radiometer*.

Thus, radiometry is optimum in the case in which both signal and noise are i.i.d. Gaussian sequences with zero means. Since in this case the presence of the signal is manifested only by an increase in energy level, it is intuitively obvious that radiometry is the only way of detecting the signal's presence. More generally, when the signal is correlated, the quadratic function (Equation (6.13)) exploits both the increased energy level and the correlation structure introduced by the presence of the signal. For example, if the signal is a narrowband Gaussian process, then the quadratic function (Equation (6.13)) acts as a narrowband radiometer with bandpass characteristic that approximately matches that of the signal. In general, the quadratic detector will make use of whatever spectral properties the signal exhibits.

If the signal is random but not Gaussian, then its optimum detection (described by Equation (6.6)) typically requires more complicated nonlinear processing than the quadratic processing of Equation (6.13) in order to exploit the distributional differences between signal and noise. This type of processing is often not practical for implementation, and thus approximations to the optimum detector are typically used. An interesting family of such detectors uses cubic or quartic functions of the measurements, which exploit the higher-order spectral properties of the signal [Nikias and Petropulu, 1993]. As with deterministic signals, random signals can be parametrized. In this case, however, it is the distribution of the signal that is parametrized. For example, the power spectrum of the signal of interest may be known only up to a set of unknown parameters. Generalized likelihood ratio detectors (Equation (6.9)) are often used to detect such signals.

Deciding Among Multiple Signals

The preceding results have been developed under the model (Equations 6.1–6.2) that there is a single signal that is either present or absent. In digital communications applications, it is more common to have the situation in which we wish to decide between the presence of two (or more) possible signals in a given set of measurements. The foregoing results can be adapted straightforwardly to such problems. This can be seen most easily in the case of deciding among known signals. In particular, consider the problem of deciding between two alternatives:

$$Y_k = N_k + s_k^{(0)}, \quad k = 1, 2, \ldots, n \tag{6.16}$$

and

$$Y_k = N_k + s_k^{(1)}, \quad k = 1, 2, \ldots, n \tag{6.17}$$

where $s_1^{(0)}, s_2^{(0)}, \ldots, s_n^{(0)}$ and $s_1^{(1)}, s_2^{(1)}, \ldots, s_n^{(1)}$ are two known equi-energy signals. Such problems arise in data transmission problems, in which the two signals $s^{(0)}$ and $s^{(1)}$ correspond to the waveforms received after transmission of a logical "zero" and "one," respectively. In such problems, we are generally interested in minimizing the *average probability of error,* which is the average of the two error probabilities weighted by the prior probabilities of occurrence of the two signals. This is a Bayesian performance criterion, and the optimum decision rule is a straightforward extension of the correlation detector based on Equation (6.5). In particular, under the assumptions that the two signals are equally likely to occur prior to measurement, and that the noise is white and Gaussian, the optimum decision between Equation (6.16) and Equation (6.17) is to choose the model (Equation (6.16)) if $\sum_{k=1}^{n} s_k^{(0)} Y_k$ is larger than $\sum_{k=1}^{n} s_k^{(1)} Y_k$, and to choose the model (Equation (6.17)) otherwise.

More generally, many problems in digital communications involve deciding among M equally likely signals with $M > 2$. In this case, again assuming equi-energy signals and white Gaussian noise, the decision rule that minimizes the error probability is to choose the signal $s_1^{(j)}, s_2^{(j)}, \ldots, s_n^{(j)}$, where j is a solution of the maximization problem

$$\sum_{k=1}^{n} s_k^{(j)} Y_k = \max_{0 \leqslant m \leqslant M-1} \sum_{k=1}^{n} s_k^{(m)} Y_k \qquad (6.18)$$

There are two basic types of digital communications applications in which the problem (Equation (6.18)) arises. One is in *M-ary data transmission,* in which a symbol alphabet with M elements is used to transmit data, and a decision among these M symbols must be made in each symbol interval [Proakis, 2000]. The other type of application in which (Equation (6.18)) arises is that in which data symbols are correlated in some way because of intersymbol interference, coding, or multiuser transmission. In such cases, each of the M possible signals represents a frame of data symbols, and a joint decision must be made about the entire frame since individual symbol decisions cannot be decoupled. Within this latter framework, the problem (Equation (6.18)) is known as *sequence detection.* The basic distinction between M-ary transmission and sequence detection is one of degree. In typical M-ary transmission, the number of elements in the signaling alphabet is typically a small power of 2 (say 8 or 32), whereas the number of symbols in a frame of data could be on the order of thousands. Thus, solution of Equation (6.18) by exhaustive search is prohibitive for sequence detection, and less complex algorithms must be used. Typical digital communications applications in which sequence detection is necessary admit dynamic programming solutions to Equation (6.18) (see, e.g., Poor [2002]).

Detection of Signals in More General Noise Processes

In the foregoing paragraphs, we have described three basic detection procedures: correlation detection of signals that are completely known, envelope detection of signals that are known except for a random phase, and quadratic detection for Gaussian random signals. These detectors were all derived under an assumption of white Gaussian noise. This assumption provides an accurate model for the dominant noise arising in many communication channels. For example, the thermal noise generated in signal processing electronics is adequately described as being white and Gaussian. However, there are also many channels in which the statistical behavior of the noise is not well described in this way, particularly when the dominant noise is produced in the physical channel rather than in the receiver electronics.

One type of noise that often arises is noise that is Gaussian but not white. In this case, the detection problem Equations (6.1)–(6.2) can be converted to an equivalent problem with white noise by applying a linear filtering process known as *prewhitening* to the measurements. In particular, on denoting the noise covariance matrix by Σ, we can write

$$\Sigma = \mathbf{C}\mathbf{C}^T \qquad (6.19)$$

where $\mathbf{C}$ is an $n \times n$ invertible, lower-triangular matrix and where the superscript T denotes matrix transposition. The representation (6.19) is known as the *Cholesky decomposition.* On multiplying the measurement vector $Y \overset{\triangle}{=} (Y_1, Y_2, \ldots, Y_n)^T$ satisfying Equations (6.1)–(6.2) with noise covariance Σ, by C^{-1}, we produce an equivalent (in terms of information content) measurement vector that satisfies the model Equations (6.1)–(6.2) with white Gaussian noise and with the signal conformally transformed. This model can then be treated using the methods described previously.

In other channels, the noise can be modeled as being i.i.d. but with an amplitude distribution that is not Gaussian. This type of model arises, for example, in channels dominated by impulsive phenomena, such as certain radio channels. In the non-Gaussian case the procedures discussed previously lose their optimality as defined in terms of the error probabilities. These procedures can still be used, and they will work well under

many conditions; however, there will be a resulting performance penalty with respect to optimum procedures based on the likelihood ratio. Generally speaking, likelihood-ratio-based procedures for non-Gaussian noise channels involve more complex nonlinear processing of the measurements than is required in the standard detectors, although the retention of the i.i.d. assumption greatly simplifies this problem. A treatment of methods for such channels can be found in Kassam [1988].

When the noise is both non-Gaussian and dependent, the methodology is less well developed, although some techniques are available in these cases. An overview can be found in Poor and Thomas [1993].

Robust and Nonparametric Detection

All of the procedures outlined above are based on the assumption of a known (possibly up to a set of unknown parameters) statistical model for signals and noise. In many practical situations it is not possible to specify accurate statistical models for signals or noise, and so it is of interest to design detection procedures that do not rely heavily on such models. Of course, the parametrized models described in the foregoing paragraphs allow for uncertainty in the statistics of the observations. Such models are known as *parametric* models, because the set of possible distributions can be parametrized by a finite set of real parameters.

While parametric models can be used to describe many types of modeling uncertainty, composite models in which the set of possible distributions is much broader than a parametric model would allow are sometimes more realistic in practice. Such models are termed *nonparametric models*. For example, one might be able to assume only some very coarse model for the noise, such as that it is symmetrically distributed. A wide variety of useful and powerful detectors have been developed for signal-detection problems that cannot be parametrized. These are basically of two types: *robust* and *nonparametric*. Robust detectors are those designed to perform well despite small, but potentially damaging, nonparametric deviations from a nominal parametric model, whereas nonparametric detectors are designed to achieve constant false-alarm probability over very wide classes of noise statistics.

Robustness problems are usually treated analytically via minimax formulations that seek best worst-case performance as the design objective. This formulation has proven to be very useful in the design and characterization of robust detectors for a wide variety of detection problems. Solutions typically call for the introduction of light limiting to prevent extremes of gain dictated by an (unrealistic) nominal model. For example, the correlation detector of Figure 6.1 can be made robust against deviations from the Gaussian noise model by introducing a soft-limiter between the multiplier and the accumulator.

Nonparametric detection is usually based on relatively coarse information about the observations, such as the algebraic signs or the ranks of the observations. One such test is the *sign test,* which bases its decisions on the number of positive observations obtained. This test is nonparametric for the model in which the noise samples are i.i.d. with zero median and is reasonably powerful against alternatives such as the presence of a positive constant signal in such noise. More powerful tests for such problems can be achieved at the expense of complexity by incorporating rank information into the test statistic.

Distributed and Sequential Detection

The detection procedures discussed in the preceding paragraphs are based on the assumption that all measurements can and should be used in the detection of the signal, and moreover that no constraints exist on how measurements can be combined. There are a number of applications, however, in which constraints apply to the information pattern of the measurements.

One type of constrained information pattern that is of interest in a number of applications is a network consisting of a number of distributed or local decision makers, each of which processes a subset of the measurements, and a *fusion center,* which combines the outputs of the distributed decision makers to produce a global detection decision. The communication between the distributed decision makers and the fusion center is constrained, so that each local decision maker must reduce its subset of measurements to a summarizing local decision to be transmitted to the fusion center. Such structures arise in applications such as the testing of large-scale integrated circuits, in which data collection is decentralized, or in detection problems

involving very large data sets, in which it is desirable to distribute the computational work of the detection algorithm or to partition the data for security reasons. Such problems lie in the field of *distributed detection.* Except in some trivial special cases, the constraints imposed by distributing the detection algorithm introduce a further level of difficulty into the design of optimum detection systems. Nevertheless, considerable progress has been made on this problem, a survey of which can be found in Tsitsiklis [1993].

Another type of nonstandard information pattern that arises is that in which the number of measurements is potentially infinite, but in which there is a cost associated with taking each measurement. This type of model arises in applications such as the synchronization of wideband communication signals. In such situations, the error probabilities alone do not completely characterize the performance of a detection system, since consideration must also be given to the cost of sampling. The field of *sequential detection* deals with the optimization of detection systems within such constraints. In sequential detectors, the number of measurements taken becomes a random variable depending on the measurements themselves. A typical performance criterion for optimizing such a system is to seek a detector that minimizes the expected number of measurements for given levels of miss and false-alarm probabilities.

The most commonly used sequential detection procedure is the *sequential probability ratio test,* which operates by recursive comparison of the likelihood ratio (Equation (6.3)) to two thresholds. In this detector, if the likelihood ratio for a given number of samples exceeds the larger of the two thresholds, then the signal's presence is announced and the test terminates. Alternatively, if the likelihood ratio falls below the smaller of the two thresholds, the signal's absence is announced and the test terminates. However, if neither of the two thresholds is crossed, then another measurement is taken and the test is repeated.

Detection with Continuous-Time Measurements

Note that all of the preceding formulations have involved the assumption of discrete-time (i.e., sampled-data) measurements. From a practical point of view, this is the most natural framework within which to consider these problems, since implementations most often involve digital hardware. However, the procedures discussed herein all have continuous-time counterparts, which are of both theoretical and practical interest. Mathematically, continuous-time detection problems are more difficult than discrete-time ones, because they involve probabilistic analysis on function spaces. The theory of such problems is quite elegant, and the interested reader is referred to Poor [1994] or Kailath and Poor [1998] for more detailed exposition.

Continuous-time models are of primary interest in the front-end stages of radio frequency or optical communication receivers. At radio frequencies, continuous-time versions of the models described in the preceding paragraphs can be used. For example, one may consider the detection of signals in continuous-time Gaussian white noise. At optical wavelengths, one may consider either continuous models (such as Gaussian processes) or point-process models (such as Poisson counting processes), depending on the type of detection used (see, e.g., Snyder and Miller [1991]). In the most fundamental analyses of optical detection problems, it is sometimes desirable to consider the quantum mechanical nature of the measurements [Helstrom, 1976].

Defining Terms

Bayesian detector: A detector that minimizes the average of the false-alarm and miss probabilities, weighted with respect to prior probabilities of signal-absent and signal-present conditions.

Correlation detector: The optimum structure for detecting coherent signals in the presence of additive white Gaussian noise.

Discrete-time white Gaussian noise: Noise samples modeled as independent and identically distributed Gaussian random variables.

Envelope detector: The optimum structure for detecting a modulated sinusoid with random phase in the presence of additive white Gaussian noise.

False-alarm probability: The probability of falsely announcing the presence of a signal.

Likelihood ratio: The optimum processor for reducing a set of signal-detection measurements to a single number for subsequent threshold comparison.

Miss probability: The probability of falsely announcing the absence of a signal.
Neyman-Pearson detector: A detector that minimizes the miss probability within an upper-bound constraint on the false-alarm probability.
Quadratic detector: A detector that makes use of the second-order statistical structure (e.g., the spectral characteristics) of the measurements. The optimum structure for detecting a zero-mean Gaussian signal in the presence of additive Gaussian noise is of this form.

References

C.W. Helstrom, *Quantum Detection and Estimation Theory*, New York: Academic Press, 1976.

T. Kailath and H.V. Poor, "Detection of stochastic processes", *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 2230–2259, 1998.

S.A. Kassam, *Signal Detection in Non-Gaussian Noise*, New York: Springer-Verlag, 1988.

C.L. Nikias and A. Petropulu, *Higher-Order Spectral Analysis*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

H.V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed., New York: Springer-Verlag, 1994.

H.V. Poor, "Dynamic programming in digital communications: Viterbi decoding to turbo multiuser detection," *Journal of Optimization Theory and Applications*, vol. 115, no. 3, pp. 629–657, 2002.

H.V. Poor and J.B. Thomas, "Signal detection in dependent non-Gaussian noise", in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.

J.G. Proakis, *Digital Communications*, 4th ed., New York: McGraw-Hill, 2000.

D.L. Snyder and M.I. Miller, *Random Point Processes in Time and Space*, New York: Springer-Verlag, 1991.

J. Tsitsiklis, "Distributed detection," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.

Further Information

Except as otherwise noted in the accompanying text, further details on the topics introduced in this section can be found in the textbook:

Poor, H.V. *An Introduction to Signal Detection and Estimation,* 2nd ed., New York: Springer-Verlag, 1994.

The monthly journal, *IEEE Transactions on Information Theory,* publishes recent advances in the theory of signal detection. It is available from the Institute of Electrical and Electronics Engineers, Inc., 445 Hoes Lane, Piscataway, NJ 08854.

Papers describing applications of signal detection are published in a number of journals, including the monthly journals *IEEE Transactions on Communications, IEEE Transactions on Signal Processing,* and the *Journal of the Acoustical Society of America*. The IEEE journals are available from the IEEE, as above. The *Journal of the Acoustical Society of America* is available from the American Institute of Physics, 2 Huntington Quadrangle, Melville, NY 11747.

6.2   Noise

Carl G. Looney

The physical processing of every information signal $s(t)$ corrupts the original signal by adding new fluctuations at the output that were not present at the input. The output signal can be decomposed into two components: the original signal and the added fluctuations, which are referred to as noise. The noise is undesirable because it limits the dynamic range of $s(t)$, and therefore effort is made to reduce it to the minimum theoretical value. These undesirable signals were termed **noise** due to early measurements with sensitive audio amplifiers.

Noise sources are: (1) *intrinsic*, (2) *external*, or (3) *process induced*. Intrinsic noise arises in active devices, passive components, and conductors. The largest noise contribution is from statistical fluctuations

in current flow in active devices, the next largest is from thermal energy of electrons in resistors, and the least is from microboundaries of impurities and grains with varying potential in conductors. Intrinsic sources such as Johnson (thermal) noise and shot noise are well understood, but the physical processes for additional layers of noise on top of these at low frequencies such as 1/*f* (one over *f*) or flicker noise, popcorn or burst noise are current topics of research. External interference sources can be either electromagnetic or electrostatic fields, or in the case or radio waves, both. Electromagnetic sources are the largest contributor and include power lines, incandescent lighting dimmer controls, brush type electric motors, welding apparatus, gasoline engines, switching power supplies, and any device that causes a large and rapid rate of change in current flow. Electrostatic sources include fluorescent lightning and power lines. Other sources include electromagnetic waves from radio transmitters, lightning, cosmic rays, plasmas (charged particles) in space, and solar/stellar radiation. Reflective objects and other macroboundaries cause multiple paths of transmitted signals and lead to delayed copies of the original signal being superimposed on *s*(*t*). Process-induced errors include measurement, quantization, truncation, and signal generation errors. These all can corrupt the signal by introducing a noise floor that masks *s*(*t*) below that level.

Noise can be divided into the categories random, impulse, and periodic. While intrinsic sources are principally random, most extrinsic sources are either periodic or impulse. In the case of periodic noise, time domain filtering can achieve large reductions of noise by subtracting a copy of the undesired noise. A typical extrinsic source is power line noise and will have spectra consisting of multiples of the line frequency that can be removed by comb filtering in the frequency domain. Impulse noise can occur randomly or periodically. In either case, time domain slew rate limiting can reduce the noise. For random noise, frequency domain filtering can remove those portions of noise spectra outside the needed spectra of *s*(*t*).

Statistics of Noise

Statistics allow us to analyze the spectra of noise. We model a noise signal by a **random** (or *stochastic*) **process** $N(t)$, a function with a realized value $N(t) = x_t$ at any time instant t is chosen by the outcome of the random variable $N_t = N(t)$. $N(t)$ has a probability distribution for the values x it can assume. Any particular trajectory $\{(t, x_t)\}$ of outcomes is called a **realization** of the noise process. The first-order statistic of $N(t)$ is the expected value $\mu_t = E[N(t)]$. The *second-order statistic* is the *autocorrelation function* $R_{NN}(t, t + \tau) = E[N(t)N(t + \tau)]$, where $E[-]$ is the expected value operator. **Autocorrelation** measures the extent to which noise random variables $N_1 = N(t_1)$ and $N_2 = N(t_2)$ at times t_1 and t_2 depend on each other in an average sense.

When the first- and second-order statistics do not change over time, we call the noise a **weakly** (or *wide-sense*) **stationary process**. This means that: (1) $E[N(t)] = \mu_t = \mu$ is constant for all t, and (2) $R_{NN}(t, t+\tau) = E[N(t)N(t+\tau)] = E[N(0)N(\tau)] = R_{NN}(\tau)$ for all t (see Brown, 1983, p. 82; Gardner, 1990, p. 108; or Peebles, 1987, p. 153 for properties of $R_{NN}(\tau)$). In this case, the autocorrelation function depends only on the *offset* τ. Therefore, we assume that $\mu = 0$ (we can subtract μ, which does not change the autocorrelation). When $\tau = 0$, $R_{NN}(0) = E[N(t)N(t + 0)] = E[(N(t))^2] = \sigma_N^2$, which is the fixed variance of each random variable N_t for all t. Weakly stationary (ws) processes are the most commonly encountered cases and are the ones considered here. *Evolutionary processes* have statistics that change over time and are difficult to analyze.

Figure 6.3 shows a realization of a noise process $N(t)$, where at any particular time t the probability density function projects out of the page in a third dimension. For a ws noise, the distributions are the same for each *t*. The most mathematically tractable noises are *Gaussian* ws processes, where at each time *t* the probability distribution for the random variable $N_t = N(t)$ is Gaussian (also called *normal*). The first- and second-order statistics completely determine Gaussian distributions, and therefore ws makes the statistics of all orders stationary over time also. It is well known [Brown, 1983, p. 39] that linear transformations of Gaussian random variables are also Gaussian random variables. The probability density function for a Gaussian random variable N_t is $f_N(x) = \{1/[2\pi\sigma_N^2]^{1/2}\exp[-(x - \mu_N)^2/2\sigma_N^2]\}$, which is the familiar bell-shaped curve centered on $x = \mu_N$. The standard Gaussian probability table [Peebles, 1987, p. 314] is useful, e.g., $\Pr[-\sigma_N < N_t < \sigma_N] = 2\Pr[0 < Nt < \sigma_N] = 0.8413$ from the table.
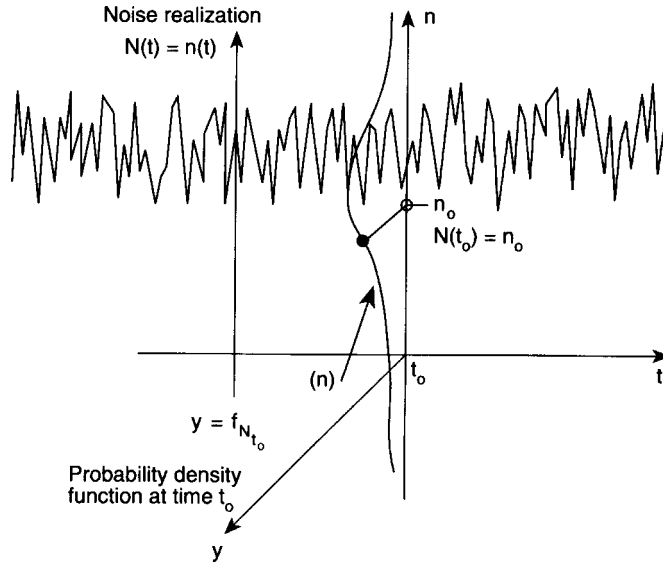
FIGURE 6.3   A noise process.

Noise Power

The noise signal $N(t)$ represents voltage, therefore the autocorrelation function at offset 0, $R_{NN}(0) = E[N(t)N(t)]$ represents expected power in volts squared or watts per ohm. When $R = 1$ Ω, then $N(t)N(t) = N(t)[N(t)/R] = N(t)I(t)$ volt-amperes = watts (where $I(t)$ is the current in a 1-Ω resistor). The Fourier transform $\mathbf{F}[R_{NN}(\tau)]$ of the autocorrelation function $R_{NN}(\tau)$ is the power spectrum, called the **power spectral density function** (psdf), $S_{NN}(w)$ in W/(rad/s). Then

$$S_{NN}(w) = \int_{-\infty}^{\infty} R_{NN}(\tau)e^{-jws}d\tau = \mathbf{F}[R_{NN}(\tau)]$$

$$R_{NN}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{NN}(w)e^{jws}dw = \mathbf{F}^{-1}[S_{NN}(w)]$$

(6.20)

The psdf at frequency f is defined as the expected power that the voltage $N(t)$, bandlimited to an incremental band df centered at f, would dissipate in a 1-Ω resistance, divided by df.

Equation (6.20), known as the *Wiener–Khinchin* relation, establishes that $S_{NN}(w)$ and $R_{NN}(\tau)$ are a Fourier transform pair for ws random processes [Brown, 1983; Gardner, 1990, p. 230; Peebles, 1987]. The psdf $S_{NN}(w)$ has units of W/(rad/s), whereas the autocorrelation function $R_{NN}(t)$ has units of watts. When $\tau = 0$ in the second integral of Equation (6.20), the exponential becomes $e^0 = 1$, so that $R_{NN}(0)(= E[N(t)^2] = \sigma_N^2)$ is the integral of the psdf $S_{NN}(w)$ over all radian frequencies, $-\infty < w < \infty$. The rms (root-mean-square) voltage is $N_{\text{rms}} = (\sigma_N$ (the standard deviation). The power spectrum in W/(rad/s) is a density that is summed up via an integral over the radian frequency band w_1 to w_2 to obtain the total power over that band.

$$P_{NN}(w_1, w_2) = \frac{1}{2\pi}\int_{w1}^{w2} S_{NN}(w) \cdot dw \text{ watts}$$

$$P_{NN} = \sigma_N^2 = E[N(t)^2] = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{NN}(w) \cdot dw \text{ watts}$$

(6.21)

The variance $\sigma_N^2 = R_{NN}(0)$ is the mean instantaneous power P_{NN} over all frequencies at any time t.

Effect of Linear Transformations on Autocorrelation and Power Spectral Density

When $h(t)$ is the impulse response function of a time-invariant linear system L, and $H(w) = \mathbf{F}[h(t)]$ is its transfer function, let the input noise signal $N(t)$ have the autocorrelation function $R_{NN}(\tau)$ and psdf $S_{NN}(w)$. Then we denote the output noise signal by $Y(t) = L[N(t)]$. The Fourier transforms $Y(w) \equiv \mathbf{F}[Y(t)]$ and $N(w) \equiv \mathbf{F}[N(t)]$ do not exist, but they are not needed. The output $Y(t)$ of a linear system is ws whenever the input $N(t)$ is ws [Gardner, 1990, p. 195; or Peebles, 1987, p. 215]. The output psdf $S_{YY}(w)$ and autocorrelation function $R_{YY}(\tau)$ are given by the following, respectively,

$$S_{YY}(w) = |H(w)|^2 S_{NN}(w), \quad R_{YY}(\tau) = \mathrm{F}^{-1}[S_{YY}(w)] \tag{6.22}$$

[see Gardner, 1990, p. 223]. The output noise power is

$$\sigma_Y^2 = P_{YY} = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{YY}(w)\,dw = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(w)|^2 S_{NN}(w)\,dw \tag{6.23}$$

White, Gaussian, and Pink Noise Models

White noise [Brown, 1983; Gardner, 1990, p. 234; or Peebles, 1987] is a theoretical model $W(t)$ of noise that is ws with zero mean. It has a constant power level with n_o overall frequencies (analogous to white light), so its psdf is $S_{WW}(w) = n_o$ W/(rad/s), $-\infty < w < \infty$. The inverse Fourier transform of this is the impulse function $R_{WW}(\tau) = (n_o)\delta(\tau)$, which is zero for all offsets except $\tau = 0$. Therefore, white noise $W(t)$ is a process that is uncorrelated over time, i.e., $E[W(t_1)W(t_2)] = 0$ for t_1 not equal to t_2. Figure 6.4(a) shows the autocorrelation and psdf for white noise where the offset is $s = \tau$. A *Gaussian white noise* is white noise such that the probability distribution of each random variable $W_t = W(t)$ is Gaussian. When two Gaussian random variables W_1 and W_2 are *uncorrelated*, i.e., $E[W_1 W_2] = 0$, they are independent [Gardner, 1990, p. 37]. We use Gaussian models because of the *central limit theorem* states that the sum of a number of random variables is approximately Gaussian.

Actual circuits attenuate signals above cutoff frequencies, and the power must be finite. However, for white noise, $P_{WW} = R_{NN}(0) = \infty$, so we often truncate the white noise spectral density (psdf) at cutoffs $-w_c$ to w_c. The result is known as band limited white noise, $P(t)$, and is usually taken to be Gaussian because linear filtering of any white noise (through the effect of the central limit theorem) tends to make the noise Gaussian [Gardner, 1990, p. 241]. Figure 6.4(b) shows the sinc function $R_{PP}(s) = \mathrm{F}^{-1}[S_{PP}(w)]$ for pink noise. Random variables P_1 and P_2 at times t_1 and t_2 are correlated only for t_1 and t_2 close.
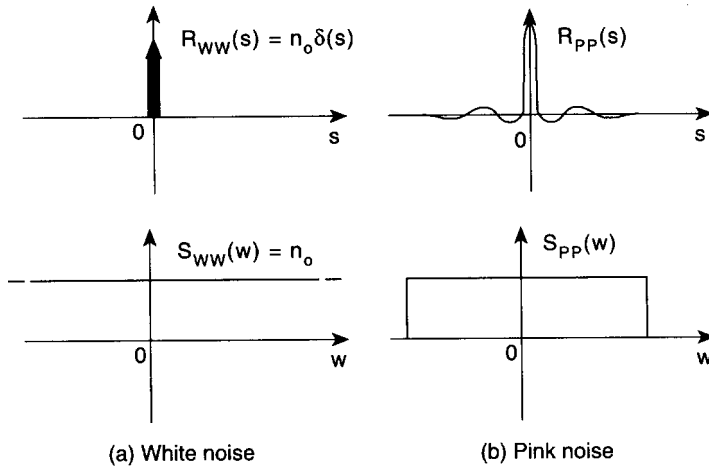
FIGURE 6.4 Power transform pairs for white and band limited white noise.

Thermal Noise as Gaussian White Noise

Brown observed in 1828 that pollen and dust particles moved randomly when suspended in liquid. In 1906, Einstein analyzed such motion based on the random walk model. Perrin confirmed in 1908 that the thermal activity of molecules in a liquid caused irregular bombardment of the much larger particles. It was predicted that charges bound to thermally vibrating molecules would generate electromotive force (emf) at the open terminals of a conductor, and that this placed a limit on the sensitivity of galvanometers. Thermal noise (also called *Johnson noise*) was first observed by J.B. Johnson at Bell Laboratories in 1927. Figure 6.5 displays white noise as seen in the laboratory on an oscilloscope.

The voltage $N(t)$ generated thermally between two points in an open circuit conductor is the sum of an extremely large number of superimposed, independent electronically, and ionically induced microvoltages at all frequencies up to $f_c = 6,000$ GHz at room temperature [Gardner 1990, p. 235] near infrared. The mean relaxation time of free electrons is $1/f_c = 0.5 \times 10^{-10}/T$ s, therefore at room temperature of $T = 290$K, it is 0.17 ps (1 picosecond $= 10^{-12}$ s). The values of $N(t)$ at different times are uncorrelated for time differences (offsets) greater than $\tau_c = 1/f_c$. The expected value of $N(t)$ is zero. The power is somewhat constant across a broad spectrum, and we cannot sample signals at picosecond periods, so we model Johnson noise $N(t)$ with Gaussian white noise $W(t)$. Although $\mu = E[W(t)] = 0$, the average power is positive at temperatures above 0K, and is $\sigma_W^2 = R_{WW}(0)$ (see the right side of Equation (6.21)). A disadvantage of the white noise model is its infinite power, i.e., $R_{WW}(0) = \sigma_W^2 = \infty$, but it is valid if band limited to B Hz, in which case its power is finite.

In 1927, Nyquist [1928] theoretically derived thermal noise power in a resistor to be

$$P_{WW}(B) = 4kTRB \text{ (watts)} \tag{6.24}$$

where R is resistance (ohms), B is the frequency bandwidth of measurement in Hz (all emf fluctuations outside of B are ignored), $P_{WW}(B)$ is the mean power over B (see Equation (6.21)), and Boltzmann's constant is $k = 1.38 \times 3 \ 10^{-23}$ J/K [Ott, 1988; Gardner, 1990, p. 288; or Peebles, 1987, p. 227]. Under external emf, the thermally induced collisions are the main source of resistance in conductors (electrons pulled into motion by an external emf at 0K meet no resistance). The rms voltage is $W_{\text{rms}} = \sigma_W = [(4kTRB)]^{1/2}$ V over a bandwidth of B Hz.

Planck's radiation law is $S_{NN}(w) = (2h|f|)/[\exp(h|f|/kT) - 1]$, where $h = 6.63 \times 10^{-34}$ J/s is Planck's constant, and f is the frequency [Gardner, 1990, p. 234]. For $|f|$ much smaller than $kT/h = 6.04 \times 10^{12}$ Hz $\approx 6,000$ GHz, the exponential above can be approximated by $\exp(h|f|/kT) = 1 + h|f|/kT$. The denominator of $S_{NN}(w)$ becomes

TIME, 200 MICROSECONDS PER DIVISION
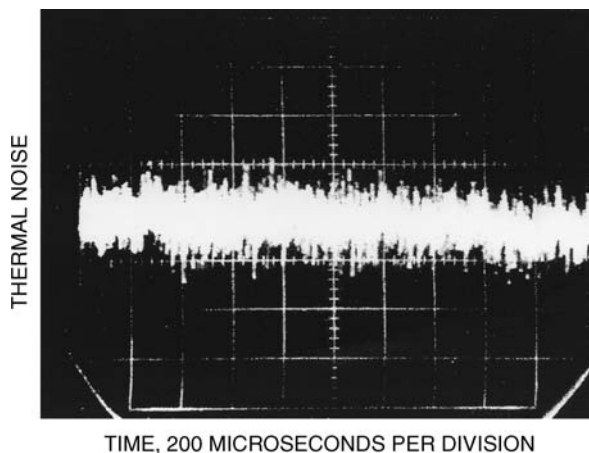
FIGURE 6.5 Thermal noise in a resistor. (*Source*: H.W. Ott, *Noise Reduction Techniques in Electronic Systems*, 2nd ed., New York: Wiley-Interscience, 1988, p. 203. With permission.)
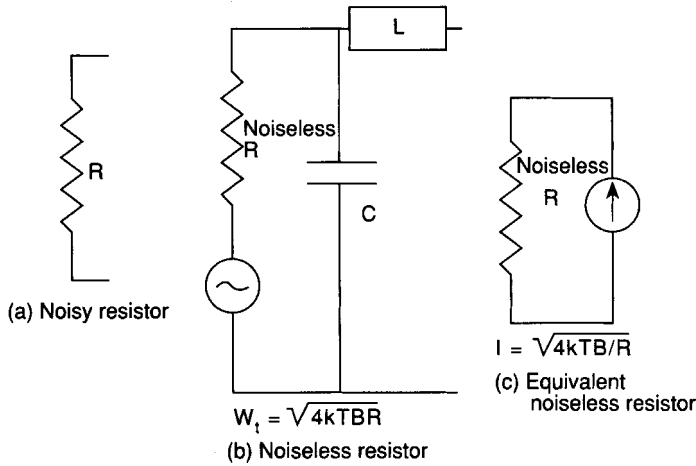
FIGURE 6.6   Thermal noise in a resistor.

$h|f|/kt$, so $S_{NN}(w) = (2h|f|)/(h|f|/kT) = 2kT$ W/Hz in a 1-Ω resistor. Over a resistance of R Ω and a bandwidth of B Hz (positive frequencies) yields the total power $P_{WW}(B) = 2BRS_{NN}(w) = 4kTRB$ W over the two-sided frequency spectrum. This is Nyquist's result.

Thermal noise is the same in a 1000-Ω carbon resistor as it is in a 1000-Ω tantalum thin-film resistor [Ott, 1988]. While the intrinsic noise may never be less, it may be higher because of other superimposed noise. We model the thermal noise in a resistor by an internal source (generator), as shown in Figure 6.6. Capacitance cannot be ignored at high f, but pure reactance (C or L) cannot dissipate energy, and therefore cannot generate thermal noise. The white noise model $W(t)$ for thermal noise $N(t)$ has a constant psdf $S_{WW}(w) = n_o$ W/(rad/s) for $-\infty < w < \infty$. By Equation (6.21), the white noise mean power over the frequency bandwidth B is

$$P_{WW}(B) = \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} S_{WW}(w) dw = n_o(4\pi B/2\pi) = 2n_oB \qquad (6.25)$$

Solving for the constant n_o, we obtain $n_o = P_{WW}(B)/2B$, which we put into Equation (6.20) to obtain the spectral density as a function of temperature and resistance using Nyquist's result above.

$$S_{WW}(w) = n_o = P_{WW}(B)/4\pi B = 4kTR2\pi B/4\pi B = 2kTR \text{ watts}/(\text{rad/s}) \qquad (6.26)$$

Some Examples

The parasitic capacitance in the terminals of a resistor may cause a roll-off of approximately 6 dB/octave in actual resistors [Brown, 1983, p. 139]. At 290K (room temperature), we have $2kT = 2 \times 1.38 \times 10^{-23} \times 290 = 0.8 \times 10^{-20}$ W/Hz due to each ohm [Ott, 1988]. For $R = 1$ MΩ (10^6 Ω), $S_{WW}(w) = 0.8 \times 10^{-14}$. Over a band of 10^8 Hz, we have $P_{WW}(B) = S_{WW}(w)B = 0.8 \times 10^{-14} \times 10^8 = 0.8 \times 10^{-6}$ W = 0.8 μW by Equation (6.24) and Equation (6.26). In practice, parasitic capacitance causes thermal noise to be bandlimited (bandlimited white noise). Now consider Figure 6.6(b), and let the temperature be 300K, $R = 10^6$ Ω, $C = 1$ pf (1 picofarad = 10^{-12} farads), and assume L is 0H. By Equation (6.21), the thermal noise power is

$$S_{WW}(w) = 2kTR = 2 \times 1.38 \times 10^{-23} \times 300 \times 10^6 = 828 \times 10^{-17} \text{ W/Hz}$$

The power across a bandwidth $B = 10^6$ is $P_{WW}(B) = S_{WW}(w)B = 8280 \times 10^{-12}$ W, so the rms voltage is $W_{\text{rms}} = [P_{WW}(B)]^{1/2} = 91$ μV.

Now let $Y(t)$ be the output voltage across the capacitor. The transfer function can be seen to be $H(w) = \{I(w)(1/jwC)\}/\{I(w)[R+(1/jwC)]\} = (1/jwC)/[R+1/jwC] = 1/[1+jwRC]$ (where $I(w)$ is the Fourier transform of the current). The output psdf (see Equation (6.22)) is

$$S_{YY}(w) = |H(w)|^2 S_{WW}(w) = (1/[1 + w^2 R^2 C^2]) S_{WW}(w)$$

Integrating $S_{YY}(w) = (1/[1+w^2 R^2 C^2]) S_{WW}(w)$ over all radian frequencies $w = 2\pi f$ (see Equation (6.21)), we obtain the antiderivative $(828 \times 10^{-17})(1/RC) \operatorname{atan}(RCw)/2\pi$. Upon substituting the limits $w = \pm\infty$, this becomes $828 \times 10^{-17}[\pi/2+\pi/2]/2\pi RC = 414 \times 10^{-17}(1/2RC) = 207 \times 10^{-17} \times 10^6 = 2070 \times 10^{-12}$ W/Hz. Then $\sigma_Y^2 = E[Y(t)^2] = P_{YY}(-\infty, \infty) = 2070 \times 10^{-12}$ W, so $Y_{\text{rms}}(t) = \sigma_Y = [P_{YY}(-\infty,\infty)]^{1/2} = 45.5~\mu$V. The half-power (cut-off) radian frequency is $w_c = 1/RC = 10^6$ rad/s, or $f_c = w_c/2\pi = 159.2$ kHz. Approximating $S_{YY}(w)$ by the rectangular spectrum $S_{YY}(w) = n_o$, $-10^6 < w < 10^6$ rad/s (0 elsewhere), we have that $R_{YY}(\tau) = (w_c/\pi)\operatorname{sinc}(w_c\tau)$, which has the first zeros at $|w_c\tau| = \pi$, that is $|\tau| = 1/(2f_c)$ (see Figure 6.4(b)). We approximate the autocorrelation by $R_{YY}(\tau) = 0$ for $|s| \geq 1/2f_c$.

Measuring Thermal Noise

In Figure 6.7, the thermal noise from a noisy resistor R is to be measured, where R_L is the measurement load. The incremental noise power in R over an incremental frequency band of width df is $P_{WW}(df) = 4kTR\,df$ W, by Equation (6.24). $P_{YY}(df)$ is the integral of $S_{YY}(w)$ over df by Equations (6.21), where $S_{YY}(w) = |H(w)|^2 S_{WW}(w)$, by Equation (6.22). In this case, the transfer function $H(w)$ is nonreactive and does not depend upon the radian frequency (we can factor it out of the integral). Thus,

$$P_{YY}(df) = \int_{-df}^{df} |H(f)|^2 (2kTR)df = \{R_L/R + R_L)^2\}(4kTR\,df)$$

To maximize the power measured, let $R_L = R$. The *incremental available power* measured is then $P_{YY}(df) = 4kTR^2 df/(4R^2) = kT\,df$ [Ott, 1988, p. 201; Gardner, 1990, p. 288; or Peebles, 1987, p. 227]. Therefore, we have the result that incremental available power over bandwidth df depends only on the temperature T.

$$P_{YY}(df) = kT\,df \quad (\text{output power over } df) \tag{6.27}$$
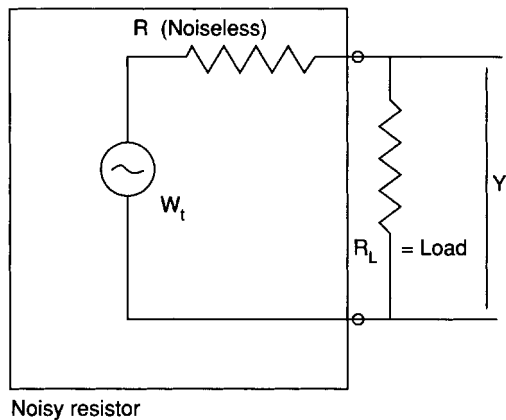
FIGURE 6.7   Measuring thermal noise voltage.

In 1906, Albert Einstein used statistical mechanics to postulate that the mean kinetic energy per degree of freedom of a particle, $(1/2)mE[v^2(t)]$, is equal to $(1/2)kT$, where: m is the mass of the particle, $v(t)$ is its instantaneous velocity in a single dimension, k is Boltzmann's constant, and T is the temperature in kelvin. A shunt capacitor C is charged by the thermal noise in the resistor (see Figure 6.6(b), where L is assumed to be zero). The average potential energy stored is $(1/2)CE[W(t)^2]$. Equating this to $1/2kT$ and solving, we obtain the mean square power

$$E[W(t)^2] = kT/C \tag{6.28}$$

For example, let $T = 300K$ and $C = 50$ pf, and recall that $k = 1.38 \times 10^{-23}$ J/K. Then $E[W(t)^2] = kT/C = 82.8 \times 10^{-12}$, so that the input rms voltage is $\{E[W(t)^2]\}^{1/2} = 9.09\ \mu V$.

Effective Noise and Antenna Noise

Let two series resistors R_1 and R_2 have respective temperatures of T_1 and T_2. The total noise power over an incremental frequency band df is $P_{\text{Total}}(df) = P_{11}(df) + P_{22}(df) = 4kT_1R_1df + 4kT_2R_2df = 4k(T_1R_1 + T_2R_2)df$. By putting

$$T_{\text{E}} = (T_1R_1 + T_2R_2)/(R_1 + R_2) \tag{6.29}$$

we can write $P_{\text{Total}}(df) = 4kT_{\text{E}}(R_1 + R_2)df$. T_{E} is called the *effective noise temperature* [Gardner, 1990, p. 289; or Peebles, 1987, p. 228]. An antenna receives noise from various sources of electromagnetic radiation, such as radio transmissions and harmonics, switching equipment (such as computers, electrical motor controllers), thermal (blackbody) radiation of the atmosphere and other matter, solar radiation, stellar radiation, and galaxial radiation (the ambient noise of the universe). To account for noise at the antenna output, we model the noise with an equivalent thermal noise using an effective noise temperature T_{E}. The incremental available power (output) over an incremental frequency band df is $P_{YY}(df) = kT_{\text{E}}\,df$, from Equation (6.27). T_{E} is often called *antenna temperature*, denoted by T_{A}. Although it varies with the frequency band, it is usually virtually constant over a small bandwidth.

Noise Factor and Noise Ratio

In reference to Figure 6.8(a), we define the *noise factor* F = (*noise power output of actual device*)/(*noise power output of ideal device*), where (noise power output of ideal device) = (power output due to
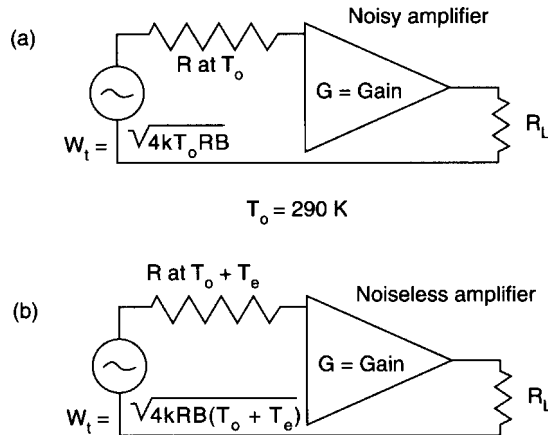
FIGURE 6.8 Equivalent input noise and noise factor.

thermal noise source). The noise source is taken to be a noisy resistor R at a temperature T, and all output noise measurements must be taken over a resistive load R_L (reactance is ignored). Letting $P_{WW}(B) = 4kTRB$ be the open circuit thermal noise power of the source resistor over a frequency bandwidth B, and noting that the gain of the device is G, the output power due to the resistive noise source becomes $G^2P_{WW}(B) = 4kTRBG^2/R_L$. Now let $Y(t)$ be the output voltage measured at the output across R_L. Then the noise factor is

$$F = (P_{YY}(B)/R_L)/(G^2P_{WW}(B)/R_L) = (P_{YY}(B))/(4kTRBG^2) \qquad (6.30)$$

F is seen to be independent of R_L, but not R. To compare two noise factors, the same source must be used. In the ideal noiseless case, $F = 1$, but as the noise level in the device increases, F increases. Because this is a power ratio, we may take the logarithm, called the *noise ratio*, which is

$$N_F = 10\log_{10}(F) = 10\log_{10}(P_{YY}(B)) - 10\log_{10}(4kTRBG^2) \qquad (6.31)$$

The noise power output $P_{YY}(B)$ of an actual device is a superposition of the amplified source thermal noise $G^2P_{WW}(B)$ and the device noise, i.e., $P_{YY}(B) = G^2P_{WW}(B) + (\text{device noise})$. The output noise across R_L can be measured by putting a single frequency (in the passband) source generator $S(t)$ as input. First, $S(t)$ is turned off, the output rms voltage $Y(t)$ is measured, and the output power $P_{Y(W)}(B)$ is recorded. This is the sum of the thermal available power and the device noise. Next, $S(t)$ is turned on and adjusted until the output power doubles, i.e., until the output power $P_{Y(W)}(B) + P_{Y(S)}(B) = 2P_{Y(W)}(B)$. This $P_{SS}(B)$ is recorded. Solving for $P_{Y(S)}(B) = P_{Y(W)}(B)$, we substitute this in $F = P_{Y(W)}(B)/(G^2P_{WW}(B))$ to obtain

$$F = P_{Y(S)}(B)/(G^2P_{WW}(B)) = (G^2P_{SS}(B))/(G^24kTRB) = P_{SS}(B)/4kTRB \qquad (6.32)$$

A better way is to input white noise $W(t)$ in place of $S(t)$ (a noise diode may be used). The disadvantages of noise factors are: (1) when the device has low noise relative to thermal noise, the noise factor has value close to 1; (2) a low resistance causes high values; and (3) increasing the source resistance decreases the noise factor while increasing the total noise in the circuit [Ott, 1988, p. 216]. Therefore, the accuracy is not adequate. For cascaded devices, the noise factors can be conveniently computed [Buckingham, 1985, p. 67; or Ott, 1988, p. 228].

Equivalent Input Noise

Shot noise (see below) and other noise can be modeled by the equivalent thermal noise that would be generated in an input resistor by increased temperature. Recall the (maximum) incremental available power (output) in a frequency bandwidth df is $P_{WW}(df) = kTdf$ from Equation (6.27). Figure 6.8(b) presents the following situation. Let the resistor be the noise source at temperature T_o with thermal noise $W(t)$. Then $E[W(t)^2] = 4kT_oRdf$, by Equation (6.24) (Nyquist's result). Let the open circuit output noise power at R_L be $E[Y(t)^2]$. The incremental available noise power $P_{YY}(df)$ at the output ($R_L = R$) can be considered to be due to the resistor R having a higher temperature and an ideal (noiseless) device, usually an amplifier. We must find a temperature T_e at which a pseudothermal noise power $E[W_e(t)^2] = 4kT_eRdf$ yields the extra "input" noise power. Let $V(t) = W(t) + W_e(t)$. Then $P_{VV}(df) = 4kT_oRdf + 4kT_eRdf = 4k(T_o + T_e)Rdf$ W, from Equation (6.24). T_e is called the *equivalent input noise temperature*. It is related to the noise factor F by $T_e = 290(F - 1)$. In cascaded amplifiers with gains $G_1, G_2, \ldots$ and equivalent input noise temperatures $T_{e1}, T_{e2}, \ldots$, the total equivalent input noise temperature is

$$T_{e(\text{Total})} = T_{e1} + T_{e2}/G_1 + T_{e3}/G_1G_2 + \ldots \qquad (6.33)$$

[Gardner, 1990, p. 289].

Other Electrical Noise

Thermal noise and shot noise, which can be modeled by thermal noise with equivalent input noise, are the main noise sources. Other noises are discussed in the following paragraphs.

Shot Noise

In a conductor under an external emf, there is an average flow of electrons, holes, photons, etc. In addition to this induced net flow and thermal noise, there is another effect. The potential differs across the boundaries of metallic grains and particles of impurities, and when the kinetic energy of electrons exceeds this potential, electrons jump across the barrier. This summed random flow is known as *shot noise* [Gardner, 1990, p. 239; Ott, 1988, p. 208]. The shot effect was analyzed by Schottky in 1918 as $I_{sh} = (2qI_{dc}B)^{1/2}$, where $q = 1.6 \times 10^{-19}$ coulombs per electron, I_{dc} = average dc current in amperes, and B = noise bandwidth (Hz).

Partition Noise

Partition noise is caused by a parting of the flow of electrons to different electrodes into streams of randomly varying density. Suppose that electrons from some source S flow to destination electrodes A and B. Let $n(A)$ and $n(B)$ be the average numbers of electrons per second that go to nodes A and B respectively, so that $n(S) = n(A) + n(B)$ is the average total number of electrons emitted per second. It is a success when an electron goes to A, and the probability of success on a single trial is p, where

$$p = n(A)/n(S), 1 - p = n(B)/n(S) \tag{6.34}$$

The current to the respective destinations is $I(A) = n(A)q$, $I(B) = n(B)q$, where q is the charge of an electron, so that $I(A)/I(S) = p$ and $I(B)/I(S) = 1 - p$. Using the binomial model, the average numbers of successes are $E[n(A)] = n(S)p$ and $E[n(B)] = n(S)(1-p)$. The variance is $\text{Var}(n(A)) = n(S)p(1-p) = \text{Var}(n(B))$ (from the binomial formula for variance). Therefore, substitution yields

$$\text{Var}(I(A)) = q^2[n(S)p(1-p)] = q^2 n(S)\{I(A)I(B)/[I(A) + I(B)]\} \tag{6.35}$$

Partition noise applies to pentodes, where the source is the cathode, A is the anode (success), and B is the grid. For transistors, the source is the emitter, A is the collector, and B represents recombination in the base. In photo devices, a photoelectron is absorbed, and either an electron is emitted (a success) or not. Even a partially silvered mirror can be considered to be a partitioner; the passing of a photon is a success and reflection is a failure. While the binomial model applies to partitions with destinations A and B, multinomial models are analogous for more than two destinations.

1/f, Burst, and Contact Noise

In 1925, J.B. Johnson first noticed that noise across thermionic gates exceeded the expected shot noise at lower frequencies. This added noise, above thermal noise, is characterized by a spectral distribution that has equal energy in each octave. This leads to a straight line on a log-log graph with a slope of −1. The corner frequency where it intercepts the thermal noise is usually in the neighborhood of 1 KHz, but improvements in semiconductor processing cleanliness have pushed it down towards 10 Hz. The physical mechanism causing 1/f noise is currently attributed to one of two competing theories: (1) surface defects trapping carriers, and (2) a bulk phenomenon involving filaments of current varying within a semiconductor channel [Lundberg p. 3]. The surface defect theory explains the current problem in semiconductor miniaturization: 1/f noise is increasing as the size of transistors become smaller. Since the ratio of surface area to volume increases as size decreases, the surface defect process theory explains that 1/f noise is increasing in more densely populated semiconductors.

The problem with a 1/f spectra is the inverse of the ultraviolet catastrophe in physics, where integrating to infinity yields an unbounded energy. Here, integrating to zero frequency leads to unbounded energy. One popular model is the Lorentzian spectra, or the response of a low-pass filter, where at some very low frequency the energy becomes constant per unit bandwidth for all lower frequencies and decreases above that point as 1/f [Poore, 2001, pp. 1–2]. For some physical processes there is often a high frequency above which the 1/f spectra

falls off as 1/f^^2 [Milotti, p. 4]. The psdf of the extra noise, called flicker noise, is

$$S(f) = I^2/\alpha f, f > 0 \tag{6.36}$$

where I is the dc current flowing through the device and f is the positive frequency. Empirical values of (a) are approximately 0.5 to 1.6 for different sources. These sources vary, but include the size irregularity of the cathode surface macroregions, impurities in the conducting channel, and generation and recombination noise in transistors. In the early days of transistors, this generation-recombination was of great concern because the materials were not high in purity. Flicker noise occurs in thin layers of metallic or semiconducting material, solid-state devices, carbon resistors, and vacuum tubes [Buckingham, 1985, p. 143]. Flicker noise may be high at low frequencies.

Contact noise is caused by fluctuating conductivity due to imperfect contact between two surfaces, especially in switches and relays.

Burst noise is also called *popcorn noise*. Audio amplifiers sound like popcorn popping in a frying pan background (thermal noise). Its characteristic is $1/f^n$ (usually $n = 2$), so its power density falls off rapidly, where f is frequency. It may be problematic at low frequencies. The cause is manufacturing defects in the junction of transistors (usually a metallic impurity).

Barkhousen and Other Noise

Barkhousen noise is due to the variations in size and orientation of the small regions of ferromagnetic material and is especially noticeable in the steeply rising region of the hysteresis loop. There is also secondary emission, photo, and collision ionization, etc.

Measurement and Quantization Noise

Measurement Error

The measurement X_t of a signal $X(t)$ at any t results in a measured value $X_t = x$ that contains error, and so is not equal to the true value $X_t = x_T$. The probability is higher that the magnitude of $e = (x - x_T)$ is closer to zero. The bell-shaped Gaussian probability density $f(e) = [1/(2\pi\sigma2]^{1/2}\exp(-e^2/2\pi\sigma)$ fits the error well. This noise process is stationary over time. The expected value is $\mu_e = 0$, the mean-square error is σ_e^2, and the rms error is σ_e. Its instantaneous power at time t is σ_e^2. To see this, the error signal $e(t) = (x - x_T)$ has instantaneous power per Ω of

$$P_i = e(t)i(t) = e(t)[e(t)/R] = e^2(t) \tag{6.37}$$

where $R = 1 \ \Omega$ and $i(t)$ is the current. The average power is the summed instantaneous power over a period of time T, divided by the time, taken in the limit as $T \to \infty$, i.e.,

$$P_{\text{ave}} = \lim_{T \to \infty} (1/T) \int_0^T e^2(t)dt$$

This average power can be determined by sampling on known signal values and then computing the sample variance (assuming ergodicity: see Gardner [1990, p. 163]). The error and signal are probabilistically independent, unless the error depends on the values of X. The signal-to-noise power ratio is computed by $S/N = P_{\text{signal}}/P_{\text{ave}}$.

Quantization Noise

Quantization noise is due to the digitization of an exact signal value $v_t = v(t)$ captured at sampling time t by an A/D converter. The binary representation is $b_{n-1}b_{n-2} \ldots b_1 b_0$ (an n-bit word). The n-bit digitization has 2^n different values possible, from zero to $2^n - 1$. Let the voltage range be R. The resolution is $dv = R/2^n$. Any voltage v_t is coded into the nearest lower binary value x_b, where the error $e = x_t - x_b$ satisfies $0 \leq e \leq dv$.

Thus, the errors e are distributed equally over the interval $[0, dv]$ in such a way that implies the uniform distribution on $[0, dv]$. The expected value of $e = e_t = e(t)$ at any time is $\mu_e = dv/2$, and the variance is $\mu_e^2 = dv^2/12$ (the variance of a uniform distribution on an interval $[a,b]$ is $\sigma = (b - a)^2/12$). Therefore, the noise is ws and the power of quantization noise is

$$\sigma_e^2 = \int_0^{dv} (e - dv/2)^2 (1/dv) \mathrm{d}e$$
$$\mathrm{d}e = (e - dv/2)^3 / 3dv|_0^{dv} = [(dv)^3 + (dv)^3]/24dv = dv^2/12 \tag{6.38}$$

We can find the signal-to-noise voltage ratio for the total range R via $R/(dv/(12)^{1/2}) = 2^n \, dv/(dv/(12)^{1/2}) = 2^n (12)^{1/2}$. The power ratio is the square of this, which is $(2^{2n})(12)$. In decibels this becomes $(S/N)_{\mathrm{dB}} = 10 \log_{10} (2^{2n} \cdot 12) = 10 \log_{10} (12) + 20n \log_{10} (2) = 10.8 + 6.02n$. Thus, the quantization S/N power ratio depends directly upon the number of bits n in that the higher S/N power ratio is better as is expected.

Coping with Noise

External interference is ubiquitous. Intrinsic noise is present up to the incremental available power at temperatures above absolute zero, and other intrinsic noises depend on material purity and connection integrity. Processing error is always introduced in some form.

External Sources

Standard defenses are:

1. 100% shielding of lines and circuits: thin conductors will stop electrostatic fields, but magnetic fields require thicker shielding and low reluctance metals.
2. Twisted wire pairs are effective against magnetic fields combined with high common mode rejection balanced inputs.
3. Coaxial cables with multiple shields for radio frequencies.
4. Short lines and leads.
5. Multilayered PCBs with internal ground planes each side of buried signal lines.
6. Digital regeneration at waypoints of digital signals.
7. Matched filtering of signals.
8. Correlation of received signals with multipaths.
9. Adaptive notch filtering to eliminate interference at known frequencies; e.g., the second harmonic of 60-Hz ac power lines may interfere with biological microvoltage measurements, but could be eliminated via adaptive notch filtering.
10. Star grounding to prevent ground loops.
11. Low pass filters on power leads between circuits.
12. Orienting magnetic components for minimum pickup.

Ferrite beads can dampen interference [Barnes, 1987]. They are used on power leads providing a low resistance to DC current while presenting high impedance to high frequencies. The core losses lead to an appreciable portion of the AC impedance being resistive, therefore avoiding resonances, digital signal processing, spectral shaping filters [Brown, 1983], and frequency-shift filters [Gardner, 1990, p. 400] that can be used to lower noise power. Kalman filtering is a powerful estimation method, and frequency-shift filtering is a newer technique for discriminating against both measurement error (e.g., in system identification applications) and extrinsic sources of both noise and interference [Gardner, 1990, p. 400].

Intrinsic Sources

Strategies for minimizing intrinsic noise are: (1) small bandwidth B, (2) small resistances R, (3) low temperature T (higher temperatures can be devastating), (4) low voltage and currents (CMOS transistors), (5) modern materials of high purity, (6) the latest generation active components with the lowest $1/f$ noise, (7) wire

wound resistors (thermal noise is the same, but other noise will be less), (8) fewer and better connections (of gold), (9) smaller circuits of lower power, (10) shunt capacitors to reduce noise bandwidth, (11) averaging to gain squareroot of N improvement in signal to noise ratio, and (12) using selected JFET devices. Currently, progress in the purity of integrated circuit materials is leading to a yearly improvement in $1/f$ and other extra noises. Better design and materials are the keys to lower noise.

Processing Sources

Processing errors can be reduced by using higher resolution of analog-to-digital converters, i.e., more bits to represent each value. This lowers the quantization error power. Measurement error can be reduced while using the same instruments by taking multiple measurements and averaging. Other estimation/correlation can yield better values (e.g., the Global Positioning System location determination can be reduced from meters to a few centimeters by a multiple measurement estimation).

Defining Terms

Autocorrelation: A function associated with a random signal $X(t)$ that is defined on pairs of time instants t_1 and t_2, and which the value is the expected value of the product of the random variables $X(t_1)$ and $X(t_2)$, i.e., $R_{XX}(t_1,t_2) = E[X(t_1)X(t_2)]$. For weakly stationary random signals, it depends only on the offset $\tau = t_2 - t_1$, so we write $R_{XX}(\tau) = E[X(t)X(t+\tau)]$.

Noise: A signal $N(t)$ in which the value at any time t is randomly selected by events beyond our control. At any time instant t, $N(t)$ is a random variable N_t with a probability distribution that determines the relative frequencies at which N_t assumes values. The statistics of the family of random variables $\{N_t\}$ may be constant (stationary) over time (in most cases) or may vary.

Power spectral density: The Fourier transform of the power $X^2(t)$ does not necessarily exist, but it does for $X_T^2(t)/2T(X_T(t) = 0$ for $|t| > T, = X(t)$ elsewhere), for any T>0. Letting T→∞, the expected value of the Fourier transforms $E[\mathrm{F}[X_T^2(t)/2T] = \mathrm{F}[E[X_T^2(t)]/2T$ goes to the limit of the average power in $X(t)$ over $-T$ to T, known as the power spectral density function $S_{xx(w)}$. Summed up over all frequencies, it gives the total power in the signal $X(t)$.

Random process: A signal that is either a noise, an interfering signal $s(t)$, or a sum of these such as $X(t) = s_1(t) + \ldots + s_m(t) + N_1(t) + \ldots + N_n(t)$.

Realization: A trajectory $\{(t, x_t): X(t) = x_t\}$ determined by the actual outcomes $\{x_t\}$ of values from a random signal $X(t)$, where $X(t) = x_t$ at each instant t. A trajectory is also called a *sample function* of $X(t)$.

Weakly stationary (ws) random process (signal):   A random signal whose first- and second-order statistics remain stationary (fixed) over time.

References

J.R. Barnes, *Electronic System Design: Interference and Noise Control*, Englewood Cliffs, N.J.: Prentice-Hall, 1987.

R.G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering*, New York: Wiley, 1983.

M.J. Buckingham, *Noise in Electronic Devices and Systems*, New York: Halstead Press, 1985.

W.A. Gardner, *Introduction to Random Processes*, 2nd ed., New York: McGraw-Hill, 1990.

J.B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.*, vol. 29, pp. 367–368, 1927.

J.B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.*, vol. 32, pp. 97–109, 1928.

K.H. Lundberg, "Noise Sources in Bulk CMOS," Unpublished paper, p. 3 http://web.mit.edu/klund/www/CMOSnoise.pdf, 2002.

E. Milotti, "1/f Noise: A Pedagoglical Review,"

H. Nyquist, "Thermal agitation of electric charge in conductors," *Phys. Rev.*, vol. 32, pp. 110–113, 1928.

H.W. Ott, *Noise Reduction Techniques in Electronic Systems*, 2nd ed., New York: Wiley-Interscience, 1988.

P.Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles*, 2nd ed., New York: McGraw-Hill, 1987.

R. Poore, "Phase Noise and Jitter," Agilent Technologies, http://eesof.tm.agilent.com/pdf/jitter_background.pdf, 2001.

Further Information

The IEEE Individual Learning Program, "Random Signal Analysis with Random Processes and Kalman Filtering," prepared by Carl G. Looney (IEEE Educational Activities Board, PO Box 1331, Piscataway, NJ 08855–1331, 1989) contains a gentle introduction to estimation and Kalman filtering.

Another recommended source is by H.M. Denny, "Getting Rid of Interference", *IEEE Video Conference*, Educational Activities Dept., Piscataway, NJ, 08855–1331, 1992.

6.3 Stochastic Processes

Carl G. Looney

Introduction to Random Variables

A *random variable* (rv) A is specified by its *probability density function* (pdf)

$$f_A(a) = \lim_{\varepsilon \to 0}(1/\varepsilon)P[a - (\varepsilon/2) < A \leqslant a + (\varepsilon/2)]$$

In other words, the rectangular area $\varepsilon \cdot f_A(a)$ approximates the probability $P[(A \leqslant a + (\varepsilon/2)] - P[a - (\varepsilon/2) \geqslant A]$. The joint pdf of two rv's A and B is specified by

$$f_{AB}(a, b) = \lim_{\varepsilon \to 0}(1/\varepsilon^2)P[a - \varepsilon < A \leqslant a + (\varepsilon/2) \text{ and } b - \varepsilon < B \leqslant b + (\varepsilon/2)]$$

A similar definition holds for any finite number of rv's.

The *expected value* $E[A]$, or *mean* μ_A, of a rv A is the first moment of the pdf, and the *variance* of A is the second centralized moment, defined respectively by

$$\mu_A = E[A] \equiv \int_{-\infty}^{\infty} af_A(a)\mathrm{d}a \qquad (6.39\text{a})$$

$$\sigma_A^2 = E[(A - \mu_A)^2] \equiv \int_{-\infty}^{\infty} (a - \mu_A)^2 f_A(a)\mathrm{d}a \qquad (6.39\text{b})$$

The square root of the variance is the *standard deviation*, which is also called the *root mean square* (rms) error. The *covariance* of two rv's A and B is the second-order centralized joint moment

$$\sigma_{AB} = E[(A - \mu_A)(B - \mu_B)] \equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (a - \mu_A)(b - \mu_B)f_{AB}(a, b)\mathrm{d}a\mathrm{d}b \qquad (6.40)$$

The noncentralized second moments are the *mean-square value* and the *correlation*, respectively,

$$E[A^2] = \int_{-\infty}^{\infty} a^2 f_A(a)\mathrm{d}a = \sigma_A^2 +, \mu_A^2, \qquad E[AB] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} abf_{AB}(a, b)\mathrm{d}a\mathrm{d}b = \sigma_{AB} + \mu_A\mu_B$$

A set of rv's *A, B,* and *C* is defined to be *independent* whenever their joint pdf factors as

$$f_{ABC}(a, b, c) = f_A(a)f_B(b)f_C(c) \tag{6.41}$$

for all *a, b,* and *c,* and similarly for any finite set of rv's. A *weak independence* holds when the second moment of the joint pdf, the correlation, factors as $E[AB] = E[A]E[B]$, so that $\sigma_{AB} = 0$, in which case the rv's are said to be uncorrelated. The covariance of *A* and *B* is a measure of how often *A* and *B* vary together (have the same sign), how often they vary oppositely (different signs), and by how much, on the average over trials of outcomes. To standardize so that units do not influence the measure of dependence, we use the correlation coefficient

$$\rho_{AB} \equiv \sigma_{AB}/\sigma_A\sigma_B$$

The accuracy of approximating a rv *A* as a linear function of another rv *B,* $A \approx cB + d$, for real coefficients *c* and *d,* is found by minimizing the mean-square error $\varepsilon = E\{[A - (cB + d)]^2\}$. Upon squaring and taking the expected values, we can obtain $\varepsilon_{\min} = \sigma_A^2(1 - |\rho_{AB}|^2)$, which shows $|\rho_{AB}|$ to be a measure of the degree of linear relationship between *A* and *B*. Because $\varepsilon_{\min} \geq 0$, this shows that $|\rho_{AB}| \leq 1$, which demonstrates the Cauchy-Schwarz inequality

$$|E[AB]| \leq \left\{E[A^2]E[B^2]\right\}^{1/2} \tag{6.42}$$

When $|\rho_{AB}| = 1$, then knowledge of one of *A* or *B* completely determines the other ($c \neq 0$), and so *A* and *B* are completely dependent, while $|\rho_{AB}| = 0$ indicates there is no linear relationship, i.e., that *A* and *B* are uncorrelated.

An important result is the *fundamental theorem of expectation*: if $g(\cdot)$ is any real function, then the expected value of the rv $B = g(A)$ is given by

$$E[B] = E[g(A)] = \int_{-\infty}^{\infty} g(a)f_A(a)\mathrm{d}a \tag{6.43}$$

Stochastic Processes

A **stochastic** (or *random*) **process** is a collection of random variables $\{X_t: t \in T\}$, indexed on an ordered set *T* that is usually a subset of the real numbers or integers. Examples are the Dow-Jones averages $D(t)$ at each time *t,* the pressure $R(x)$ in a pipe at distance *x,* or a noise voltage $N(t)$ at time *t.* A process is thus a *random function* $X(t)$ of *t* whose value at each *t* is drawn randomly from a range of outcomes for the rv $X_t = X(t)$ according to a probability distribution for X_t. A trajectory $\{x_t: t \in T\}$ of outcomes over all $t \in T$, where $X_t = x_t$ is the realized value at each *t,* is called a **sample function** (or *realization*) of the process. A stochastic process $X(t)$ has mean value $E[X(t)] = \mu(t)$ at time *t,* and **autocorrelation function** $R_{XX}(t, t + \tau) = E[X(t)X(t + \tau)]$ at times *t* and $t + \tau$, the correlation of two rv's at two times offset by τ. When $\mu(t) = 0$ for all *t,* the autocorrelation function equals the *autocovariance function* $C_{XX}(t, t + \tau) = E[(X(t) - \mu(t))(X(t + \tau) - \mu(t + \tau))]$.

A process $X(t)$ is completely determined by its joint pdf's $f_{X(t(1))...X(t(n))}(x(t_1), \ldots, x(t_n))$ for all time combinations $t_1, \ldots, t_n$ and all positive integers *n* (where $t(j) = t_j$). When the rv's $X(t)$ are *iid* (independent, identically distributed), then knowledge of one pdf yields the knowledge of all joint pdf's. This is because we can construct the joint pdf by factorization, per Equation (6.41).

Classifications of Stochastic Processes

The ordered set *T* can be continuous or discrete, and the values that $X(t)$ assumes at each *t* may also be continuous or discrete, as shown in Table 6.1.

TABLE 6.1 Continuous/Discrete Classification of Stochastic Processes

	X Values	
T Values	Continuous	Discrete
Continuous	Continuous stochastic processes	Discrete valued stochastic processes
Discrete	Continuous random sequences	Discrete valued random sequences

In another classification, a stochastic process $X(t)$ is *deterministic* whenever an entire sample function can be determined from an initial segment $\{x_t : t \leq t_1\}$ of $X(t)$. Otherwise, it is *nondeterministic* [Brown, 1983, p. 79; or Gardner, 1990, p. 304].

Stationarity of Processes

A stochastic process is *nth order (strongly) stationary* whenever all joint pdf's of n and fewer rv's are independent of all translations of times $t_1, \ldots, t_n$ to times $\tau + t_1, \ldots, \tau + t_n$. The case of $n = 2$ is very useful. Another type of process is called **weakly stationary** (ws), or *wide-sense stationary*, and is defined to have first- and second-order moments that are independent of time. These satisfy (1) $\mu(t) = \mu$ (constant) for all t, and (2) $R_{XX}(t, t + \tau) = R_{XX}(t + s, t + s + \tau)$ for all values of s. For $s = -t$, this yields $R_{XX}(t, t + \tau) = R_{XX}(0, 0 + \tau)$, which is abbreviated to $R_{XX}(\tau)$. $X(t)$ is *uncorrelated* whenever $C_{XX}(\tau) = 0$ for τ not zero [we say $X(t)$ has *no memory*]. If $X(t)$ is correlated, then $X(t_1)$ depends on values $X(t)$ for $t \neq t_1$ [$X(t)$ has *memory*].

Some properties of autocorrelation functions for ws processes follow. First, $R_{XX}(\tau)| \leq R_{XX}(0), -\infty \geq \tau \geq \infty$, as can be seen from Equation (6.42) with $|R_{XX}(\tau)|^2 = E[X(0)X(\tau)] \leq E[X(0)^2]E[X(\tau)^2] = R_{XX}(0)R_{XX}(\tau)$. Next, $R_{XX}(\tau)$ is real and even, i.e., $R_{XX}(-\tau) = R_{XX}(\tau)$, which is evident from substituting $s = t - \tau$ in $E[X(s)X(s + \tau)]$ and using time independence. If $X(t)$ has a periodic component, then $R_{XX}(\tau)$ will have that same periodic component, which follows from the definition. Finally, if $X(t)$ has a nonzero mean m and no periodic components, then the variance goes to zero (the memory fades) and so $\lim_{\tau \to \infty} R_{XX}(\tau) \to 0 + \mu^2 = \mu^2$.

Gaussian and Markov Processes

A process $X(t)$ is defined to be *Gaussian* if for every possible finite set $\{t_1, \ldots, t_n\}$ of times, the rv's $X(t_1), \ldots, X(t_n)$ are *jointly Gaussian*, which means that every linear combination $Z = a_1 X(t_1) + \ldots + a_n X(t_n)$ is a Gaussian rv, defined by the Gaussian pdf

$$f_Z(z) = \left[1/(\sigma_Z \sqrt{2\pi})\right] \exp\{-(z - \mu_Z)^2/2\sigma_Z^2\} \tag{6.44}$$

In case the n rv's are *linearly independent*, i.e., $Z = 0$ only if $a_1 = \ldots = a_n = 0$, the joint pdf has the Gaussian form [see Gardner, 1990, pp. 39–40]

$$f_{X(t(1))}\cdots_{X(t(n))}(x_1, \ldots, x_n) = \left[1/(2\pi)^{n/2}|C|^{1/2}\right]\exp\{-(x - \boldsymbol{\mu})^t C^{-1}(x - \boldsymbol{\mu})\} \tag{6.45}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is a column vector, $\mathbf{x}^t$ is its transpose, $\mu = (\mu_1, \ldots, \mu_n)$ is the mean vector, C is the *covariance matrix*

$$C = \left\langle \begin{matrix} \sigma_1^2 & .. & \sigma_{ln} \\ : & : & \\ \sigma_{ln} & .. & \sigma_n^2 \end{matrix} \right\rangle \tag{6.46}$$

and $|C|$ is the determinant of C. If $X(t_1),\ldots, X(t_n)$ are linearly dependent, then the joint pdf takes on a form similar to Equation (6.45), but contains impulses [Gardner, 1990, p. 40].

A weakly stationary Gaussian process is strongly stationary to all orders n: all Gaussian joint pdf's are completely determined by their first and second moments by Equation (6.45), and those moments are time independent by weak stationarity, and so all joint pdf's are also. Every second-order strongly stationary stochastic process $X(t)$ is also weakly stationary because the time translation independence of the joint pdf's determines the first and second moments to have the same property. However, non-Gaussian weakly stationary processes need not be strongly second-order stationary.

Rather than with pdf's, a process $X(t)$ may be specified in terms of conditional pdf's

$$f_{X(t(1))\ldots X(t(n))}(x_1,\ldots,x_n) = f_{X(t(n))|X(t(n-1))}(x_n|x_{n-1}) \cdot \ldots \cdot f_{X(t(2))|X(t(1))}(x_2|x_1) f_{X(t(1))}(x_1)$$

by successive applications of Bayes' law, for $t_1 < t_2 < \ldots < t_n$. The conditional pdf's satisfy

$$f_{A|B}(a|b) = f_{AB}(a, b)/f_B(b) \tag{6.47}$$

The conditional factorization property may satisfy

$$f_{X(t(n))|X(t(n-1))\ldots X(t(1))}(x_n|x_{n-1},\ldots,x_1) = f_{X(t(n))|X(t(n-1))}(x_n|x_{n-1}) \tag{6.48}$$

which indicates that the pdf of the process at any time t_n, given values of the process at any number of previous times $t_{n-1},\ldots, t_1$, is the same as the pdf at t_n given the value of the process at the most recent time t_{n-1}. Such an $X(t)$ is called a *first-order Markov process*, in which case we say the process remembers only the previous value (the previous value has influence). In general, an *nth-order Markov* process remembers only the n most recent previous values. A first-order Markov process can be fully specified in terms of its first-order conditional pdf's $f_{X(t)|X(s)}(x_t, x_s)$ and its unconditional first-order pdf at some initial time t_0, $f_{X(t(0))}(x_0)$.

Examples of Stochastic Processes

Figure 6.9 shows two sample functions of nonstationary processes. Now consider the discrete time process $X(k) = A$, for all $k \geq 0$, where A is a rv (a *random initial condition*) that assumes a value 1 or -1 with respective probabilities p and $1 - p$ at $k = 0$. This value does not change, once the initial random draw is done at $k = 0$. This stochastic sequence has two sample functions only, the constant sequences $\{-1\}$ and $\{1\}$. The expected value of $X(k)$ at any time k is $E[X(k)] = E[A] = p1 + (1 - p)(-1) = 2p - 1$, which is independent of k. The autocorrelation function is, by definition, $E[X(k)X(k + m)] = E[AA] = E[A^2] = p1^2 + (1 - p)(-1)^2 = 1$ which is also independent of time k. Thus $X(k)$ is perfectly correlated for all time (the process has *infinite memory*). This process is deterministic.

For another example, put $X(t) = (c)\cos(wt + \Phi)$, where Φ is the uniform rv on $(-\pi, \pi)$. Then $X(t)$ is a function of the rv Φ (as well as t), so by use of Equation (6.39a), we obtain

$$E[X(t)] = c \cdot \int_{-\pi}^{\pi} \cos(wt + \phi) f_\Phi(\phi)\mathrm{d}\phi = (c/2\pi)\sin(wt + \phi)|_{-\pi}^{\pi} = 0$$
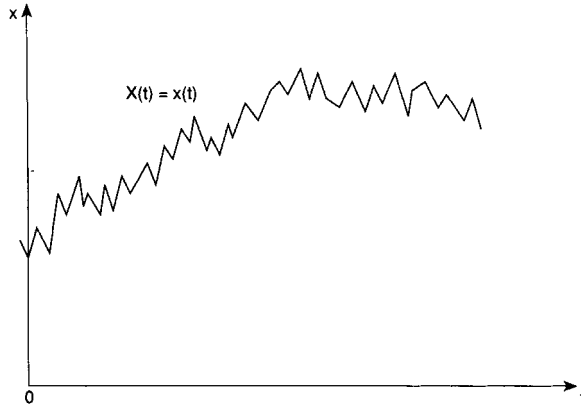
(a) Sample function with nonstationary mean and stationary variance
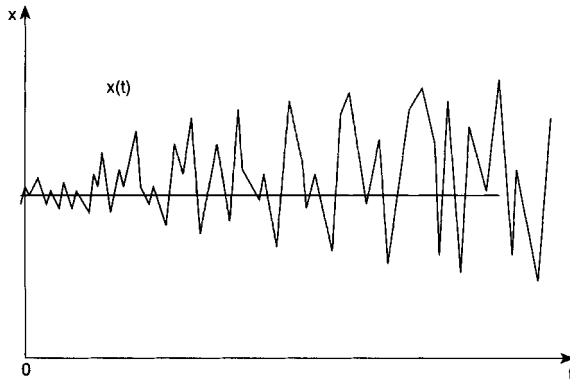
(b) Sample function with stationary mean and nonstationary (increasing) variance

FIGURE 6.9 Examples of nonstationary processes.

Therefore, the mean does not vary with time t. The autocorrelation is

$$R_{XX}(t, t + \tau) = E[(c)\cos(wt + \Phi)(c)\cos(wt + w\tau + \Phi)]$$

$$= c^2 E[\cos(wt + \Phi)\cos(wt + w\tau + \Phi)]$$

$$= c^2 \int_{-\pi}^{\pi} \cos(wt + \phi)\cos(wt + w\tau + \Phi)f_{\Phi}(\phi)d\phi$$

$$= (c^2/2) \int_{-\pi}^{\pi} \{\cos(2wt + 2\phi + w\tau) + \cos(w\tau)\}(1/2\pi)d\phi$$

$$= (c^2/4\pi) \cdot \{\sin(\Theta + 2\pi) + \cos(w\tau) \cdot 2\pi\}$$

$$= (c^2/4\pi) \cdot \{\cos(w\tau) \cdot 2\pi\} = (c^2/2)\cos(w\tau)$$

[using $\cos(x)\cos(y) = 1/2\{\cos(x + y) + \cos(x - y)\}$ and letting $\Theta = 2wt + 2\Phi + wt$]. Therefore, $X(t)$ is ws. The autocorrelation is periodic in the offset variable τ.

Now consider the example $X(t) = A \cos(2\pi f_0 t)$ for each t, where f_0 is a constant frequency, and the amplitude A is a random initial condition as given above. There are only two sample functions here: (1) $x(t) = \cos(2\pi f_0 t)$ and (2) $x(t) = -\cos(2\pi f_0 t)$. A related example is $X(t) = A \cos(2\pi f_0 t + \Phi)$, where A is given above, the phase Φ is the uniform random variable on $[0, \pi]$, and A and Φ are independent. Again, Φ and A do not

depend on time (initial random conditions). Thus, the sample functions for $X(t)$ are $x(t) = \pm\cos(2\pi ft + \phi)$, where $\Phi = \phi$ is the value assumed initially. There are infinitely many sample functions because of the phase. Equation (6.39b) and the independence of A and Φ yield

$$E[X(t)] = E[A\cos(2\pi ft + \Phi)] = E[A]E[g(\Phi)] = \mu_A \int_0^\pi \cos(2\pi ft + \phi)(1/\pi)\mathrm{d}\phi$$

$$= (\mu_A/\pi)\sin(2\pi t + \phi)|_{\phi=0}^\pi = (\mu_A/\pi)[\sin(2\pi t + \pi) - \sin(2\pi t)]$$

$$= (\mu_A/\pi)[\sin(-2\pi t) - \sin(2\pi t)] = (-2\mu_A/\pi)\sin(2\pi t)$$

which is dependent upon time. Thus, $X(t)$ is not ws.

Next, let $X(t) = [a + S(t)]\cos[2\pi ft + \Phi]$, where the signal $S(t)$ is a nondeterministic stochastic process. This is an amplitude-modulated sine wave carrier. The carrier $\cos[2\pi ft + \Phi]$ has random initial condition Φ and is deterministic. Because $S(t)$ is nondeterministic, $X(t)$ is also. The expected value $E[X(t)] = E[a + S(t)]E[\cos(2\pi ft + \Phi)]$ can be found as above by independence of $S(t)$ and Φ.

Finally, let $X(t)$ be uncorrelated ($E[X(t)X(t + \tau)] = 0$ for τ not zero) such that each rv $X(t) = X_t$ is Gaussian with zero mean and variance $\sigma^2(t) = t$, for all $t < 0$. Any realized sample function $x(t)$ of $X(t)$ cannot be predicted in any average sense based on past values (uncorrelated Gaussian random variables are independent). The variance grows in an unbounded manner over time, so $X(t)$ is neither stationary nor deterministic. This is called the *Wiener* process.

A useful model of a ws process is that for which $\mu = 0$ and $R_{XX}(\tau) = \sigma_X^2\exp(-a|\tau|)$. If this process is also Gaussian, then it is strongly stationary and all of its joint pdf's are fully specified by $R_{XX}(\tau)$. In this case it is also a first-order Markov process and is called the *Ornstein-Uhlenbeck* process [Gardner, 1990, p. 102]. Unlike white noise, many real-world ws stochastic processes are correlated ($|R_{XX}(t, t + \tau)| \leqslant 0$) for $|\tau| \leqslant 0$. The autocorrelation either goes to zero as τ goes to infinity, or else it has periodic or other nondecaying memory. We consider ws processes henceforth (for nonstationary processes, see Gardner, 1990). We will also assume without loss of generality that $\mu = 0$.

Linear Filtering of Weakly Stationary Processes

Let the ws stochastic process $X(t)$ be the input to a linear time-invariant stable filter with impulse response function $h(t)$. The output of the filter is also a ws stochastic process and is given by the convolution

$$Y(t) = h(t) * X(t) = \int_{-\infty}^\infty h(s)X(t - s)\mathrm{d}s \tag{6.49}$$

The mean of the output process is obtained by using the linearity of the expectation operator [see Gardner, 1990, p. 32]

$$\mu_Y = E[Y(t)] = E\left[\int_{-\infty}^\infty h(s)X(t - s)\mathrm{d}s\right] = \int_{-\infty}^\infty h(s)E[X(t - s)]\mathrm{d}s$$

$$= \int_{-\infty}^\infty h(s)\mu_X\mathrm{d}s = \mu_X \int_{-\infty}^\infty h(s)\mathrm{d}s = \mu_X \cdot H(0) \tag{6.50}$$

where $H(f) = \int_{-\infty}^\infty h(t)e^{-j2\pi ft}\mathrm{d}t$ is the filter transfer function and $H(0)$ is the dc gain of the filter.

The autocorrelation of the output process, obtained by using the linearity of $E[\cdot]$, is

$$
\begin{aligned}
R_{YY}(\tau) = E[Y(t)Y(t+\tau)] &= E\left[\int_{\infty}^{\infty} h(v)X(t-v)\mathrm{d}v \int_{-\infty}^{\infty} h(u)X(t+\tau-u)\mathrm{d}u\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[X(t-v)X(t+\tau-u)]h(v)h(u)\mathrm{d}v\mathrm{d}u \\
&= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} R_{XX}(\tau-u+v)h(u)\mathrm{d}u\right\}h(v)\mathrm{d}v \\
&= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} R_{XX}([\tau-(-v)]-u)h(u)\mathrm{d}u\right\}h(-v)\mathrm{d}v \\
&= \int_{-\infty}^{\infty}\left\{R_{XX}(\tau+v)\star h(\tau+v)\right\}h(-v)\mathrm{d}v \\
&= [R_{XX}(\tau)\star h(\tau)]\star h(-\tau) = R_{XX}(\tau)\star[h(\tau)\star h(-\tau)] = R_{XX}(\tau)\star r_h(\tau)
\end{aligned}
$$

(6.51)

where $r_h(\tau)=\int_{-\infty}^{\infty} h(\tau+u)h(u)\mathrm{d}u$. However, $r_h(\tau)$ has Fourier transform $H(f)H\star(f)=|H(f)|^2$, because the Fourier transform of the convolution of two functions is the product of their Fourier transforms, and the Fourier transform of $h(-\tau)$ is the complex conjugate $H^\star(f)$ of the Fourier transform $H(f)$ of $h(\tau)$. Thus, the Fourier transform of $R_{YY}(\tau)$, denoted by $F\{R_{YY}(t)\}$, is

$$
\mathbf{F}\{R_{YY}(\tau)\} = \mathbf{F}[R_{XX}(\tau)\star h(\tau)\star h(-\tau)] = \mathbf{F}\{R_{XX}(\tau)\}\cdot H(f)H\star(f) = \mathbf{F}\{R_{XX}(\tau)\}\cdot|H(f)|^2
$$

Upon defining the functions

$$
S_{XX}(f)\equiv \mathbf{F}\{R_{XX}(\tau)\},\ S_{YY}(f)\equiv \mathbf{F}\{R_{YY}(\tau)\}
$$

(6.52)

we can also determine $R_{YY}(\tau)$ via the two steps

$$
S_{YY}(f) = S_{XX}(f)\cdot|H(f)|^2
$$

(6.53)

$$
R_{YY}(\tau) = \mathbf{F}^{-1}\{S_{YY}(f)\} = \int_{-\infty}^{\infty} S_{YY}(f)\mathrm{e}^{j2\pi f\tau}\mathrm{d}f
$$

(6.54)

Equations (6.52) define the *power spectral density functions* (psdf's) $S_{XX}(f)$ for $X(t)$ and $S_{YY}(f)$ for $Y(t)$. Thus, $R_{XX}(\tau)$ and $S_{XX}(f)$ are Fourier transform pairs, as are $R_{YY}(\tau)$ and $S_{YY}(f)$ (see Equation (6.58)). Further, the psdf $S_{XX}(f)$ of $X(t)$ is a power spectrum (in an average sense). If $X(t)$ is the voltage dropped across a 1-Ω resistor, then $X^2(t)$ is the instantaneous power dissipation in the resistance. Consequently, $R_{XX}(0) = E[X^2(t)]$ is the expected power dissipation over all frequencies, i.e., by Equation (6.54) with $\tau = 0$, we have

$$
R_{XX}(0) = \int_{-\infty}^{\infty} S_{XX}(f)\mathrm{d}f
$$

We want to show that when we pass $X(t)$ through a narrow bandpass filter with a bandwidth δ centered at the frequency $\pm f_0$, the expected power at the output terminals, divided by the bandwidth δ, is $S_{XX}(f_0)$ in the limit as $\delta \to 0$. This shows that $S_{XX}(f)$ is a density function (whose area is the total expected power over all frequencies, just as the area under a pdf is the total probability). This result that $R_{XX}(\tau)$ and $S_{XX}(f)$ are a Fourier transform pair is known as the *Wiener-Khinchin* relation [see Gardner, 1990, p. 230].

To verify this relation, let $H(f)$ be the transfer function of an ideal bandpass filter, where

$$H(f) = 1, |f - f_0| \geqslant \delta/2; \qquad H(f) = 0, \text{ otherwise}$$

Let Y(t) be the output of the filter. Then Equation (6.54) and Equation (6.53) provide

$$E[Y^2(t)] = R_{YY}(0) = \int_{-\infty}^{\infty} S_{YY}(f)df = \int_{-\infty}^{\infty} S_{XX}(f)|H(f)|^2 \, df$$
$$= \int_{f_0-\delta/2}^{f_0+\delta/2} S_{XX}(f)df + \int_{-f_0-\delta/2}^{-f_0+\delta/2} S_{XX}(f)df$$

Dividing by 2δ and taking the limit as $\delta \to 0$ yields $(1/2)S_{XX}(f_0) + (1/2)S_{XX}(-f_0)$, which becomes $S_{XX}(f_0)$ when we use the fact that psdf's are even and real functions (because they are the Fourier transforms of autocorrelation functions, which are even and real).

For example, let $X(t)$ be white noise, with $S_{XX}(f) = N_0$, being put through a first-order linear time-invariant system with respective impulse response and transfer functions

$$h(t) = \exp\{-\alpha t\}, t \geqslant 0; h(t) = 0, t < 0 \qquad H(f) = 1/[\alpha + j2\pi f], \text{ all } f$$

The temporal correlation of $h(t)$ with itself is $r_h(\tau) = (1/2\alpha)\exp\{-\alpha|\tau|\}$, so the power transfer function is $|H(f)|^2 = 1/[\alpha^2 + (2\pi f)^2]$. The autocorrelation for the input $X(t)$ is

$$R_{XX}(\tau) = \int_{-\infty}^{\infty} N_0 e^{j2\pi f\tau} df = N_0\delta(\tau)$$

which is an impulse. It follows (see Equation (6.22)) that the output $Y(t)$ has respective autocorrelation and psdf

$$R_{YY}(\tau) = [N_0\delta(\tau)] * [(1/2\alpha)e^{-\alpha|\tau|}] = (N_0/2\alpha)e^{-\alpha|\tau|}, \quad S_{YY}(f) = N_0/[\alpha^2 + (2\pi f)^2]$$

The output expected power $E[Y^2(t)]$ can be found from either one of

$$E[Y^2(t)] = R_{YY}(0) = N_0/2\alpha \quad \text{or} \quad E[Y^2(t)] = \int_{-\infty}^{\infty} S_{YY}(f)df = N_0/2\alpha$$

If the input $X(t)$ to a linear system is Gaussian, then the output will also be Gaussian [see Brown, 1983; or Gardner, 1990]. Thus, the output of a first-order linear time-invariant system driven by Gaussian white noise is the Ornstein–Uhlenbeck process, which is also a first-order Markov process.

For another example, let $X(t) = A\cos(w_0 t + \Theta)$, where the random amplitude A has zero mean, the random phase Θ is uniform on $[-\pi, \pi]$, and A and Θ are independent. As before, we obtain $R_{XX}(\tau) = \sigma_A^2 \cos(w_0\tau)$, from which it follows that $S_{XX}(f) = (\sigma_A^2/2)[\delta(f - w_0/2\pi) + \delta(f + w_0/2\pi)]$. These impulses in the psdf, called *spectral lines*, represent positive amounts of power at discrete frequencies.

Cross-Correlation of Processes

The *cross-correlation function* for two random processes $X(t)$ and $Y(t)$ is defined via

$$R_{XY}(t, t + \tau) \equiv E[X(t)Y(t + \tau)] \tag{6.55}$$

Let both processes be ws with zero means, so the covariance coincides with the correlation function. We say that two ws processes $X(t)$ and $Y(t)$ are *jointly ws* whenever $R_{XY}(t, t + \tau) = R_{XY}(\tau)$. In case $Y(t)$ is the output of

a filter with impulse response $h(t)$, we can find the cross-correlation $R_{XY}(\tau)$ between the input and output via

$$R_{XY}(\tau) = E[X(t)Y(t+\tau)] = E[X(t)\int_{-\infty}^{\infty} h(u)X(t+\tau-u)\mathrm{d}u]$$

$$= \int_{-\infty}^{\infty} h(u)E[X(t)X(t+\tau-u)]\mathrm{d}u \qquad (6.56)$$

$$= \int_{-\infty}^{\infty} h(u)R_{XX}(\tau-u)\mathrm{d}u = R_{XX}(\tau)h(\tau)$$

Cross-correlation functions of ws processes satisfy (1) $R_{XY}(-\tau) = R_{YX}(\tau)$, (2) $|R_{XY}(\tau)|^2 \le R_{XX}(0)R_{YY}(0)$, and (3) $|R_{XY}(\tau)| \le (1/2)[R_{XX}(0) + R_{YY}(0)]$. The first follows from the definition, while the second comes from expanding $E[\{Y(t+\tau) - \alpha X(t)\}^2] \ge 0$. The third comes from the fact that the geometric mean cannot exceed the arithmetic mean [see Peebles, 1987, p. 154].

Taking the Fourier transform of the leftmost and rightmost sides of Equation (6.56) yields

$$S_{XY}(f) = S_{XX}(f)H(f) \qquad (6.57)$$

The Fourier transform of the cross-correlation function is the *cross-spectral density function*

$$S_{XY}(f) = \int_{-\infty}^{\infty} R_{XY}(\tau)\mathrm{e}^{-\mathrm{j}2\pi f\tau}\mathrm{d}\tau \qquad (6.58)$$

According to Gardner [1990, p. 228], this is a *spectral correlation density function* that does not represent power in any sense.

Equation (6.57) suggests a method for identifying a linear time-invariant system. If the system is subjected to a ws input $X(t)$ and the power spectral density of $X(t)$ and the cross-spectral density of $X(t)$ and the output $Y(t)$ are measured, then the ratio yields the system transfer function

$$H(f) = S_{XY}(f)/S_{XX}(f) \qquad (6.59)$$

In fact, it can be shown that this method gives the best linear time-invariant model of the (possibly time varying and nonlinear) system in the sense that the time-averaged mean-square error between the outputs of the actual system and of the model, when both are subjected to the same input, is minimized [see Gardner, 1990, pp. 282–286].

As an application, suppose that an undersea sonar-based device is to find the range to a target, as shown in Figure 6.10, by transmitting a sonar signal $X(t)$ and receiving the reflected signal $Y(t)$. If v is the velocity of the sonar signal, and τ_o is the offset that maximizes the cross-correlation $R_{XY}(\tau)$, then the range (distance) d can be determined from $d = v\tau_o/2$ (note that the signal travels twice the range d).

Coherence

When $X(t)$ and $Y(t)$ have no spectral lines at f, the finite spectral correlation $S_{XY}(f)$ is actually a spectral covariance and the two associated variances are $S_{XX}(f)$ and $S_{YY}(f)$. We can normalize $S_{XY}(f)$ to obtain a *spectral correlation coefficient* $Y_{XY}(f)$ defined by

$$Y_{XY}(f)^2 = |S_{XY}(f)|^2/S_{XX}(f)S_{YY}(f) \qquad (6.60)$$

We call $Y_{XY}(f)$ the *coherence function*. It is a measure of the power correlation of $X(t)$ and $Y(t)$ at each frequency f. When $Y(t) = X(t) * h(t)$, it has a maximum: by Equation (6.53), Equation (6.59), and
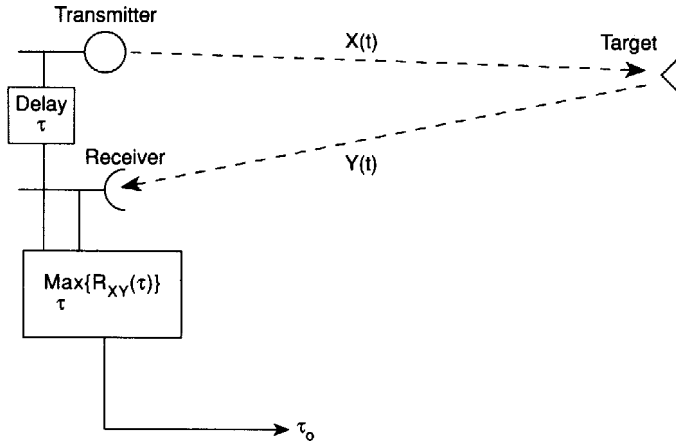
FIGURE 6.10   A sonar range finder.

Equation (6.60), $|Y_{XY}(f)|^2 = |S_{XX}(f) \cdot H(f)|^2 / [S_{XX}(f) \cdot S_{XX}(f) \cdot |H(f)|^2] = 1$. In the general case we have

$$|S_{XY}(f)| \leqslant [S_{XX}(f)S_{YY}(f)]^{1/2} \tag{6.61}$$

Upon minimizing the mean-square error $\varepsilon = E[(Y(t) - X(t)h(t))^2]$ over all possible impulse response functions $h(t)$, the optimal one, $h_{\mathrm{o}}(t)$, has transfer function

$$H_{\mathrm{o}}(f) = S_{XY}(f)/S_{XX}(f) \tag{6.62}$$

Further, the resultant minimum value is given by

$$\varepsilon_{\min} = \int_{-\infty}^{\infty} S_{YY}(f)[1 - |Y_{XY}(f)|^2]\mathrm{d}f$$

[see Gardner, 1990, pp. 434–436; or Bendat and Piersol, 1986]. At frequencies f where $|Y_{XY}(f)| \approx 1, \varepsilon_{\min} \approx 0$. Thus $1 - |Y_{XY}(f)|^2$ is the mean-square proportion of $Y(t)$ not accounted for by $X(t)$, while $|Y_{XY}(f)|^2$ is the proportion due to $X(t)$. When $Y(t) = X(t)h(t), \varepsilon_{\min} = 0$

The optimum system $H_{\mathrm{o}}(f)$ of Equation (6.62) is known as the *Wiener filter* for minimum mean-square error estimation of one process $Y(t)$ using a filtered version of another process $X(t)$ [see Gardner, 1990; or Peebles, 1987, p. 262].

Ergodicity

When the **time average**

$$\lim_{T\to\infty}(1/T) \int_{-T/2}^{T/2} X(t)\mathrm{d}t$$

exists and equals the corresponding expected value $E[X(t)]$, then the process $X(t)$ is said to possess an *ergodic property associated with the mean*. There are ergodic properties associated with the mean, autocorrelation (and power spectral density), and all finite-order joint moments, as well as finite-order joint pdf's. If a process has all possible ergodic properties, it is said to be an *ergodic process*.

Let $Y(t) = g[X(t + t_1), \ldots, X(t + t_n)]$, where $g[\cdot]$ is any nonrandom real function, so that $Y(t)$ is a function of a finite number of time samples of a strongly stationary process. For example, let (1) $Y(t) = X(t + t_1)X(t + t_2)$, $E[Y(t)] = R_{XX}(t_1 - t_2)$ and (2) $Y(t) = 1$ if $X(t) < x$, $Y(t) = 0$, otherwise, so that

$$E[Y(t)] = 1 \cdot P(X(t) < x) + 0 \cdot P(X(t) < x) = P(X(t) < x) = \int_{-\infty}^{1} f_{X(t)}(z)\mathrm{d}z$$

We want to know under what conditions the mean-square error between the time average

$$\langle Y(t) \rangle_T \equiv (1/T) \int_{-T/2}^{T/2} Y(t)\mathrm{d}t$$

and the expected value $E[Y(t)]$ will converge to zero. It can be shown that a necessary and sufficient condition for the mean-square ergodic property

$$\lim_{T \to \infty} E[\{\langle Y(t) \rangle - E[Y(t)]\}^2 = 0 \tag{6.63}$$

to hold is that

$$\lim_{T \to \infty} (1/T) \int_{0}^{T} C_{YY}(\tau)\mathrm{d}\tau = 0 \tag{6.64}$$

For example, if $C_{YY}(\tau) \to 0$ as $\tau \to \infty$, then Equation (6.64) will hold, and thus Equation (6.63) will also, where $C_{YY}(\tau)$ is the covariance function of $Y(t)$. As long as the two sets of rv's $\{X(t + t_1), \ldots, X(t + t_n)\}$ and $\{X(t + t_1 + \tau), \ldots, X(t + t_n + \tau)\}$ become independent of each other as $\tau \to \infty$, the above condition holds, so Equation (6.63) holds [see Gardner, 1990, pp. 163–174].

In practice, if $X(t)$ exhibits ergodicity associated with the autocorrelation, then we can estimate $R_{XX}(\tau)$ using the time average

$$\langle X(t)X(t + \tau) \rangle_T \equiv (1/T) \int_{-T/2}^{T/2} X(t)X(t + \tau)dt \tag{6.65}$$

In this case the mean-square estimation error $E[\{\langle X(t)X(t + \tau)\rangle_T - R_{XX}(\tau)\}^2]$ will converge to zero as T increases to infinity, and the power spectral density $S_{XX}(f)$ can also be estimated via time averaging [see Gardner, 1990, pp. 230–231].

Defining Terms

Autocorrelation function: A function $R_{XX}(t, t + \tau) = E[X(t)X(t + \tau)]$ that measures the degree to which any two rv's $X(t)$ and $X(t + \tau)$, at times t and $t + \tau$, are correlated.

Coherence function: A function of frequency f that provides the degree of correlation of two stochastic processes at each f by the ratio of their cross-spectral density function to the product of their power spectral density functions.

Power spectral density function: The Fourier transform of the autocorrelation function of a stochastic process $X(t)$, denoted by $S_{XX}(f)$. The area under its curve between f_1 and f_2 represents the total power over all t in $X(t)$ in the band of frequencies f_1 to f_2. Its dimension is watts per Hz.

Sample function: A real-valued function $x(t)$ of t where at each time t the value $x(t)$ at the argument t was determined by the outcome of a rv $X_t = x(t)$.

Stochastic process: A collection of rv's $\{X_t : t \in T\}$, where T is an ordered set such as the real numbers or integers [$X(t)$ is also called a random function, on the domain T].

Time average: Any real function $g(t)$ of time has average value g_{ave} on the interval $[a,b]$ such that the rectangular area $g_{ave}(b - a)$ is equal to the area under the curve between a and b, i.e., $g_{ave} = [1/(b-a)] \int_a^b g(t) dt$. The time average of a sample function $x(t)$ is the limit of its average value over $[0,T]$ as T goes to infinity.

Weakly stationary: The property of a stochastic process $X(t)$ whose mean $E[X(t)] = \mu(t)$ is a fixed constant μ over all time t, and whose autocorrelation is also independent of time in that $R_{XX}(t, t + \tau) = R_{XX}(s + t, s + t + \tau)$ for any s. Thus, $R_{XX}(t, t + \tau) = R_{XX}(0, \tau) = R_{XX}(\tau)$.

Acknowledgment

References

J.S. Bendat and A.G. Piersol, *Random Data: Analysis and Measurement,* 2nd ed., New York: Wiley-Interscience, 1986.

R.G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering,* New York: Wiley, 1983.

W.A. Gardner, *Introduction to Random Processes,* 2nd ed., New York: McGraw-Hill, 1990.

P.Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles,* 2nd ed., New York: McGraw-Hill, 1987.

Further Information

The IEEE Individual Learning Package, *Random Signal Analysis with Random Processes and Kalman Filtering*, prepared for the IEEE in 1989 by Carl G. Looney, IEEE Educational Activities Board, PO Box 1331, Piscataway, NJ 08855–1331.

R. Iranpour and P. Chacon, *Basic Stochastic Processes: The Mark Kac Lectures,* New York: Macmillan, 1988.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* 3rd ed., New York: Macmillan, 1991.

6.4   The Sampling Theorem

Robert J. Marks II

Most signals originating from physical phenomena are analog. Most computational engines, on the other hand, are digital. Transforming from analog to digital is straightforward: we simply sample. Regaining the original signal from these samples and assessing the information lost in the sampling process are the fundamental questions addressed by the **sampling theorem**.

The fundamental result of the sampling theorem is, remarkably, that a bandlimited signal is uniquely specified by its sufficiently close equally spaced samples. Indeed, the sampling theorem illustrates how the original signal can be regained from knowledge of the samples and the sampling rate at which they were taken.

Popularization of the sampling theorem is credited to Shannon [1948] who, in 1948, used it to show the equivalence of the information content of a bandlimited signal and a sequence of discrete numbers. Shannon was aware of the pioneering work of Whittaker [1915] and Whittaker's son [1929] in formulating the sampling theorem. Kotel'nikov's [1933] independent discovery in the then Soviet Union deserves mention. Higgins [1985] credits Borel [1897] with first recognizing that a signal could be recovered from its samples.

Surveys of sampling theory are in the widely cited paper of Jerri [1977] and in two books by the author [1991, 1993]. Marvasti [1987] has written a book devoted to nonuniform sampling.

The Cardinal Series

If a signal has finite energy, the minimum **sampling rate** is equal to two samples per period of the highest frequency component of the signal. Specifically, if the highest frequency component of the signal is B Hz, then

the signal, $x(t)$, can be recovered from the samples by

$$x(t) = \frac{1}{\pi} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \frac{\sin[\pi(2Bt - n)]}{2Bt - n} \tag{6.66}$$

The frequency B is also referred to as the signal's bandwidth and, if B is finite, $x(t)$ is said to be bandlimited. The signal, $x(t)$, is here being sampled at a rate of $2B$ samples per second. If sampling were done at a lower rate, the replications would overlap and the information about $X(\omega)$ [and thus $x(t)$] is irretrievably lost. Undersampling results in *aliased* data. The minimum sampling rate at which **aliasing** does not occur is referred to as the **Nyquist rate** which, in our example, is $2B$. Equation (6.66) was dubbed the **cardinal series** by the junior Whittaker [1929].

A signal is bandlimited in the low-pass sense if there is a $B > 0$ such that

$$X(\omega) = X(\omega)\Pi\left(\frac{\omega}{4\pi B}\right) \tag{6.67}$$

where the gate function $\Pi(\xi)$ is one for $\xi \leq 1/2$ and is otherwise zero, and

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \tag{6.68}$$

is the **Fourier transform** of $x(t)$. That is, the spectrum is identically zero for $|\omega| \leq 2\pi B$. The B parameter is referred to as the signal's bandwidth. The inverse Fourier transform is

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x(t) e^{j\omega t} d\omega \tag{6.69}$$

The sampling theorem reduces the normally continuum infinity of ordered pairs required to specify a function to a countable—although still infinite—set. Remarkably, these elements are obtained directly by sampling.

How can the cardinal series interpolate uniquely the bandlimited signal from which the samples were taken? Could not the same samples be generated from another bandlimited signal? The answer is no. Band-limited functions are smooth. Any behavior deviating from smooth would result in high-frequency components which in turn invalidates the required property of being bandlimited. The smoothness of the signal between samples precludes arbitrary variation of the signal there.

Let's examine the cardinal series more closely. Evaluate Equation (6.74) at $t = m/2B$. Since sinc(n) is one for $n = 0$ and is otherwise zero, only the sample at $t = m/2B$ contributes to the interpolation at that point. This is illustrated in Figure 6.11, where the reconstruction of a signal from its samples using the cardinal series is shown. The value of $x(t)$ at a point other than a sample location [e.g., $t = (m+1/2)/2B$] is determined by all of the sample values.
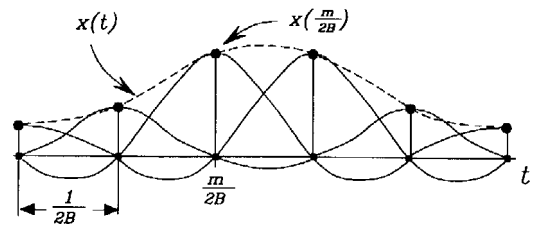
FIGURE 6.11 Illustration of the interpolation that results from the Cardinal series. A sinc function, weighted by the sample, is placed at each sample bottom. The sum of the sincs exactly generates the original bandlimited function from which the samples were taken.

Proof of the Sampling Theorem

Borel [1897] and Shannon [1948] both discussed the sampling theorem as the Fourier transform dual of the Fourier series. Let $x(t)$ have a bandwidth of B. Consider the periodic signal

$$Y(\omega) = \sum_{n=-\infty}^{\infty} X(\omega - 4\pi nB) \tag{6.70}$$

The function $Y(\omega)$ is a periodic function with period $4\pi B$. From Equation (6.67) $X(\omega)$ is zero for $\omega > 2\pi B$ and is thus finite in extent. The terms in Equation (6.70) therefore do not overlap. Periodic functions can be expressed as a Fourier series.

$$Y(\omega) = \sum_{n=-\infty}^{\infty} \alpha_n \exp\left(\frac{-jn\omega}{2B}\right) \tag{6.71}$$

where the Fourier series coefficients are

$$\alpha_n = \frac{1}{4\pi B} \int_{-2\pi B}^{2\pi B} Y(\omega) \exp\left(\frac{jn\omega}{2B}\right) d\omega$$

or

$$\alpha_n = \frac{1}{2B} x\left(\frac{n}{2B}\right) \tag{6.72}$$

where we have used the inverse Fourier transform in Equation (6.69). Substituting into the Fourier series in Equation (6.71) gives

$$Y(\omega) = \frac{1}{2B} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \exp\left(\frac{-jn\omega}{2B}\right) \tag{6.73}$$

Since a period of $Y(\omega)$ is $X(\omega)$, we can get back the original spectrum by

$$X(\omega) = Y(\omega)\Pi\left(\frac{\omega}{4\pi B}\right)$$

Substitute Equation (6.73) and inverse transforming gives, using Equation (6.69),

$$x(t) = \frac{1}{4\pi B} \int_{-2\pi B}^{2\pi B} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \exp\left(\frac{-jn\omega}{2B}\right) e^{j\omega t} d\omega$$

or

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \sin c(2Bt - n) \tag{6.74}$$

where

$$\sin c(t) = \frac{\sin \pi t}{\pi t}$$

is the inverse Fourier transform of $\Pi(\omega/2\pi)$. Equation (6.74) is, of course, the cardinal series.

The sampling theorem generally converges uniformly, in the sense that

$$\lim_{N \to \infty} |x(t) - x_N(t)|^2 = 0$$

where the truncated cardinal series is

$$x_N(t) = \sum_{n=-N}^{N} x\left(\frac{n}{2B}\right) \operatorname{sin} c(2Bt - n) \qquad (6.75)$$

Sufficient conditions for uniform convergence are [Marks, 1991]

1. the signal, $x(t)$, has finite energy, E,

$$E = \int_{-\infty}^{\infty} |x(t)|^2 \mathrm{d}t < \infty$$

2. or $X(\omega)$ has finite area,

$$A = \int_{-\infty}^{\infty} |X(\omega)| \mathrm{d}\omega < \infty$$

Care must be taken in the second case, though, when singularities exist at $\omega = \pm 2\pi B$. Here, sampling may be required to be strictly greater than $2B$. Such is the case, for example, for the signal, $x(t) = \sin(2\pi Bt)$. Although the signal is bandlimited, and although its Fourier transform has finite area, all of the samples of $x(t)$ taken at $t = n/2B$ are zero. The cardinal series in Equation (6.74) will thus interpolate to zero everywhere. If the sampling rate is a bit greater than $2B$, however, the samples are not zero and the cardinal series will uniformly converge to the proper answer.

The Time-Bandwidth Product

The cardinal series requires knowledge of an infinite number of samples. In practice, only a finite number of samples can be used. If most of the energy of a signal exists in the interval $0 \leq t \leq T$, and we sample at the Nyquist rate of $2B$ samples per second, then a total of $S = \langle 2BT \rangle$ samples are taken. ($\langle \theta \rangle$ denotes the largest number not exceeding θ.) The number S is a measure of the degrees of freedom of the signal and is referred to as its **time-bandwidth product**. A 5-min single-track audio recording requiring fidelity up to 20,000 Hz, for example, requires a minimum of $S = 2 \times 20,000 \times 5 \times 60 = 12$ million samples. In practice, audio sampling is performed well above the Nyquist rate.

Sources of Error

Exact interpolation using the cardinal series assumes that (1) the values of the samples are known exactly, (2) the sample locations are known exactly, and (3) an infinite number of terms are used in the series. Deviation from these requirements results in interpolation error due to (1) data noise, (2) jitter, and (3) truncation, respectively. The effect of data error on the restoration can be significant. Some innocently appearing sampling theorem generalizations, when subjected to performance analysis in the presence of data error, are revealed as ill-posed. In other words, a bounded error on the data can result in unbounded error on the restoration [Marks, 1991].

Data Noise

The source of data noise can be the signal from which samples are taken, or from round-off error due to finite sampling precision. If the noise is additive and random, instead of the samples

$$x\left(\frac{n}{2B}\right)$$

we must deal with the samples

$$x\left(\frac{n}{2B}\right) + \xi\left(\frac{n}{2B}\right)$$

where $\xi(t)$ is a stochastic process. If these noisy samples are used in the cardinal series, the interpolation, instead of simple $x(t)$, is

$$x(t) + \eta(t)$$

where the interpolation noise is

$$\eta(t) = \sum_{n=-\infty}^{\infty} \xi\left(\frac{n}{2B}\right) \sin c(2Bt - n)$$

If $\xi(t)$ is a zero mean process, then so is the interpolation noise. Thus, the noisy interpolation is an unbiased version of $x(t)$. More remarkably, if $\xi(t)$ is a zero-mean (wide-sense) stationary process with uncertainty (variance) σ^2, then so is $\eta(t)$. In other words, *the uncertainty at the sample point locations is the same as at all points of interpolation* [Marks, 1991].

Truncation

The truncated cardinal series is in Equation (6.75). A signal cannot be both bandlimited and of finite duration. Indeed, a bandlimited function cannot be identically zero over any finite interval. Thus, other than the rare case where an infinite number of the signal's zero crossings coincide with the sample locations, truncation will result in an error.

The magnitude of this **truncation error** can be estimated through the use of Parseval's theorem for the cardinal series that states

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt = \frac{1}{2B} \sum_{-\infty}^{\infty} \left| x\left(\frac{n}{2B}\right) \right|^2$$

The energy of a signal can thus be determined directly from either the signals or the samples. The energy associated with the truncated signal is

$$E_N = \frac{1}{2B} \sum_{-N}^{N} \left| x\left(\frac{n}{2B}\right) \right|^2$$

If $E - E_N \ll E$, then the truncation error is small.

Jitter

Jitter occurs when samples are taken near to but not exactly at the desired sample locations. Instead of the samples $x(n/2W)$, we have the samples

$$x\left(\frac{n}{2W} - \sigma_n\right)$$

where σ_n is the jitter offset of the nth sample. For jitter, the σ_n's are not known. If they were, an appropriate nonuniform sampling theorem [Marks, 1993; Marvasti, 1987] could be used to interpolate the signal.

Using the jittered samples in the cardinal series results in an interpolation that is not an unbiased estimate of $x(t)$. Indeed, if the probability density function of the jitter is the same at all sample locations, the expected

value of the jittered interpolation is the convolution of $x(t)$ with the probability density function of the jitter. This bias can be removed by inverse filtering at a cost of decreasing the signal-to-noise ratio of the interpolation [Marks, 1993].

Generalizations of the Sampling Theorem

There exist numerous generalizations of the sampling theorem [Marks, 1991; Marks, 1993].

1. **Stochastic processes.** A wide-sense stationary stochastic process, $\chi(t)$, is said to be bandlimited if its autocorrelation, $R_\chi(t)$, is a bandlimited function. The cardinal series

$$\hat{\chi}(t) = \sum_{n=-\infty}^{\infty} \chi\left(\frac{n}{2B}\right) \sin c(2Bt - n)$$

   converges to $\chi(t)$ in the sense that

$$E[|\hat{\chi}(t) - \chi(t)|^2] = 0$$

   where E denotes expectation.

2. **Nonuniform sampling.** There exist numerous scenarios wherein interpolation can be performed from samples that are not spaced uniformly. Marvasti [1987] devotes a book to the topic.

3. **Kramer's generalization.** Kramer generalized the sampling theorem to integral transforms other than Fourier, for example, to Legendre and Laguerre transforms.

4. **Papoulis' generalization.** Shannon noted that a bandlimited signal could be restored when sampling was performed at half the Nyquist rate if, at every sample location, a sample of the signal's derivative were also taken. Recurrent nonuniform sampling is where P samples are spaced the same in every P Nyquist intervals. Another sampling scenario is when a signal and its Hilbert transform are both sampled at half their respective Nyquist rates. Restoration of the signal from these and numerous other sampling scenarios are subsumed in an eloquent generalization of the sampling theorem by Papoulis.

5. **Lagrangian interpolation.** Lagrangian interpolation is a topic familiar in numerical analysis. An Nth order polynomial is fit to $N+1$ arbitrarily spaced sample points. If an infinite number of samples are equally spaced, then Lagrangian interpolation is equivalent to the cardinal series.

6. **Trigonometric polynomials.** All periodic bandlimited signals can be expressed as trigonometric polynomials (i.e., a Fourier series with a finite number of terms). If the series has M terms, then the signal has M degrees of freedom which can be determined from M samples taken within a single period.

7. **Multidimensional sampling theorems.** Multidimensional signals, such as images, require dimensional extensions of the sampling theorem. The sampling of the signal now requires geometrical interpretation. Uniform sampling of an image, for example, can either be done on a rectangular or hexagonal grid. The minimum sampling density for one geometry may differ from that of another. The smallest sampling density that does not result in aliasing can be achieved, in many cases, with a number of different uniform sampling geometries and is referred to as the Nyquist density. Interestingly, sampling can sometimes be performed below the Nyquist density with nonuniform sampling geometries such that the multidimensional signal can be restored. Such is not the case for one dimension.

8. **Continuous sampling.** When a signal is known on one or more disjoint intervals, it is said to have been continuously sampled. Divide the time line into intervals of T. Periodic continuous sampling assumes that the signal is known on each interval over an interval of αT where α is the duty cycle. Continuously sampled signals can be accurately interpolated even in the presence of aliasing. Other continuously

sampled cases, each of which can be considered as a limiting case of continuously periodically sampled restoration, include

    a. **Interpolation.** The tails of a signal are known and we wish to restore the middle.

    b. **Extrapolation.** We wish to generate the tails of a function with knowledge of the middle.

    c. **Prediction.** A signal for $t > 0$ is to be estimated from knowledge of the signal for $t < 0$.

Final Remarks

Since its popularization in the late 1940s, the sampling theorem has been studied in depth. More than 1000 papers have been generated on the topic [Marks, 1993]. Its understanding is fundamental in matching the largely continuous world to digital computation engines.

Defining Terms

Aliasing: A phenomenon that occurs when a signal is undersampled. High-frequency information about the signal is lost.

Cardinal series: The formula by which samples of a bandlimited signal are interpolated to form a continuous time signal.

Fourier transform: The mathematical operation that converts a time-domain signal into the frequency domain.

Jitter: A sample is temporally displaced by an unknown, usually small, interval.

Kramer's generalization: A sampling theory based on other than Fourier transforms and frequency.

Lagrangian interpolation: A classic interpolation procedure used in numerical analysis. The sampling theorem is a special case.

Nyquist rate: The minimum sampling rate that does not result in aliasing.

Papoulis' generalization: A sampling theory applicable to many cases wherein signal samples are obtained either nonuniformly and/or indirectly.

Sampling rate: The number of samples per second.

Sampling theorem: Samples of a bandlimited signal, if taken close enough together, exactly specify the continuous time signal from which the samples were taken.

Signal bandwidth: The maximum frequency component of a signal.

Time bandwidth product: The product of a signal's duration and bandwidth approximates the number of samples required to characterize the signal.

Truncation error: The error that occurs when a finite number of samples are used to interpolate a continuous time signal.

References

E. Borel, "Sur l'interpolation," *C.R. Acad. Sci. Paris,* vol. 124, pp. 673–676, 1897.

J.R. Higgins, "Five short stories about the cardinal series," *Bull. Am. Math. Soc.,* vol. 12, pp. 45–89, 1985.

A.J. Jerri, "The Shannon sampling theorem—its various extension and applications: a tutorial review," *Proc. IEEE,* vol. 65, pp. 1565–1596, 1977.

V.A. Kotel'nikov, "On the transmission capacity of 'ether' and wire in electrocommunications," *Izd. Red. Upr. Svyazi RKKA* (Moscow), 1933.

R.J. Marks II, *Introduction to Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1991.

R.J. Marks II, Ed., *Advanced Topics in Shannon Sampling and Interpolation Theory*, New York: Springer-Verlag, 1993.

F.A. Marvasti, *A Unified Approach to Zero-Crossing and Nonuniform Sampling*, Oak Park, Ill.: Nonuniform, 1987.

C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379, 623, 1948.

E.T. Whittaker, "On the functions which are represented by the expansions of the interpolation theory," *Proc. Royal Society of Edinburgh*, vol. 35, pp. 181–194, 1915.

J.M. Whittaker, "The Fourier theory of the cardinal functions," *Proc. Math. Soc. Edinburgh*, vol. 1, pp. 169–176, 1929.

A.I. Zayed, Advances in Shannon's Sampling Theory, Boca Raton, Fla.: CRC Press, 1993.

Further Information

An in-depth study of the sample theorem and its numerous variations is provided in R. J. Marks II, Ed., *Introduction to Shannon Sampling and Interpolation Theory*, New York:Springer-Verlag, 1991.

In-depth studies of modern sampling theory with over 1000 references are available in R. J. Marks II, Ed., *Advanced Topics in Shannon Sampling and Interpolation Theory*, New York: Springer-Verlag, 1993.

The specific case of nonuniform sampling is treated in the monograph by F. A. Marvasti, *A Unified Approach to Zero-Crossing and Nonuniform Sampling*, Oak Park, Ill:Nonuniform, 1987.

The sampling theorem is treated generically in the *IEEE Transactions on Signal Processing*. For applications, topical journals are the best source of current literature.

6.5   Channel Capacity

Sergio Verdú

Information Rates

Tens of millions of users access the Internet daily via standard telephone lines. **Modems** operating at data rates of up to 28,800 bits per second enable the transmission of text, audio, color images, and even low-resolution video. The progression in modern technology for the standard telephone channel shown in Figure 6.12 exhibits, if not the exponential increases ubiquitous in computer engineering, then a steady lope of abut 825 bits per second per year.

Few technological advances can result in as many time-savings for worldwide daily life as advances in modem information rates. However, modem designers are faced with a fundamental limitation in the maximum transmissible information rate. Every communication channel has a number associated with it called **channel capacity**, which determines the maximum information rate that can flow through the channel regardless of the complexity of the transmitting and receiving devices. Thus, the progression of modem rates shown in Figure 6.12 is sure to come to a halt. But, at what rate? Answering this question for any communication channel model is one of the major goals of information theory—a discipline founded in 1948 by Claude E. Shannon [Shannon, 1948].

Communication Channels

The communication channel is the set of devices and systems that connects the transmitter to the receiver. The transmitter and receiver consist of an **encoder** and **decoder**, respectively, which translate the information stream produced by the source into a signal suitable for channel transmission and vice versa (Figure 6.13). For example, in the case of the telephone line, two communication channels (one in each direction) share the same physical channel that connects the two modems. That physical channel usually consists of twisted copper wires at both ends and a variety of switching and signal processing operations that occur at the telephone exchanges. The modems themselves are not included in the communication channel. A microwave radio link is another example of a communication channel that consists of an amplifier and an antenna (at both ends) and a certain portion of the radio spectrum. In this case, the communication channel model does not fully correspond with the physical channel. Why not, for example, view the antenna as part of the transmitter rather than the channel (Figure 6.13)? Because considerations other than the optimization of the efficiency of the link are likely to dictate the choice of antenna size. This illustrates that the boundaries encoder–channel and channel–decoder in Figure 6.13 are not always uniquely defined. This suggests an alternative definition of a channel as that part of the communication system that the designer is unable or unwilling to change.

Bits per second
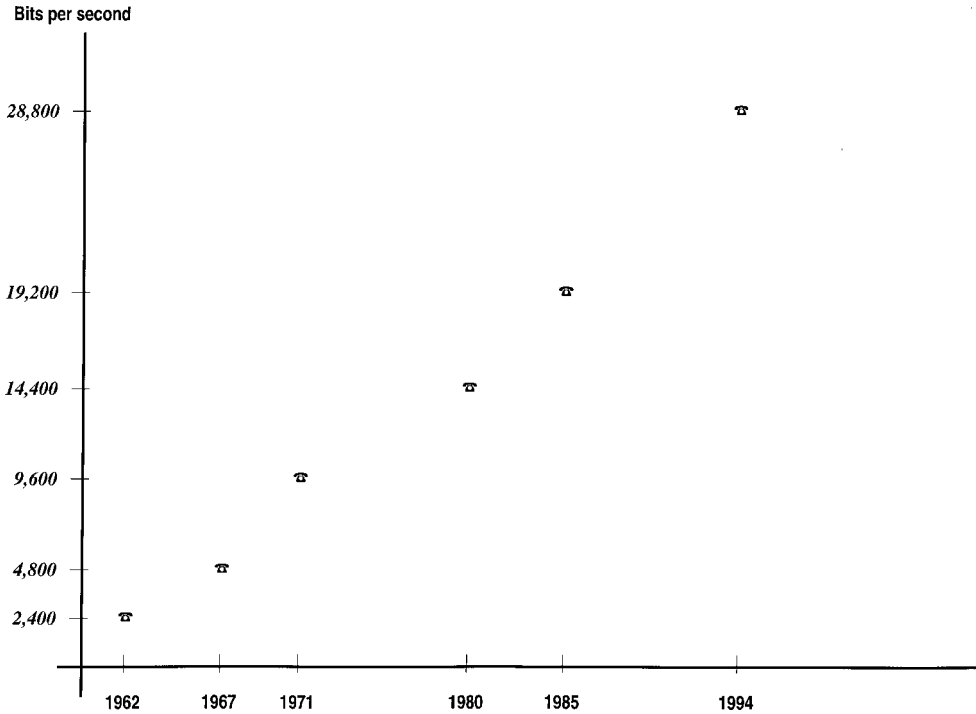
FIGURE 6.12   Information rates of modems for telephone channels.
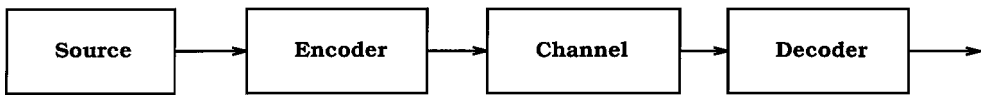
FIGURE 6.13   Elements of a communication system.

A channel is characterized by the probability distributions of the output signals given every possible input signal. Channels are divided into (1) **discrete-time channels** and (2) **continuous-time channels** depending on whether the input/output signals are sequences or functions of a real variable. Some examples are as follows.

Example 6.1.   Binary Symmetric Channel

A discrete-time **memoryless channel** with binary inputs and outputs (Figure 6.14) where the probabilities that 0 and 1 are received erroneously are equal.

Example 6.2.   Z-Channel

A discrete-time memoryless channel with binary inputs and outputs (Figure 6.15) where 0 is received error-free.

Example 6.3.   Erasure Channel

A discrete-time memoryless channel with binary inputs and ternary outputs (Figure 6.16). The symbols 0 and 1 cannot be mistaken for each other but they can be "erased".

Example 6.4.   White Gaussian Discrete-Time Channel

A discrete-time channel whose output sequence is given by

$$y_i = x_i + n_i \tag{6.76}$$

where x_i is the input sequence and n_i is a sequence of independent Gaussian random variables with equal variance.
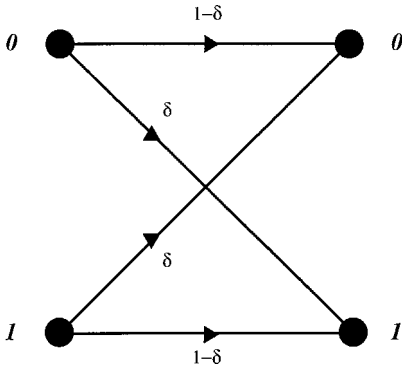
FIGURE 6.14 Binary symmetric channel.
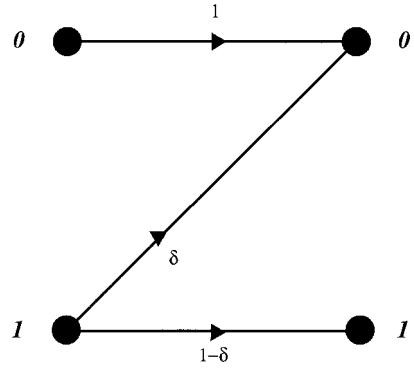
FIGURE 6.15 Z-channel.
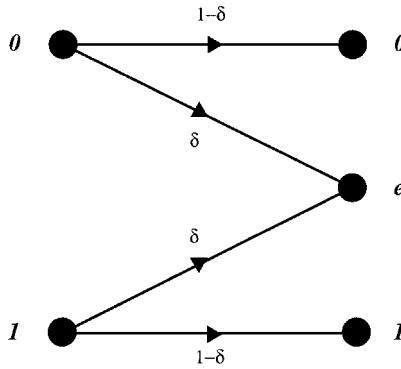
FIGURE 6.16 Erasure channel.

Example 6.5. Linear Continuous-Time Gaussian Channel

A continuous-time channel whose output signal is given by (Figure 6.17)

$$y(t) = h(t) * x(t) + n(t) \tag{6.77}$$

where $x(t)$ is the input signal, $n(t)$ is a stationary Gaussian process, and $h(t)$ is the impulse response of a linear time-invariant system. The telephone channel is typically modeled by Equation (6.77).

The goal of the encoder Figure 6.13 is to convert strings of binary data (messages) into channel-input signals. Source strings of m bits are translated into channel input strings of n symbols (with $m \leq n$) for discrete
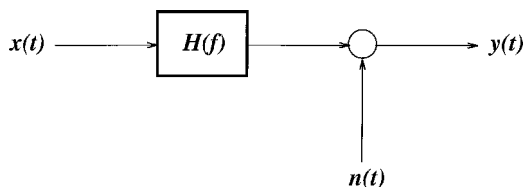
FIGURE 6.17 Linear continuous-time Gaussian channel.

channels, and into continuous-time signals of duration T for continuous-time channels. The channel code (or more precisely the codebook) is the list of 2^m **codewords** (channel input signals) that may be sent by the encoder. The **rate** of the code is equal to the logarithm of its size divided by the duration of the codewords. Thus, the rate is equal to

$$\frac{m}{n} \quad \text{bits per channel use}$$

for a discrete-time channel, whereas it is equal to

$$\frac{m}{T} \quad \text{bits per second}$$

for a continuous-time channel.

Once a codeword has been chosen by the encoder, the channel probabilistic mechanisms govern the distortion suffered by the transmitted signal. The role of the decoder is to recover the transmitted binary string (message) upon reception of the channel-distorted version of the transmitted codeword. To that end, the decoder knows the codebook used by the encoder. For most channels (including those above) there is a nonzero probability that the best decoder (maximum likelihood decoder) selects the wrong message. Thus, for a given channel the two figures of merit and of interest are the rate and the probability of error. The higher the tolerated probability of error, the higher the allowed rate; however, computing the exact tradeoff is a formidable task unless the code size either is very small or tends to infinity. The latter case was the one considered by Shannon and treated in the following section.

Reliable Information Transmission: Shannon's Theorem

Shannon [1948] considered the situation in which the codeword duration grows without bound. Channel capacity is the maximum rate for which encoders and decoders exist whose probability of error vanishes as the codewords get longer and longer.

Shannon's Theorem [Shannon, 1948] The capacity of a discrete memoryless channel is equal to

$$C = \max_X I(X; Y) \tag{6.78}$$

where $I(X;Y)$ stands for the input-output mutual information, which is a measure of the dependence of the input and the output defined as the divergence between the joint input/output distribution and the product of its marginals, $D(P_{XY} \| P_X P_Y)$. For any pair of probability mass functions P and Q defined on the same space, divergence is an asymmetric measure of their similarity:

$$D(P \| Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}. \tag{6.79}$$

Divergence is zero if both distributions are equal; otherwise it is strictly positive. The maximization in Equation (6.78) is over the set of input distributions. Although, in general, there is no closed-form solution for that optimization problem, an efficient algorithm was obtained by Blahut and Arimoto in 1972 [Blahut, 1987]. The distribution that attains the maximum in Equation (6.78) determines the statistical behavior of optimal codes and, thus, is of interest to the designer of the encoder. For the discrete memoryless channels mentioned above, the capacity is given by the following examples.
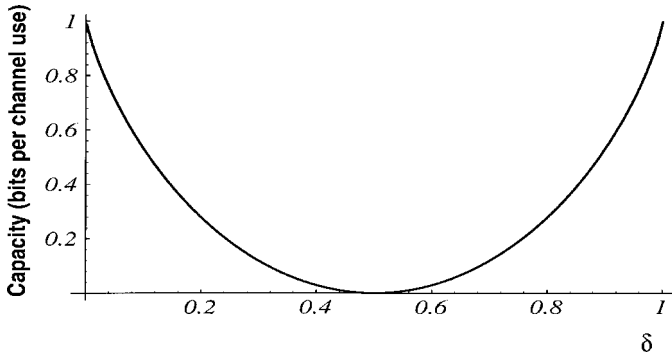
FIGURE 6.18   Capacity of the binary symmetric channel as a function of crossover probability.
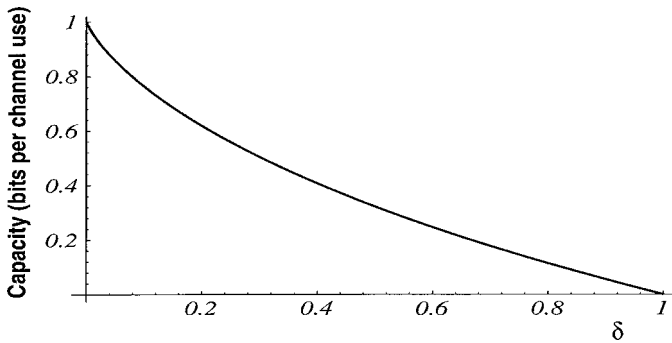
FIGURE 6.19   Capacity of the Z-channel.

Example 6.6.   Binary Symmetric Channel

$$C = 1 - \delta\log\frac{1}{\delta} - (1-\delta)\log\frac{1}{1-\delta}$$

attained by an equiprobable distribution and shown in Figure 6.18.

Example 6.7.   Z-Channel

$$C = \log(1 - \delta^{1/1-\delta} + \delta^{\delta/1-\delta})$$

attained for a distribution whose probability mass at 0 ranges from 1/2 ($\delta = 0$) to $1/e(\delta \rightarrow 1)$ (Figure 6.19).

Example 6.8.   Erasure Channel

$$C = 1 - \delta$$

attained for equiprobable inputs.

Oftentimes the designer is satisfied with not exceeding a certain fixed level of bit error rate, ϵ, rather than the more stringent criterion of vanishing probability of selecting the wrong block of data. In such a case, it is
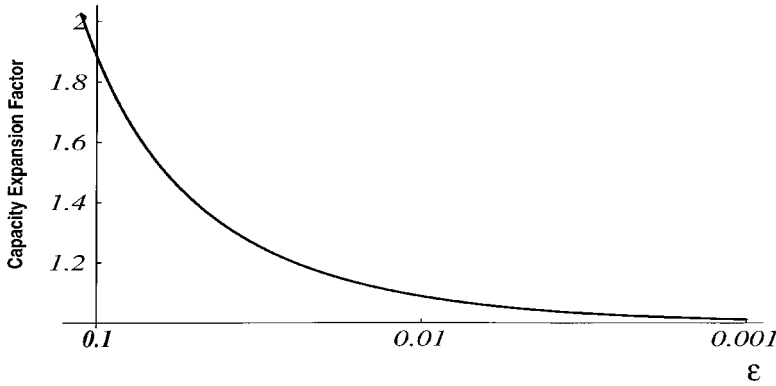
FIGURE 6.20   Capacity expansion factor as a function of bit-error-rate.

possible to transmit information at a rate equal to capacity times

$$\left(1 - \varepsilon \log \frac{1}{\varepsilon} - (1 - \varepsilon)\log \frac{1}{1 - \varepsilon}\right)^{-1}$$

which is shown in Figure 6.20.

   If, contrary to what we have assumed thus far, the message source in Figure 6.13 is not a source of pure bits, the significance of capacity can be extended to show that as long as the source entropy (see Chapter 6.6 on Data Compression) is below the channel capacity, an encoder/decoder pair exists that enables arbitrarily reliable communication. Conversely, if the source entropy is above capacity, then no such encoder/decoder pair exists.

Bandwidth and Capacity

The foregoing formulas for discrete channels do not lead to the capacity of continuous-time channels such as Example 5. We have seen that in the case of the telephone channel whose bandwidth is approximately equal to 3 kHz, capacity is lower bounded by 28,800 bits per second. How does bandwidth translate into capacity? The answer depends on the noise level and distribution. For example, if in the channel of Example 5, the noise is absent, capacity is infinite regardless of bandwidth. We can encode any amount of information as the binary expansion of a single scalar, which can be sent over the channel as the amplitude or phase of a single sinusoid; knowing the channel transfer function, the decoder can recover the transmitted scalar error-free. Clearly, such a transmission method is not recommended in practice because it hinges on the non-physical scenario of noiseless transmission.

   In the simplest special case of Example 5, the noise is white, the channel has an ideal flat transfer function with bandwidth B (in Hz), and the input power is limited. Then, the channel capacity is equal to

$$C = B \, \text{SNR}^{\text{dB}} \log_2 10^{0.1} \quad \text{bits per second} \tag{6.80}$$

where $\log_2 10^{0.1} = 0.33$ and SNR^{dB} is equal to the optimum signal-to-noise ratio (in dB) of a linear estimate of a flat input signal given the channel output signal. Such an optimum signal-to-noise ratio is equal to one plus the power alloted to the input divided by the noise power in the channel band, i.e.,

$$\text{SNR}^{\text{dB}} = 10 \log_{10}\left(1 + \frac{P}{BN_0}\right)$$

It is interesting to notice that as the bandwidth grows, the channel capacity does not grow without bound. It tends to

$$\frac{P}{N_0} \log_2 e \quad \text{bits per second}$$

where $\log_2 e = 1.44$. This means that the energy per bit necessary for reliable communication is equal to 0.69 times the noise power spectral density level. When the channel bandwidth is finite, the energy necessary to send one bit of information is strictly larger. The energy required to send one bit of information reliably can be computed for other (non-Gaussian) channels even in cases where expressions for channel capacity are not known [Verdú, 1990].

When the channel transfer function $H(f)$ and/or noise spectral density $N(f)$ are not flat, the constant in Equation (6.80) no longer applies. The so-called water-filling formula [Shannon, 1949] gives the channel capacity as

$$C = \frac{1}{2} \int \log\left(1 + \frac{\max\{0, w - M(f)\}}{M(f)}\right) df$$

where w is chosen so that

$$\int \max\{0, w - M(f)\} df = P$$

and

$$M(f) = \frac{N(f)}{|H(f)|^2}$$

The linear Gaussian-noise channel is a widely used model for space communication (in the power limited region) and for the telephone channel (in the bandwidth limited region). Thanks to the prevalence of digital switching and digital transmission in modern telephone systems, not only signal-to-noise ratios have improved over time but the Gaussian-noise model in Example 5 becomes increasingly questionable because quantization is responsible for a major component of the channel distortion. Therefore, future improvements in modem speeds are expected to arise mainly from finer modeling of the channel.

Due to the effect of time-varying received power (fading), several important channels fall outside the scope of Example 5 such as high-frequency radio links, tropospheric scatter links, and mobile radio channels.

Channel Coding Theorems

In information theory, the results that give a formula for channel capacity in terms of the probabilistic description of the channel are known as channel coding theorems. They typically involve two parts: an achievability part, which shows that codes with vanishing error probability exist for any rate below capacity; and a converse part, which shows that if the code rate exceeds capacity, then the probability of error is necessarily bounded away from zero. Shannon gave the first achievability results in [Shannon, 1948] for discrete memory channels. His method of proof, later formalized as the method of "typical sequences" (e.g., [Cover and Thomas, 1991]), is based on showing that the average probability of error of a code chosen at random vanishes with blocklength. Other known achievability proofs such as

Feinstein's [1954], Gallager's [1968], and the method of types [Csiszar and Korner, 1981] are similarly non-constructive. The discipline of coding theory deals with constructive methods to design codes that approach the Shannon limit. The first converse channel coding theorem was not given by Shannon, but by Fano in 1952. A decade after Shannon's pioneering paper, several authors obtained the first channel coding theorems for channels with memory [Dobrushin, 1963]. The most general formula for channel capacity known to date can be found in [Verdú and Han, 1994]. The capacity of channels with feedback was first considered by Shannon in 1961 [Shannon, 1961], with later developments for Gaussian channels summarized in [Cover and Thomas, 1991]. In his 1961 paper [Shannon, 1961], Shannon founded the discipline of multiuser information theory by posing several challenging channels with more than one transmitter and/or receiver. In contrast to the **multiaccess channel** (one receiver) which has been solved in considerable generality, the capacities of channels involving more than one receiver, such as **broadcast channels** [Cover, 1972] and **interference channels** remain unsolved except in special cases.

Channel capacity has been shown to have a meaning outside the domain of information transmission [Han and Verdú, 1993]: it is the minimum rate of random bits required to generate any input random process so that the output process is simulated with arbitrary accuracy.

Defining Terms

Blocklength: The duration of a codeword, usually in the context of discrete-time channels.

Channel capacity: The maximum rate for which encoders and decoders exist whose probability of error vanishes as the codewords get longer and longer.

Codeword: Channel-input signal chosen by the encoder to represent the message.

Communication channel: Set of devices and systems that connect the transmitter to the receiver, not subject to optimization.

Broadcast channel: A communication channel with one input and several outputs each connected to a different receiver such that possibly different messages are to be conveyed to each receiver.

Discrete memoryless channel: A discrete-time memoryless channel where each channel input and output takes a finite number of values.

Discrete-time channel: A communication channel whose input/output signals are sequences of values. Its capacity is given in terms of bits per "channel use".

Continuous-time channel: A communication channel whose input/output signals are functions of a real variable (time). Its capacity is given in terms of bits per second.

Interference channel: A channel with several inputs/outputs such that autonomous transmitters are connected to each input and such that each receiver is interested in decoding the message sent by one and only one transmitter.

Memoryless channel: A channel where the conditional probability of the output given the current input is independent of all other inputs or outputs.

Multiaccess channel: A channel with several inputs and one output such that autonomous transmitters are connected to each input and such that the receiver is interested in decoding the messages sent by all the transmitters.

Decoder: Mapping from the set of channel-output signals to the set of messages.

Maximum-likelihood decoder: A decoder which selects the message that best explains the received signal, assuming all messages are equally likely.

Encoder: Mapping from the set of messages to the set of input codewords.

Modem: Device that converts binary information streams into electrical signals (and vice-versa) for transmission through the voiceband telephone channel.

Rate: The rate of a code is the number of bits transmitted (logarithm of code size) per second for a continuous-time channel or per channel use for a discrete-time channel.

References

R.E. Blahut, *Principles of Information Theory.* Reading, Mass.: Addison-Wesley, 1987.

T.M. Cover, "Broadcast channels," *IEEE Trans. on Information Theory*, pp. 2–14, Jan. 1972.

T. M. Cover and J. A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York: Academic Press, 1981.

R.L. Dobrushin, *General Formulation of Shannon's Main Theorem in Information Theory*, American Mathematical Society Translations, pp. 323–438, 1963.

A. Feinstein, "A new basic theorem of information theory," *IRE Trans. PGIT,* pp. 2–22, 1954.

R.G. Gallager, *Information Theory and Reliable Communication,* New York: Wiley, 1968.

T.S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. on Information Theory*, IT-39, 752–772, May 1993.

C.E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.,* 27, 379–423, 623–656, July–Oct. 1948.

C.E. Shannon, "Communication in the presence of noise," *Proc. Institute of Radio Engineers,* 37, 10–21, 1949.

C.E. Shannon, "Two-way communication channels," *Proc. 4th. Berkeley Symp. Math. Statistics and Prob.*, pp. 611–644, 1961.

S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Information Theory,* IT-36(5), 1019–1030, Sept. 1990.

S. Verdú and T.S. Han, "A general formula for channel capacity," *IEEE Trans. on Information Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

Further Information

The premier journal and conference in the field of information theory are the *IEEE Trans. on Information Theory* and the *IEEE International Symposium on Information Theory,* respectively. *Problems of Information Transmission* is a translation of a Russian-language journal in information theory. The newsletter of the IEEE Information Theory Society regularly publishes expository articles.

6.6   Data Compression

Joy A. Thomas and Thomas M. Cover

Data compression is a process of finding the most efficient representation of an information source in order to minimize communication or storage. It often consists of two stages—the first is the choice of a (probabilistic) model for the source, and the second is the design of an efficient coding system for the model. Here, we will concentrate on the second aspect of the compression process, though we will touch on some common sources and models in the last section.

Thus, a data compressor (sometimes called a source coder) maps an information source into a sequence of bits, with a corresponding decompressor that, given these bits, provides a reconstruction of the source. Data compression systems can be classified into two types: lossless, where the reconstruction is exactly equal to the original source, and lossy, where the reconstruction is a distorted version of the original source. For lossless data compression, the fundamental lower bound on the rate of the data compression system is given by the entropy rate of the source. For lossy data compression, we have a tradeoff between the rate of the compressor and the distortion we incur, and the fundamental limit is given by the rate distortion function, which is discussed later in this section.

Shannon [1948] was the first to distinguish the probabilistic model that underlies an information source from the semantics of the information. An information source produces one of many possible messages; the goal of communication is to transmit an unambiguous specification of the message so that the receiver can

reconstruct the original message. For example, the information to be sent may be the result of a horse race. If the recipient is assumed to know the names and numbers of the horses, then the only data that must be transmitted is the number of the horse that won. In a different context, the same number might mean something quite different, e.g., the price of a barrel of oil. The significant fact is that the difficulty in communication depends only on the length of the representation. Thus, finding the best (shortest) representation of an information source is critical to efficient communication.

When the possible messages are all equally likely, then it makes sense to represent them by strings of equal length. For example, if there are 32 possible equally likely messages, then each message can be represented by a binary string of 5 bits. However, if the messages are not equally likely, then it is more efficient on the average to allot short strings to the frequently occurring messages and longer strings to the rare messages. For example, the Morse code allots the shortest string (a dot) to the most frequent letter (E), and allots long strings to the infrequent letters (e.g., dash, dash, dot, dash for **Q**). The minimum average length of the representation is a fundamental quantity called the entropy of the source, which is defined in the next section.

Entropy

An information source will be represented by a random variable, X, which takes on one of a finite number of possibilities, $i \in \chi$, with probability, $p_i = \Pr(X = i)$. The *entropy* of the random variable X is defined as

$$H(X) = -\sum_{i \in \chi} p_i \log p_i \qquad (6.81)$$

where the log is to base 2 and the entropy is measured in *bits*. We will use logarithms to base 2 throughout this chapter.

Example 6.9.   Let X be a random variable that takes on a value, 1, with probability, θ, and takes on the value 0 with probability, $1 - \theta$. Then, $H(X) = -\theta \log \theta - (1 - \theta)\log(1 - \theta)$. In particular, the entropy of a fair coin toss is 1 bit.

This definition of entropy is related to the definition of entropy in thermodynamics. It is the fundamental lower bound on the average length of a code for the random variable.

A *code* for a random variable, X, is a mapping from χ, the range of X, to the set of finite length binary strings. We will denote the codeword corresponding to i by $C(i)$, and the length of the codeword by l_i. The average length of the code is then, $L(C) = \sum_i p_i l_i$.

A code is said to be *instantaneous* or *prefix-free* if no codeword is a prefix of any other codeword. This condition is sufficient (but not necessary) to allow a sequence of received bits to be parsed unambiguously into a sequence of codewords.

Example 6.10.   Consider a random variable, X, taking on the values {1,2,3} with probabilities (0.5,0.25,0.25). An instantaneous code for this random variable might be (0,10,11). Thus, a string 01001110 can be uniquely parsed into 0,10,0,11,10, which decodes to the string $x = (1,2,1,3,2)$. Note that the average length of the code is 1.5 bits, which is the same as the entropy of the source.

The following property, called the *Kraft inequality,* holds for any instantaneous code

$$\sum_i 2^{-l_i} \leqslant 1 \qquad (6.82)$$

where $l_i, i = 1, 2, \ldots$ are the lengths of the codewords. Conversely, it can be shown that, given a set of lengths that satisfies the Kraft inequality, we can find a set of prefix free codewords with those lengths.

The problem of finding the best source code then reduces to finding the optimal set of lengths that satisfies the Kraft inequality and minimizes the average length of the code. Simple calculus can then be used to show [Cover and Thomas, 1991] that the average length of any instantaneous code is larger than the entropy of the

random variable, i.e., the minimum of $\sum p_i l_i$ over all l_i satisfying $\sum 2^{-l_i} \leqslant 1$ is $-\sum p_i \log p_i$. Also, if we take into account the integer constraint and let $l_i = \left\lceil \log \frac{1}{p_i} \right\rceil$ (where $\lceil t \rceil$ denotes the smallest integer greater than t), we can verify that this choice of lengths satisfies the Kraft inequality and that

$$L(C) = \sum_i p_i \left\lceil \log \frac{1}{p_i} \right\rceil \geqslant \sum_i p_i \left(\log \frac{1}{p_i} + 1 \right) = H(X) + 1 \tag{6.83}$$

Therefore, we have the following theorem:

Theorem 6.1. Let $L^\star$ be the average length of the optimal instantaneous code for a random variable, X. Then

$$H(X) \leqslant L^\star(X) + 1. \tag{6.84}$$

This theorem is one of the fundamental theorems of information theory. It identifies the entropy as the fundamental limit for the description of a discrete information source, and shows that we can find representations with average length within 1 bit of the entropy.

The Huffman Algorithm

The choice of codeword lengths, $l_i = \left\lceil \log \frac{1}{p_i} \right\rceil$ (called the *Shannon codelengths*), is close to optimal, but not necessarily optimal, in terms of average codeword length. We will now describe an algorithm (the *Huffman algorithm*) that produces an instantaneous code of minimal average length for a random variable with the distribution $p_1, p_2, \ldots, p_m$. The algorithm is a greedy algorithm for building the tree from the bottom up.

Step 1. Arrange the probabilities in decreasing order so that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_m$.
Step 2. Form a subtree by combining the last two probabilities, p_{m-1} and p_m, to a single node of weight, $p'_{m-1} = p_{m-1} + p_m$.
Step 3. Recursively execute Steps 1 and 2, decreasing the number of nodes each time, until a single node is obtained.
Step 4. Use the tree constructed above to allot codewords.

The algorithm for tree construction is illustrated for a source with distribution $(0.5, 0.2, 0.2, 0.1)$ in Figure 6.21. After constructing the tree, the leaves of the tree (which correspond to the symbols of $\mathcal{X}$) can be assigned codewords that correspond to the paths from the root to the leaf. We will not give a proof of the optimality of the Huffman algorithm; the reader is referred to [Gallager, 1968] or [Cover and Thomas, 1991] for details.
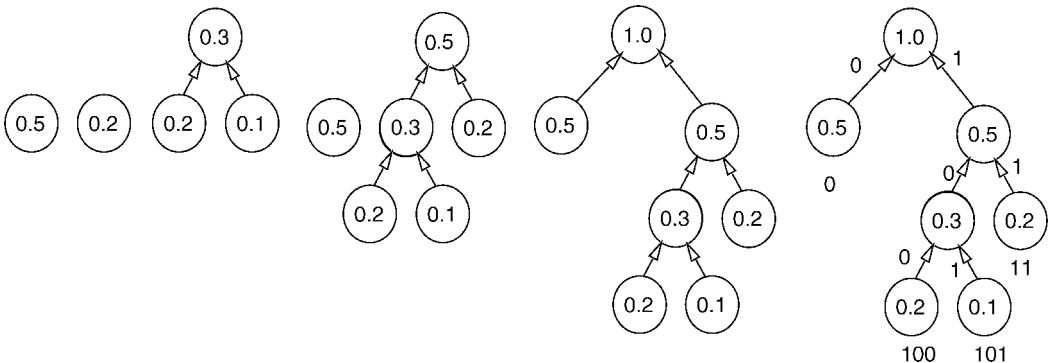
FIGURE 6.21 Example of the Huffman Algorithm.

Entropy Rate

The entropy of a sequence of random variables, $X_1, X_2, \ldots, X_n$, with joint distribution, $p(x_1, x_2, \ldots, x_n)$, is defined analogously to the entropy of a single random variable as

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1}\sum_{x_2} \cdots \sum_{x_n} p(x_1, x_2, \ldots, x_n) \log p(x_1, x_2, \ldots, x_n) \qquad (6.85)$$

For a stationary process, $X_1, X_2, \ldots$, we define the *entropy rate, $\mathcal{H}(\chi)$*, of the process as

$$\mathcal{H}(\chi) = \lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \qquad (6.86)$$

It can be shown [Cover and Thomas, 1991] that the limit exists for all stationary processes. In particular, if $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed random variables, then $H(X_1, X_2, \ldots, X_n) = nH(X_1)$, and $\mathcal{H}(\chi) = H(X_1)$.

In the previous section, we showed the existence of a prefix-free code having an average length within 1 bit of the entropy. Now instead of trying to represent one occurrence of the random variable, we can form a code to represent a block of n random variables. In this case, the average code length is within 1 bit of $H(X_1, X_2, \ldots, X_n)$. Thus the average length of the code per input symbol satisfies

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} < \frac{L_n^*}{n} < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n} \qquad (6.87)$$

Since $\dfrac{H(X_1, X_2, \ldots, X_n)}{n} \to \mathcal{H}(\chi)$, we can get arbitrarily close to the entropy rate by using longer and longer block lengths. Thus, the entropy rate is the fundamental limit for data compression for stationary sources, and we can achieve rates arbitrarily close to this limit by using long blocks.

All the above assumes that we know the probability distribution that underlies the information source. In many practical examples, however, the distribution is unknown or too complex to be used for coding. There are various ways to handle this situation:

- Assume a simple distribution and design an appropriate code for it. Use this code on the real source. If an estimated distribution, $\hat{p}$, is used when in fact the true distribution is p, then the average length of the code increases to $H(X) + \sum_x p(x) \log \dfrac{p(x)}{\hat{p}(x)}$. The second term, which is denoted $D(p\|\hat{p})$, is called the *relative entropy* or the *Kullback–Leibler distance* between the two distributions.
- Estimate the distribution empirically from the source, and adapt the code to the distribution. For example, with *adaptive Huffman coding*, the empirical distribution of the source symbols is used to design the Huffman code used for the source.
- Use a *universal coding algorithm* like the Lempel-Ziv algorithm.

Arithmetic Coding

In the previous sections, it was shown how to construct a code for a source that achieves an average length within 1 bit of the entropy. For small source alphabets, however, we have efficient coding only if we use long blocks of source symbols. For example, if the source is binary, and we code each symbol separately, we must use 1 bit per symbol irrespective of the entropy of the source. If we use long blocks, we can achieve an expected length per symbol close to the entropy rate of the source.

Therefore, it is desirable to have an efficient coding procedure that works for long blocks of source symbols. Huffman coding is not ideal for this situation, since it is a bottom-up procedure with a complexity that grows
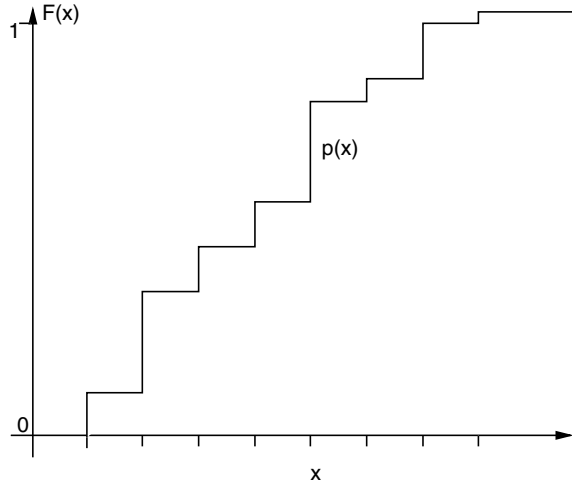
FIGURE 6.22 Cumulative distribution function for sequences *x*.

rapidly with the block length. Arithmetic coding is an incremental coding algorithm that works efficiently for long block lengths and achieves an average length within 1 bit of the entropy of the block.

The essential idea of arithmetic coding is to represent a sequence, $x^n = x_1 x_2 \ldots x_n$, by the cumulative distribution function, $F(x^n) = \sum_{\tilde{x}^n \leqslant x^n} p(\tilde{x}^n)$, expressed to an appropriate accuracy. The cumulative distribution function for x^n is illustrated in Figure 6.22. We can use any real number in the interval, $(F(x^n) - p(x^n), F(x^n)]$, as the code for x^n. Expressing $F(x^n)$ to an accuracy of $\left\lceil \log \frac{1}{p(x^n)} \right\rceil$ will give us a code for the source. The receiver can draw the cumulative distribution function, draw a horizontal line corresponding to the truncated value, $\lfloor F(x^n) \rfloor$, that was sent, and read off the corresponding x^n. (This code is not prefix-free but can be easily modified to construct a prefix-free code [Cover and Thomas, 1991].) To implement arithmetic coding, however, we need efficient algorithms to calculate $p(x^n)$ and $F(x^n)$ to the appropriate accuracy based on a probabilistic model for the source. Details can be found in [Langdon, 1984].

Lempel–Ziv Coding

The Lempel–Ziv algorithm [Ziv and Lempel, 1978] is a universal coding procedure that does not require knowledge of the source statistics but is asymptotically optimal. The basic idea of the algorithm is to construct a table or dictionary of frequently occurring strings and to represent new strings by pointing to their prefixes in the table. We first parse the string into sequences that have not appeared so far. For example, the binary string 11010011011100 is parsed into 1,10,100,11,0,111,00, etc. Then instead of sending the bits of each phrase, we send a pointer to its prefix and the value of the last bit. Thus, if we use three bits for the pointer, we will represent this string by (000,1), (001,0), (010,0), (001,1), (000,0), (100,1), (101,0), etc. For this short example, the algorithm has not compressed the string — it has in fact expanded it.

But the very surprising fact is that, as Lempel and Ziv have shown, the algorithm is asymptotically optimal for any stationary ergodic source. This fact is expressed in the following theorem [Ziv and Lempel, 1978; Cover and Thomas, 1991]:

Theorem 6.2 Let L_n be the length of the Lempel–Ziv code for n symbols drawn from a stationary ergodic process, $X_1, X_2, \ldots, X_n$, with entropy rate $\mathcal{H}(\chi)$. Then,

$$\frac{L_n}{n} \to \mathcal{H}(\chi) \quad \text{with probability 1} \tag{6.88}$$

Thus, for long enough block lengths, the Lempel-Ziv algorithm (which does not make any assumptions about the distribution of the source) does as well as if we knew the distribution in advance and designed the optimal code for this distribution.

The algorithm described above is only one of a large class of similar adaptive dictionary-based algorithms, which are all rather loosely called Lempel–Ziv. These algorithms are simple and fast, and have been implemented in both software and hardware, e.g., in the *compress* command in UNIX and the PKZIP command on PCs. On ASCII text files, the Lempel–Ziv algorithms achieve compressions of the order of 50%. It has also been implemented in hardware and has been used to "double" the capacity of data storage media or main memory, or to "double" the effective transmission rate of a modem. Many variations on the basic algorithm can be found in [Bell et al., 1990].

Rate Distortion Theory

An infinite number of bits are required to describe an arbitrary real number, and therefore it is not possible to perfectly represent a continuous random variable with a finite number of bits. How "good" can the representation be? First, define a distortion measure, which is a measure of the distance between the random variable and its representation. Then, consider the tradeoff between the number of bits used to represent a random variable and the distortion incurred. This tradeoff is represented by the rate distortion function, $R(D)$, which represents the minimum rate required to represent a random variable with a distortion, D.

We will consider a discrete information source that produces random variables, $X_1, X_2, \ldots, X_n$, that are drawn independently and identically distributed (i.i.d.) according to $p(x)$. (The results are also valid for continuous sources.) The encoder of the rate distortion system of rate, R, will encode a block X^n of n outputs as an index, $f(X^n) \in \{1, 2, \ldots, 2^{nR}\}$. (Thus the index will require R bits/input symbol.) The decoder will calculate a representation, $\hat{X}^n(f(X^n))$ of X^n. Normally, the representation alphabet, $\hat{\chi}$, of the representation is the same as the alphabet source, χ, but that need not be the case.

Definition 6.1. *A distortion function or distortion measure is a mapping*

$$d : \chi \times \hat{\chi} \rightarrow R^+ \tag{6.89}$$

from the set of source alphabet-reproduction alphabet pairs into the set of non-negative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by the symbol $\hat{x}$.

Examples of common distortion functions are

- *Hamming (probability of error) distortion. The Hamming distortion is given by*

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases} \tag{6.90}$$

  and thus, $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$; the distortion is the probability of error.
- *Squared error distortion.* The squared error distortion,

$$d(x, \hat{x}) = (x - \hat{x})^2 \tag{6.91}$$

  is the most popular distortion measure used for continuous alphabets. Its advantages are its simplicity and its relationship to least squares prediction. However, for information sources such as images and speech, the squared error might not be an appropriate measure for distortion as perceived by a human observer.

The distortion between sequences, x^n and $\hat{x}^n$, of length n is defined by

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{6.92}$$

For a rate distortion system, the expected distortion D is defined as

$$D = Ed(X^n, \hat{X}^n(f(X^n))) = \sum_{x^n} p(x^n) d(x^n, \hat{X}^n(f(x^n))) \tag{6.93}$$

Definition 6.2.   *The rate distortion pair (R,D) is said to be achievable if there exists a rate distortion code of rate, R, that has expected distortion, D. The rate distortion function, R(D), is the infimum of rates, R, such that (R,D) is achievable for a given D.*

Definition 6.3.   *The mutual information, $I(X; \hat{X})$, between random variables, X and $\hat{X}$, with joint probability distribution function, $p(x, \hat{x})$, and marginal probability distribution functions, p(x) and p($\hat{x}$), is defined as*

$$I(X; \hat{X}) = \sum_{x \in \chi} \sum_{\hat{x} \in \hat{\chi}} p(x, \hat{x}) \log \frac{p(x, \hat{x})}{p(x) p(\hat{x})} \tag{6.94}$$

The mutual information is a measure of the amount of information that one random variable carries about another.

The main result of rate distortion theory is contained in the following theorem, which provides a characterization of the rate distortion function in terms of the mutual information of joint distributions that satisfy the expected distortion constraint

Theorem 6.3   The rate distortion function for an i.i.d. source, X, with distribution, *p(x)*, and distortion function, d(x, $\hat{x}$), is

$$R(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leqslant D} I(X; \hat{X}) \tag{6.95}$$

We can construct rate distortion codes that can achieve distortion, D, at any rate greater than $R(D)$, and we cannot construct such codes at any rate below $R(D)$.

The proof of this theorem uses ideas of random coding and long block lengths, as in the proof of the channel capacity theorem. The basic idea is to generate a codebook of 2^{nR} reproduction codewords, $\hat{X}^n$, at random and show that, for long block lengths, for any source sequence, it is very likely that there is at least one codeword in this codebook that is within distortion, D, of that source sequence. See [Gallager, 1968] or [Cover and Thomas, 1991] for details of the proof.

Example 6.11.   (*Binary source.*) The rate distortion function for a Bernoulli(p) source (a random variable that takes on values {0,1} with probabilities $p, 1-p$) with Hamming distortion is given by

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leqslant D \leqslant \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\} \end{cases} \tag{6.96}$$

where $H(p) = -p \log p - (1-p) \log(1-p)$ is the binary entropy function.

Example 6.12. (*Gaussian source.*) The rate distortion function for a Gaussian random variable with variance σ^2 and squared error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2 \end{cases} \tag{6.97}$$

Thus with nR bits, we can describe n i.i.d. Gaussian random variables, $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ with a distortion of $\sigma^2 2^{-2R}$ per symbol.

Quantization and Vector Quantization

The rate distortion function represents the lower bound on the rate that is needed to represent a source with a particular distortion. We now consider simple algorithms that represent a continuous random variable with a few bits. Suppose we want to represent a single sample from a continuous source. Let the random variable to be represented be X and let the representation of X be denoted as $\hat{X}(X)$. If we are given R bits to represent X, then the function $\hat{X}$ can take on 2^R values. The problem of optimum quantization is to find the optimum set of values for $\hat{X}$ (called the reproduction points or codepoints) and the regions that are associated with each value, $\hat{X}$, so to minimize the expected distortion.

For example, let X be a Gaussian random variable with mean, 0, and variance, σ^2, and assume a squared error distortion measure. In this case, we wish to find the function $\hat{X}(X)$ such that $\hat{X}$ takes on at most 2^R values and minimizes $E(X - \hat{X}(X))^2$. If we are given 1 bit to represent X, it is clear that the bit should distinguish whether $X > 0$ or not. To minimize squared error, each reproduced symbol should be at the conditional mean of its region. If we are given 2 bits to represent the sample, the situation is not as simple. Clearly, we want to divide the real line into four regions and use a point within each region to represent the sample. We can state two simple properties of optimal regions and reconstruction points for the quantization of a single random variable

- Given a set of reconstruction points, the distortion is minimized by mapping a source random variable, X, to the representation, $\hat{X}(\omega)$, that is closest to it (in distortion). The set of regions of χ defined by this mapping is called a Voronoi or Dirichlet partition, defined by the reconstruction points.
- Given a set of reconstruction regions, the reconstruction points should be chosen to minimize the conditional expected distortion over their respective assignment regions.

These two properties enable construction of a simple algorithm to find a "good" quantizer: start with a set of reconstruction points, find the optimal set of reconstruction regions (which are the nearest neighbor regions with respect to the distortion measure), then find the optimal reconstruction points for these regions (the centroids of these regions if the distortion is squared error), and then repeat the iteration for this new set of reconstruction points. The expected distortion is decreased at each stage in the algorithm, so the algorithm will converge to a local minimum of the distortion. This algorithm is called the *Lloyd algorithm*.

It follows from the arguments of rate distortion theory that we will do better if we encode long blocks of source symbols rather than encoding each symbol individually. In this case, we will consider a block of n symbols from the source as a vector-valued random variable, and we will represent these n-dimensional vectors by a set of 2^{nR} codewords. This process is called *vector quantization (VQ)*. We can apply the Lloyd algorithm to design a set of representation vectors (the codebook) and the corresponding nearest neighbor regions. Instead of using the probability distribution for the source to calculate the centroids of the regions, we can use the empirical distribution from a training sequence. Many variations of the basic vector quantization algorithm are described in [Gersho and Gray, 1992].

Common information sources like speech produce continuous waveforms, not discrete sequences of random variables as in the models we have been considering so far. But by sampling the signal at twice the maximum frequency present (the Nyquist rate), we convert the continuous time signal into a set of discrete samples from which the original signal can be recovered (the sampling theorem). We can then apply the theory of rate distortion and vector quantization to such waveform sources as well.

Kolmogorov Complexity

In the 1960s, the Russian mathematician Kolmogorov considered the question, "What is the intrinsic descriptive complexity of a binary string?" From the discussion above, it follows that if the binary string is a sequence of independent and identically distributed random variables, $X_1, X_2, \ldots, X_n$, then on average it would take $nH(x)$ bits to represent the sequence. But what if the bits were the first million bits of the binary expansion of π? In that case, the string appears random, but can be generated by a simple computer program. So if we wanted to send these million bits to another location that has a computer, we could instead send the program and ask the computer to generate these million bits. Thus, the descriptive complexity of π is quite small.

Motivated by such considerations, Kolmogorov defined the complexity of a binary string to be the length of the shortest program for a universal computer that generated that string. (This concept was also proposed independently and at about the same time by Chaitin and Solomonoff.)

Definition 6.3. *The Kolmogorov complexity, $K_u(x)$, of a string, x, with respect to a universal computer, $\mathcal{U}$, is defined as*

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p) \tag{6.98}$$

the minimum length over all programs that print x and halt. Thus, $K_{\mathcal{U}}(x)$ is the shortest description length of x over all descriptions interpreted by computer $\mathcal{U}$.

A universal computer can be thought of as a Turing machine that can simulate any other universal computer. At first sight, the definition of Kolmogorov complexity seems to be useless, since it depends on the particular computer that we are talking about. But using the fact that any universal computer can simulate any other universal computer, any program for one computer can be converted to a program for another computer by adding a constant length "simulation program" as a prefix. Thus we can show that for any two universal computers, $\mathcal{U}$ and $\mathcal{A}$,

$$\left| K_{\mathcal{U}}(x) - K_{\mathcal{A}}(x) \right| \tag{6.99}$$

where the constant, c, though large, does not depend on the string, x, under consideration. Thus, Kolmogorov complexity is universal in that it does not depend on the computer (up to a constant additive factor).

Kolmogorov complexity provides a unified way to think about problems of data compression. It is also the basis of principles of inference (Occam's Razor: "The simplest explanation is the best") and is closely tied with the theory of computability.

Data Compression in Practice

The previous sections discussed the fundamental limits to compression for a stochastic source. We will now consider the application of these algorithms to some practical sources, namely, text, speech, images, and video. In real applications, the sources may be neither stationary nor ergodic, and the distributions underlying the source are often unknown. Also, in addition to the efficiency of the algorithm, important considerations in practical applications include the computational speed and memory requirements of the algorithm, the perceptual quality of the reproductions to a human observer, etc. There is a considerable amount of research and engineering that has gone into the development of these algorithms, and many issues are only now being explored. We will not go into the details, but simply list some popular algorithms for the different sources.

Text

English text is normally represented in ASCII, which uses 8 bits/character. There is considerable redundancy in this representation (the entropy rate of English is about 1.3 bits/character). Popular compression algorithms include variants of the Lempel–Ziv algorithm, which compress text files by about 50% (to about 4 bits/character).

Speech

Telephone-quality speech is normally sampled at 8 KHz and quantized at 8 bits/sample (a rate of 64 Kbits/sec) for uncompressed speech. Simple compression algorithms like ADPCM (Adaptive Differential Pulse Code Modulation) [Jayant and Noll, 1984] use the correlation between adjacent samples to reduce the number of bits used by a factor of 2 to 4 or more with almost imperceptible distortion. Much higher compression ratios can be obtained with algorithms like LPC (Linear Predictive Coding), which model speech as an autoregressive process, and send the parameters of the process as opposed to sending the speech itself. With LPC based methods, it is possible to code speech at less than 4 Kbits/sec. However, at very low bit rates, the reproduced speech sounds "synthetic."

Music

Music on CDs is stored as an uncompressed stream at a sampling frequency of 44,100 samples/sec, 16 bits/ sample in 2 channels with an effective bit rate of about 1.4 Mb/sec. There is much redundancy in the signal, and strong correlation between the two channels. Therefore, it is possible to compress the signal with very little loss of quality. The most popular standard for compressing audio signals is derived from a standard from the MPEG group called MP3, which achieves compressions ratios of about 10:1 with near CD quality, enabling users to easily download songs over the Internet.

Images

A single high quality color image of 1024 by 1024 pixels with 24 bits per pixel represents about 3 MB of storage in an uncompressed form. It is therefore very important to use compression to save storage and communication capacity for images. There are many different algorithms that have been proposed for image compression, and standards are still being developed for compression of images. For example, the popular GIF standard uses Lempel–Ziv coding, and the JPEG standard developed by the Joint Photographic Experts Group uses an 8 by 8 discrete cosine transform(DCT) followed by quantization (the quality of which can be chosen by the user) and Huffman coding. The compression ratios achieved by these algorithms are very dependent on the image being coded. The lossless compression methods achieve compression ratios of up to about 3:1, whereas lossy compression methods achieve ratios up to 50:1, with very little perceptible loss of quality. The JPEG standard is now widely used for images on the Web and within digital cameras to compress images before they are stored in the camera's memory.

Video

Video compression methods exploit the correlation in both space and time of the sequence of images to improve compression. There is a very high correlation between successive frames of a video signal. This correlation can be exploited along with methods similar to those used for coding images to achieve compression ratios up to 200:1 for high quality lossy compression. Standards for full-motion video and audio compression are being developed by the Moving Pictures Experts Group (MPEG), including MPEG-1 (used for Video CDs), MPEG-2 (used in digital TVs and DVDs), MPEG-4 (used for multimedia on the Web), and MPEG-7 (a standard that allows for description of the content to enable search). Applications of video compression techniques include videoconferencing, multimedia CD-ROMs, and HDTV.

   A fascinating and very readable introduction to different sources of information, their entropy rates, and different compression algorithms can be found in the book by Lucky [1989]. Implementations of popular data compression algorithms, including adaptive Huffman coding, arithmetic coding, Lempel–Ziv, and the JPEG algorithm can be found in Nelson and Gailly [1995] and Sayood [2000].

Defining Terms

Code: A mapping from a set of messages into binary strings.
Entropy: A measure of the average uncertainty of a random variable. For a random variable with probability distribution, $p(x)$, the entropy, $H(X)$, is defined as $\sum_x -p(x)\log p(x)$.

Huffman coding: A procedure that constructs the code of minimum average length for a random variable.

Kolmogorov complexity: The minimum length description of a binary string that would enable a universal computer to reconstruct the string.

Lempel–Ziv coding: A dictionary-based procedure for coding that does not use the probability distribution of the source and is nonetheless asymptotically optimal.

Quantization: A process by which the output of a continuous source is represented by one of a set of discrete points.

Rate distortion function: The minimum rate at which a source can be described to within a given average distortion.

Vector quantization: Quantization applied to vectors or blocks of outputs of a continuous source.

References

T. Bell, J. Cleary, and I. Witten, *Text Compression*, Englewood Cliffs, N.J.: Prentice Hall, 1990.

T.M. Cover and J.A. Thomas, *Elements of Information Theory*, New York: John Wiley, 1991.

R. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968.

A. Gersho and R. Gray, *Vector Quantization and Source Coding*, Boston: Kluwer Academic Publishers, 1992.

B. Haskell, P. Howard, Y. Le Cun, A. Puri, J. Ostermann, M. Civanlar, L. Rabiner, L. Bottou, and P. Haffner, "Image and video coding: emerging standards and beyond," *CirSysVideo*, vol. 8, no. 7, p. 814, 1998.

S. Verdu, Ed., "IEEE IT October, special issue: the first fifty years," *IEEE Transactions on Information Theory,* vol. IT-44, 1998.

N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Englewood Cliffs, N.J.: Prentice Hall, 1984.

G. Langdon, "An introduction to arithmetic coding," *IBM J Res. Dev.*, vol. 28, pp. 135–149, 1984.

R. Lucky, *Silicon Dreams: Information, Man and Machine,* New York: St. Martin's Press, 1989.

M. Nelson and J. Gailly, *The Data Compression Book*, 2nd ed., New York: M & T Books, 1995.

K. Sayood, *Introduction to Data Compression*, 2nd ed., San Francisco, CA: Morgan Kauffman, 2000.

C.E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, no. 379–423, pp. 623–656, 1948.

J. Ziv and A. Lempel, "Compression of individual sequences by variable rate coding," *IEEE Trans. Info. Theory*, IT-24, 530–536, 1978.

Further Information

Discussion of various data compression algorithms for sources like speech and images can be found in the *IEEE Transactions on Communications* and the *IEEE Transactions on Acoustics, Speech and Signal Processing*, while the theoretical underpinnings of compression algorithms are discussed in the *IEEE Transactions on Information Theory*, including a special issue [IEEE IT October, 1998] that surveys the development of data compression algorithms over the last 50 years.

Some of the latest developments in the areas of image and video compression are described in the *IEEE Transactions on Circuits and Systems for Video Technology*. It includes a survey [Haskell et al., 1998] of current work on image compression, including various image and video compression standards.

This page intentionally left blank

7
Satellites and Aerospace

Daniel F. DiFonzo
*Planar Communications
Corporation*

7.1   Introduction

The impact of satellites on world communications since commercial operations began in the mid-1960s is such that we now take for granted many services such as: worldwide TV; reliable communications to remote areas; mobile communications to ships, aircraft, and moving vehicles; wide area data networks; direct satellite TV and audio broadcast to homes and moving vehicles; position determination; Earth observation (weather and mapping); and personal communications such as hand-held mobile telephones.

Communications satellites function as line-of-sight microwave relays in orbits high above the Earth, where they can see large areas of the Earth's surface. Because of this unique feature, satellites are particularly well suited to communications over wide coverage areas for such applications as broadcasting, mobile communications, and point-to-multipoint communications. Satellite systems can also provide cost-effective access for many locations where the high investment cost of terrestrial facilities might not be warranted.

7.2   Satellite Applications

Figure 7.1 depicts several kinds of satellite links and orbits. The geostationary Earth orbit (GEO) is in the equatorial plane at an altitude of 35,786 km with a period of one sidereal day (23 h 56m 4.09s). This orbit is sometimes called the Clarke orbit in honor of Arthur C. Clarke, who first described its usefulness for communications in 1945 [Clarke, 1945]. GEO satellites appear to be almost stationary from the ground (subject to small perturbations), and the Earth antennas pointing to these satellites may need only limited
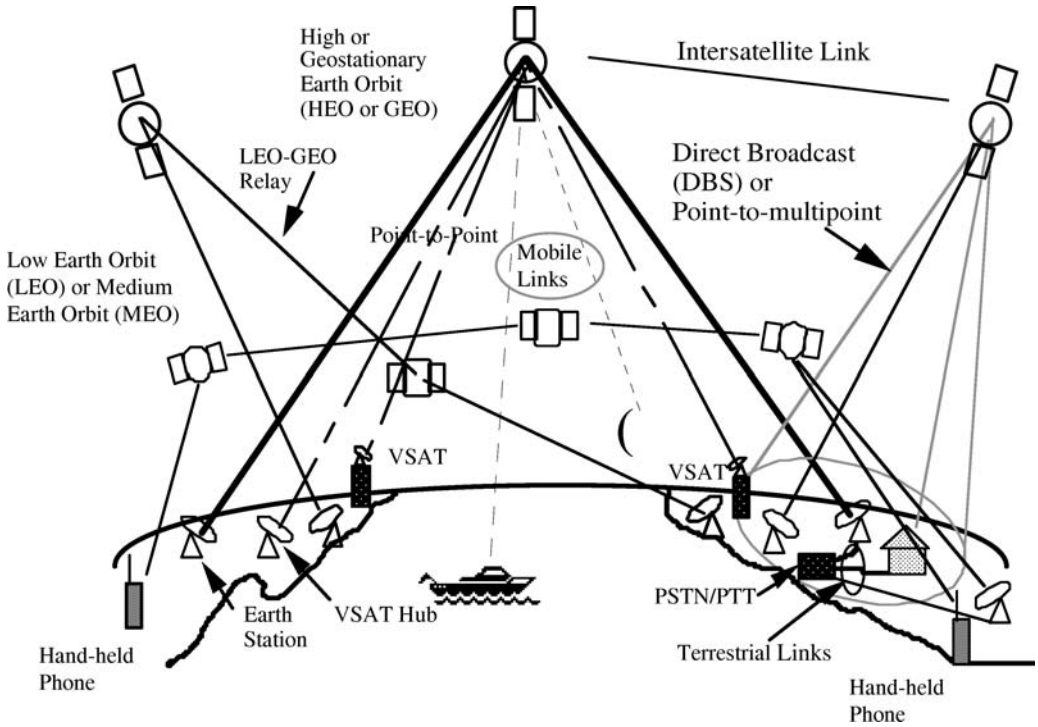
FIGURE 7.1   Several types of satellite links. Illustrated are point-to-point, point-to-multipoint, VSAT, direct broadcast, mobile, personal communications, and intersatellite links.

or no tracking capability. Satellites having orbit periods that are synchronous with the Earth's rotation such as 12 h or 6 h, but are not generally in the equatorial plane, are called geosynchronous orbits (GSO).

An orbit for which the highest altitude (apogee) is greater than GEO and is sometimes referred to as a High Earth orbit (HEO). An orbit with altitude between a few hundred km and about 2000 km is considered to be in Low Earth orbit (LEO). A Medium Earth orbit (MEO) occupies an intermediate altitude.

LEO systems include those of Iridium Satellite LLC (www.iridium.com) with 66 operational satellites at an altitude of 780 km and Globalstar™ with 40 satellites at an altitude of 1414 km for voice and data communications by satellite to hand held telephones and pagers (www.globalstar.com). Iridium uses intersatellite links to relay information beyond the coverage of a single satellite, whereas Globalstar uses an extensive network of Earth terminals for signal relay in multiple "hops" and for interconnection to terrestrial phone networks. MEO systems include the U.S. Department of Defense Global Positioning System (GPS) with 24 operational satellites at an altitude of 20,200 km and the proposed Galileo positioning system with 27 operational satellites at 23,616 km.

GEO satellites are the most prevalent. Initially, communications satellites were used primarily for point-to-point traffic in the GEO Fixed Satellite Service (FSS), e.g., for telephony across the oceans and for point-to-multipoint TV distribution to cable *head end* stations. Large Earth station antennas with high-gain narrow beams and high uplink powers were needed to compensate for limited satellite power. This type of system, exemplified by the early global network of Intelsat, which was formerly an international treaty organization and now is a private company (www.intelsat.com). The early Intelsat system used large Earth terminal antennas with 30 m diameters. Inmarsat, another former treaty organization that is now a private company, was formed to offer mobile communications services to ships and has evolved to offer medium speed data services to small portable terminals (www.inmarsat.com). Many other satellite organizations

have been formed around the world to provide international, regional, and domestic services [Rees, 1990; Roddy, 2001; Maral and Bousquet, 2002].

As satellites have grown in power and sophistication, the average size of Earth terminals has been reduced. High gain satellite antennas and relatively high power satellite transmitters have led to *very small aperture* Earth *terminals* (VSAT) with diameters of less than 2 m, modest powers of less than 10 W [Maral, 2003] and to even smaller diameters typically less than one meter. As depicted in Figure 7.1, VSAT terminals may be placed atop urban office buildings, permitting private networks of hundreds or thousands of terminals, which bypass terrestrial lines. A new development of in-motion VSATs will allow two-way high-speed data and Internet communication to and from moving vehicles (www.raysat.com). VSATs are typically incorporated into *star* or *hub-and-spoke* networks where the small terminals communicate through the satellite with a larger *hub* terminal. The hub retransmits through the satellite to another small terminal, or it can connect to a *gateway* that routes the signal to terrestrial facilities. Star connections between VSATs require two *hops* with attendant time delays. With high-gain satellite antennas and relatively narrow-band digital signals, direct single-hop *mesh* interconnections of VSATs may be used. Descriptions of Earth terminals and their properties may be found in Ebert [2000].

7.3 Satellite Functions

The traditional satellite function is that of a bent pipe quasilinear repeater in space. As shown in Figure 7.2, *uplink* signals from Earth terminals directed at the satellite are received by the satellite's antennas, amplified, translated to a different *downlink* frequency band, channelized into *transponder channels*, further amplified to relatively high power and retransmitted toward the Earth. Transponder channels are generally rather broad (e.g., bandwidths from 24 MHz to more than 100 MHz) and each may contain many individual or user channels.

The functional diagram in Figure 7.2 is appropriate to a satellite using frequency-division duplex (FDD), which refers to the fact that the satellites use separate frequency bands for the uplink and
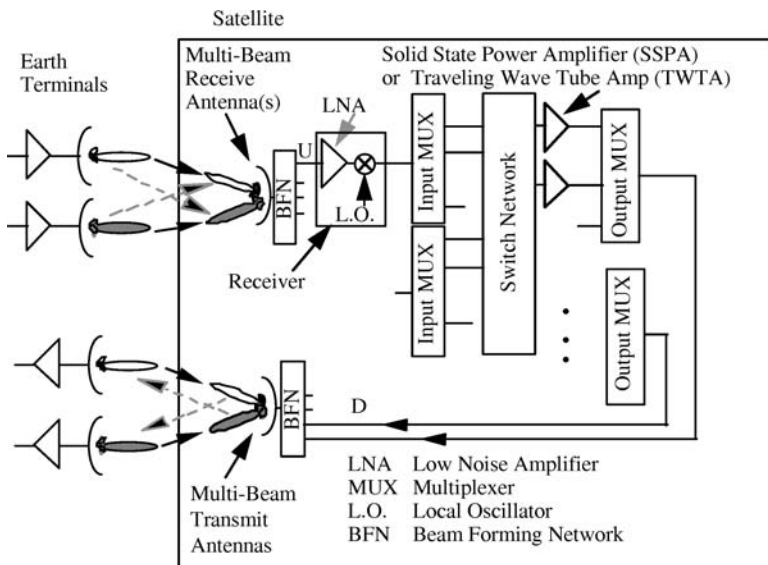
FIGURE 7.2 A satellite repeater receives uplink signals (U), translates them to a downlink frequency band (D), channelizes, amplifies to high power, and retransmits to Earth. Multiple beams allow reuse of the available band. Interference (dashed lines) can limit performance. Downconversion may also occur after the input multiplexers. Several intermediate frequencies and downconversions may be used.

downlink and where both links operate simultaneously. This diagram also illustrates a particular *multiple access* technique, known as frequency-division multiple access (FDMA), which has been prevalent in mature satellite systems.

Multiple access allows many different user signals to utilize the satellite's resources of power and bandwidth without interfering with each other. Multiple access techniques to segregate users include: frequency division (FDMA), where each user is assigned a specific frequency channel; space-division multiple access (SDMA) that reuses the same frequencies on multiple spatially isolated beams; time-division multiple access (TDMA), where each user signal occupies an entire allocated frequency band but for only part of the time; polarization-division (PD), where frequencies may be reused on spatially overlapping but orthogonally polarized beams; and code-division multiple access (CDMA), where different users occupy the same frequency band, but use spread spectrum signals that contain orthogonal signaling codes [Roddy, 2001; Sklar, 2001; Richharia, 1999].

Frequency modulation (FM) was widely used, but advances in digital voice, music and video compression have led to the widespread use of digital modulation methods such as quadrature phase shift keying (QPSK) and quadrature amplitude modulation (QAM) [Schwartz 1990, Sklar 2001].

Some satellite architectures incorporate on-board demodulation of the uplink digitally modulated signals. The demodulated baseband bits are routed by digital processors, switched to an appropriate downlink antenna beam, and then remodulated prior to downlink transmission. Such *regenerative repeaters* or *onboard processors* permit flexible routing of the user signals and can improve the overall communications link by separating the uplink noise from that of the downlink. The baseband signals may be those of individual users or they may represent frequency-division multiplexed (FDM), or time-division multiplexed (TDM) aggregate signals from many users.

Examples include the experimental NASA ACTS Ka-band (20 and 30 GHz) satellite, the Ka-band Spaceway satellites for data networks and direct TV broadcast (www.directtv.com), and Iridium. Many new Ka-band satellite systems were proposed in the 1990s for high-speed data networks and TV distribution. Their deployment has been slower than expected for data networks but Ka-band satellites are expected to be used for TV broadcast.

The 66 Iridium LEO satellites operate with on-board processing and time-division duplex (TDD), using the same 1.6-GHz L-band frequencies for transmission and reception, but only receiving or transmitting for somewhat less than half the time each. Globalstar operates at 1.6 GHz for the uplink and 2.5 GHz for the downlink. Both systems were developed at huge cost and are operational after having emerged from bankruptcy [Finkelstein, Inkpen et al., 2000].

High-power *direct broadcast satellites* (DBS) provide TV to millions of users worldwide. In the United States, these systems operate in the Broadcast Satellite Service (BSS) Ku-band (12.2–12.7 GHz) and provide hundreds of TV channels directly to more than 25 million subscribers (as of 2004) with parabolic dish antennas as small as 45 cm. DBS service providers include: U.S. companies DirecTV and EchoStar (Dish Network); Canadian provider Bell ExpressVu; and many organizations in Europe and Asia. The U.S. BSS satellites have 32 transponder channels, each with 24-MHz bandwidth, and each transponder carries typically about 12 NTSC TV channels or 1–2 High Definition TV (HDTV) channels using the Motion Picture Experts Group MPEG 2 compression standard (http://www.chiariglione.org/mpeg/). MPEG 4 compression (www.mpeg4.net) is expected to further increase the number of TV channels that each transponder can carry. Digital Audio Radio Service (DARS) is now providing direct nationwide high quality audio broadcast in the United States from high power satellites operating in the S-band (2.3 GHz). XM Radio uses two high power GEO satellites to provide more than 100 channels of highly compressed near-CD quality audio to vehicles in the United States (www.xmradio.com). Sirius Satellite Radio provides a competing DARS S-band service using a constellation of three satellites with inclined elliptical orbits (www.Sirius.com). Both companies supplement their satellite coverages with terrestrial repeaters in urban areas.

Links between LEO satellites (or the NASA Shuttle) and GEO satellites are used for data relay, for example, through the NASA tracking and data relay satellite system (TDRSS). Iridium uses intersatellite links (ISL) to improve the interconnectivity and minimize the need for ground terminals to route signals over a wide-area network. ISL systems may typically operate at frequencies such as 23 GHz, 60 GHz, or even use optical links.

As of 2005, the usage of satellites has evolved considerably, such that it is now dominated by video services such as TV broadcast and cable head end distribution, followed by data network and Internet trunking services. Voice services rank behind these applications. By 2010 it is expected that DBS will represent nearly 60% of U.S. Satellite capacity. (www.futron.com) [Futron Corporation, January 2005]

7.4   Satellite Orbits and Pointing Angles

Reliable communication to and from a satellite requires knowledge of its position and velocity relative to a location on the Earth. Details of the relevant astrodynamic formulas for satellite orbits are given in Griffin and French [1991] and Morgan and Gordon [1989]. Launch vehicles needed to deliver the satellites to their intended orbits are described in Isakowitz [1994].

A satellite, having mass m, in orbit around the Earth, having mass M_e, traverses an elliptical path such that the centrifugal force due to its acceleration is balanced by the Earth's gravitational attraction, leading to the equation of motion for two bodies

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} + \frac{\mu}{r^3}\mathbf{r} = 0 \tag{7.1}$$

where $\mathbf{r}$ is the radius vector joining the satellite and the Earth's center and $\mu = G(m + M_e) \approx GM_e = 398,600.5 km^3/s^2$ is the product of the gravitational constant and the mass of the Earth. Because $m << M_e$, the center of rotation of the two bodies may be taken as the Earth's center, which is at one of the focal points of the orbit ellipse.

Figure 7.3 depicts the orbital elements for a geocentric right-handed coordinate system where the x-axis points to the first point of Aries, which is the fixed position against the stars where the sun's apparent path around the Earth crosses the Earth's equatorial plane while traveling from the southern toward the northern hemisphere at the vernal equinox. The z-axis points to the north and the y-axis is in the equatorial plane and points to the winter solstice. The elements shown are: longitude or right ascension of the ascending node Ω
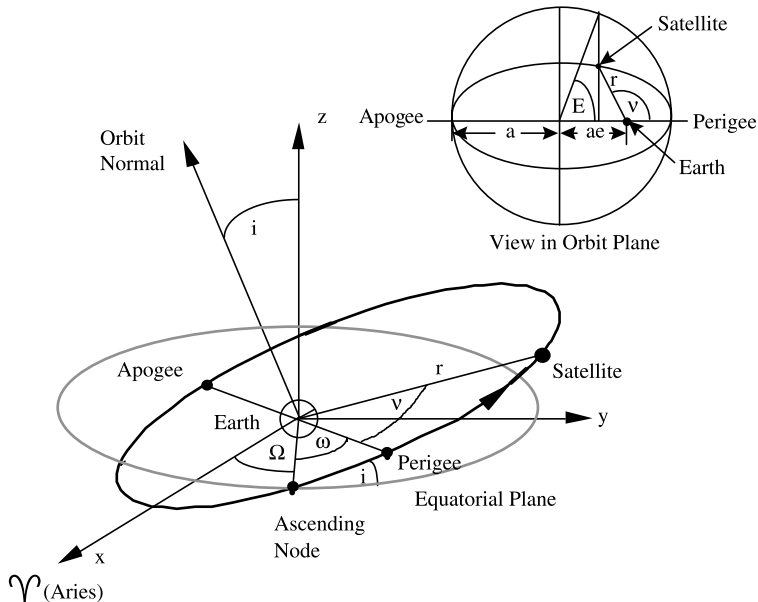
FIGURE 7.3   Orbital elements.

measured in the equatorial plane, the orbit's inclination angle i relative to the equatorial plane; the ellipse semimajor axis length a, the ellipse eccentricity e, the argument (angle) of perigee ω measured in the orbit plane from the ascending node to the satellite's closest approach to the Earth; and the true anomaly (angle) in the orbit plane from the perigee to the satellite v.

The mean anomaly M is the angle from perigee that would be traversed by a satellite moving at its mean angular velocity n. Given an initial value M_0, usually taken as zero for a particular epoch (time) at perigee, the mean anomaly at time t is $M = M_0 + n(t - t_0)$, where $n = \sqrt{\mu/a^3}$. The eccentric anomaly E may then be found from Kepler's transcendental equation $M = E - e \sin E$, which must be solved numerically by, for example, guessing an initial value for E and using a root finding method. For small eccentricities, the series approximation $E \approx M + e \sin M + (e^2/2) \sin 2M + (e^3/8)(3 \sin 3M - \sin M)$ yields good accuracy [Morgan and Gordon, 1989, p. 806]. Other useful quantities include the orbit radius, r; the period, P; of the orbit, [i.e., for $n(t - t_0) = 2\pi$]; the velocity, V; and the radial velocity, V_r

$$r = a(1 - e \cos E) \tag{7.2}$$

$$P = 2\pi\sqrt{a^3/\mu} \tag{7.3}$$

$$V^2 = \mu\left(\frac{2}{r} - \frac{1}{a}\right) \tag{7.4}$$

$$V_r = \frac{e(\mu a)^{1/2} \sin E}{a(1 - e \cos E)} \tag{7.5}$$

Figure 7.4 and Figure 7.5 depict Earth-centered geometrical quantities useful for communications links. In Figure 7.4, the satellite position vector for altitude h above the equator at longitude l_s is $\mathbf{r}_0 = (r_e + h)\hat{x}$, where $r_e = 6378.14$ km is the mean Earth radius and $\hat{x}$ is a unit vector. The position vector of an Earth terminal at north latitude Ψ and east longitude l_e is $\mathbf{r}_e = r_e(\cos\psi \cos\delta\hat{x} + \cos\psi \sin\delta\hat{y} + \sin\psi\hat{z})$, where $\delta = l_e - l_s$.

Figure 7.5 shows the geometry in the plane formed by a satellite, the terminal point on the Earth's surface and the Earth's center, where γ is the Earth central or core angle, θ is the nadir angle relative to the subsatellite axis, and ε is the elevation or angle above the local horizon to the satellite.

Note that $\theta + \varepsilon + \gamma = 90°$. The normalized orbit radius is

$$k = \frac{(r_e + h)}{r_e} = \frac{\cos\varepsilon}{\sin\theta} \tag{7.6}$$
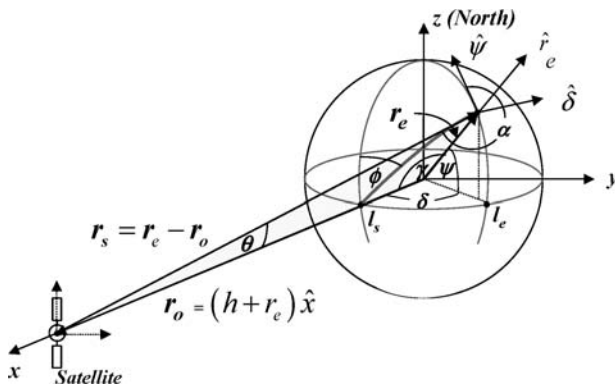
FIGURE 7.4   Geometry for a satellite above the equator.
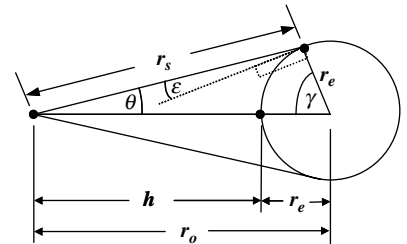
FIGURE 7.5   Geometry in the plane of the satellite, Earth terminal, and Earth center.

Then,

$$\text{Slant range :} \quad r_s = |r_e - r_0| = r_e\sqrt{1 + k^2 - 2k\cos\gamma} \tag{7.7}$$

$$\text{Nadir angle :} \quad \theta = \tan^{-1}\left(\frac{\sin\gamma}{k - \cos\gamma}\right) \tag{7.8}$$

$$\text{Elevation angle :} \quad \varepsilon = \tan^{-1}\left(\frac{\cos\gamma - 1/k}{\sin\gamma}\right) \tag{7.9}$$

For the specific case of an equatorial orbit (inclination $= 0$) as shown in Figure 7.4, the Earth central angle is given by $\cos\gamma = \cos\psi\cos\delta$, and the tilt of the satellite-Earth terminal-Earth center plane is

$$\phi = \tan^{-1}\left(\frac{\sin\delta}{\tan\psi}\right) \tag{7.10}$$

The Earth terminal azimuth angle, α, measured clockwise from north viewed from local zenith as shown in Figure 7.4 is given first by computing $\beta = \tan^{-1}(\tan\delta/\sin\psi)$. If the terminal is in the northern hemisphere ($\psi > 0°$), then $\alpha = \beta + 180°$. If the terminal is in the southern hemisphere, then $\alpha = \beta + 360°$ if it is east of the satellite ($\delta > 0°$) and $\alpha = \beta$ if it is west of the satellite.

The fraction of the Earth's surface area covered by the satellite for a given elevation angle, ε, and the corresponding Earth central angle, γ, is

$$A_c/A_e = (1 - \cos\gamma)/2 \tag{7.11}$$

For a GEO satellite, $k = 6.6107$, $r_0 = 42164$ km, $h = 35786$ km and the maximum nadir angle, from Equation (7.6) when $\varepsilon = 0$ is $\sin\theta = 1/k$ for which $\theta = 8.7°$ and $\gamma = 81.3°$ that represents the maximum latitude that can be covered from GEO. For this case, a GEO satellite *sees* about 42% of the Earth's surface. However, since they are not typically operated below approximately $\varepsilon \approx 10°$, the practical relative Earth coverage area is in the region of 34%.

For a GEO satellite transmitting (or receiving) linear polarization, a useful quantity is the polarization tilt as seen by the Earth terminal when it is aimed toward the satellite. The general method is developed from the vector relationships among the satellite antenna polarization vector, the direction of propagation, and the Earth terminal location (refer to Figure 7.4). For a general linear polarization unit vector $\hat{e}$ at the satellite, a unit vector component normal to the direction of propagation, $\hat{r}_s$, is $\hat{e}_p = (\hat{r}_s \times \hat{e}) \times \hat{r}_s = \hat{e} - (\hat{e} \cdot \hat{r})\hat{r}_s$. The tilt angle may be taken as the angle between this vector and a vector in a reference plane containing the Earth center, satellite, and Earth terminal. This plane, shown in Figure 7.4, appears tilted at angle ϕ as viewed from the satellite. A unit vector normal to this plane is given by $(\hat{r}_e \times r_s)/r_s$ and the tilt angle relative to this plane is

$$\tau = \sin^{-1}(\hat{e}_p \cdot (\hat{r}_e \times r_s)/r_s) \tag{7.12}$$

For the special case of a "vertical" polarization vector in the x–z or north–south plane as seen at the satellite, the tilt angle relative to the local "vertical" or up-and down direction as seen from behind the terminal looking at the satellite is just ϕ (see Equation (7.10)).

7.5   Communications Link

Figure 7.6 illustrates the elements of the radio frequency (RF) link between a satellite and Earth terminals. The overall link performance is determined by computing the link equation for the uplink and downlink separately and then combining the results along with interference and intermodulation effects.

For a radio link with only thermal noise, the received carrier-to-noise power ratio is

$$\left(\frac{c}{n}\right) = (p_t\, g_t)\left(\frac{1}{4\pi r_s^2}\right)\left(\frac{g_r}{T}\right)\left(\frac{1}{k}\right)\left(\frac{\lambda^2}{4\pi}\right)\left(\frac{1}{a}\right)(\rho)\left(\frac{1}{b}\right) \qquad (7.13a)$$

The same quantities expressed in dB are

$$(C/N) = EIRP - 10\log\left(4\pi r_s^2\right) + (G_r - 10\log T) + 228.6 - 10\log\left(4\pi/\lambda^2\right) - A + \Gamma - B$$

$$(7.13b)$$

where $\lambda = c/f$ is the wavelength and $c = 2.9979\times10^8$ m/s is the velocity of light. Subscripts refer to transmit (t) and receive (r). Lower case terms are the actual quantities in watts, meters, etc., and the capitalized terms in Equation (7.13b) correspond to the decibel (dB) versions of the parenthesized quantities in Equation (7.13a). For example, $EIRP = P + G = 10\log p + 10\log g$ decibels relative to 1 W (dBW) and the expression (C/N) should be interpreted as $10\log c - 10\log n$. The uplink and downlink equations have identical form with the appropriate quantities substituted in Equation (7.13). The relevant quantities are described below.

The ratio of received carrier power to noise power c/n and its corresponding decibel value $(C/N) = 10\log(c/n)$ dB is the primary measure of link quality. The product of transmit power p_t (W) and the transmit antenna gain g_t, or equivalently, P_t (dBW) $+$ G_t (dBi), is where the antenna gain is expressed in decibel relative to an isotropic antenna, is called the equivalent isotropically radiated power (EIRP), and its unit is dBW because the antenna gain is dimensionless. The antenna gain is that *in the direction of the link*, i.e., it is not necessarily the antenna's peak gain. The received thermal noise power is $n = kTB$ W, where $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant and $10\log(k) = -228.6$ dBW/K/Hz. T is the system noise temperature in kelvins (K), and B is the bandwidth in dB Hz. Then, $G - 10\log T$ dB/K is a figure of merit for the receiving system. It is usually written as G/T and read as *gee over tee*. The antenna gain and noise temperature must be defined at the same reference point, e.g., at the receiver's input port or at the antenna terminals.

The spreading factor $4\pi r_s^2$ is independent of frequency and depends *only* on the slant range distance r_s. The gain of an antenna with an effective aperture area of 1 m^2 is $10\log(4\pi/\lambda^2)$. The dB sum of the spreading factor and the gain of a 1 m^2 antenna is the frequency-dependent *path loss*. "A" is the signal attenuation due to dissipative losses in the propagation medium. $B = 10\log(b)$ is the bandwidth in dB Hz, where b is the bandwidth in Hz.

The polarization mismatch factor between the incident wave and the receive antenna, is given by $\Gamma = 10\log\rho$ where $0\le\rho\le1$. This factor may be obtained from the voltage axial ratio of the incident wave r_w, the voltage axial ratio of the receive antenna's polarization response r_a, and the difference
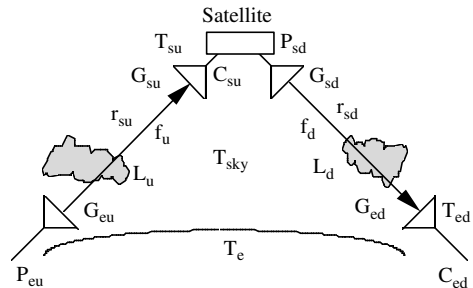
FIGURE 7.6 Quantities for a satellite RF link. $P =$ transmit power (dBW). $G =$ antenna gain (dBi). $C =$ received carrier power (dBW). $T =$ noise temperature (K). $L =$ dissipative loss (dB). $r_s =$ slant range (m). $f =$ frequency (Hz). u $=$ uplink. d $=$ downlink. e $=$ Earth. s $=$ satellite.

in tilt angles of the wave and antenna polarization ellipses $\Delta\tau = \tau_w - \tau_a$, as follows

$$\rho = \frac{1}{2} + \frac{4r_w r_a + (r_w^2 - 1)(r_a^2 - 1)\cos(2\Delta\tau)}{2(r_w^2 + 1)(r_a^2 + 1)} \tag{7.14}$$

where the axial ratios are each signed quantities, having a positive sign for right-hand sense and a negative sign for left-hand sense. Therefore, if the wave and antenna have opposite senses, the sign of $4r_w r_a$ is negative. The axial ratio in dB is $R = 20 \log|r|$. The maximum polarization coupling occurs when the wave and antenna are copolarized, have identical axial ratios, and their polarization ellipses are aligned ($\Delta\tau = 0$). It is minimum when the axial ratios are identical, the senses are opposite, and the tilt angles differ by 90°.

7.6 System Noise Temperature and G/T

The system noise temperature T incorporates contributions to the noise power radiated into the receiving antenna from the sky, ground, and galaxy, as well as the noise temperature due to circuit and propagation losses and the noise figure of the receiver. The clear sky antenna temperature for an Earth station depends upon the elevation angle since the antenna's sidelobes will receive a small fraction of the thermal noise power radiated by the Earth, which has a noise temperature $T_{\text{Earth}} \approx 290\text{K}$. At 11 GHz, the clear sky *antenna* noise temperature, T_{Earth}, ranges from 5 to 10 K at zenith ($el = 90°$) to more than 50 K at $el = 5°$ [Pratt and Bostian, 1986].

As shown in Figure 7.7, the system noise temperature is developed from the standard formula for the equivalent temperature of tandem elements including: the antenna in clear sky, propagation (rain) loss of $A = 10 \log(a)$ dB, circuit losses between the aperture and receiver of L_c dB, and receiver noise figure of F dB (corresponding to receiver noise temperature T_r K). The system noise temperature referred to the antenna aperture is approximated by the following equation, where $T_{\text{rain}} \approx 280\text{K}$ is a reasonable approximation for the physical temperature of the rain [Pratt et al., 2000].

$$T = T_{\text{aclear}} + T_{\text{rain}}(a - 1)/a + T_r l_c + 290(l_c - 1) \tag{7.15}$$
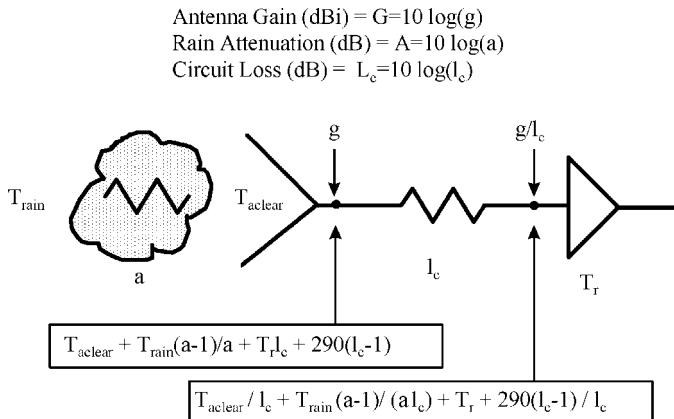
FIGURE 7.7 Tandem connection of antenna, loss elements such as waveguide, and receiver front end. The noise temperature depends on the reference plane but g/T is the same for both points shown.

The system noise temperature is defined at a specific reference point such as the antenna aperture or the receiver input. However, G/T is independent of the reference point when G correctly accounts for circuit losses. The satellite's noise temperature is generally higher than an Earth terminal's under clear sky conditions because the satellite antenna sees a warm Earth temperature of $\approx$150–300K, depending on the proportion of clouds, oceans, and land in the satellite antenna's beam; whereas a directive Earth antenna generally sees cold sky, and the sidelobes generally receive only a small fraction of noise power from the warm Earth. Furthermore, a satellite receiving system typically has a higher noise temperature due to circuit losses in the beam forming networks, protection circuitry, and extra components for redundancy.

Figure 7.8 illustrates the link loss factors, maximum nadir angle, θ, Earth central angle, γ, and Earth-space propagation time delay as a function of satellite altitude. The delay for a signal *hop* between two Earth locations includes the delays for the Earth-space path, the space-Earth path, and all circuit delays. The path losses are shown for several satellite frequencies in use. The variation in path loss and Earth central angle is substantial. For example, L-band LEO personal communications systems to low-cost hand-held telephones with low gain (e.g., $G \approx -2$ to $+3$dBi) need less link power than for MEO or GEO. On the other hand, more satellites are needed from LEO constellations to provide full Earth coverage since each satellite sees a much smaller fraction of the Earth compared with higher orbits.

The design for a constellation of satellites to serve communications needs such as the number of satellites, their orbital parameters, the satellite G/T and *EIRP*, etc., are topics related to mission analysis and design and these involve trades of many factors such as total communications capacity, link margins, space segment and Earth segment costs, reliability, interconnectivity, availability and cost of launch vehicles, mission lifetime, and system operations [Wertz and Larson, 1991].
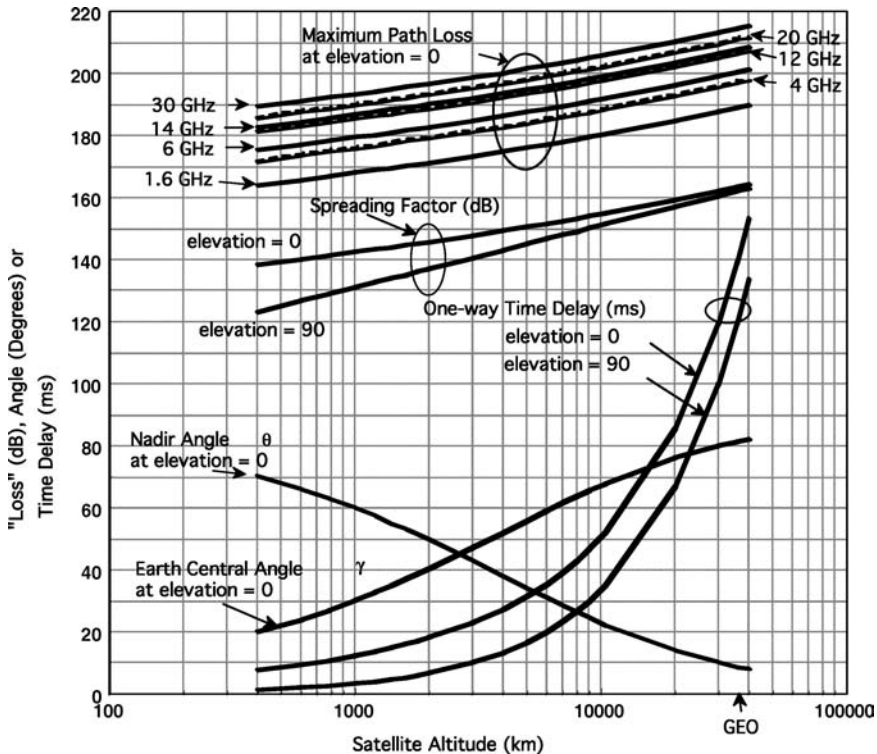
FIGURE 7.8   Satellite link losses, spreading factors, maximum nadir angle, θ_{max}, Earth central angle, γ, and one-way time delay vs. satellite altitude, h km.

7.7   Digital Links

For digital modulation systems, the bit error rate (BER) is related to the dimensionless ratio (dB difference) of energy per bit, E_b dB J, to the total noise power density $N_0 = 10\log(kT)$ dB J [Sklar, 2001]. For a system with only thermal noise N_0,

$$(E_b/N_0) = (C/N) + B - R = (C/N_0) - R \text{ dB} \tag{7.16}$$

where $R = 10\log$ (bit rate in bit/s), B is the bandwidth (dB Hz), and (C/N_0) is dB ratio of *carrier power to thermal noise power density*, that is, (C/N) normalized to unit bandwidth. Curves relating the communications performance measure of (BER) vs (E_b/N_0) for different modulations may be found in Sklar [2001]. The link equation may then be expressed in terms of (E_b/N_0) and transmission data rate, R, without explicit reference to the bandwidth

$$(E_b/N_0) = EIRP + (G/T) + 228.6 - 20\log(4\pi r_s/\lambda) - A + \Gamma - R \text{ dB} \tag{7.17}$$

where the appropriate quantities are substituted depending on whether the uplink or downlink is being considered.

7.8   Interference

A complete transponder link analysis must include the contributions of the uplink, downlink, and also the power sum of all interference signals due, for example, to intermodulation products that are generated in the output stages of the amplifiers, external interference from other systems, and intra-system interference from reusing the same frequency band on spatially isolated or dual-polarized antenna beams to increase communications capacity. For most applications the total interference power may taken as the power sum of interfering signals as long as they are not correlated with the desired carrier. The values for the interfering signals due to effects outside the satellite, for example, *frequency reuse* cross-polarization, multiple beam interferers, and interference power received from other systems must be obtained by carefully constructing the link equation for each case, taking into account the antenna gains for each polarization and beam direction of concern.

For an interference power i w and carrier power c w, the interference ratio c/i must be combined with the uplink and downlink c/n values to yield the total c/n. Here, the ratios are written in lower case to indicate they are *numerical power ratios*.

$$\left(\frac{c}{n}\right)_{total} = \frac{1}{\left(\frac{c}{n}\right)_u^{-1} + \left(\frac{c}{n}\right)_d^{-1} + \left(\frac{c}{i}\right)_u^{-1} + \left(\frac{c}{i}\right)_d^{-1} + \left(\frac{c}{i}\right)_{other}^{-1}} \tag{7.18}$$

Equation (7.18) applies to a *bent pipe* satellite. If on-board signal regeneration is used for digital transmission, the uplink signal is demodulated and a *clean* set of baseband bits is remodulated. This has the effect of separating the accumulation of uplink and downlink noise contributions by causing the uplink noise to be effectively modulated onto the downlink carrier with the desired signal [Gagliardi, 1991]. In that case, only the uplink or the downlink term in the denominator of Equation (7.18) would be used as appropriate. Remodulation is also useful for intersatellite links. In each case, a saving in power or antenna size may be obtained at the expense of circuit and processing complexity.

The degradation to a digital link from interference follows a form similar to that of Equation (7.18) in terms of e_b/n_0, where the lower case quantities refer to numerical ratios. For a link that is subject to a *given* additive white noise-like interference power expressed as a ratio of desired signal power to interference power c/i, and

assuming digital modulation with m bits per symbol,

$$\frac{e_b}{i_0} = \frac{1}{m}\frac{c}{i}$$ (7.19)

The ratio of energy per bit to total thermal noise plus interference power density is

$$\left(\frac{e_b}{n_0 + i_0}\right) = \frac{1}{\left(\frac{e_b}{n_0}\right)^{-1} + \left(\frac{e_b}{i_0}\right)^{-1}}$$ (7.20)

For a system employing frequency reuse via dual polarizations, the polarization coupling factor Γ between a wave and antenna determines the interference power. The (C/I) due to polarization is the ratio of desired (copolarized) receive power and undesired (cross polarized) powers and is called the *polarization isolation*. It may be found by application of Equation (7.14) to copolarized and cross-polarized cases.

7.9   Some Particular Orbits

A *geosynchronous* orbit has a period that is a multiple of the Earth's rotation period, but it is not necessarily circular, and it may be inclined. Therefore, a *geostationary* Earth orbit (GEO) is a special case of a geosynchronous orbit where: $e = 0$, $i = 0$, $k = (r_e + h/r_e) = 6.61$, and $h = 35,786$ km. When $el = 0$, the maximum nadir angle $\theta = 8.7°$, the maximum slant range is 41,680 km, and from Equation (7.8), $\gamma = 81.3°$. Therefore, a GEO satellite cannot see the Earth above 81.3° latitude.

Molniya and Tundra orbits are critically inclined ($i = 63.4°$) and have orbital periods of 12 h and 24 h, respectively. These highly inclined elliptical orbits (HIEO) cause the satellite's subsatellite ground trace to dwell at apogee at the same place each day, allowing several satellites to be phased to offer quasi-stationary satellite service with high elevation angles at high latitudes [Maral and Bousquet, 2002]. The Sirius Satellite Radio constellation of three satellites is inclined at 63.4° and has an orbit period of 24 h with apogee at orbit radius of 42164 km (the same as a GEO satellite). However, the elliptical orbit has a perigee of 24,469 km, so that the satellites *dwell* over the northern hemisphere, providing high elevation angles seen from the ground (see www.ils.launch.com). Sirius claims that this reduces the need for terrestrial repeaters. For full Earth coverage from a constellation of LEO satellites, circular polar constellations [Adams and Rider, 1987] and constellations of orbit planes with different inclinations, e.g., Walker Orbits [Walker, 1977] have received attention.

The oblateness of the Earth causes the right ascension of the ascending node Ω (Figure 7.3) to move in time with the equatorial plane in the opposite direction to the satellite's motion as seen from above the ascending node. This is called regression of the nodes. For inclination $i < 90°$ (prograde orbit), the ascending node rotates westward. For $i > 90°$ (retrograde orbit), the ascending node rotates eastward. For $i = 90°$, the regression is zero. The orbit parameters may be chosen such that the nodal regression is $360°/365.24 = 0.9856$ degrees eastward per day. In that case, the orbit plane will maintain a constant angle with the sun. The local solar time for the line of nodes is constant, that is, the satellite crosses a given latitude at the same solar time and solar lighting conditions each day. This *sun-synchronous* orbit has advantages for certain applications such as weather and surveillance satellites [Roddy, 2001].

Table 7.1 compares the geometry, coverage, and some parameters relevant to the communications links for some LEO, MEO, and GEO systems. Reference should be made to Figure 7.4 for the geometry and to the given equations for geometrical and link parameters.

TABLE 7.1   Comparison of Orbit and Link Parameters for LEO, MEO, and GEO for the Particular Case of Circular Orbits (eccentricity, $e = 0$) and for Elevation Angle, $el = 10°$

Orbit	LEO	MEO	GEO
Example System	Iridium®	ICO*	Intelsat
Inclination, i (deg.)	86.4	±45	0
Altitude, h (km)	780	10,400	35,786
Semi-major axis radius, a (km)	7159	16,778	42,164
Orbit Period (minutes)	100.5	360.5	1,436.1
$(r_e + h)/r_e$	1.1222	2.6305	6.6107
Earth Central Angle, γ(deg.)	18.658	58.015	71.433
Nadir Angle, θ (deg.)	61.3	22	8.6
Nadir Spread Factor:			
$10 \log(4\pi h^2)$ (dB m^2)	128.8	151.3	162.1
Slant Range, r_s (km)	2,325	14,450	40,586
One-way time delay (ms)	2.6	51.8	139.1
Maximum Spread Factor:			
$10 \log(4\pi r_s^2)$ (dB m^2)	138.3	154.2	163.2
$20 \log(r_s/h)$ (dB)	9.5	2.9	1.1
Ground Coverage Area (km^2)	13.433×10^6	120.2×10^6	174.2×10^6
Fraction of Earth Area	0.026	0.235	0.34

Note: Earth radius, $r_e = 6378.14$ km, Earth surface area, $a_e = 511.2 \times 10^6$ km^2, elevation, $e = 10°$. The MEO system never materialized and is shown here for comparison.

7.10   Access and Modulation

Satellites act as central relay nodes, which are visible to a large number of users who must efficiently use the limited resources of power and bandwidth. For detailed discussions of access issues, see Gagliardi [1991], Pritchard et al., [1993], Miya [1985], Roddy [2001], and Shimbo [1988]. A brief summary of issues specific to satellite systems is now given.

Frequency-division multiple access (FDMA) has been the most prevalent access for satellite systems until recently. Individual users assigned a particular frequency band may communicate at any time. Satellite filters subdivide a broad frequency band into a number of *transponder channels*. For example, the 500 MHz uplink FSS band from 5.925 to 6.425 GHz may be divided into 12 transponder channels of 36 MHz bandwidth plus guard bands. This limits the interference among adjacent channels in the corresponding downlink band of 3.7 to 4.2 GHz.

FDMA implies that several individual carriers coexist in the transmit amplifiers. In order to limit intermodulation products caused by nonlinearities, the amplifiers must be operated in a *backed off* condition relative to their saturated output power. For example, to limit third-order intermodulation power for two carriers in a conventional traveling wave tube (TWT) amplifier to about -20 dB relative to the carrier, its input power must be reduced (*input backoff*) by about 10 dB relative to the power that would drive it to saturation. The output power of the carriers is reduced by about 4 to 5 dB (*output backoff*). Amplifiers with fixed bias levels will consume power even if no carrier is present. Therefore, DC-to-RF efficiency degrades as the operating point is backed off. For amplifiers with many carriers, the intermodulation products have a noise-like spectrum and the noise-power ratio is a better measure of multicarrier performance.

When reusing the available frequency spectrum by multiple spatially isolated beams (SDMA), interference can result if the sidelobes of one beam receives or transmits substantial energy in the direction of the other beams. Two beams that point in the same direction may reuse frequencies provided that they are orthogonally polarized, for example, vertical and horizontal linear polarizations or right- and left-hand circular polarizations. Typical values of sidelobe or polarization isolation among beams reusing the same frequency bands are from 27 to 35 dB.

Time-division multiple access (TDMA) users share a common frequency band and are each assigned a unique time slot for their digital transmissions. At any instant the DC-RF efficiency is high because there is

only one carrier in the transmit amplifier, which may be operated near saturation. A drawback is the system complexity required to synchronize widely dispersed users in order to avoid intersymbol interference caused by more than one signal appearing in a given time slot. Also, the total transmission rate in a TDMA satellite channel must be essentially the sum of the users' rates, including overhead bits such as for framing, synchronization and clock recovery, and source coding.

In *code-division multiple access* (CDMA) each carrier is modulated with a unique pseudo-random code, usually by means of either a direct sequence or frequency hopping spread spectrum modulation. Because the CDMA users occupy the same frequency band at the same time, the aggregate signal in the satellite amplifier is noise-like. Individual signals are extracted at the receiver by correlation processes. CDMA tolerates noise-like interference but does not tolerate large deviations from average loading conditions. One or more very strong carriers could violate the noise-like interference condition and generate strong intermodulation signals. Careful power control of each user's signal is usually required in CDMA systems.

User access is through assignments of a frequency, time slot, or code. Fixed assigned channels allow a user unlimited access. However, this may result in poor utilization efficiency for the satellite resources and may imply higher user costs (analogous to a leased terrestrial line). Other assignment schemes include *demand assigned multiple access* (DAMA) and *random access* (e.g., for the Aloha concept). DAMA systems require the user to first send a channel request over a common control channel. The network controller (at another Earth station) seeks an empty channel and instructs the sending unit to tune to it either in frequency or time slot. A link is maintained for the call duration and then released to the system for other users to request. Random access is economical for lightly used burst traffic such as data. It relies on random time of arrival of data packets, and protocols are in place for repeat requests in the event of collisions [Gagliardi, 1991].

In practice, combinations of multiplexing and access techniques may be used. A broad band may be channelized or *frequency-division multiplexed* (FDM), and FDMA may be used in each sub-band, for example, FDM/FDMA.

7.11   Frequency Allocations

Table 7.2 contains a partial list of frequency allocations for satellite communications. The International Telecommunications Union (ITU) holds periodic World Radiocommunications Conferences (WRC) related to frequency allocations. A useful chart of U.S. allocations may be obtained at http://www.ntia.doc.gov/osmhome/allochrt.pdf.

TABLE 7.2   A Partial List of Satellite Frequency Allocations — Frequencies in GHz

Band	Uplink	Downlink	Satellite Service
	1.613.8–1.6605	1.6138–1.6265	Mobile
	1.215–1.240		Radio navigation
		1.227	GPS
S-Band	1.980–2.010	2.170–2.200	MSS
		2.32–2.345	Broadcast Satellite (U.S.)
		2.4835–2.500	Mobile
C-Band	5.85–7.075	3.4–4.2	Fixed (FSS)
	7.250–7.300	4.5–4.8	FSS
X-Band	7.9–8.4	7.25–7.75	FSS
Ku-Band	12.75–13.25	10.7–12.2	FSS
	14.0–14.8	12.2–12.7	Direct Broadcast (BSS) (U.S.)
Ka-Band	17.3–17.78		FSS (BSS in U.S.)
			22.55–23.55 Intersatellite
	27–31	17–21	FSS
Q	42.5–43.5,	37.5–40.5	FSS, MSS
	47.2–50.2		

7.12   Satellite Subsystems

The major satellite subsystems are described in, for example, Griffin and French [2004]. They are: propulsion, power, antenna, communications repeater, structures, thermal, **attitude** determination and control, telemetry, tracking, and command. Thermal control is described in Gilmore [1994].

The satellite *antennas* typically are offset-fed paraboloids. Typical sizes are constrained by launch vehicles and have ranged from less than 1 m to more than 20 m for some applications. The XM satellites use two 5-m shaped reflectors, each with a single feed and high power transmitters to produce downlink EIRP of more than 67 dBW over the contiguous United States (CONUS). A more traditional approach to beam shaping were the Intelsat 6 satellites, which used a D = 3.2-m diameter antenna at 4 GHz. Multiple feeds in the focal region. Each feed produces a narrow *component beam* whose beamwidth is $\approx$65λ/D and whose directions are established by the displacement of the feeds from the focal point. These beams are combined to produce a shaped beam with relatively high gain over a geographical region. Multiple beams are also used to reuse frequencies on the satellite. Figure 7.2 suggests that a satellite may have several beams for frequency reuse. In that case, the carriers occupying the same frequencies must be isolated from each other by either polarization orthogonality or antenna sidelobe suppression. As long as the sidelobes of one beam do not radiate strongly in the direction of another, both may use the same frequency band and increase the satellite's capacity.

The *repeaters* include the following main elements (see Figure 7.2): a low noise amplifier (LNA) amplifies the received signal and establishes the uplink noise. The G/T of the satellite receiver includes the effect of losses in the satellite antenna, the noise figure of the LNA, and the noise temperature of the Earth seen from space (from 150 to 290 K depending on the percentage of the beam area over oceans, land and clouds). In a conventional repeater, the overall frequency band is down-converted by a local oscillator (LO) and mixer from the uplink band to the downlink band. It is channelized by an input multiplexer into a number (e.g., 12) of transponder channels. A separate high-power amplifier is typically used for each of these channelized signals. Traveling wave tube amplifiers (TWTA) are typically used for very high powers, e.g., up to >200 W for a DBS. Solid-state amplifiers at can provide more than 15 W at C- and Ku-Bands.

The *attitude determination and control system* (ADCS) must maintain the proper angular orientation of the satellite in its orbit in order to keep the antennas pointed to the Earth and the solar arrays aimed toward the sun (for example). The two prevalent stabilization methods are spin stabilization and body stabilization. In the former, the satellite body spins, and the angular momentum maintains gyroscopic stiffness. The latter uses momentum wheels to keep the spacecraft body orientation fixed. Components of this subsystem include the momentum wheels, torquers (which interact with the Earth's magnetic field), gyros, sun and Earth sensors, and thrusters to maintain orientation.

The *telemetry tracking and command* (TT&C) subsystem receives data from the ground and enables functions on the satellite to be activated by appropriate codes transmitted from the ground. This system operates with low data rates and requires omni-directional antennas to maintain ground contact in the event the satellite loses its orientation.

The *power* subsystem comprises batteries and a solar array. The solar array must provide enough power to drive the communications electronics as well as the housekeeping functions, and it must also have enough capacity to charge the batteries that power the satellite during eclipse, that is, when it is shadowed and receives no power from the sun [Richharia, 1999; Griffin and French, 2004]. Typical battery technology uses Nickel-Hydrogen cells, which can provide a power density of more than 50 W-h/kg. Silicon solar cells can yield more than 170 W/m^2 at a satellite's beginning of life (BOL). Gallium Arsenide solar cells (GaAs) yield more than 210 W/m^2. However, they are more expensive than silicon cells.

The space environment including radiation, thermal, and debris issues are described in Wertz and Larson [1991], Griffin and French [2004], Committee on Space Debris 1995, and Johnson and McKnight [1991]. The structure must support all the functional components and withstand the rigors of the launch environment. The thermal subsystem must control the radiation of heat to maintain a required operating temperature for critical electronics [Gilmore, 1994].

7.13   Trends

Satellite communications must now be viewed as a mature industry. Satellites lost competitiveness for point-to-point voice traffic compared with fiber and terrestrial mobile systems. Satellites do best when they exploit their unique wide view of the Earth for such applications as broadcast, data services to wide areas, and mobile communications. Video will represent a large component of satellite capacity demand, and as HDTV increases, direct broadcast is expected to be the largest user of satellite capacity.

During the 1990s and early 2000s, considerable speculation of expensive systems such as Iridium, Globalstar, ICO, and Teledesic, along with an excess of GEO capacity relative to demand, forced the industry into a downturn. This was partly fueled by the growing perception that satellites were risky, thereby resulting in lack of favor among investors. Ka-band systems for broadband data did not materialize as expected and were found to be costly relative to emerging cable and DSL. Yet, some services can be best provided by satellite systems, and satellites will find applications, even for broadband data services such as satellite Internet using the Ku- and Ka-bands, perhaps even at higher frequencies. Satellite construction has evolved from a craft industry with extensive custom design, long lead times, long test programs, and high cost toward reduced design and construction cycle times, although care must be taken to ensure reliability. Technology advances include development of light-weight small satellites for economical provision of data, communications services at low cost, and improved components.

Defining Terms

Attitude: The angular orientation of a satellite in its orbit, characterized by roll (R), pitch (P), and yaw (Y). The roll axis points in the direction of flight, the yaw axis points toward the Earth's center, and the pitch axis is perpendicular to the orbit plane such that $R \times P \rightarrow Y$. For a GEO satellite, roll motion causes north-south beam pointing errors, pitch motion causes east-west pointing errors, and yaw causes a rotation about the subsatellite axis.

Backoff: Amplifiers are not linear devices when operated near saturation. To reduce intermodulation products for multiple carriers, the drive signal is reduced or backed off. Input backoff is the decibel difference between the input power required for saturation and that employed. Output backoff refers to the reduction in output power relative to saturation.

Beam and polarization isolation: Frequency reuse allocates the same bands to several independent satellite transponder channels. The only way these signals can be kept separate is to isolate the antenna response for one reuse channel in the direction or polarization of another. The beam isolation is the coupling factor for each interfering path and is always measured at the receiving site, that is, the satellite for the uplink and the Earth terminal for the downlink.

Bus: The satellite bus is the ensemble of all the subsystems that support the antennas and payload electronics. It includes subsystems for electrical power, attitude control, thermal control, TT&C, and structures.

Frequency reuse: A way to increase the effective bandwidth of a satellite system when available spectrum is limited. Dual polarizations and multiple beams pointing to different Earth regions may utilize the same frequencies as long as, for example, the gain of one beam or polarization in the directions of the other beams or polarization (and vice versa) is low enough. Isolations of 27 to 35 dB are typical for reuse systems.

References

W.S. Adams and L. Rider, "Circular polar constellations providing continuous single or multiple coverage above a specified latitude," *J. Astronaut. Sci.*, vol. 35, no. 2, pp. 155–192, 1987.

G. Björnstrom, "Digital payloads: enhanced performance through signal processing," *ESA J.*, 17, 1–29, 1993.

R.D. Briskman, "Satellite radio technology," in *16th Int. Commun. Satell. Syst. Conf.*, Washington, DC: American Institute of Aeronautics and Astronautics, Febrauary 25–29, 1996, pp. 821–825.

A. Chobotov, *Orbital Mechanics*, 2nd ed., Washington, DC: American Institute of Aeronautics and Astronautics, 1991.

A.C. Clarke, "Extra terrestrial relays," *Wireless World*, October, 1945.

S. De Gaudenzi, F. Gianetti, and M. Luise, "Advances in satellite CDMA transmission for mobile and personal communications," *Proc. IEEE*, vol. 84, no. 1, pp. 18–39, 1996.

Committee on Space Debris, National Research Council, *Orbital Debris*, Washington, DC: National Academy Press, 1995.

B. Ebert, *The Satellite Communication Ground Segment and Earth Station Handbook*, Wiley, 2000.

K. Feher, *Digital Communications: Satellite/Earth Station Engineering*, Englewood Cliffs, NJ: Prentice-Hall, 1983.

K. Feher, *Advanced Digital Communications*, Englewood Cliffs, NJ: Prentice Hall, 1987.

S. Finkelstein and S.H. Sanford, "Learning from corporate mistakes: The rise and fall of iridium," *Org. Dynamics*, 29(2): 138–148, 2000.

Futron Corporation, "The Transformation of the Satellite Services Industry," January 2005, www.futron.com

M. Gagliardi, *Satellite Communications*, New York: Van Nostrand Reinhold, 1991.

D.G. Gilmore, Ed., *Satellite Thermal Control Handbook*, El Segundo, CA: The Aerospace Corporation Press, 1994.

G. Gordon and W. Morgan, *Principles of Communications Satellites*, New York: Wiley, 1993.

M.D. Griffin and J.R. French, *Space Vehicle Design*, 2nd ed., Reston, VA: American Institute of Aeronautics and Astronautics, 2004.

A. Inkpen, M. Martin, and I. Fas-Pacheo, The Rise and Fall of Iridium, Thunderbird, the American Graduate School of International Management: 2000.

L.J. Ippolito, *Radiowave Propagation in Satellite Communications*, New York: Van Nostrand Reinhold, 1986.

L.J. Ippolito, *Propagation Effects Handbook for Satellite Systems Design*, NASA, 2000.

J. Isakowitz, *International Reference Guide to Space Launch Systems*, 3rd ed., Reston, VA: American Institute of Aeronautics and Astronautics, 1999.

ITV Handbook on Satellite Communications, International Communications Union, 2002.

K.G. Johannsen, "Mobile P-service satellite system comparison," *Int. J. Satell. Commun.*, 13, 453–471, 1995.

N.L. Johnson and D.S. McKnight, *Artificial Space Debris*, Malabar, FL: Krieger Publishing Co., 1991.

G. Maral, *VSAT Networks*, 2nd ed., New York: Wiley, 2003.

G. Maral and M. Bousquet, *Satellite Communication Systems*, 4th ed., John Wiley & Sons, 2002.

K. Miya, Ed., *Satellite Communications Technology*, Tokyo: KDD Engineering and Consulting, Inc., 1985.

W.L. Morgan and G.D. Gordon, *Communications Satellite Handbook*, New York: Wiley, 1989.

Office of Science and Technology Policy, Interagency Report on Orbital Debris, (Library of Congress Catalog No. 95-72164), November, 1995.

J.J. Pocha, *An Introduction to Mission Design for Geostationary Satellites*, Dordrecht, The Netherlands: D. Reidel, 1987.

T. Pratt, C.W. Bostian, and J.E. Allnut, *Satellite Communications*, 2nd ed., New York: Wiley, 2000.

W.L. Pritchard, H.G. Suyderhoud, and R.A. Nelson, *Satellite Communications Systems Engineering*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1993.

D.W.E. Rees, *Satellite Communications: The First Quarter Century of Service*, New York: Wiley, 1990.

M. Richharia, *Satellite Communications Systems*, 2nd ed., New York: McGraw-Hill, 1999.

D. Roddy, *Satellite Communications*, 3rd ed., New York: McGraw-Hill, 2001.

M. Schwartz, *Information Transmission, Modulation, and Noise*, New York: McGraw-Hill, 1990.

A. Scott, *Understanding Microwaves*, New York: Wiley, 1993.

O. Shimbo, *Transmission Analysis in Communications Systems*, vol. 1 & 2, New York: Computer Science Press, 1988.

B. Sklar, *Digital Communications*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 2001.

J.G. Walker, *Continuous Whole-Earth Coverage by Circular Orbit Satellite Patterns*, Technical Report 77044, Royal Aircraft Establishment, Farnborough, Hants, U.K. 1977.

J.R. Wertz, Ed., *Spacecraft Attitude Determination and Control*, Dordrecht, The Netherlands: D. Reidel Publishing Co., 1978.

J.R. Wertz and W.J. Larson, Eds., *Space Mission Analysis and Design,* 3rd ed., Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999.

Further Information

Extensive summaries of satellites, frequencies, and orbital locations may be found at www.lyngsat.com. Satmaster, a popular commercial link budget program may be purchased from Arrowe Technical Services at www.satmaster.com. A convenient Web calculator for Earth terminal pointing and polarization tilt adjustment may be found at http://www.sadoun.com/sat/installation/satellite-heading-calculator.htm. For a brief history of satellite communications see *Satellite Communications: The First Quarter Century of Service*, by D. Reese, Wiley, 1990. Descriptions of many satellites can be found in D.H. Martin, *Communications Satellites*, 4th ed., Reston, VA, American Institute of Aeronautics and Astronautics, 2000. An overview of trends may be found in T. Iida et al., *Satellite Communications in the 21st Century: Trends and Technologies*, Reston, VA, American Institute of Aeronautics and Astronautics, 2003. Basic information and analyses for satellite communications may be found in *The ITV Handbook of Satellite Communications*. The emerging digital video broadcast standard, DVB S2 is described in the European Telecommunications Standards Institute (ETSI) http://webapp.etsi.org/action/pv/20050215/tr_102376v0101.

Propagation issues are summarized in *Propagation Effects Handbook for Satellite Systems Design*, 2000, NASA. Many of the organizations mentioned can be accessed through the Internet: (www.nasa.gov); International Telecommunications Union (ITU) (www.itu.ch); Inmarsat FCC (www.fcc.gov); Galileo: www.esa.int/export/esaNA/GGGMX650NDC_index_0.html.

Sirius: http://www.ilslaunch.com/launches/cbin/Mission_Overview/proton/sirius3_mo.pdf and www.sirius. com, XM: www.xmradio.com. For satellite communications to low profile in-motion terminals on vehicles, see www.raysat.com.

8

Digital Video Processing

Todd R. Reed
University of Hawaii

8.1   Introduction

Rapid increases in performance and decreases in cost of computing platforms and digital image acquisition and display subsystems have made digital images ubiquitous. Continued improvements promise to make digital video as widely used, opening a broad range of new application areas. In this chapter, some of the key aspects of this evolving data type are examined.

A Historical Perspective

The use of image sequences substantially predates modern video displays (see, e.g., [1]). As might be expected, the primary initial motivation for using these sequences was the depiction of motion. One of the earlier approaches to motion picture display was invented by the mathematician William George Horner in 1834. Originally called the Daedaleum (after Daedalus, who was supposed to have made figures of men that seemed to move), it was later called the Zoetrope (life turning) or the Wheel of Life. The Daedaleum works by presenting a series of images, one at a time, through slits in a circular drum, as the drum is rotated.

   Although this device is very simple, it illustrates some important concepts. First and foremost, the impression of motion conveyed by a sequence of images is illusory. It is the result in part of a property of the human visual system (HVS) referred to as persistence of vision. An image is perceived to remain for a period of time after it has been removed from view. This illusion is the basis of all motion picture displays. When the

drum in the device is rotated slowly, the images appear (as they are) a disjoint sequence of still images. As the speed of rotation increases (the images are displayed at a higher rate), a point is reached at which motion is perceived, even though the images appear to "flicker." Further increasing the speed of rotation, a point is reached at which flicker is no longer perceived (the critical fusion frequency). Finally, the slits in the drum illustrate a critical aspect of this illusion. In order to perceive motion from a sequence of images, the stimulus the individual images represent must be removed for a period of time between each presentation. If not, the sequence of images simply merges into a blur, and no motion is perceived.

These concepts (rooted in the nature of human visual motion perception) are fundamental, and are reflected in all motion picture acquisition and display systems.

Video

Unlike image sequences on film, video is represented as a 1-D signal, derived by scanning the camera sensor. The fact that the signal is derived by scanning imposes a particular signal structure, an example of which is shown in Figure 8.1 for a noninterlaced system.

In principle, scanning can be done in many ways. The simplest in concept is noninterlaced line continuous scanning (which yields the video signal just discussed). This approach is also referred to as progressive scanning. Viewed in the 2-D plane (either at the camera or display), this approach appears as shown in Figure 8.2.

The bandwidth of the resulting video signal is relatively high. Transmitting a frame of 485 lines,[1] with a 4:3 aspect ratio (NTSC resolution), at 60 frames per second requires roughly twice the available channel bandwidth (6 MHz). Sixty updates per second are needed to avoid wide area flicker, dictated by the temporal response of the HVS. One approach to reducing the signal bandwidth is to send half as many samples (lines). This cannot be accomplished by reducing the frame rate to 30 frames per second, because an unacceptable degree of flicker is introduced. Reducing the spatial resolution of each frame results in unacceptable blurring. Interlaced scanning is a compromise between the two approaches.
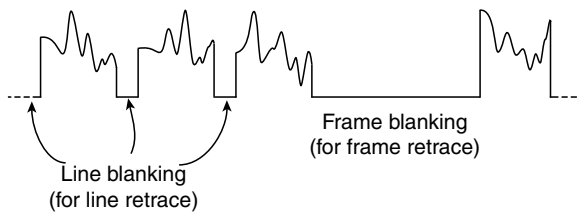
FIGURE 8.1   A noninterlaced video signal.
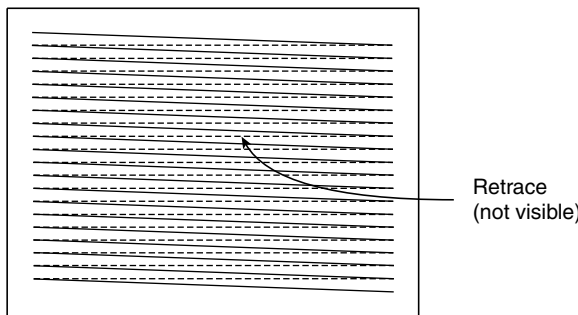
FIGURE 8.2   A noninterlaced scanning raster.

---

[1]NTSC consists of 525 lines, but only ~485 lines are active.
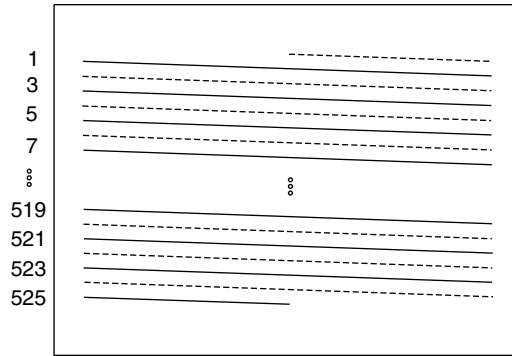
FIGURE 8.3 An NTSC frame, formed by interlacing two fields (2:1 interlace).

Direction
of motion

Serration due
to interlace
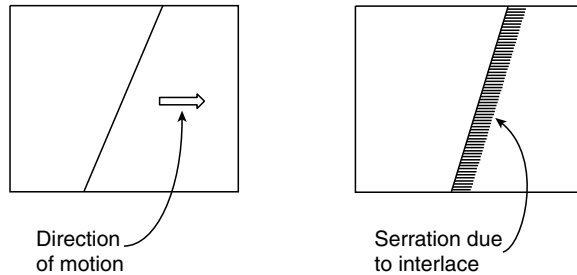
FIGURE 8.4 The effect of interlace on an edge in horizontal motion.

As used in NTSC television, each complete scan (a frame) contains 525 lines and occurs every 1/30 s. The frame consists of two fields (even and odd), 262½ lines each. These fields are interlaced to form the frame. Fields are scanned every 1/60 s (reducing flicker). Because two fields are interlaced to form one frame, this is called 2:1 interlace. Two interlaced fields (NTSC) are shown in Figure 8.3.

Image acquisition and display via scanning has several disadvantages. Nonideal aspects of the scanning system (e.g., nonzero spot size), and under some circumstances the act of scanning itself, lead to a reduction in vertical resolution below that predicted by the sampling theorem. The ratio of the actual to ideal resolution is called the Kell factor k, $0 \leq k \leq 1$. Typical values of k are $.6 < k < .8$, with interlaced systems having lower k. Scanning also causes distortion when objects in the scene are in motion. For example, a vertical line in motion will result in a tilted scanned image (not due to the tilt of the scan line, but because points on the line at the bottom of the screen are reached later than points at the top). Finally, different points in space within the frame do not correspond to the same point in time. Viewed in the spatiotemporal volume, each frame is tilted, with the upper left corner of the frame corresponding to a significantly earlier time than the lower right corner. This can make the accurate analysis of the image sequence difficult.

Interlaced scanning has additional disadvantages. Interlaced display systems suffer from interline flicker (particularly in regions of the image with nearly horizontal structure). Interlacing results in reduced vertical resolution, which increases aliasing. It also increases the complexity of subsequent processing or analysis (such as motion estimation). Interoperability with other systems, such as computer workstations (which use noninterlaced displays), is made difficult. Still images extracted from interlaced video ("freeze frames") are generally of poor quality.

Often only "freeze fields" are provided. This last point can be seen by considering the case of an edge in horizontal motion (Figure 8.4). To merge two fields to get a still image of reasonable quality, or to get a good progressively scanned sequence from an interlaced one, is a nontrivial problem.

FIGURE 8.5 An image sequence represented as a spatiotemporal volume, raytraced to exhibit its internal structure.

Image Sequences as Spatiotemporal Data

As discussed previously, the scanning process makes the precise specification of an image sequence difficult (since every spatial point exists at a different time). Interlace complicates matters further. In the remainder of this chapter, the simplifying assumption will be made that each point in a frame corresponds to the same point in time. This is analogous to the digitization of motion picture film, or the sequence which results from a CCD camera with a shutter. It is a reasonable assumption in progressive or interlaced video systems when scene motion is slow compared to the frame rate. The series of frames are no longer tilted in the spatiotemporal domain and can be "stacked" in a straightforward way to form a spatiotemporal volume (see Figure 8.5).

8.2   Some Fundamentals

Following are some notational conventions and basic principles used in the balance of this chapter. A continuous sequence is denoted as $u(x, y, t)$, $v(x, y, t)$, etc., where x, y are the continuous spatial variables and t is the continuous temporal variable. Similarly, a discrete sequence is denoted as $u(m, n, p)$, $v(m, n, p)$, etc., where m, n are the discrete (integer) spatial variables and p is the discrete (integer) temporal variable.

A 3-D System

As in 1-D and 2-D, a 3-D discrete system can be defined as

$$y(m, n, p) = \mathcal{H}[x(m, n, p)] \tag{8.1}$$

where $\mathcal{H}$ is the system function. In general, this function need be neither linear nor shift invariant. If the system is both linear and shift invariant (LSI), it can be characterized in terms of its impulse response $h(m, n, p)$. The linear shift invariant system response can then be written as

$$y(m, n, p) = \sum_{m'=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} \sum_{p'=-\infty}^{\infty} x(m', n', p')h(m - m', n - n', p - p')$$

$$\equiv x(m, n, p) * h(m, n, p) \tag{8.2}$$

where '*' denotes (discrete) convolution. Similarly, for the continuous case,

$$g(x, y, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x', y', t')h(x - x', y - y', t - t')dx'dy'dt' \qquad (8.3)$$

The 3-D Fourier Transform

The 3-D continuous Fourier transform can be expressed as

$$F(\xi_x, \xi_y, \xi_t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, t)e^{-j2\pi(x\xi_x + y\xi_y + t\xi_t)}dx\,dy\,dt \qquad (8.4)$$

where ξ_x, ξ_y, and ξ_t are the spatiotemporal frequency variables and $f(x, y, t)$ is a continuous spatiotemporal signal. As in the 2-D case, the 3-D Fourier transform is separable:

$$F(\xi_x, \xi_y, \xi_t) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x, y, t)e^{-j2\pi x\xi_x}dx \right] e^{-j2\pi y\xi_y}dy \right] e^{-j2\pi t\xi_t}dt \qquad (8.5)$$

Also as in the 1-D and 2-D cases, if

$$g(x, y, t) = h(x, y, t) * f(x, y, t) \qquad (8.6)$$

then

$$G(\xi_x, \xi_y, \xi_t) = H(\xi_x, \xi_y, \xi_t)F(\xi_x, \xi_y, \xi_t) \qquad (8.7)$$

If $h(x, y, t)$ is the LSI system impulse response, then $H(\xi_x, \xi_y, \xi_t)$ is the frequency response of the system.
   The spatiotemporal discrete Fourier transform is defined as

$$v(h, k, l) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \sum_{p=0}^{N-1} u(m, n, p)W_N^{hm}W_N^{kn}W_N^{lp} \qquad (8.8)$$

where $0 \le h,\ k,\ l \le N-1$ and $W_N = e^{-j2\pi/N}$.
   The inverse transform is

$$u(m, n, p) = \frac{1}{N^3} \sum_{h=0}^{N-1} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} v(h, k, l)W_N^{-hm}W_N^{-kn}W_N^{-lp} \qquad (8.9)$$

where $0 \le m,\ n,\ p \le N-1$.

Moving Images in the Frequency Domain

Following the discussion in Ref. [2], a moving monochrome image can be represented by an intensity distribution $f(x, y, t)$. The image is static if $f(x, y, t) = f(x, y, 0)$ for all t. The velocity of the image can be expressed via the image velocity vector

$$\vec{r} = (r_x, r_y) \qquad (8.10)$$

If the (initially static) image translates at a constant velocity $\vec{r}$, then

$$f_r(x, y, t) = f(x - r_x t, y - r_y t, t) \tag{8.11}$$

Consider the case of a simple 2-D "image" $f(x, t)$. Let

$$\vec{a} = \begin{pmatrix} x \\ t \end{pmatrix} \quad \text{and} \quad \vec{b} = \begin{pmatrix} \xi_x \\ \xi_t \end{pmatrix} \tag{8.12}$$

where ξ_x and ξ_t are the spatial and temporal frequency variables. Then the transform pair can be written as

$$f(\vec{a}) \xrightarrow{\mathcal{F}} (\vec{b}) \tag{8.13}$$

Now, translation can be represented as a coordinate transformation

$$\vec{a}' = \begin{pmatrix} x - r_x t \\ t \end{pmatrix} = \mathbf{A}\,\vec{a} \tag{8.14}$$

where

$$A = \begin{bmatrix} 1 & -r_x \\ 0 & 1 \end{bmatrix} \tag{8.15}$$

and r_x is the horizontal speed.

Using the expression for the Fourier transform after an affine coordinate transformation (any combination of scaling, rotation, and translation),

$$f\left(\vec{a}'\right) \xrightarrow{\mathcal{F}} [(A^{-1})^T \vec{b}] \tag{8.16}$$

where

$$(A^{-1})^T = \begin{bmatrix} 1 & 0 \\ r_x & 1 \end{bmatrix} \tag{8.17}$$

so that

$$f(x - r_x t, t) \xrightarrow{\mathcal{F}} (\xi_x, \xi_t + r_x \xi_x) \tag{8.18}$$

Example

Consider a simple static image with only two components (Figure 8.6). As the image undergoes translation with horizontal speed r_x, all temporal frequencies are shifted by $-r_x \xi_x$. Spatial frequency coordinates remain unchanged. That is, all frequency components of an image moving with velocity r_x lie on a line through the origin, with slope $-r_x$.

Extending the analysis to the 3-D case ($f(x, y, t)$), let the velocity $\vec{r} = (r_x, r_y)$. Then

$$f(x - r_x t, y - r_y t, t) \xrightarrow{\mathcal{F}} (\xi_x, \xi_y, \xi_t + r_x \xi_x + r_y \xi_y) \tag{8.19}$$
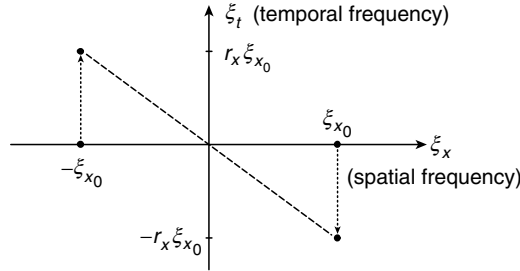
FIGURE 8.6 A two-component, 1-D signal in translational motion.
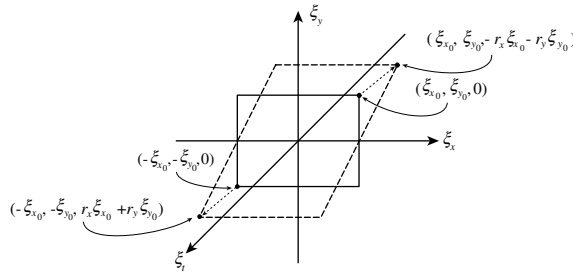
FIGURE 8.7 A two-component, 2-D signal in translational motion.

Each temporal frequency is shifted by the dot product of the spatial frequency vector $\vec{s} = (\xi_x, \xi_y)$ and the image velocity vector $\vec{r} = (r_x, r_y)$. If the image was originally static, then

$$\xi_t = -\vec{r} \cdot \vec{s} = -(r_x \xi_x + r_y \xi_y) \tag{8.20}$$

Geometrically, the image motion changes the static image transform (which lies in the (ξ_x, ξ_y) plane) into a spectrum in a plane with slope $-r_y$ in the (ξ_y, ξ_t) plane and $-r_x$ in the (ξ_x, ξ_t) plane. As in the 2-D case, the shifted points lie on a line through the origin. Note that this represents a relatively sparse occupation of the frequency domain (of interest for compression applications). A 3-D volume of data has been "compressed" into a plane. This compactness is not observed in the spatiotemporal domain.

In summary, the spectrum of a stationary image lies in the (ξ_x, ξ_y) plane. When the image undergoes translational motion, the spectrum occupies an oblique plane which passes through the origin. The orientation of the plane indicates the speed and direction of the motion. It is, therefore, possible to associate energy in particular regions of the frequency domain with particular image velocity components. By filtering specific regions in the frequency domain, these image velocity components can be detected. As will be seen shortly, other effects (such as the visual impact of temporal aliasing) can also be understood in the frequency domain.

3-D Sampling

In its simplest form (regular sampling on a rectangular grid, the method used here), 3-D sampling is a straightforward extension of 2-D (or 1-D) sampling (Figure 8.8). Given a bandlimited sequence

$$f(x, y, t) \xrightarrow{\mathcal{F}} (\xi_x, \xi_y, \xi_t) \tag{8.21}$$

with

$$F(\xi_x, \xi_y, \xi_t) = 0 \quad \text{whenever} \quad |\xi_x| > \xi_{x_0}, |\xi_y| > \xi_{y_0}, \quad \text{or} \quad |\xi_t| > \xi_{t_0} \tag{8.22}$$

the continuous sequence can be reconstructed from a discrete set of samples whenever

$$\xi_{x_s} > 2\xi_{x_0}, \quad \xi_{y_s} > 2\xi_{y_0}, \quad \text{and} \quad \xi_{t_s} > 2\xi_{t_0} \tag{8.23}$$

where $\xi_{x_s}, \xi_{y_s}, \xi_{z_s}$ are the sampling frequencies. Equivalently, the sequence can be reconstructed if the intervals
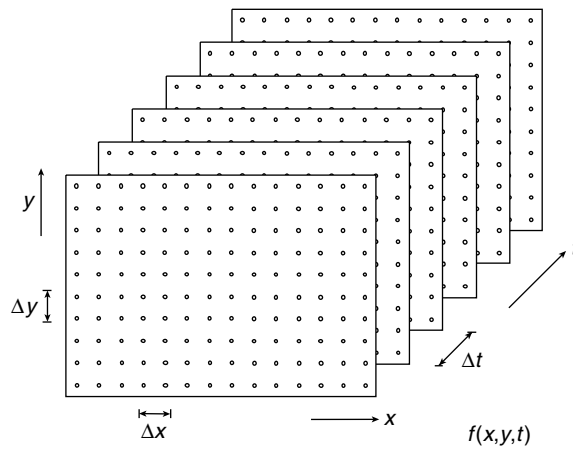
FIGURE 8.8 A sampled spatiotemporal signal (image sequence).
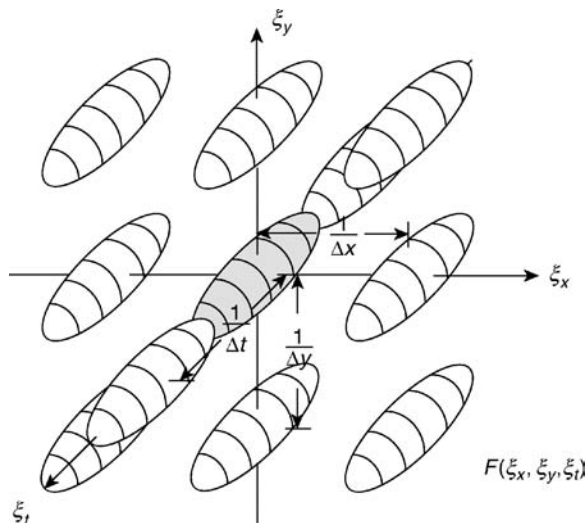
FIGURE 8.9 An image sequence with insufficiently high sampling in the temporal dimension.
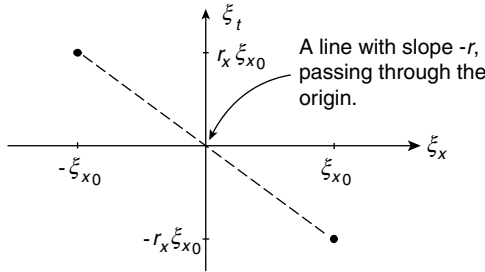
FIGURE 8.10 A continuous, two-component, 1-D signal in translational motion.
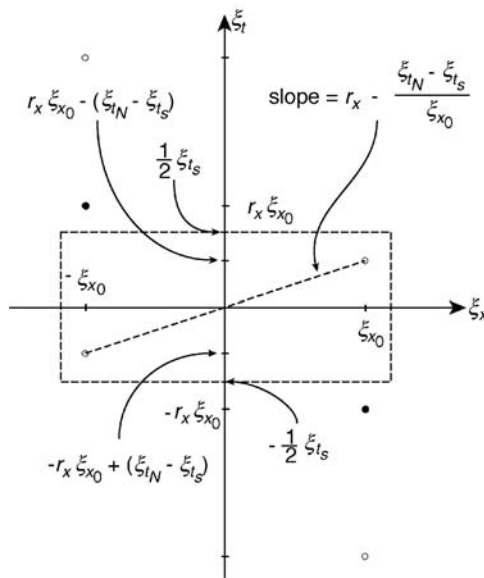
FIGURE 8.11 A reconstruction of a sampled 1-D signal with temporal aliasing.

between samples are such that

$$\Delta x < \frac{1}{2\xi_{x_0}}, \quad \Delta y < \frac{1}{2\xi_{y_0}}, \quad \text{and} \quad \Delta t < \frac{1}{2\xi_{t_0}} \tag{8.24}$$

If any of the sampling frequencies fall below the specified rates, the neighboring spectra (replications of the continuous spectrum, produced by the sampling process) overlap, and aliasing results. A case for which the temporal sampling frequency is too low is shown in Figure 8.9. The appearance of aliasing in the spatial domain, where it commonly manifests as a jagged approximation of smooth high contrast edges, is relatively familiar and intuitive. The effect of sampling at too low a rate temporally is perhaps less so.

Consider the earlier simple example of a 1-D image with only two components, moving with velocity r_x. The continuous case, as derived previously, is shown in Figure 8.10. ξ_{x_0} is the frequency of the static image. Suppose this image is sampled along the temporal dimension at a sampling frequency ξ_{t_s} less than the Nyquist rate ($\xi_{t_N} = 2r_x\xi_{x_0}$), and the image is reconstructed via an ideal lowpass filter with temporal cutoff frequencies at plus and minus half the sampling frequency (Figure 8.11). What is the visual effect of the aliased components?

As seen previously, the velocity of motion is reflected in the slope of the line connecting the components. For the situation shown, a sinusoidal grid (of the same frequency as the original) moving in the *opposite direction*, with speed $r_x \xi_{x_0} - (\xi_{t_N} - \xi_{t_s})$ is observed. As the sampling frequency drops, the velocity decreases, eventually reaching zero. Continued reduction in ξ_{t_s} results in motion in the *same direction* as the original image, increasing in velocity until (at $\xi_{t_s} = 0$) the velocities of the two components are identical.

In the simple example just considered, the image was spatially homogeneous, so that the effects of aliasing were seen throughout the image. In general, this is not the case. As in the 1-D and 2-D cases, the temporal aliasing effect is seen in regions of the sequence with sufficiently high temporal frequency components to alias. Circumstances leading to high temporal frequencies include high velocity (large values of r_x in our simple example) and high spatial frequency components with some degree of motion (high ξ_{x_0} in our example). Higher spatial frequency components require slower speeds to cause aliasing.

A well-known example of temporal aliasing is the so-called "wagon wheel" effect, in which the wheels of a vehicle appear to move in a direction opposite to that of the vehicle itself. The wheels have both high spatial frequency components (due to their spokes) and relatively high rotational velocity. Hence, aliasing occurs (the wheels appear to rotate in reverse). The vehicle itself, however, which is moving more slowly and is also generally composed of lower spatial frequency components, moves forward (does not exhibit aliasing effects).

8.3   The Perception of Visual Motion

Visual perception can be discussed at a number of different levels: the anatomy or physical structure of the visual system; the physiology or basic function of the cells involved; and the psychophysical behavior of the system (the response of the system to various stimuli). Following is a brief discussion of visual motion perception. A more extensive treatment can be found in Ref. [3].

Anatomy and Physiology of Motion Perception

The retina (the hemispherical surface at the back of the eye) is the sensor surface of the visual system, consisting of two major types of sensor elements. The rods are long and thin structures, numbering approximately 120 million. They provide scotopic ("low-light") vision and are highly sensitive to motion. The cones are shorter and thicker, and substantially fewer in number (approximately 6 million per retina). They are less sensitive than the rods, providing photopic ("high-light") and color vision. The cones are much less sensitive to motion.

The rods and cones are arranged in a roughly hexagonal array. However, they are not uniformly distributed over the retina. The cones are packed in the fovea (hence color vision is primarily foveal). The rods are primarily outside the fovea. As a result, motion sensitivity is higher outside the fovea, corresponding to the periphery of the visual field.

Visual information leaves each eye via the optic nerve. The nerves from each eye split at the optic chiasma, pass through the lateral geniculate nucleus, and continue to the visual cortex. Information is retinotopically mapped on the cortex (organized as in the original scene, but reversed). Note, however, that the mapping is not one-to-one (one retinal rod or cone to one cortical cell). As mentioned previously, approximately 120 million rods and 6 million cones are found in each eye, but only 1 million fibers in the associated optic nerve. This 126:1, apparently visually lossless compression, is one of the motivations for studying perceptually inspired image and video compression techniques, as discussed later in this chapter.

To achieve this compression, each cortical cell receives information from a set of rods and/or cones. This set makes up the *receptive field* for that cell. The response of a cortical cell to stimuli at different points in this field can be measured (e.g., via a moving spot of light) and plotted just as one might plot the impulse response of a 2-D filter.

Physiologically, nothing mentioned so far seems specifically adapted to the detection (or measurement) of motion. It might be reasonable to expect to find cells, which respond selectively to, e.g., the direction of motion. There appear to be no such cells in the human retina (although other species do have retinal
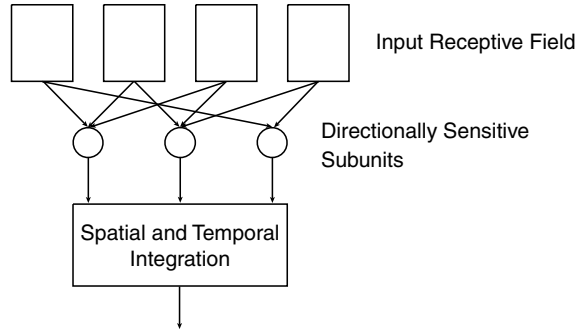
FIGURE 8.12   A common organizational structure for modeling complex cell behavior.
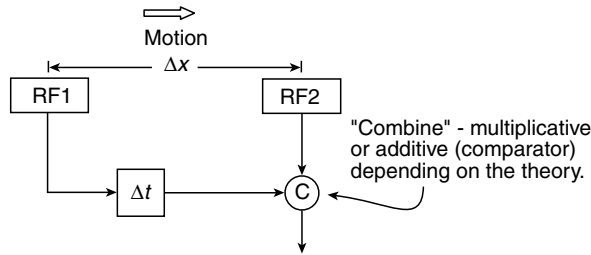
FIGURE 8.13   A mechanism for the directionally sensitive detection of motion.

cells that respond in this way); however, cells in the mammalian striate cortex exhibit this behavior (the complex cells).

How these cells come to act this way remains under study. However, most current theories fit a common organizational structure [4], shown in Figure 8.12. The input receptive fields are sensitive both to the spatial location and spatial frequency of the stimulus. The role, if any, of orientation is not widely agreed upon. The receptive field outputs are combined, most likely in a nonlinear fashion, in the directionally sensitive subunits to produce an output highly dependent on the direction and/or velocity of the stimulus. The output of these subunits are then integrated both spatially and temporally.

Consider the hypothetical directionally sensitive mechanism in more detail for the case of rightward moving patterns (Figure 8.13). For example, suppose the receptive fields are symmetric, and C is a comparator which requires both inputs to be high to output a high value. If a pattern, which stimulates receptive field 1 (RF1), moves a distance Δx in time Δt so that it falls within receptive field 2 (RF2), then the comparator will "fire."

Although it is simple, such a model establishes a basic link between moving patterns on the retina and the perception of motion. Additional insight can be obtained by considering the problem from a systems perspective.

The Psychophysics of Motion Perception

Spatial Frequency Response

In the case of spatial vision, much can be understood by modeling the visual system as shown in Figure 8.14. The characteristics of the filter $H(\xi_x, \xi_y)$ have been estimated by determining the threshold visibility of sine wave gratings. The resulting measurements indicate visual sensitivity as a function of spatial frequency that is
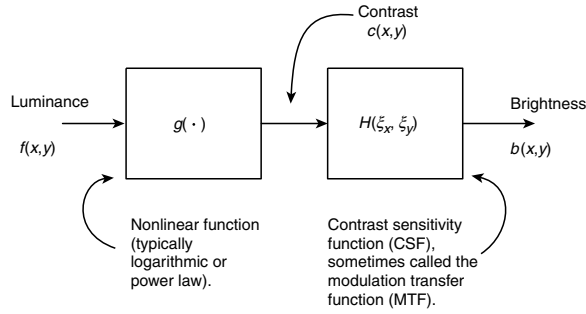
FIGURE 8.14   A simple block diagram modeling spatial vision.

approximately lowpass in nature. The response peaks in the vicinity of 5 cycles/degree, and falls off rapidly beyond 10 cycles/degree.

If it were separable (that is, $H(\xi_x, \xi_y)$ could be determined by finding $H(\xi_x)$ and $H(\xi_y)$ independently), with $H(\xi_x) = H(\xi_y)$, or isotropic, the spatial response could be characterized via a single 1-D function. Although the assumption of separability is often useful, the spatial CSF of the human visual system is not, in fact, separable. It has been shown that visual sensitivity is reduced at orientations other than vertical and horizontal. This may be due to the predominance of vertical and horizontal structures in the visual environment, leading to the development or evolution of the visual system to be particularly sensitive at (or conversely, less sensitive away from) these orientations. This is referred to as the "oblique effect."

Temporal Frequency Response

The most straightforward approach to extending the above spatial vision model to include motion is to modify the CSF to include temporal frequency sensitivity, so that $H(\xi_x, \xi_y)$ becomes $H(\xi_x, \xi_y, \xi_t)$.

One way to estimate the temporal frequency response of the visual system is to measure the flicker response. Although the flicker response varies with intensity and with the spatial frequency of the stimulus, it is again generally lowpass, with a peak in response in the vicinity of 10 Hz. The attenuation of the response above 10 Hz increases rapidly, so that at 60 Hz (the field rate of NTSC television) the flicker response is very low.

It is natural, as in the 2-D case, to ask whether the spatiotemporal frequency response $H(\xi_x, \xi_y, \xi_t)$ is separable with respect to the temporal frequency. There is evidence to believe that this is not the case. The flicker response curves for high and low spatial frequency patterns do not appear consistent with a separable spatiotemporal response.

Reconstruction Error

To a first approximation, the data discussed above indicate that the HVS behaves as a 3-D lowpass filter, with bandlimits (for bright displays) at 60 cycles/degree along the spatial frequency axes, and 70 Hz temporally. This approximation is useful in understanding errors, which may occur in reconstructing a continuous spatiotemporal signal from a sampled one. Consider the case of an image undergoing simple translational motion. This spatiotemporal signal occupies an oblique plane in the frequency domain. With sampling, the spectrum is replicated (with periods determined by the sampling frequencies along the respective dimensions) to fill the infinite 3-D volume. The spectrum of a sufficiently sampled (aliasing free) image sequence produced in this way is shown in Figure 8.15.

The 3-D lowpass reconstruction filter (the spatiotemporal CSF) can be approximated as an ideal lowpass filter, as shown in Figure 8.16. As long as the cube in Figure 8.16 completely encloses the spectrum centered at DC, without including neighboring spectra, there is no reconstruction error. This case included no aliasing. If aliasing is included (the sample rate during acquisition is too low), the aliased components will be visible only if they fall within the passband of the CSF filter.
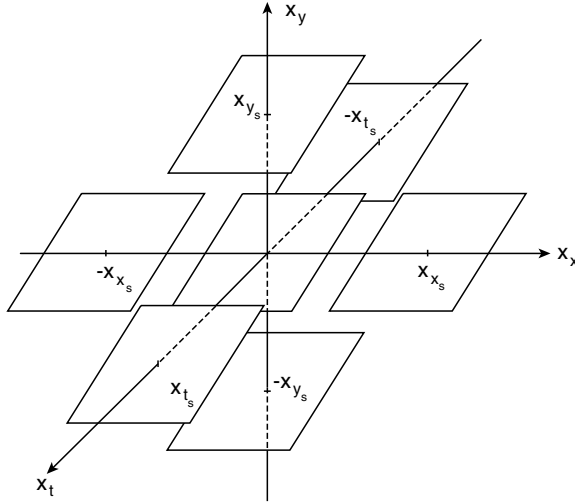
FIGURE 8.15 The spectrum of a sampled image undergoing uniform translational motion.
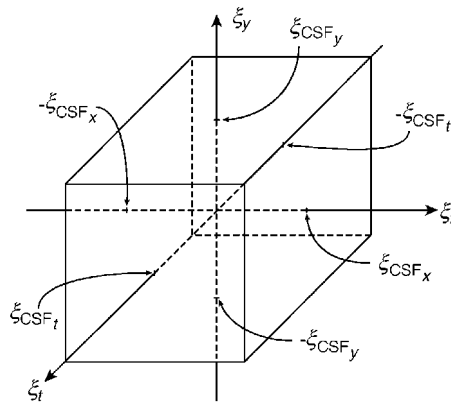
FIGURE 8.16 An ideal 3-D, lowpass reconstruction filter, with cutoff frequencies determined by the spatiotemporal contrast sensitivity funtion.

The above frequency domain analysis explains some important aspects of human visual motion perception. Other observations are not as easily explained in this way, however. As observed in Ref. [5], perceived motion is local (different motions can be seen in different areas of the visual field) and spatial-frequency specific (individual motion sensors respond differently (selectively) to different spatial frequencies). These two observations suggest an underlying representation that is local in both the spatiotemporal and spatiotemporal-frequency domains. Examples of such representations will be discussed in the following subsection.

The Effects of Eye Motion

The analysis of motion perception described previously assumed a "passive" view. That is, any change in the pattern of light on the retinal surface is due to motion in the scene. That this is not the case can be seen by

considering the manner in which static images are viewed. They are not viewed as a whole, but in a series of "jumps" from position to position. These "jumps" are referred to as *saccades* (French for "jolt" or "jerk").

Even at the positions where the eye is "at rest" it is not truly static. It undergoes very small motions (microsaccades) of 1–2 min of arc. In fact, the eye is essentially never at rest. It has been shown that if the eye is stabilized, vision fades away after about a second. The relevance of this to the current discussion is that although the eye is in constant motion, so that the intensity patterns on the retina are constantly changing, when viewing a static scene no motion is perceived. Similar behavior is observed when viewing dynamic scenes [6]. Obviously, however, in the case of dynamic scenes motion is often perceived (even though the changes in intensity patterns on the retina are not necessarily greater than for static images).

Two hypotheses might explain these phenomena. The first is that the saccades are so fast that they are not sensed by the visual system; however, this does not account for the fact that motion is seen in dynamic scenes, but not static ones. The second is that the motion sensing system is "turned off" under some circumstances (the theory of corollary discharge). The basic idea is that the motor signals that control eye movement are also involved in the perception of motion, so that when intensity patterns on the retina change and there is a motor signal present, no motion is perceived. When intensity patterns change but there is no motor signal, or if there is no change in intensity patterns but there is a motor signal, motion is perceived. The latter situation corresponds to the tracking of moving objects (smooth pursuit). The first hypothesis (the less plausible of the two) can be easily modeled with temporal linear filters. The second, more interesting behavior can be modeled with a simple comparator network.

8.4 Image Sequence Representation

What Does "Representation" Mean?

The term "representation" may require some explanation. Perhaps the best way to do so is to consider some examples of familiar representations. For simplicity, 2-D examples will be used. Extension to 3-D is relatively straightforward.

The Pixel Representation

The pixel representation is so common and intuitive that it is usually considered to be "the image." More precisely, however, it is a linear sum of weighted impulses:

$$u(m, n) = \sum_{m'=0}^{N-1} \sum_{n'=0}^{N-1} u(m', n')\delta(m - m', n - n') \tag{8.25}$$

where $u(m, n)$ is the image, $u(m', n')$ are the coefficients of the representation (numerically equal to the pixel values in this case), and the $\delta(m - m', n - n')$ play the role of basis functions.

The DFT

The next most familiar representations (at least to engineers) is the DFT, in which the image is expressed in terms of complex exponentials:

$$u(m, n) = \frac{1}{N^2} \sum_{h=0}^{N-1} \sum_{k=0}^{N-1} v(h, k) W_N^{-hm} W_N^{-kn} \tag{8.26}$$

where $0 \leq m,\ n \leq N-1$ and

$$W_N = e^{-j2\pi/N} \tag{8.27}$$

In this case $v(h, k)$ are the coefficients of the representation and the 2-D complex exponentials $W_N^{-hm} W_N^{-kn}$ are the basis functions.

The choice of one representation over the other (pixel vs. Fourier) for a given application depends on the image characteristics that are of most interest. The pixel representation makes the spatial organization of intensities in the image explicit. Because this is the basis of the visual stimulus, it seems more "natural." The Fourier representation makes the composition of the image in terms of complex exponentials ("frequency components") explicit. The two representations emphasize their respective characteristics (spatial vs. frequency) to the exclusion of all others. If a mixture of characteristics is desired, different representations must be used.

Spatial/Spatial-Frequency Representations

A natural mixture is to combine frequency analysis with spatial location. An example of a 1-D representation of this type (a time/frequency representation) is a musical score. The need to know not only what the frequency content of a signal is, but where in the signal the frequency components exist is common to many signal, image, and image sequence processing tasks [7]. A variety of approaches [8,9] can be used to develop a representation to facilitate these tasks. The most intuitive approach is the finite-support Fourier transform.

The Finite-Support Fourier Transform

This approach to local frequency decomposition has been used for many years for the analysis of time varying signals. In the 2-D continuous case,

$$F_{x,y}(\xi_x, \xi_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{x,y}(x', y') e^{-j2\pi(\xi_x x' + \xi_y y')} dx' dy' \tag{8.28}$$

where

$$f_{x,y}(x', y') = f(x', y') h(x - x', y - y') \tag{8.29}$$

$f(x', y')$ is the original image, and $h(x - x', y - y')$ is a window centered at (x, y).

The properties of the transform depend a great deal on the properties of the window function. Under certain circumstances (i.e., for certain windows) the transform is invertible. The most obvious case is for nonoverlapping (e.g., rectangular) windows.

The windowed transform idea can, of course, be applied to other transforms, as well. An example that is of substantial practical interest is the discrete cosine transform, with a rectangular nonoverlapping window:

$$F(h, k) = \alpha(h)\alpha(k) \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} f(m, n) \cos\left(\frac{(2m+1)h\pi}{2N}\right) \cos\left(\frac{(2n+1)k\pi}{2N}\right) \tag{8.30}$$

where $h, k = 0, 1, \ldots, N-1$,

$$\alpha(h) = \begin{cases} \sqrt{\dfrac{1}{N}} & \text{for} \quad h = 0 \\[2em] \sqrt{\dfrac{2}{N}} & \text{otherwise} \end{cases} \tag{8.31}$$

$\alpha(k)$ is defined similarly, and the window dimensions are $N \times N$. This transform is the basis for the well known JPEG and MPEG compression algorithms.

The Gabor Representation

This representation was first proposed for 1-D signal analysis by Dennis Gabor in 1946 [10]. In 2-D [11], an image can be represented as the weighted sum of functions of the form

$$g(x, y) = \hat{g}(x, y)e^{j2\pi[\xi_{x_0}(x-x_0)+\xi_{y_0}(y-y_0)]} \tag{8.32}$$

where

$$\hat{g}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2 + \left(\frac{y-y_0}{\sigma_y}\right)^2\right]} \tag{8.33}$$

is a 2-D Gaussian function, σ_x and σ_y determine the extent of the Gaussian along the respective axes, (x_0, y_0) is the center of the function in the spatial domain, and (ξ_{x_0}, ξ_{x_0}) is the center of support in the frequency domain. A representative example of a Gabor function is shown in Figure 8.17.

Denoting the distance between spatial centers as D and the distance between their centers of support in the frequency domain as *W*, the basis is complete if $WD = 2\pi$. These functions have a number of interesting aspects. They achieve the lower limits of the Heisenburg uncertainty inequalities:

$$\Delta x \Delta \xi_x \geqslant \frac{1}{4\pi}, \quad \Delta y \Delta \xi_y \geqslant \frac{1}{4\pi} \tag{8.34}$$

where Δ_x, Δ_y, $\Delta\xi_x$, and $\Delta\xi_y$ are the effective widths of the functions in the spatial and spatial-frequency domains. By this measure, then, these functions are optimally local. Their real and imaginary parts also agree reasonably well with measured receptive field profiles. The basis is not orthogonal, however. Specifically, the Gabor transform is not equivalent to the finite-support Fourier transform with a Gaussian window. For a cross-section of the state of the art in Gabor transform-based analysis, see [12].

The Derivative of Gaussian Transform

In 1987, Young [13] proposed a receptive field model based on the Gaussian and its derivatives. These functions, like the Gabor functions, are spatially and spectrally local and consist of alternating regions of
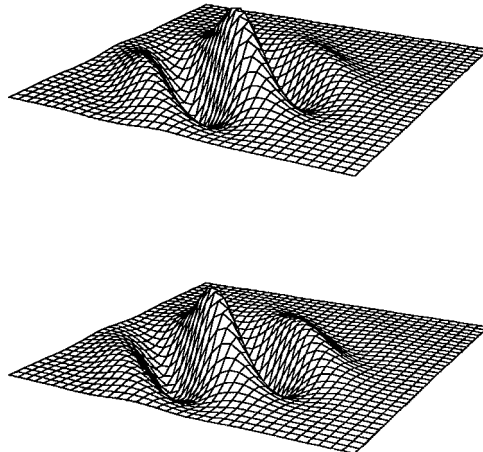
FIGURE 8.17   The real (top) and imaginary (bottom) parts of a representative 2-D Gabor function.

excitation and inhibition in a decaying envelope. Young showed that Gaussian derivative functions more accurately model the measured receptive field data than do the Gabor functions [14].

In Ref. [15], a spatial/spatial-frequency representation based on shifted versions of the Gaussian and its derivatives was introduced (the derivative of Gaussian transform (DGT)). As with the Gabor transform, although this transform is nonorthogonal, with a suitably chosen basis it is invertible. The DGT has significant practical advantage over the Gabor transform in that both the basis functions and coefficients of expansion are real-valued.

The family of 2-D separable Gaussian derivatives centered at the origin can be defined as

$$
\begin{aligned}
g_{0,0}(x, y) &= g_0(x)g_0(y) \\
&= e^{-(x^2 + y^2)/2\sigma^2}
\end{aligned}
\tag{8.35}
$$

$$
\begin{aligned}
g_{m,n}(x, y) &= g_m(x)g_n(y) \\
&= \frac{d^{(m)}}{dx^{(m)}}g_0(x)\frac{d^{(n)}}{dy^{(n)}}g_0(y)
\end{aligned}
\tag{8.36}
$$

This set can then be shifted to any desired location. The variance σ defines the extent of the functions in the spatial domain. There is an inverse relationship between the spatial and spectral extents, and the value of this variable may be constant or may vary with context.

The 1-D Gaussian derivative function spectra are bimodal (except for that of the original Gaussian, which is itself a Gaussian) with modes centered at $\pm\Omega_m$ rad/pixel:

$$
\Omega_m = \frac{\sqrt{m}}{\sigma}
\tag{8.37}
$$

where m is the derivative order. The order of derivative necessary to center a mode at a particular frequency is therefore

$$
m = (\Omega_m\sigma)^2
\tag{8.38}
$$

The Wigner Distribution

The previous examples indicate that a local frequency representation need not have an orthogonal basis. In fact, it need not even be linear. The Wigner distribution was introduced by Eugene Wigner in 1932 [16] for use in quantum mechanics (in 1-D). In 2-D, the Wigner distribution can be written as

$$
W_f(x, y, \xi_x, \xi_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f\left(x + \frac{\alpha}{2}, y + \frac{\beta}{2}\right)f^*\left(x - \frac{\alpha}{2}, y - \frac{\beta}{2}\right)e^{-j2\pi(\alpha\xi_x + \beta\xi_y)}d\alpha d\beta
\tag{8.39}
$$

where the asterisk denotes complex conjugation. The Wigner distribution is real valued, so does not have an explicit phase component (as seen in, e.g., the Fourier transform). A number of discrete approximations to this distribution (sometimes referred to as pseudo-Wigner distributions) have also been formulated.

Spatial/Scale Representations (Wavelets)

Scale is a concept that has proven very powerful in many applications, and may under some circumstances be considered as fundamental as frequency. Given a set of (1-D) functions

$$
W_{jk}(x) = W(2^j x - k)
\tag{8.40}
$$

where the indices j and k correspond to dilation (change in scale) and translation, respectively, a signal decomposition

$$f(x) = \sum_j \sum_k b_{jk} W_{jk}(x) \tag{8.41}$$

emphasizes the scale (or resolution) characteristics of the signal (specified by j) at specific points along x (specified by k), yielding a multiresolution description of the signal.

A class of functions $W_{jk}(x)$ that have proven extremely useful are referred to as wavelets. A detailed discussion of wavelets is beyond the scope of this chapter (see [17–19] for excellent treatments of this topic); however, an important aspect of any representation (including wavelets) is the resolution of the representation, and how it can be measured.

Resolution

In dealing with joint representations, resolution is a very important issue. It arises in a number of ways. In discussing the Gabor representation, it was noted that the functions minimized the uncertainty inequalities, e.g.,

$$\Delta x \Delta \xi_x \geq \frac{1}{4\pi} \tag{8.42}$$

Note that it is the product that is minimized. Arbitrarily high resolution cannot be achieved in both domains simultaneously, but can be traded between the two domains at will. The proper balance depends on the application. It should be noted that the "effective width" measures Δx, $\Delta \xi_x$, etc. (normalized second moment measures) are not the only way to define resolution. For example, the degree of energy concentration could be used (leading to a different "optimal" set of functions, the prolate spheroidal functions). The appropriateness of the various measures again depends on the application. Their biological (psychophysical) relevance remains to be determined.

All the previously mentioned points are relevant for both spatial/spatial-frequency and spatial/scale representations (wavelets). Wavelets, however, present some special considerations. Suppose one wishes to compare the resolutions of time/frequency and wavelet decompositions? Specifically, what is the resolution of a multiresolution method? This question can be illustrated by considering the 1-D case, and examining the behavior of the two methods in the time-frequency plane (Figure 8.18).

In the time/frequency representation, the dimensions Δt and $\Delta \xi_t$ remain the same throughout the time-frequency plane. In wavelet representations the dimensions vary, but their product remains constant. The resolution characteristics of wavelets may lead one to believe that the uncertainty of a wavelet decomposition may fall below the bound in Equation (8.42). This is not the case. The tradeoff between Δt and $\Delta \xi_t$ simply varies. The fundamental limit remains.
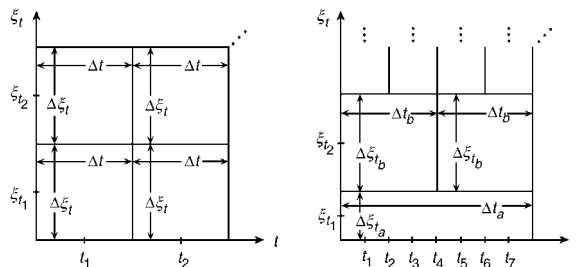
FIGURE 8.18 The resolution of a time/frequency representation and a wavelet representation in the time-frequency plane.

A final point relates more specifically to the representation of image sequences. The HVS has a specific (bandlimited) spatiotemporal frequency response. Beyond indicating the maximum perceivable frequencies (setting an upper bound on resolution) it seems feasible to exploit this point further, to achieve a more efficient representation. Recalling the relationship between motion and temporal frequency, a surface with high spatial frequency components, moving quickly, has high temporal frequency components. When it is static, it does not. The characteristics of the spatiotemporal CSF may lead us to the conclusions that static regions of an image require little temporal resolution, but high spatial resolution, and that regions in an image undergoing significant motion require less spatial resolution (due to the lowered sensitivity of the CSF), but require high temporal resolution (for smooth motion rendition).

The first conclusion is essentially correct (although not trivial to exploit). The second conclusion, however, neglects eye tracking. If the eye is tracking a moving object, the spatiotemporal frequency characteristics experienced by the viewer are very similar to those in the static case, i.e., visual sensitivity to spatial structure is not reduced significantly.

8.5   The Computation of Motion

Many approaches are used for the computation of motion (or, more precisely, the estimation of motion based on image data). Before examining some of these approaches in more detail, it is worthwhile to review the relationship between the motion in a scene and the changes observed in an image of the scene.

The Motion Field

The motion field [20] is determined by establishing a correspondence between the motion of points in the scene (the real world) and the motion of points in the image plane. This correspondence is found geometrically, and is independent of the brightness patterns in the scene (e.g., the presence or absence of surface textures, changes in luminance, etc.).

Consider the situation in Figure 8.19. At a particular instant in time, a point P_{image} in the image corresponds to some point P_{object} on the surface of an object. The two points are related via the perspective projection equation. Now, suppose the object point P_{object} has velocity (v_x, v_y, v_z) relative to the camera. The result is a velocity (v'_x, v'_y) for the point P_{image} in the image plane. The relationship between the velocities can be found by differentiating the perspective projection equation with respect to time. In this way, a velocity vector can be assigned to each image point, yielding the motion field.

Optical Flow

Usually, the intensity patterns in the image move as the objects to which they correspond move. Optical flow is the motion of these intensity patterns. Ideally, optical flow and the motion field correspond; but this is not
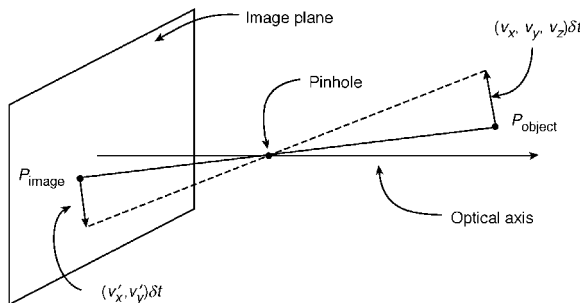
FIGURE 8.19   The motion field based on a simple pinhole camera model.

always the case. For a perfectly uniform sphere rotating in front of an imaging system, there is shading over the surface of the sphere (due to the shape of the sphere), but it does not change with time. The optical flow is zero everywhere, while the motion field is not. For a fixed sphere illuminated by a moving light source, the shading changes with time, although the sphere is not in motion. The optical flow is nonzero, while the motion field is zero.

Furthermore, optical flow is not uniquely determined by local information in the changing image. Consider, for example, a region with uniform brightness which does not vary with time. The "most likely" optical flow value is zero, but (as long as there are corresponding points of equal brightness in both images) there are many "correct" flow vectors. What we would like is the motion field, but what we have access to is optical flow. Fortunately, the optical flow is usually not too different from the motion field.

The Calculation of Optical Flow

Wide variety of approaches are used for the calculation of optical flow. The first, below, is a conceptually simple yet very widely used method. This approach is particularly popular for video compression, and is essentially that used in MPEG-1 and 2.

Optical Flow by Block Matching

The calculation of optical flow by block-matching is the most commonly used motion estimation technique. The basic approach is as follows. Given two successive images from a sequence, the first image is partitioned into nonoverlapping blocks (e.g., 8×8 pixels in size, Figure 8.20(left)). To find the motion vector for each block, the similarity (e.g., via mean-squared error) between the block and the intensities in the neighborhood of that block in the next frame (Figure 8.20(right)) is calculated. The location that shows the best match is considered the location to which the block has moved. The motion vector for the block is the vector connecting the center of the block in frame n to the location of the best match in frame $n+1$.

The approach is simple, but a number of things must be consider. The size of the search neighborhood must be established, which in turn determines the maximum velocity that can be estimated. The search strategy must be decided, including the need to evaluate every potential match location and the precision with which the match locations must be determined (e.g., is each pixel a potential location? Is subpixel accuracy required?). The amount of computation time/power available is a critical factor in these decisions. Even at its simplest, block matching is computationally intensive. If motion estimates must be computed at frame rate (1/30 s) this will have a strong effect on the algorithm design. A detailed discussion of these and related issues can be found in Ref. [21].

Optical Flow via Intensity Gradients

The calculation of optical flow via intensity gradients, as proposed by Horn and Shunck [22], is a classical approach to motion estimation.

Search neighborhood
for the block in the
previous frame

Frame *n*   Frame *n* + 1
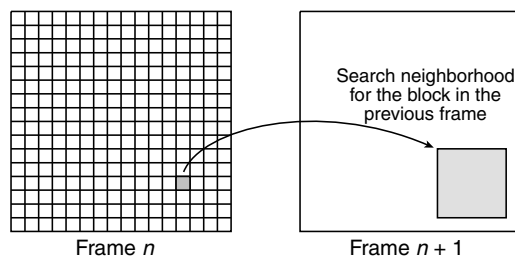
FIGURE 8.20   Motion estimation by block matching.

Let $f(x, y, t)$ be the intensity at time t for the image point (x, y), and let $r_x(x, y)$ and $r_y(x, y)$ be the x and y components of the optical flow at that point. Then for a small time interval δt,

$$f(x + \underbrace{r_x\delta t}_{\delta x}, y + \underbrace{r_y\delta t}_{\delta y}, t + \delta t) = f(x, y, t) \tag{8.43}$$

This single equation is not sufficient to determine r_x and r_y. It can, however, provide a constraint on the solution. Assuming that intensity varies smoothly with x, y, and t, the left hand side of Equation (8.43) can be expanded in a Taylor's series:

$$f(x, y, t) + \delta x \frac{\partial f}{\partial x} + \delta y \frac{\partial f}{\partial y} + \delta t \frac{\partial f}{\partial t} + \text{higher-order terms} = f(x, y, t) \tag{8.44}$$

Ignoring the higher order terms, canceling $f(x, y, t)$, dividing by δt and letting $\delta t \rightarrow 0$,

$$\frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial t} = 0 \tag{8.45}$$

or

$$f_x r_x + f_y r_y + f_t = 0 \tag{8.46}$$

where f_x, f_y, and f_t are estimated from the image sequence.

This equation is called the *optical flow constraint equation*, since it constrains r_x and r_y of the optical flow. The values of (r_x, r_y) which satisfy the constraint equation lie on a straight line in the (r_x, r_y) plane. A local brightness measurement can identify the constraint line, *but not a specific point on the line.* Note that this problem cannot really be *solved* via, e.g., adding an additional constraint. It is a fundamental aspect of the image data. A "true" solution cannot be guaranteed, but a solution can be found.

To view this limitation in another way, the constraint equation can be rewritten in vector form, as

$$(f_x, f_y) \cdot (r_x, r_y) = -f_t \tag{8.47}$$

so that the component of optical flow in the direction of the intensity gradient $(f_x, f_y)^{\text{T}}$ is

$$\frac{f_1}{\sqrt{f_x^2 + f_y^2}} \tag{8.48}$$

However, the component of the optical flow perpendicular to the gradient (along isointensity contours) cannot be determined. This is a manifestation of the *aperture problem*. If the motion of an oriented element is detected by a unit that is small compared with the size of the moving element, the only information that can be extracted is the component of motion perpendicular to the local orientation of the element. For example, looking at a moving edge through a small aperture (Figure 8.21), it is impossible to tell whether the actual motion is in the direction of *a* or of *b*.

One way to work around this limitation is to impose an explicit smoothness constraint. Motion was implicitly assumed smooth earlier, when a Taylor's expansion was used and when the higher order terms were ignored. Following this approach, an iterative scheme for finding the optical flow for the image sequence
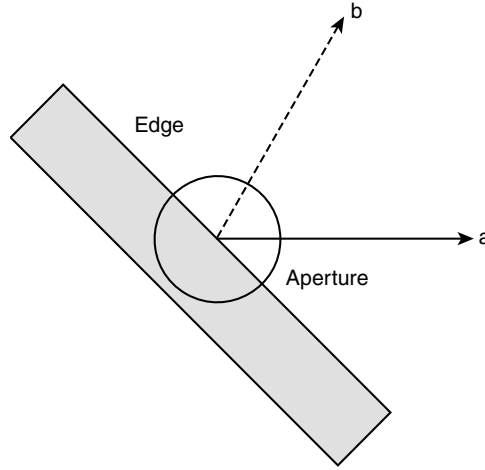
FIGURE 8.21   An instance of the aperture problem.

can be formulated:

$$r_x(k, l)^{n+1} = \overline{r_x(k, l)}^{\,n} - \frac{\lambda f_x^2 \overline{r_x(k, l)}^{\,n} + \lambda f_x f_y \overline{r_y(k, l)}^{\,n} + \lambda f_x f_t}{1 + \lambda(f_x^2 + f_y^2)}$$

$$= \overline{r_x(k, l)}^{\,n} - \lambda f_x \frac{f_x \overline{r_x(k, l)}^{\,n} + f_y \overline{r_y(k, l)}^{\,n} f_t}{1 + \lambda(f_x^2 + f_y^2)} \tag{8.49}$$

and

$$r_y(k, l)^{n+1} = \overline{r_y(k, l)}^{\,n} - \lambda f_y \frac{f_x \overline{r_x(k, l)}^{\,n} + f_y \overline{r_y(k, l)}^{\,n} f_t}{1 + \lambda(f_x^2 + f_y^2)} \tag{8.50}$$

where the superscripts n and $n+1$ indicate the iteration number, λ is a parameter allowing a tradeoff between smoothness and errors in the flow constraint equation, and $\overline{r_x(k, l)}$ and $\overline{r_y(k, l)}$ are local averages of r_x and r_y. The updated estimates are thus the average of the surrounding values, minus an adjustment (which in velocity space is in the direction of the intensity gradient).

The previous discussion relied heavily on smoothness of the flow field. However, there are places in image sequences where discontinuities *should* occur. In particular, the boundaries of moving objects should exhibit discontinuities in optical flow. One approach taking advantage of smoothness but allowing discontinuities is to apply segmentation to the flow field. In this way, the boundaries between regions with smooth optical flow can be found, and the algorithm can be prevented from smoothing over these boundaries. Because of the "chicken-and-egg" nature of this method (a good segmentation depends on a good optical flow estimate, which depends on a good segmentation …), it is best applied iteratively.

Spatiotemporal-Frequency-Based Methods

It was shown in section "Some Fundamentals" that motion can be considered in the frequency domain, as well as in the spatial domain. A number of motion estimation methods have been developed with this in mind. If the sequence to be analyzed is very simple (has only a single motion component, for example) or if motion detection alone is required, the Fourier transform can be used as the basis for motion analysis, as examined

in Refs. [23–25]; however, due to the global nature of the Fourier transform, it cannot be used to determine the location of the object in motion. It is also poorly suited for cases in which multiple motions exist (i.e., when the scene of interest consists of more than one object moving independently), since the signatures of the different motions are difficult (impossible, in general) to separate in the Fourier domain. As a result, although Fourier analysis can be used to illustrate some interesting phenomena, it cannot be used as the basis of motion analysis methods for the majority of sequences of practical interest.

To identify the locations and motions of objects, frequency analysis localized to the neighborhoods of the objects is required. Windowed Fourier analysis has been proposed for such cases [26], but the accuracy of a motion analysis method of this type is highly dependent on the resolution of the underlying transform, in both the spatiotemporal and spatiotemporal-frequency domains. It is known that the windowed Fourier transform does not perform particularly well in this regard. Filterbank-based approaches to this problem have also been proposed, as in Ref. [27]. The methods examined below each exploit the frequency domain characteristics of motion, and provide spatiotemporally localized motion estimates.

Optical Flow via the 3-D Wigner Distribution. Jacobson and Wechsler [28] proposed an approach to spatiotemporal-frequency, based derivation of optical flow using the 3-D Wigner distribution (WD). Extending the 2-D definition given earlier, the 3-D WD can be written as

$$W_f(x, y, t, \xi_x, \xi_y, \xi_t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f\left(x + \frac{\alpha}{2}, y + \frac{\beta}{2}, t + \frac{\tau}{2}\right) \cdot f^*\left(x - \frac{\alpha}{2}, y - \frac{\beta}{2}, t - \frac{\tau}{2}\right)$$

$$\times e^{-j2\pi(\alpha\xi_x + \beta\xi_y + \tau\xi_t)} d\alpha d\beta d\tau \tag{8.51}$$

It can be shown that the WD of a linearly translating image with velocity $\vec{r} = (r_x, r_y)$ is

$$W_f(x, y, t, \xi_x, \xi_y, \xi_t) = \delta(r_x\xi_x + r_y\xi_y + \xi_t) \cdot W_f(x - r_x t, y - r_y t, \xi_x, \xi_y) \tag{8.52}$$

which is nonzero only when $r_x\xi_x + r_y\xi_y + \xi_t = 0$.

For a linearly translating image, then, the local spectra $W_{f_{x,y,t}}(\xi_x, \xi_y, \xi_t)$ contain energy only in a plane (as in the Fourier case) the slope of which is determined by the velocity. Jacobson and Wechsler proposed to find this plane by integrating over the possible planar regions in these local spectra (via a so-called "velocity polling function"), using the plane of maximum energy to determine the velocity.

Optical Flow Using 3-D Gabor Filters. Heeger [29] proposed the use of 3-D Gabor filters to determine this slope. Following the definition discussed for 2-D, a 3-D Gabor filter has the impulse response

$$g(x, y, t) = \hat{g}(x, y, t) e^{j2\pi[\xi_{x_0}(x-x_0) + \xi_{y_0}(y-y_0) + \xi_{t_0}(t-t_0)]} \tag{8.53}$$

where

$$\hat{g}(x, y, t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t} e^{-\frac{1}{2}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2 + \left(\frac{y-y_0}{\sigma_y}\right)^2 + \left(\frac{t-t_0}{\sigma_t}\right)^2\right]} \tag{8.54}$$

To detect motion in different directions, a family of these filters is defined, as shown in Figure 8.22.

In order to capture velocities at different scales (high velocities can be thought of as occurring over large scales, because a large distance is covered per unit time), these filters are applied to a Gaussian pyramidal decomposition of the sequence. Given the energies of the outputs of these filters, which can be thought of as sampling spatiotemporal/spatiotemporal-frequency space, the problem is analogous to that shown in Figure 8.23. The slope of the line (corresponding to the slope of the plane which characterizes motion) must be found via a finite set of observations. In this method, this problem is solved under the
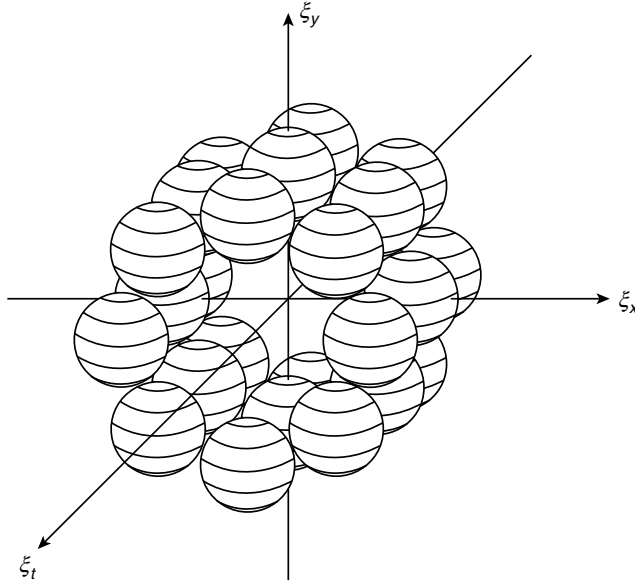
FIGURE 8.22   The (stylized) power spectra of a set of 3-D Gabor filters.
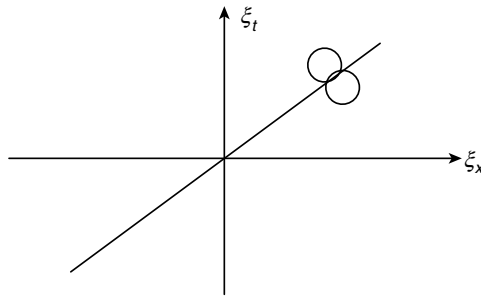
FIGURE 8.23   Velocity estimation in the frequency domain via estimation of the slope of the spectrum.

assumption of a random texture input (the plane in the frequency domain consists of a single constant value).

Optical Flow via the 3-D Gabor Transform.   One shortcoming of a filterbank approach (if the filters are not orthogonal or do not provide a complete basis) is the possibility of loss. Using the 3-D Gabor functions as the basis of a transform resolves this problem. A sequence of dimension $N \times M \times P$ can then be expressed at each discrete point (x_m, y_n, t_p) as

$$f(x_m, y_n, t_p) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} \sum_{s=0}^{S-1} c_{x_q, y_r, t_s, \xi_{x_j}, \xi_{y_k}, \xi_{t_l}} \cdot g_{x_q, y_r, t_s, \xi_{x_j}, \xi_{y_k}, \xi_{t_l}}(x_m, y_n, t_p) \tag{8.55}$$

where $J \cdot K \cdot L \cdot Q \cdot R \cdot S = N \cdot M \cdot P$ for completeness, the functions $g_{x_q, y_r, t_s, \xi_{x_j}, \xi_{y_k}, \xi_{t_l}}(x_m, y_n, t_p)$ denote the Gabor basis functions with spatiotemporal and spatiotemporal-frequency centers of (x_q, y_r, t_s) and $(\xi_{x_j}, \xi_{y_k}, \xi_{t_l})$ respectively, and $c_{x_q, y_r, t_s, \xi_{x_j}, \xi_{y_k}, \xi_{t_l}}$ are the associated coefficients. Note that these coefficients are not found by

convolving with the Gabor functions, since the functions are not orthogonal. See [30] for a survey and comparison of methods for computing this transform.

In the case of uniform translational motion, the slope of the planar spectrum is sought, yielding the optical flow vector $\vec{r}$. A straightforward approach to estimating the slope of the local spectra [31,32] is to form vectors of the ξ_x, ξ_y, and ξ_t coordinates of the basis functions that have significant energy for each point in the sequence at which basis functions are centered. From equation 20, the optical flow vector and the coordinate vectors $\vec{\xi}_x$, $\vec{\xi}_y$, and $\vec{\xi}_t$ at each point are related as

$$\vec{\xi}_t = -(r_x \vec{\xi}_x + r_y \vec{\xi}_y) = -S\,\vec{r} \tag{8.56}$$

where $S = (\vec{\xi}_x \mid \vec{\xi}_y)$. An LMS estimate of the optical flow vector at a given point can then be found using the pseudo inverse of S:

$$\vec{r}_{\mathrm{est}} = -(S^{\mathrm{T}}S)^{-1}S^{\mathrm{T}}\vec{\xi}_t \tag{8.57}$$

In addition to providing a means for motion estimation, this approach has also proven useful in predicting the apparent motion reversal associated with temporal aliasing [33].

Wavelet-Based Methods. A number of wavelet-based approaches to this problem have also been proposed. In Refs. [34–37], 2-D wavelet decompositions are applied frame-by-frame to produce multi-scale feature images. This view of motion analysis exploits the multiscale properties of wavelets, but does not seek to exploit the frequency domain properties of motion. In Ref. [38], a spatiotemporal (3-D) wavelet decomposition is employed, so that some of these frequency domain aspects can be utilized. Leduc et al. explore the estimation of translational, accelerated, and rotational motion via spatiotemporal wavelets in Refs. [39–44]. Decompositions designed and parameterized specifically for the motion of interest (e.g., rotational motion) are tuned to the motion to be estimated.

8.6   Image Sequence Compression

Image sequences represent an enormous amount of data (e.g., a 2-hour movie at the US HDTV resolution of 1280×720 pixels, 60 frames/second progressive, with 24 bits/pixel results in 1194 Gbytes of data). This data is highly redundant, and much of it has minimal perceptual relevance. One approach to reducing this volume of data is to apply still image compression to each frame in the sequence (generally referred to as intraframe coding). For example, the JPEG still image compression algorithm can be applied frame by frame (sometimes referred to as Motion-JPEG or M-JPEG). This method, however, does not take advantage of the substantial correlation, which typically exists between frames in a sequence. Compression techniques which seek to exploit this temporal redundancy are referred to as interframe coding methods.

Motion Compensated Prediction/Transform Coders

Predictive coding is based on the idea that to the degree that all or part of a frame in a sequence can be predicted, that information need not be transmitted. As a result, it is usually the case that the better the prediction, the better the compression that can be achieved. The simplest possible predicator is to assume that successive frames are identical (differential coding); however, the optical flow, which indicates the motion of intensity patterns in the image sequence, can be used to improve the predictor. *Motion compensated prediction* uses optical flow information, together with a reconstruction of the previous frame, to predict the content of the current frame.

Quantization (and the attendant loss of information) is inherent to lossy compression techniques. This loss, if introduced strategically, can be exploited to produce a highly compressed sequence, with good visual
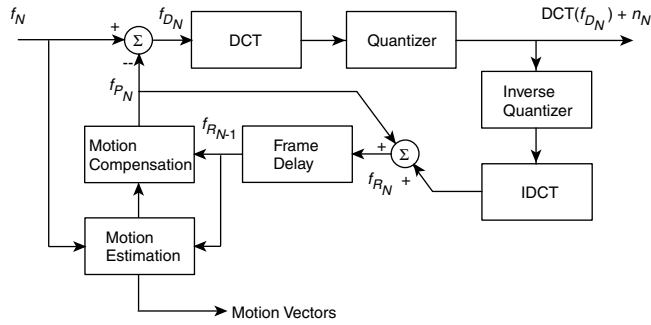
FIGURE 8.24   A hybrid (predictive/transform) encoder.
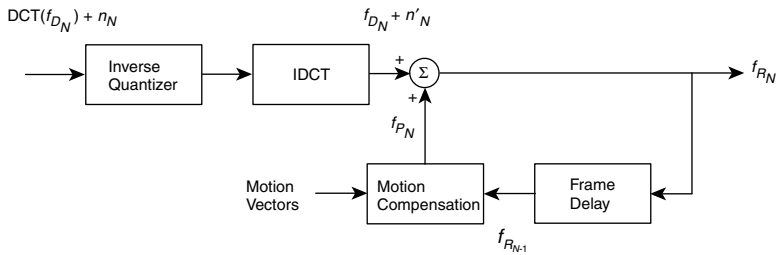
FIGURE 8.25   A predictive/transform decoder.

quality. Transforms (e.g., the DCT), followed by quantization, provide a convenient mechanism to introduce (and control) this loss. Following this approach, a hybrid (motion compensated prediction/ transform) encoder and decoder are shown in Figures 8.24 and 8.25. This hybrid algorithm (with the addition of entropy coders and decoders) is the essence of the H.261, MPEG-1, MPEG-2, and US HDTV compression methods [45].

Perceptually Based Methods

Although algorithms such as MPEG exploit the properties of visual perception (principally in the formulation of quantization matrices), it is not especially central to the structure of the algorithm. There is, for example, no explicit model of vision underlying the MPEG-1 and 2 algorithms. In perceptually-based (sometimes called second generation) methods, knowledge of the HVS takes a much more central role. This view of the problem is particularly effective (and necessary) when designing compression algorithms intended to operate at very high compression ratios (e.g., over 200:1).

The methods in this subsection are inspired by specific models of visual perception. The first is an approach based on a very comprehensive vision model, performing spatial and temporal frequency decomposition via filters designed to reflect properties of the HVS. The second and third are techniques using visually relevant transforms (the Gabor and derivative of Gaussian transforms, respectively) in an otherwise conventional hybrid (predictive/transform) framework. Finally, a method based on spatiotemporal segmentation (following the contour/texture model of vision) will be discussed.

The Perceptual Components Architecture

The perceptual components architecture [46] is a framework for the compression of color image sequences based on the processing thought to take place in the early HVS. It consists of the following steps. The input RGB image sequence is converted into an opponent color space (white/black (WB), red/green (RG), and blue/

yellow (BY)). The sequence is filtered spatially with a set of frequency and orientation selective filters, inspired by the frequency and orientation selectivity of the HVS. Filters based on the temporal frequency response of the visual system are applied along the temporal dimension. The filtered sequences are then subsampled using a hexagonal grid, and subsampled by a factor of two in the temporal dimension. Uniform quantization is applied within each subband, with higher frequency subbands quantized more coarsely. The WB (luminance) component is quantized less coarsely overall than the RG and BY (chrominance) components. The first-order entropy of the result provides an estimate of the compression ratio.

Note that there is no prediction or motion compensation. This is a 3-D subband coder, where temporal redundancy is exploited via the temporal filters. For a 256×256, 8 frame segment of the "football" sequence (a widely used test sequence depicting a play from an American football game), acceptable image quality was achieved for about 1 bit/pixel (from 24 bits/pixel). Although this is not very high compression, the sequence used is more challenging than most. Another contributing factor is that the subsampled representation is 8/3 the size (in terms of bits) of the original, which must be overcome before any compression is realized.

Very-Low-Bit-Rate Coding Using the Gabor Transform

In discussing the Gabor transform previously, it was stated that the basis functions of this transform are optimally (jointly) local. In the context of coding, there are three mechanisms that can be exploited to achieve compression, all of which depend on locality: the local correlation between pixels in the sequence; the bounded frequency response of the human visual system (as characterized by the CSF); and visual masking (the decrease in visual sensitivity near spatial and temporal discontinuities). To take advantage of local spatial correlation, the image representation upon which a compression method is based must be spatially local (which is why images are partitioned into blocks in JPEG, MPEG-1&2, H.261, etc.). If the CSF is to be exploited (e.g., by quantizing high frequency coefficients coarsely) localization in the spatial-frequency domain is required. To exploit visual masking, spatial locality (of a fairly high degree) is required.

The Gabor transform is inherently local in space, so the partitioning of the image into blocks is not required (hence no blocking artifacts are observed at high compression ratios). Its spatial locality also provides a mechanism for exploiting visual masking, while its spatial-frequency locality allows the bandlimited nature of the HVS to be utilized.

An encoder and decoder based on this transform are shown in Figures 8.26 and 8.27 [47].

Note that they are in the classic hybrid (predictive/transform) form. This codec does not include motion compensation, and is for monochrome image sequences.

Applying this method to a 128-by-128, 8 bit/pixel version of the Miss America sequence resulted in reasonable image quality at a compression ratio of approximately 335:1.[1] At 24 frames per second, the associated bit rate is 9.4 kbits/s (a bitrate consistent, e.g., with wireless videotelephony).

Video Coding Using the Derivative of Gaussian Transform

As mentioned previously, the derivative of Gaussian transform (DGT) has properties similar to the Gabor transform, but with the practical advantage that it is real-valued. This makes it particularly well-suited to video compression. In Ref. [48] the hybrid codec structure shown in Figures 8.24 and 8.25 is adapted to the DGT, replacing the DCT (and IDCT), and adapting the quantization scheme to fit the visibility of the DGT basis, via a simple quantization mask.

Comparable results to those of the standard H.261 (DCT-based) codec are obtained for bitrates around 320 kbits/s (5 channels in the p*64 model).

Object-Based Coding by Split and Merge Segmentation

Object-based coding reflects the fact that scenes are largely composed of distinct objects, and that these objects are perceived as boundaries surrounding fields of shading or texture (the contour/texture theory of vision).

---

[1]Not including the initial frame, which is intracoded to 9.1 kbits (a compression ratio of 14).
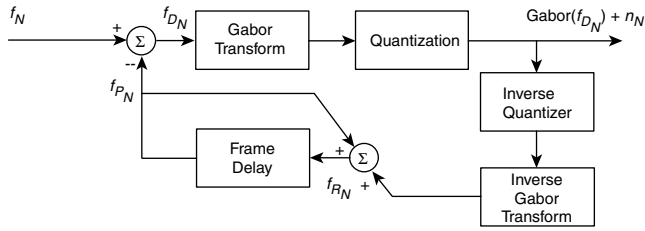
FIGURE 8.26   A Gabor transform-based video encoder.
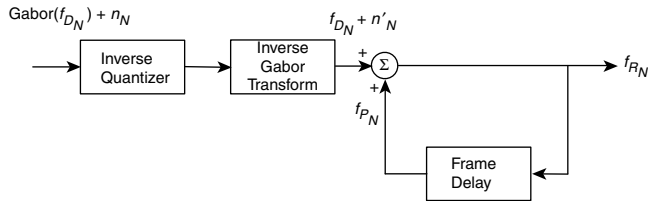
FIGURE 8.27   The associated Gabor transform-based decoder.
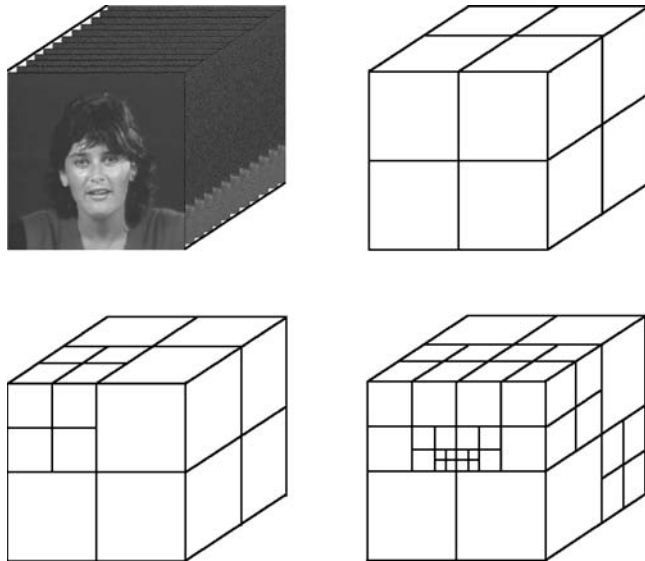
FIGURE 8.28   The split phase.

Encoding an image or sequence in this way requires segmentation to identify the constituent objects. This view of compression, which also facilitates interaction and editing, underlies the MPEG-4 video compression standard [49]. Although the method that will be described is different in detail from MPEG-4, as one of the earliest documented object-based systems, it illustrates many important aspects of such systems.

In this approach [50], 3-D (spatiotemporal) segmentation is used to reduce the redundant information in a sequence (essentially identifying objects within the sequence), while retaining information critical to the

human observer. The sequence is treated as a single 3-D data volume, the voxels of which are grouped into regions via split and merge. The uniformity criterion used for the segmentation is the goodness-of-fit to a 3-D polynomial. The sequence is then encoded in terms of region boundaries (a binary tree structure) and region interior intensities (the coefficients of the 3-D polynomial).

The data volume is first split such that each region is a parallelepiped over which the gray level variation can be approximated within a specified mean squared error (Figure 8.28). Regions are split by quadrants, following the octree strategy. A region adjacency graph is constructed, with nodes corresponding to each region and links between the nodes assigned a cost indicating the similarity of the regions. A high cost indicates low similarity. Regions are merged, starting with regions with the lowest cost, and the region adjacency graph is updated. The resulting regions are represented using a pyramidal (binary tree) structure, with the regions labeled so that adjacent regions have different labels.

Using 16 frames from the "Secretary" sequence, the compression ratio achieved was 158:1 (a bitrate of 83 kbits/s). A total of 5740 parallelepipeds (1000 regions) were used.

8.7   Conclusions

In this chapter, some of the fundamental aspects and algorithms in the processing of digital video were examined. Continued improvements in computing performance make many methods that previously required specialized platforms (or were primarily of research interest due to computational requirements) practical. In addition to bringing high-end applications to the desktop, numerous new applications are thus enabled, in areas as diverse as medical imaging, entertainment, and human-computer interaction.

References

1.   J. Wyver, *The Moving Image — An International Histoy of Film, Television and Video*, London: BFI Publishing, 1989.
2.   A.B. Watson and A.J. Ahumada, "A look at motion in the frequency domain," *SIGGRAPH/SIGART Interdisciplinary Workshop MOTION: Representation and Perception*, Toronto, Canada, 1983, pp. 1–10.
3.   A.T. Smith and R.J. Snowden, Eds., *Visual Detection of Motion*, San Diego, CA: Academic Press, 1994.
4.   K. Nakayama, "Biological image motion processing: a review," *Vision Res.*, vol. 25, no. 5, pp. 625–660, 1985.
5.   A.B. Watson and A.J. Ahumada, "Model of human visual-motion sensing," *J. Opt. Soc. Am. A*, vol. 2, no. 2, 1985.
6.   L.B. Stelmach, W.J. Tam, and P. Hearty, "Static and dynamic spatial resolution in image coding: an investigation of eye movements," *Proc. SPIE/SPSE Symp. Electron. Imaging Sci. Technol.*, San Jose, CA, 1995.
7.   T.R. Reed, "Local frequency representations for image sequence processing and coding," in *Digital Images and Human Vision*, A.B. Watson, Ed., Cambridge, MA: MIT Press, 1993.
8.   L. Cohen, *Time–Frequency Analysis*, Englewood Cliffs, NJ: Prentice-Hall, 1995.
9.   R. Tolimieri and M. An, *Time–Frequency Representations*, Boston, MA: Birkhäuser, 1998.
10. D. Gabor, "Theory of communication," *Proc. Inst. Electr. Eng.*, vol. 93, no. 26, pp. 429–457, 1946.
11. J.G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 7, pp. 1169–1179, 1988.
12. H.G. Feichtinger and T. Strohmer Eds., *Gabor Analysis and Algorithms*, Boston, MA: Birkhäuser, 1998.
13. R.A. Young, "The Gaussian derivative model for spatial vision: I. Retinal mechanisms," *Spatial Vision*, vol. 2, pp. 273–293, 1987.
14. R.A. Young, "Oh say can you see? The physiology of vision," *Proc. SPIE-Human Vision, Vis. Process. Digit. Display II*, vol. 1453, 1991, pp. 92–123.
15. J.A. Bloom and T.R. Reed, "A Gaussian derivative-based transform," *IEEE Trans. Image Process.*, vol. 5, no. 3, pp. 551–553, 1996.
16. E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, 40, 749–759, 1932.

17. M. Vetterli and J. Kovačvic, *Wavelets and Subband Coding*, Englewood Cliffs, NJ: Prentice-Hall, 1995.
18. G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley, MA: Wellesley-Cambridge Press, 1996.
19. S. Mallat, *A Wavelet Tour of Signal Processing*, San Diego, CA: Academic Press, 1998.
20. B.K.P. Horn, *Robot Vision*, Cambridge, MA: MIT Press, 1986.
21. G. Tziritas and C. Labit, *Motion Analysis for Image Sequence Coding*, Amsterdam: Elsevier, 1994.
22. B.K.P. Horn and B.G. Shunck, "Determining optical flow," *Artif. Int.*, vol. 17, no. 1–3, pp. 185–203, 1981.
23. H. Gafni and Y.Y. Zeevi, "A model for separation of spatial and temporal information in the visual system," *Biol. Cybern.*, 28, 73–82, 1977.
24. A. Kojima, N. Sakurai, and J. Kishigami, "Motion detection using 3D-FFT spectrum," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,* Minneapolis, MN, vol. V, 1993, pp. 213–216.
25. B. Porat and B. Friedlander, "A frequency domain algorithm for multiframe detection and estimation of dim targets," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 12, no. 4, pp. 398–401, 1990.
26. H. Gafni and Y.Y. Zeevi, "A model for processing of movement in the visual system," *Biol. Cybern.*, 32, 165–173, 1979.
27. D.J. Fleet and A.D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vision*, vol. 5, no. 1, pp. 77–104, 1990.
28. L. Jacobson and H. Wechsler, "Derivation of optical flow using a spatiotemporal-frequency approach," *Comput. Vision, Graph. Image Process.*, 38, 29–65, 1987.
29. D.J. Heeger, "Optical flow using spatiotemporal filters," *Int. J. Comput. Vision*, vol. 1, no. 4, pp. 279–302, 1988.
30. T.T. Chinen and T.R. Reed, "A performance analysis of fast Gabor transforms," *Graph. Models Image Process.*, vol. 59, no. 3, pp. 117–127, 1997.
31. T.R. Reed, "The analysis of motion in natural scenes using a spatiotemporal/spatiotemporal-frequency representation," *Proc. IEEE Int. Conf. Image Process.,* Santa Barbara, CA, 1997, pp. I-93–I-96.
32. T.R. Reed, "On the computation of optical flow using the 3-D Gabor transform," *Multidimens. Sys. Signal Process.*, vol. 9, no. 4, pp. 115–120, 1998.
33. T.R. Reed, "A spatiotemporal/spatiotemporal-frequency interpretation of apparent motion reversal," *Proc. 16th Int. Joint Conf. Artif. Int.,* Stockholm, Sweden, vol. 2, 1999, pp. 1140–1145.
34. J. Magarey and N. Kingsbury, "Motion estimation using a complex-valued wavelet transform," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 1069–1084, 1998.
35. Y.-T. Wu, T. Kanade, J. Cohn, and C.-C. Li, "Optical flow estimation using wavelet motion model," *Proc. IEEE Int. Conf. Comput. Vis.,* Bombay, India, 1998, pp. 992–998,.
36. G.Van der Auwera, A. Munteanu, G. Lafruit, and J. Cornelis, "Video coding based on motion estimation in the wavelet detail image," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Seattle, WA, vol. 5, 1998, pp. 2801–2804.
37. C.P. Bernard, "Discrete wavelet analysis: a new framework for fast optic flow computation," *Proc. fifth Eur. Conf. Comput. Vision*, Freiburg, Germany, vol. 2, 1998, pp. 354–368.
38. T.J. Burns, S.K. Rogers, M.E. Oxley, and D.W. Ruck, "Discrete, spatiotemporal, wavelet multiresolution analysis method for computing optical flow," *Opt. Eng.*, vol. 33, no. 7, pp. 2236–2247, 1994.
39. J.-P. Leduc, "Spatio-temporal wavelet transforms for digital signal analysis," *Signal Process.*, vol. 60, no. 1, pp. 23–41, 1997.
40. J.-P. Leduc, J. Corbett, M. Kong, V. Wickerhauser, and B. Ghosh, "Accelerated spatio-temporal wavelet transforms: an iterative trajectory estimation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* Seattle, WA, vol. 5, 1998, pp. 2781–2784.
41. J.-P. Leduc, J.R. Corbett, and M.V. Wickerhauser, "Rotational wavelet transforms for motion analysis, estimation and tracking," *Proc. IEEE Int. Conf. Image Process.,* Chicago, IL, vol. 2, 1998, pp. 195–199.
42. J.-P. Leduc and J.R. Corbett, "Spatio-temporal continuous wavelets for the analysis of motion on manifolds," *Proc. IEEE-SP Int. Symp. Time–Freq. Time–Scale Anal.,* Pittsburgh, PA, 1998, pp. 57–60.

43. M. Kong, J.-P. Leduc, B.K. Ghosh, J. Corbett, and V.M. Wickerhauser, "Wavelet-based analysis of rotational motion in digital image sequences," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Seattle, WA, vol. 5, 1998, pp. 2777–2780.

44. J. Corbett, J.-P. Leduc, and M. Kong, "Analysis of deformational transformations with spatio-temporal continuous wavelet transforms," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Phoenix, AZ, vol. 6, 1999, pp. 3189–3192.

45. A.N. Netravali and B.G. Haskell, *Digital Pictures—Representation, Compression, and Standards*, New York: Plenum Press, 1995.

46. A.B. Watson and C.L.M. Tiana, "Color motion video coded by perceptual components," in *SID1992 Digest Tech. Pap.*, vol. XXIII, 1992, pp. 314–317.

47. T.R. Reed and A.E. Soohoo, "Very-low-bit-rate coding of image sequences using the Gabor transform," *J. Soc. Inf. Display*, vol. 3, no. 2, 1995.

48. J.A. Bloom and T.R. Reed, "On the compression of video using the derivative of Gaussian transform," *Proc. 32nd Ann. Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, 1998, pp. 865–869.

49. R. Koenen, Ed., *MPEG-4 Overview*, ISO/IEC JTC1/SC29/WG11 N3747, La Baule, 2000.

50. P. Willemin, T.R. Reed, and M. Kunt, "Image sequence coding by split and merge," *IEEE Trans. Commun.*, vol. 39, no. 12, 1991.

This page intentionally left blank

9

Low Sample Support Adaptive Parameter Estimation and Packet-Data Detection for Mobile Communications[1]

Haoli Qian
Marvell Semiconductor Inc.

Stella N. Batalama
SUNY at Buffalo

Dimitri Kazakos
University of Idaho

In this chapter, we consider the problem of designing low sample support (packet-rate) adaptive receivers for mobile communications. We examine both disjoint and joint configurations where estimations of receiver parameters and packet-data detection are performed either in two separate stages or in a coupled, joint manner, respectively. For the first case, we focus on minimum mean square error/minimum variance distortionless response (MMSE/MVDR)-type receivers, while for the second case, we develop generalized likelihood ratio test (GLRT)-type detectors. Complete adaptive antenna-array DS-CDMA receiver structures are developed. Receiver performance is measured in terms of output signal-to-interference-plus-noise-ratio and bit-error-rate.

9.1   Introduction

The effectiveness of a receiver designed for a rapidly changing multiple-access (multiuser) communications environment depends on the following design attributes: (i) system adaptivity under limited data support, (ii) multiple-access-interference resistance, and (iii) 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, (ii) and (iii), 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 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. Packet throughput improvements can be achieved as a result of bit-error-rate (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. 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 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 (DS-CDMA) radio frequency (RF) system. In such a system, the transmitted signal is a spread-spectrum (SS) signal that is 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 undergoes a process known as multipath-fading that is 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. This general DS-CDMA signal model example will be revisited many times throughout our discussion, and a complete adaptive antenna-array DS-CDMA 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 re-evaluated (estimated) every time there is a significant change in the statistics of the environment. Packet-rate adaptive receivers that aim at the detection of the information bits within a data packet and assume limited or no knowledge of the time varying environment can be categorized as follows: (i) Receivers that perform estimation and detection disjointly, and (ii) receivers that perform estimation and packet-data detection jointly. In either case, both estimation and detection are based on the same data packet (it is assumed that the data packet size is commensurate to the rate of change of the environment, which implies that within a data packet the environment remains unchanged).

For class (i), the design of the receiver building blocks 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 received data packet, to generate an *estimate* of the optimum solution. The short-data-record performance of an estimation algorithm directly determines the performance of the receiver in a time varying channel. In class (ii), receivers are implemented by solving a joint, two-step, coupled optimization problem. For example, optimization with respect to the unknown parameters is carried out first, and results in a solution that is parameterized by the data bits. Then, the parameterized solution is substituted into the original criterion function. Optimization is now performed with respect to the data bits only. Such an approach avoids a separate estimation step for the unknown parameters, but often results in a receiver algorithm that exhibits higher computational complexity. We note that the time varying nature of the channel frequently necessitates fast (short-data-record) adaptive optimization through the use of small input data sets that can "catch up" with the channel variations. In the following sections, we will study the short-data-record performance of various adaptive receiver structures that perform either disjoint or joint parameter estimation and packet-data detection. Examples include sample-matrix-inversion (SMI), least-mean-square (LMS), recursive-least-square (RLS) and auxiliary-vector (AV)-type adaptive linear receivers, as well as generalized likelihood ratio test (GLRT)-type packet data receivers.

   To explain further the developments presented in this chapter, let us consider a DS-CDMA system with 5-element antenna-array reception, system processing gain 64, and 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). For such a system, we will see later that jointly optimal space-time (S-T) processing at the receiver under the the minimum-mean-square-error (MMSE) criterion requires processing in the $5(64 + 3 - 1) = 330$ space-time product space. That is, optimization needs to be carried out in the complex C^{330} vector space. We know that adaptive SMI implementations of the MMSE receiver require data samples many times the space-time product to approach the performance characteristics of their ideal counterpart (RLS/LMS implementations behave similarly). In fact, theoretically, system optimization with data samples less than the space-time product may not even be possible, as we will explain later in our discussion. With CDMA chip rates at 1.25 MHz, processing gain 64, and typical fading rates of 70 Hz or more for vehicle mobiles, the fading channel fluctuates decisively at least every 280 data symbols. In this context, conventional SMI/RLS/LMS adaptive filter optimization in the C^{330} vector space becomes an unrealistic objective.

   The goal of our presentation is to introduce and then elaborate on the underlying principles of short-data-record adaptive receivers. Through illustrative examples from the mobile communications literature, we will observe that well-designed short-data-record (packet-rate) disjoint or joint parameter estimation and packet-data detection schemes result 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.

9.2   Basic Signal Model

For illustration purposes, we consider throughout this presentation a multiuser communications system where binary antipodal information symbols from *user 0, user 1, ⋯, 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 bandlimited 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

$$H_1 : x(t) = (+1)\sqrt{E_0}v_0(t) + z(t) \quad \text{and} \quad H_0 : x(t) = (-1)\sqrt{E_0}v_0(t) + z(t), \quad 0 \leqslant t \leqslant T \quad (9.1)$$

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: (*i*) MAI, (*ii*) inter-symbol interference (ISI), and (*iii*) 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 **x** denote such a discrete-time complex, in general, received signal vector in C^P. That is,

$$H_1 : \mathbf{x} = +\sqrt{E_0}v_0 + \mathbf{z} \quad \text{and} \quad H_0 : \mathbf{x} = -\sqrt{E_0}\mathbf{v}_0 + \mathbf{z}, \quad \mathbf{x}, \mathbf{v}_0, \mathbf{z} \in C^P \tag{9.2}$$

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 **z** denotes the discrete-time comprehensive disturbance vector [2]. Our objective is to detect b_0 (that is, to decide in favor of H_1 or H_0) by means of a linear filter **w** as follows:

$$\hat{b}_0 = \text{sgn } (\text{Re}\{\mathbf{w}^H\mathbf{x}\}) \tag{9.3}$$

where sgn($\cdot$) is the ± 1 hard-limiter, 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 Figure 9.1. Our discussion will be focused on the design of the linear filter **w** according to the MMSE or minimum-variance-distortionless-response (MVDR) optimization criteria that are presented in the following section. We note that under the assumption of perfectly known input statistics and colored Gaussian noise (parameter **z** in Equation (9.2)), Figure 9.1 that utilizes the ideal MMSE/MVDR filter **w** is indeed the optimum ML scheme.

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) 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 tap delay line), and is assumed to be bandlimited 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 low-pass 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 super-vector known as space-time (S-T) received vector. This way the data are prepared for processing by the linear filter, **w**, extraction of the real part of the filter output, and finally, sign detection as shown in Figure 9.1 — a process termed "one-shot" detection, i.e. 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 discrete-time, S-T complex received vector **x** is of dimension, $M(L + N - 1)$; where $L + N - 1$ is exactly the length of what we referred to earlier as
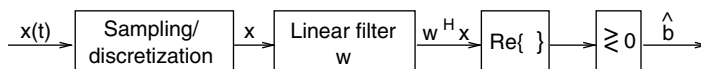
FIGURE 9.1 General receiver structure for the (one-shot) detection of binary antipodal information symbols of the user of interest.

the multipath-extended symbol period. "Inside" $\mathbf{x}$ in Equation (9.2), $\mathbf{v}_0$ corresponds to the channel-processed (also known as "effective") S-T 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:

$$\mathbf{v}_0 = \sqrt{\frac{E_0}{L}} \sum_{n=0}^{N-1} c_{0,n} \left[\underbrace{0 \ldots 0}_{n} \, \mathbf{d}_0^T \, \underbrace{0 \ldots 0}_{N-n-1} \right]^T \odot \mathbf{a}_{0,n} \tag{9.4}$$

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 (which exhibit Rayleigh distributed amplitude and a uniformly distributed phase that fits experimental measurements) and assumed to remain constant over the entire packet duration. In a realistic environment, the coefficients may vary approximately every 300 symbols [3]. Thus, keeping the packet size less than 300 validates the assumption of constant multipath coefficients over the duration of a packet. In Equation (9.4), $\mathbf{d}_0 = [\mathbf{d}_0[0], \ldots, \mathbf{d}_0[L-1]]^T$ 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 Kronecker tensor product. The array response vector of the nth path of the user of interest is defined by

$$\mathbf{a}_{0,n}(m) = e^{j2\pi(m-1)\frac{d}{\lambda}\sin\theta_{0,n}}, \quad m = 1, 2, \ldots, M \tag{9.5}$$

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 $d = \frac{\lambda}{2}$). More details on the DS-CDMA S-T received signal model in Equation (9.4) and the operational characteristics of an antenna-array system can be found in Refs. [4,5]. 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 Equation (9.4), yet with different in general energy values, signature vectors, multipath coefficients, and angles of arrival.

9.3 Data Processing with Known Input Statistics

Optimum MMSE/MVDR Filter

MVDR (minimum-variance-distortionless-response) 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, $\mathbf{x} \in C^P$, that is processed by a P-tap filter, $\mathbf{w} \in C^P$, then the filter output variance is $E\{|\mathbf{w}^H\mathbf{x}|^2\} = \mathbf{w}^H\mathbf{R}\mathbf{w}$, 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 set-up, MVDR filtering is a standard linear constraint optimization problem. The conventional Lagrange multipliers constraint optimization technique leads to the solution (the Lagrange multipliers optimization technique is presented in detail in Ref. [5])

$$\mathbf{w}_{\text{MVDR}} = \rho^\star \frac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0} \tag{9.6}$$

where $(\cdot)^\star$ denotes conjugation. Extensive tutorial treatments of MVDR filtering can be found in many sources, for example in Ref. [5].

MVDR filtering has long been a workhorse for blind (unsupervised) communications and signal processing applications where a desired (pilot) scalar filter output, $y \in C$, cannot be identified or cannot be assumed

available for each input, $\mathbf{x} \in C^P$. 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 and the input vector, then the MMSE filter obtained by minimizing the mean square (MS) error between the filter output and the desired output is given by

$$c\mathbf{R}^{-1}\mathbf{v}_0, \quad c > 0 \tag{9.7}$$

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 [5].

Conventionally, the computation of the MMSE/MVDR filter in Equation (9.6) or Equation (9.7) 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 [6]. 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 preferred 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: (i) The generalized sidelobe canceller (GSC) and its variations, and (ii) the auxiliary-vector (AV) filter. The relative performance of the above methods in limited data support environments is examined in Section "Auxiliary-Vector (AV) Filters".

Generalized Sidelobe Canceller (GSC)

For a given (not necessarily normalized) constraint vector, $\mathbf{v}_0 \in C^P$, any "distortionless" linear filter $\mathbf{w} \in C^P$, that satisfies $\mathbf{w}^H \mathbf{v}_0 = \rho$ can be expressed/decomposed as $\mathbf{w} = \dfrac{\rho^\star}{\|\mathbf{v}_0\|^2}\mathbf{v}_0 - \mathbf{u}$ for some $\mathbf{u} \in C^P$ such that $\mathbf{v}_0^H \mathbf{u} = 0$, as shown in Figure 9.2 (this decomposition is an application of the projection theorem in linear algebra). There are two general approaches for the design of the filter part, $\mathbf{u}$: *(i)* eigen-decomposition based approaches and *(ii)* non-eigen-decomposition-based approaches.

Algorithmic eigen-decomposition 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, beam-space partially adaptive processors, or generalized sidelobe cancellers (GSC). Recent developments have been influenced by principal component analysis 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. The approximation is of the general form (Figure 9.3):

$$\mathbf{u}_{P \times 1} \simeq \mathbf{B}_{P \times (P-1)} \mathbf{T}_{(P-1) \times p} \mathbf{w}^{GSC}_{p \times 1} \tag{9.8}$$

where $\mathbf{B}$ is a matrix that satisfies $\mathbf{B}^H \mathbf{v}_0 = 0_{P-1}$ 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
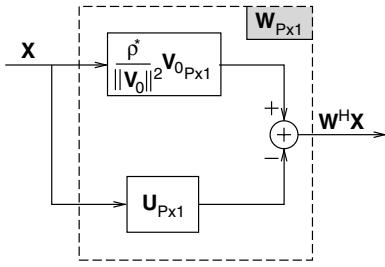
FIGURE 9.2 General decomposition of a linear filter, $\mathbf{w}$, that satisfies $\mathbf{w}^H\mathbf{v}_0 = \rho$ to two orthogonal components, $(\mathbf{u}^H\mathbf{v}_0 = 0)$.
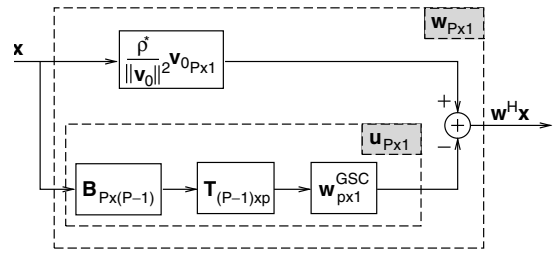
FIGURE 9.3 Generalized sidelobe canceller structure.

a $P \times P$ orthogonal projection matrix such as $\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}$ where $\mathbf{I}$ is the identity matrix). $\mathbf{T}$ is the rank reducing matrix with $1 \leqslant p \leqslant P - 1$ 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 $\mathbf{w}$, $E\left\{\left|\left(\dfrac{\rho^\star}{\|\mathbf{v}_0\|^2}\mathbf{v}_0 - \mathbf{u}\right)^H \mathbf{x}\right|^2\right\}$.

The solution to the latter optimization problem (assuming that $\mathbf{T}$ is given) is

$$\mathbf{w}^{\mathrm{GSC}} = \frac{\rho^\star}{\|\mathbf{v}_0\|^2}[\mathbf{T}^H\mathbf{B}^H\mathbf{R}\mathbf{B}\mathbf{T}]^{-1}\mathbf{T}^H\mathbf{B}^H\mathbf{R}\mathbf{v}_0 \qquad (9.9)$$

We note that the rank-reducing matrix $\mathbf{T}$ "reduces" the dimension of the linear filter (number of parameters to be designed) from P (dimension of filter, $\mathbf{w}$) to p (dimension of 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. This choice is mean-square (MS) optimum under the assumption that the disturbance-only eigenvectors are not rotated by the blocking matrix being used (i.e., when the disturbance subspace is orthogonal to the constraint vector, $\mathbf{v}_0$), which is not true 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, $\mathbf{B}^H\mathbf{R}\mathbf{B}$. 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 $\mathbf{B}^H\mathbf{R}\mathbf{B}$ with corresponding eigenvalues λ_i, that maximize the ratio, $\dfrac{|\mathbf{v}_0^H\mathbf{R}\mathbf{B}_{q_i}|^2}{\lambda_i}$, $i = 1, \ldots, p$, [7] (also known as "cross-spectral metric").

9.4 Auxiliary-Vector (AV) Filters

Auxiliary-vector (AV)-filters are non-eigen-decomposition based filters that approximate the optimum MMSE/MVDR solution [4,8,11]. 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 Figure 9.2, 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, *non-orthogonal*, to each other). The number of auxiliary vectors used to approximate the filter part, $\mathbf{u}$, in Figure 9.2 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 criterion, respectively, as explained in detail below). An important characteristic of the AV-algorithm (besides the non-orthogonality 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 [9].

A pictorial presentation of generation of the sequence of AV-filters is given by Figure 9.4(a). The sequence is initialized at the appropriately scaled constraint vector, $\mathbf{w}_0 = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, 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 $\mathbf{g}_1 \in C^P - \{0\}$, $\mu_1 \in C$, 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. 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 $\mathbf{v}_0$" (that is if $\mathbf{R}\mathbf{w}_0 = (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$ or equivalently $(\mathbf{I} - \mathbf{v}_0\mathbf{v}_0^H)\mathbf{R}\mathbf{w}_0 = 0$), 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 $\mathbf{g}_{k+1}$ (Figure 9.4.(a)), 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
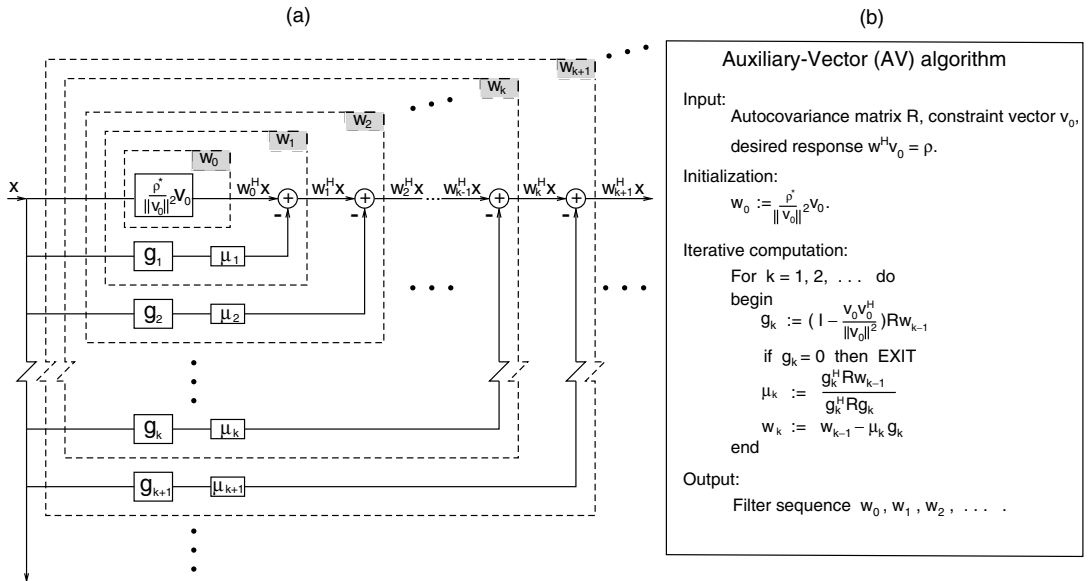
FIGURE 9.4 (a) Block diagram representation and (b) algorithmic description/generation of the auxiliary-vector (AV) filter sequence, $\mathbf{w}_1, \mathbf{w}_2, \ldots$.

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 Figure 9.4(b) (in Figure 9.4(b), 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 $\mathbf{g}_k$):

(i) 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

$$\mu_{k+1} = \frac{\mathbf{g}_{k+1}^H \mathbf{R} \mathbf{w}_k}{\mathbf{g}_{k+1}^H \mathbf{R} \mathbf{g}_{k+1}}, \quad k = 0, 1, 2, \ldots \tag{9.10}$$

(ii) Suppose that $\left(\mathbf{I} - \dfrac{\mathbf{v}_0 \mathbf{v}_0^H}{\|\mathbf{v}_0\|^2} \right) \mathbf{R} \mathbf{w}_k \neq 0$ ($\mathbf{w}_k \neq \mathbf{w}_{\text{MVDR}}$). Then, the auxiliary vector

$$\mathbf{g}_{k+1} = \frac{\mathbf{R} \mathbf{w}_k - \dfrac{\mathbf{v}_0^H \mathbf{R} \mathbf{w}_k}{\|\mathbf{v}_0\|^2} \mathbf{v}_0}{\left\| \mathbf{R} \mathbf{w}_k - \dfrac{\mathbf{v}_0^H \mathbf{R} \mathbf{w}_k}{\|\mathbf{v}_0\|^2} \mathbf{v}_0 \right\|}, \quad k = 0, 1, 2, \ldots \tag{9.11}$$

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 $\rho^\star \dfrac{\mathbf{R}^{-1} \mathbf{v}_0}{\mathbf{v}_0^H \mathbf{R}^{-1} \mathbf{v}_0}$, we can show [9]

(i) The generated sequence of auxiliary-vector weights, $\{\mu_k\}$, $k = 1, 2, \ldots$, is real-valued, positive, and bounded: $0 < \dfrac{1}{\lambda_{\max}} \leq \mu_k \leq \dfrac{1}{\lambda_{\min}}$, $k = 1, 2, \ldots$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the corresponding maximum and minimum eigenvalues of $\mathbf{R}$;

(ii) The sequence of auxiliary vectors, $\{\mathbf{g}_k\}$, $k = 1, 2, \ldots$, converges to the 0 vector, i.e., $\lim_{k \to \infty} \mathbf{g}_k = 0$;

(iii) The sequence of AV filters, $\{\mathbf{w}_k\}$, $k = 1, 2, \ldots$, converges to the MVDR filter, i.e., $\lim_{k \to \infty} \mathbf{w}_k = \rho^\star \dfrac{\mathbf{R}^{-1} \mathbf{v}_0}{\mathbf{v}_0^H \mathbf{R}^{-1} \mathbf{v}_0}$.

We conclude this section with a few comments on AV filtering. From a general input space synthesis/decomposition point of view, the key features of the AV algorithm are: *non-orthogonal AV synthesis* and *conditional statistical optimization*. Non-orthogonal 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 9.5 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 DS-CDMA signal model, ($P = L$). The BER performance of the receiver, $\mathbf{w}_3$, (which 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) receiver and the "cross-spectral-metric" (CSM) 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 Figure 9.6, 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).
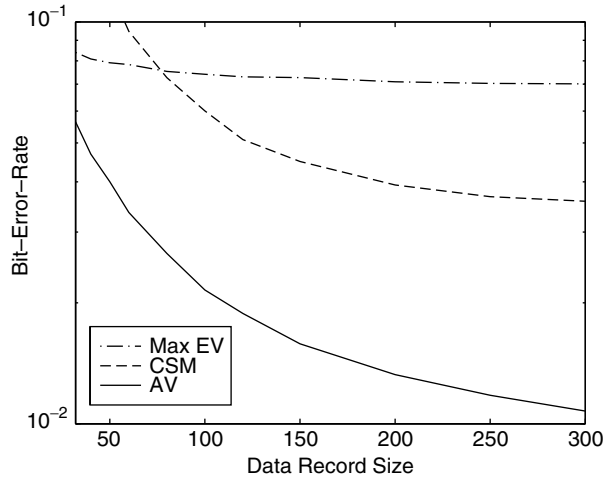
FIGURE 9.5   BER as a function of the data record size for the AV, "max, eigenvector" (MaxEV), and "cross-spectral metric" (CSM) receivers of the same order (3 auxiliary vectors, 3 eigenvectors). *Operational environment*: Synchronous DS-CDMA system, user of interest at 12 dB, 12 interferers at 10–14 dB, processing gain $L = 32$, and arbitrary normalized signatures (cross-correlation with the signature of the user of interest about 0.2).
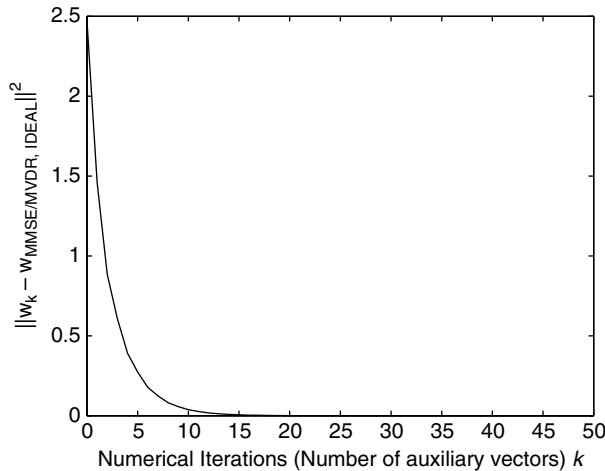
FIGURE 9.6   Convergence of the AV filter sequence to the ideal MMSE/MVDR solution as a function of the number of iterations k in Figure 9.4(b). The sequence of *conditionally optimized* AV filters (which utilize *non-orthogonal* auxiliary vectors) converges to the $\mathbf{w}_{\text{MMSE/MVDR}}$ optimum solution for a perfectly known input autocorrelation matrix $\mathbf{R}$.

9.5   Disjoint Parameter Estimation and Packet-Data Detection

Known Channel

SMI, GSC, and Auxiliary-Vector 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 the first part (part A) of this section 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 $\mathbf{x}_0, \mathbf{x}_1, \ldots \mathbf{x}_{J-1}$

$$\hat{\mathbf{R}}(J) = \frac{1}{J} \sum_{j=0}^{J-1} \mathbf{x}_j \mathbf{x}_j^{\mathrm{H}} \tag{9.12}$$

Throughout the first part of this section, $\mathbf{v}_0$ is assumed to be known (since $\mathbf{v}_0$ is a function of the channel parameters, we label part A 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 second part (Part B: Unknown Channel).

When $\mathbf{R}$ is unknown, the most widely used MMSE/MVDR filter estimator is obtained from Equation (9.6) 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 "Data Processing with Known Input Statistics" and utilize the sample average estimate of the autocorrelation matrix $\hat{\mathbf{R}}(J)$ instead of $\mathbf{R}$ in Equation (9.9), Equation (9.10), and Equation (9.11), we obtain a GSC and AV-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 GSC, 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. Based on 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 $k \rightarrow \infty$, to the SMI filter [9]

$$\hat{\mathbf{w}}_k(J) \xrightarrow[k \rightarrow \infty]{} \hat{\mathbf{w}}_\infty(J) = \hat{\mathbf{w}}_{\mathrm{SMI}} = \rho^\star \frac{[\hat{\mathbf{R}}(J)]^{-1} \mathbf{v}_0}{\mathbf{v}_0^{\mathrm{H}} [\hat{\mathbf{R}}(J)]^{-1} \mathbf{v}_0} \tag{9.13}$$

Properties of the Sequence of AV Estimators

The AV-filter sequence of estimators begins with $\hat{\mathbf{w}}_0(J) = \frac{\rho^\star}{\|\mathbf{v}_0\|^2} \mathbf{v}_0$, which is a *0*-variance, fixed-valued, 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 non-zero 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. Equation (9.13)). To quantify these remarks, we plot in Figure 9.7 the norm-square bias, $\{\hat{\mathbf{w}}_k(J)\} - \mathbf{w}_{\mathrm{MMSE/MVDR}}\|^2$, and the trace of the covariance matrix, $E\{[\hat{\mathbf{w}}_k(J) - E\{\hat{\mathbf{w}}_k(J)\}][\hat{\mathbf{w}}_k(J) - E\{\hat{\mathbf{w}}_k(J)\}]^{\mathrm{H}}\}$, as a function of the iteration step (filter index) k, for the same signal model as in Figure 9.6 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.

Since $\hat{\mathbf{w}}_\infty(J)$ 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 $\mathbf{v}_0$). From the results in Figure 9.7 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 ideal set-ups, the AV-filter estimates, $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \hat{\mathbf{w}}_3(J), \ldots$, have been viewed

---

[1]The SMI estimator is unbiased for multivariate elliptically contoured input distributions $E\{\hat{\mathbf{w}}_\infty(J)\} = \mathbf{w}_{\mathrm{MMSE/MVDR}} = \rho^\star \frac{\mathbf{R}^{-1} \mathbf{v}_0}{\mathbf{v}_0^{\mathrm{H}} \mathbf{R}^{-1} \mathbf{v}_0}$.
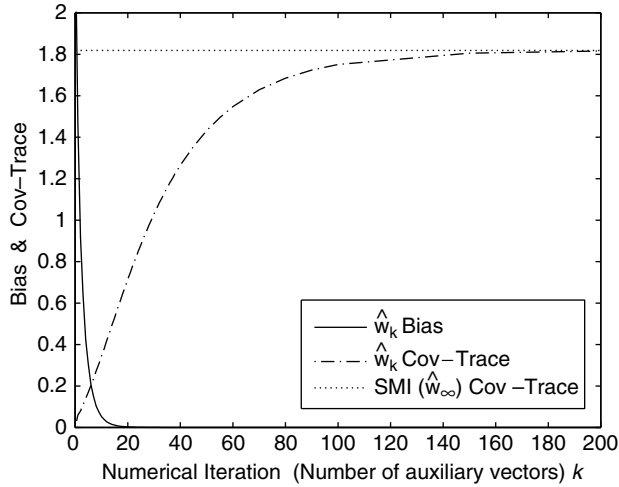
FIGURE 9.7   Norm-square bias and covariance trace for the sequence of estimators, $\hat{\boldsymbol{w}}_k(J)$, $k = 0, 1, \ldots$. The signal model is as in Figure 9.6; data record size $J = 256$.
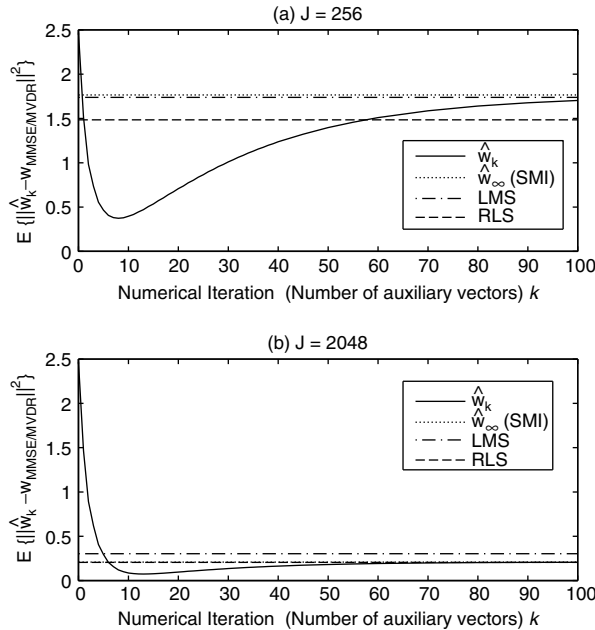
FIGURE 9.8   MS estimation error for the sequence of estimators, $\hat{\boldsymbol{w}}_k(J)$, $k = 0, 1, \ldots$. (a) Data record size $J = 256$. (b) $J = 2048$.

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 Equation (9.6) and Equation (9.7). 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 Figure 9.8, we plot the MS estimation error as a function of the iteration step k (or filter index) for the same signal model as in Figure 9.6, 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. The constraint-LMS estimator is given by the following recursion

$$\hat{\mathbf{w}}_{\text{LMS}}(j) = \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^{\text{H}}}{\|\mathbf{v}_0\|^2}\right)[\hat{\mathbf{w}}_{\text{LMS}}(j-1) - \mu\mathbf{x}_j\mathbf{x}_j^{\text{H}}\hat{\mathbf{w}}_{\text{LMS}}(j-1)] + \frac{\rho^\star}{\|\mathbf{v}_0\|^2}\mathbf{v}_0, \quad j = 1,\dots,J \quad (9.14)$$

with $\hat{\mathbf{w}}_{\text{LMS}}(0) = \frac{\rho^\star}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$ and some $\mu > 0$. The recursion of the RLS estimator can be obtained from the SMI formula in Equation (9.13) by utilizing the following iterative estimation of $\mathbf{R}^{-1}$ that is based on the matrix-inversion-lemma

$$\hat{\mathbf{R}}^{-1}(j) = \hat{\mathbf{R}}^{-1}(j-1) - \frac{\hat{\mathbf{R}}^{-1}(j-1)\mathbf{x}_j\mathbf{x}_j^{\text{H}}\hat{\mathbf{R}}^{-1}(j-1)}{1 + \mathbf{x}_j^{\text{H}}\hat{\mathbf{R}}^{-1}(j-1)\mathbf{x}_j}, \quad j = 1,\dots,J \quad (9.15)$$

where $\hat{\mathbf{R}}^{-1}(0) = \frac{1}{\epsilon_0}\mathbf{I}$ for some $\epsilon_0 > 0$. Theoretically, the LMS gain parameter $\mu > 0$ has to be less than $\frac{1}{2 \cdot \lambda_{\text{max}}^{\text{blocked}}}$, where $\lambda_{\text{max}}^{\text{blocked}}$ is the maximum eigenvalue of the "blocked-data" autocorrelation matrix $\left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^{\text{H}}}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^{\text{H}}}{\|\mathbf{v}_0\|^2}\right)$. 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 $\epsilon_0 > 0$ is necessary to achieve acceptable performance (in our study we set $\mu = \frac{1}{200 \cdot \lambda_{\text{max}}^{\text{blocked}}}$ and $\epsilon_0 = 20$, respectively).

This data specific tuning frequently results in misleading, over-optimistic 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 Figure 9.4(b) 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 Figure 9.8(a), for $J = 256$ all estimators, $\hat{\mathbf{w}}_k(J)$, from $k = 2$ up to about $k = 55$ outperform in MS-error their RLS, LMS, and SMI ($\hat{\mathbf{w}}_\infty(J)$) counterparts. $\hat{\mathbf{w}}_8(J)$ ($k = 8$ auxiliary vectors) has the least MS-error of all (best bias/variance trade-off). When the data record size is increased to $J = 2048$ (Figure 9.8(b)), we can afford more iterations and $\hat{\mathbf{w}}_{13}(J)$ offers the best bias/variance trade-off (lowest MS-error). 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 ($\hat{\mathbf{w}}_\infty(J)$) MS-error are almost identical. Figure 9.9 offers a
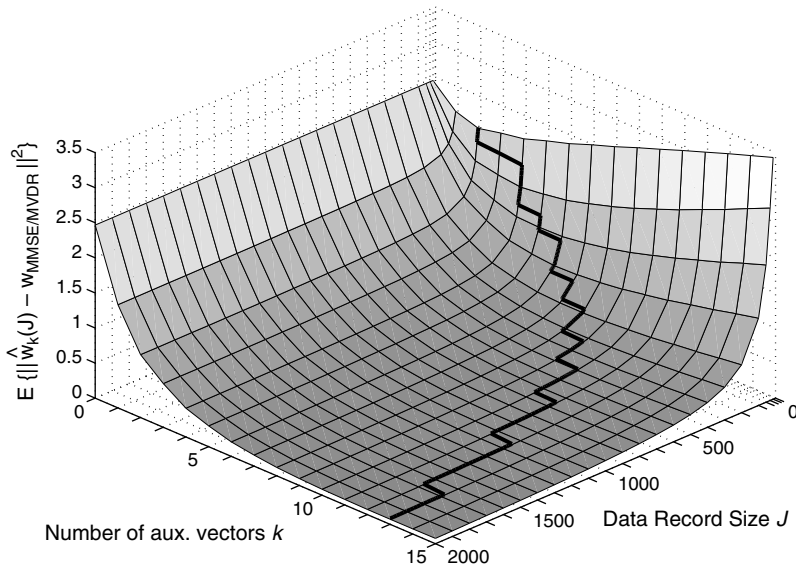
FIGURE 9.9 MS estimation error versus number of auxiliary vectors, k and sample support, J (the signal model is the same as in Figure 9.6).

three-dimensional plot of the MS 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.

How to Choose the Best AV Estimator

We recall that, when the auto-covariance matrix is sample-average estimated, the sequence of AV estimators converges to the SMI filter. Evidently, early, non-asymptotic elements of the sequence offer favorable bias/variance balance characteristics and outperform in mean-square filter estimation error the unbiased SMI filter-estimator, as well as the (constraint)-LMS, and RLS. In the context of digital wireless communication receivers, superior mean-square filter estimation error translates to superior bit-error-rate 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 [12]. 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 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 (CV-MOV)* 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 that is estimated based on 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. The following criterion outlines the CV-MOV AV-filter selection process.

Criterion 9.1   *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, i.e.,*

$$k_1 = \arg\min_k \left\{ \sum_{j=1}^{J} \hat{\mathbf{w}}_k(J \backslash j)^{\mathrm{H}} \mathbf{x}_j \mathbf{x}_j^{\mathrm{H}} \hat{\mathbf{w}}_k(J \backslash j) \right\} \tag{9.16}$$

where $(J \backslash j)$ identifies the AV-filter estimator that is evaluated from the available data record after removing the jth sample. □

While the CV-MOV criterion can be applied to general filter estimation problems, the second criterion, i.e., the *maximum 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 J-divergence distance between f_0 and f_1 is defined as the sum of the Kullback–Leibler (K–L) distances between f_0 and f_1

$$\mathcal{D}(f_0, f_1) \stackrel{\Delta}{=} \mathcal{KL}(f_1, f_0) + \mathcal{KL}(f_0, f_1) \tag{9.17}$$

where the K–L distance of f_1 from f_0 is defined as $\mathcal{KL}(f_1, f_0) \stackrel{\Delta}{=} \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_0(x)} \mathrm{d}x$.

The choice of the output 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

$$P_e \geq \pi_0 \pi_1 \exp\{-D(f_0, f_1)/2\} \tag{9.18}$$

where π_0, π_1 are the a priori probabilities of H_0 and H_1, respectively. The right-hand side of Equation (9.18) is a monotonically decreasing function of the 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, Equation (9.18) is satisfied with equality. The latter implies that the larger the J-divergence the smaller the probability of error or, equivalently, the larger the J-divergence the easier the detection problem. Thus, maximization of the J-divergence implies minimization of the probability of error. Due to the above properties and its relationship with the probability of error of the optimum detector, 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 $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots$. Then, the 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

$$\mathcal{D}(k) \approx \frac{4E^2\{b_0 \mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}]\}}{\mathrm{Var}\{b_0 \mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}]\}} \tag{9.19}$$

where $\mathrm{Var}(\cdot)$ denotes variance. The following criterion outlines the J-divergence AV-filter selection process.

Criterion 9.2 *For a given data packet (data record) of size J, the J-divergence AV-filter selection rule chooses the AV-filter estimator $\hat{\mathbf{w}}_{k_2}(J)$ that maximizes the estimated J-divergence $\hat{\mathcal{D}}(k)$ between the AV-filter output conditional distributions, i.e.,*

$$k_2 = \arg\max_k\{\hat{\mathcal{D}}(k)\} \tag{9.20}$$

$\square$

If we substitute b_0 in Equation (9.19) by $\hat{b}_0 = \mathrm{sgn}(\mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}])$ and evaluate expectations by sample averaging, then we can obtain a blind estimate of the J-divergence

$$\hat{\mathcal{D}}_B(k) = \frac{4\left[\frac{1}{J}\sum_{j=1}^{J}|\mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}_j]|\right]^2}{\frac{1}{J}\sum_{j=1}^{J}\left|\mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}_j]\right|^2 - \left[\frac{1}{J}\sum_{j=1}^{J}|\mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}_j]|\right]^2} \tag{9.21}$$

where the subscript "B" identifies the blind version of the J-divergence function. Then, we can evaluate $k_2 = \arg\max\{\hat{\mathcal{D}}_B(k)\}$. We recall that in Equation (9.21) $\mathbf{x}$ denotes the received signal vector of the general form $\mathbf{x} = b_0\sqrt{E_0}\mathbf{v}_0 + \mathbf{z}$ where (cf. Equation (9.2)) $\mathbf{v}_0 \in C^P$ is a known deterministic signal vector, $E_0 > 0$ represents the unknown energy scalar, $\mathbf{z} \in C^P$ 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. Equation (9.3)) 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, $\mathrm{Re}[\hat{\mathbf{w}}_k^{\mathrm{H}}(J)\mathbf{x}]$, where $\hat{\mathbf{w}}_k(J)$ is the AV estimator that utilizes k auxiliary vectors.

Performance Illustrations

We illustrate the overall short-data-record adaptive filter performance in Figure 9.10 and Figure 9.11 for a multipath fading DS-CDMA 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 one. The maximum cross-correlation between the assigned user signatures reaches 30%. The total SNRs (over the three paths) of the 19 interferers are set at $SNR_{2-6} = 6dB$, $SNR_{7-8} = 7dB$, $SNR_{9-13} = 8dB$, $SNR_{14-15} = 9dB$, $SNR_{16-20} = 10dB$. The space-time product (filter length) is $P = (31 + 2)5 = 165$. The experimental results are averages over 1000 runs (100 different channel realizations and 10 independent data record generations per channel). In Figure 9.10, we plot the BER[1] of the AV estimators $\hat{\mathbf{w}}_{k_1}(J)$ 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 $\hat{\mathbf{w}}_{k_{opt}}(J)$ as well as the corresponding curves of the ideal MMSE/MVDR filter $\mathbf{w}_{\mathrm{MMSE/MVDR}}$, the SMI filter estimator $\hat{\mathbf{w}}_{\infty}(J)$, and the S-T RAKE matched-filter (MF) $\hat{\mathbf{w}}_0(J)$. 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 and the matched filter. We also observe that for moderate to high SNR of the user of interest, the J-divergence selection rule is slightly superior to the CV-MOV selection rule. Figure 9.11 repeats the study of Figure 9.10 as a function of the data record size. The SNR of the user of interest is fixed at 8dB.

Concluding our discussion in the first part of this section, 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 *non-orthogonal* 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 short-data-record adaptation, early, non-asymptotic elements of the
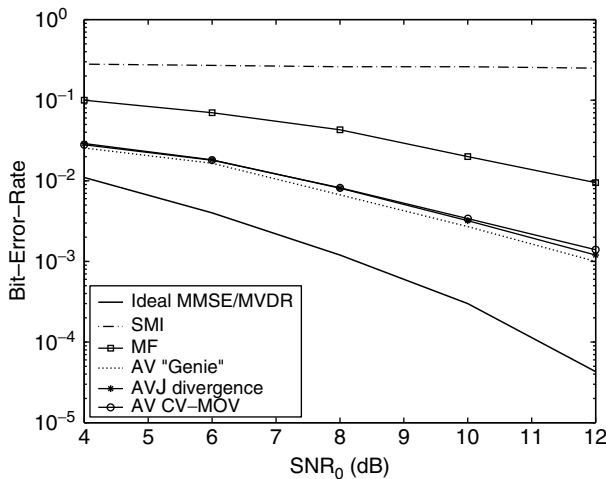
FIGURE 9.10   BER versus SNR for the user signal of interest for a multipath-fading antenna-array received signal model ($L = 31, K = 20, M = 5, N = 3$) with $P = 165$ and $J = 230$.

---

[1]The BER of each filter under consideration is approximated by $Q(\sqrt{\mathrm{SINR_{out}}})$, since the computational complexity of the BER expression for this antenna array CDMA system prohibits exact analytic evaluation [13].
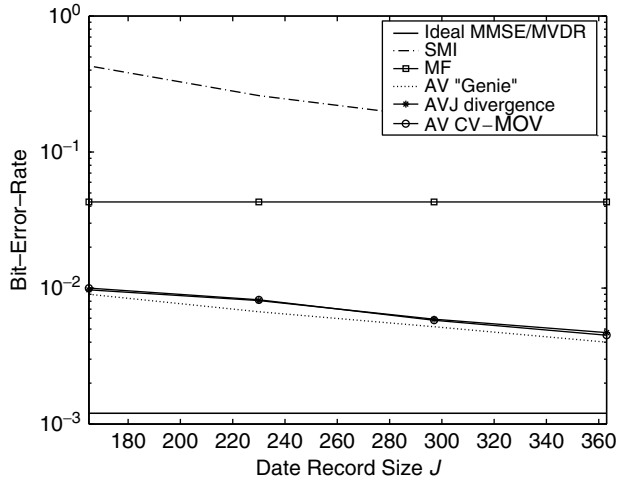
FIGURE 9.11 BER versus data record size (the signal model is the same as in Figure 9.10, $SNR_0 = 8dB$).

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. As a few concluding notes, an online version of the AV algorithm is presented in Ref. [14]. Application of AV-filtering to the problem of rapid synchronization and combined demodulation of DS-CDMA signals is considered in Ref. [15]. Detailed results on data record size requirements to achieve a given output SINR (or BER) performance level can be found in Ref. [16].

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 S-T DS-CDMA signal model described by Equation (9.1) to Equation (9.5). 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 S-T RAKE filter of the user of interest (*user 0*), defined by $\mathbf{v}_0 \overset{\Delta}{=} E_{b_0}\{\mathbf{x}b_0\}$ 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 S-T matched filter multiplied by the corresponding channel coefficients (cf. Equation (9.4)):[1]

$$\mathbf{v}_0 = \sum_{n=0}^{N-1} c_{0,n} \left[\underbrace{0 \ldots 0}_{n} \, \mathbf{d}_0^T \, \underbrace{0 \ldots 0}_{N-n-1} \right]^T \odot \mathbf{a}_{0,n} \tag{9.22}$$

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, $\theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,N-1}$,

---

[1]For the sake of mathematical accuracy, $\mathbf{v}_0 = \sqrt{\frac{E_0}{L}}\sum_{n=0}^{N-1} c_{0,n} \left[\underbrace{0 \ldots 0}_{n} \, \mathbf{d}_0^T \, \underbrace{0 \ldots 0}_{N-n-1} \right]^T \odot \mathbf{a}_{0,n}$. The positive multiplier $\sqrt{\frac{E_0}{L}}$ is dropped in Equation (9.22) as inconsequential.

(cf. Equation (9.5)). While the spreading sequence is assumed to be known to the receiver, the channel coefficients and the angles of arrival are unknown.

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, $\theta_0 \triangleq [\theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,N-1}]^T$, for the user of interest, *user 0*, can be estimated by subspace-based techniques from the S-T data packet (data record), $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$. 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 q th 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 Figure 9.12, where the J_p pilot bits appear as a *mid-amble* in the transmitted packet [17]. 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, i.e., $T_s = JT$ 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

  (i) No truncation: Data dimension $= M(L + N - 1)$, $2Q + 1 \leqslant r_s \leqslant 3Q$.
  (ii) One-sided truncation: Data dimension $= ML$, $2Q \leqslant r_s \leqslant 3Q - 1$.
  (iii) Two-sided truncation: Data dimension $= M(L - N + 1)$, $Q \leqslant r_s \leqslant 2Q - 1$.

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 [iii]) as shown in Figure 9.13, and we form the "truncated" received vector $\mathbf{x}^{tr}$ of length $M(L - N + 1)$ as follows

$$\mathbf{x}^{tr} = \begin{bmatrix} \mathbf{x}((N-1)T_c) \\ \mathbf{x}(NT_c) \\ \vdots \\ \mathbf{x}((L-1)T_c) \end{bmatrix} \tag{9.23}$$

Then, with respect to the j th information bit of *user 0*, $\mathbf{x}_j^{tr}$ can be expressed as

$$\mathbf{x}_j^{tr} = b_0(j)\frac{\sqrt{E_0}}{L}\mathbf{A}_0\mathbf{B}(\theta_0)\mathbf{c}_0 + \mathrm{MAI}_j + \mathbf{n}_j \tag{9.24}$$
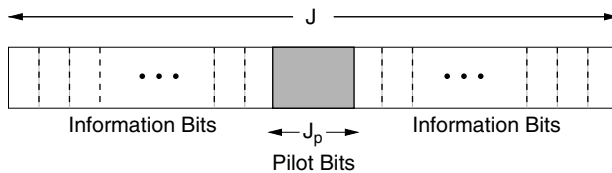
FIGURE 9.12 Data packet structure of total length J bits that contains a mid-amble of J_p pilot bits.
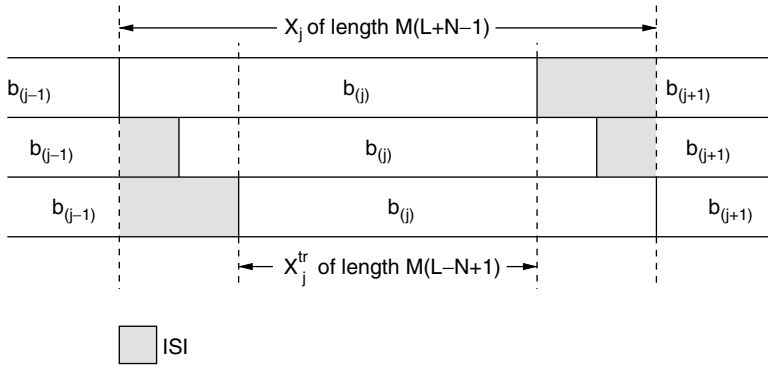
FIGURE 9.13 Data collection and ISI trimming.

where MAI_j accounts comprehensively for multiple-access-interference of rank $r_s - 1$, $\mathbf{B}(\theta_0)$ is a block diagonal matrix of the form $\mathbf{B}(\theta_0) \stackrel{\Delta}{=} \text{diag}(\mathbf{a}_{0,0}, \mathbf{a}_{0,1}, \ldots, \mathbf{a}_{0,N-1})$, and $\mathbf{A}_0 = \mathbf{A}_0^s \odot \mathbf{I}_M$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix and

$$
\mathbf{A}_0^s = \begin{bmatrix}
d_0[N-1] & d_0[N-2] & \ldots & d_0[0] \\
d_0[N] & d_0[N-1] & \ldots & d_0[1] \\
\vdots & \vdots & & \vdots \\
d_0[L-1] & d_0[L-2] & \ldots & d_0[L-N]
\end{bmatrix}
\tag{9.25}
$$

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

$$
\hat{\mathbf{R}}_{\text{tr}} = \frac{1}{J} \sum_{j=0}^{J-1} \mathbf{x}_j^{\text{tr}} \mathbf{x}_j^{\text{tr}\,H}
\tag{9.26}
$$

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 θ_0 indirectly through an estimate of the $MN \times 1$ vector

$$
\mathbf{h}_0 \stackrel{\Delta}{=} \mathbf{B}(\theta_0)\mathbf{c}_0
\tag{9.27}
$$

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 $\hat{\mathbf{U}}_n$

$$
\hat{\mathbf{h}}_0 = \arg \min_{\mathbf{h}_0} \left\| (\mathbf{A}_0 \mathbf{h}_0)^H \hat{\mathbf{U}}_n \right\| \quad \text{subject to} \quad \|\hat{\mathbf{h}}_0\| = 1
\tag{9.28}
$$

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 $\hat{\theta}_0$ by applying least-squares (LS) fitting to $\hat{\mathbf{h}}_0$. Then, the estimate $\hat{\mathbf{v}}_0$ is completely defined by Equation (9.22).

Since the above channel estimation method 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.

Phase Recovery

Without loss of generality, let $\tilde{\mathbf{v}}_0$ denote a phase ambiguous version of $\mathbf{v}_0$, i.e.,

$$\tilde{\mathbf{v}}_0 e^{j\psi} = \mathbf{v}_0 \tag{9.29}$$

where ψ is the unknown phase. We consider the class of linear filters, $\mathbf{w} \in C^{M(L+N-1)}$ that are functions of the S-T RAKE vector, $\mathbf{v}_0$, and share the following property

$$\mathbf{w}(\mathbf{v}_0) = \mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi} \tag{9.30}$$

Such filters include: *(i)* the S-T RAKE filter itself, $\mathbf{v}_0$, *(ii)* the S-T MMSE/MVDR filter of Equation (9.6), and *(iii)* the auxiliary-vector sequence of S-T filters, $\{\mathbf{w}_k\}$.

As seen by Equation (9.30), 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. We choose the value of the phase correction parameter ψ that minimizes the mean-square-error (MSE) between 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 (Figure 9.14)

$$\hat{\psi} = \arg\min_{\psi} E\{|[\mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi}]^H\mathbf{x} - b_0|^2\}, \quad \psi \in [-\pi, \pi) \tag{9.31}$$

The optimum phase correction value according to the above criterion is given by

$$\hat{\psi} = \text{angle}\{\mathbf{w}(\tilde{\mathbf{v}}_0)^H E\{\mathbf{x}b_0\}\} \tag{9.32}$$

Essentially, Equation (9.32) suggests projecting the phase ambiguous $\mathbf{w}(\tilde{\mathbf{v}}_0)$ filter onto the ideal S-T 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

$$\mathbf{w}(\hat{\mathbf{v}}_0, \hat{\mathbf{R}}), \quad \hat{\psi} = \text{angle}\left\{\mathbf{w}(\hat{\mathbf{v}}_0, \hat{\mathbf{R}})^H\left[\sum_{j=1}^{J_p}\mathbf{x}_j b_0(j)\right]\right\} \tag{9.33}$$

Since j represents the packet size of the DS-CDMA system and J_p is the number of mid-amble pilot information bits per packet, then the ratio $\frac{J_p}{J}$ quantifies the wasted bandwidth due to the use of the pilot bit
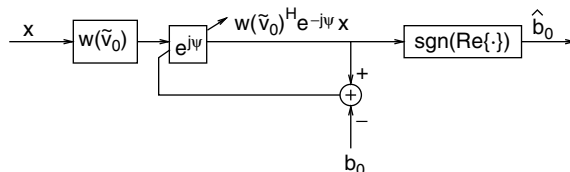
FIGURE 9.14 Supervised (pilot-assisted) phase correction for the S-T linear filter, $\mathbf{w}(\tilde{\mathbf{v}}_0)$.

sequence. Ideally, $\frac{J_p}{J}$ 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 $\frac{J_p}{J} \simeq 2\%$ only.

9.6 Joint Parameter Estimation and Packet-Data Detection

For convenience in notation we introduce the variable $\mathbf{S}_0$ to denote the $P \times N$ known signal waveform matrix of the user of interest. For a single antenna receiver, $\mathbf{S}_0$ takes the form of $\mathbf{A}_0^s$ given in Equation (9.25), while for an antenna-array receiver it is given by $\mathbf{A}_0\mathbf{B}(\theta_0)$ (θ_0 is assumed known in this section). Thus, the j-th received vector of the user of interest within the received data packet can be written as follows

$$\mathbf{x}_j = b_j\sqrt{E_0}\mathbf{S}_0\mathbf{c}_0 + \mathbf{z}_j, \quad j = 1, \ldots, J \tag{9.34}$$

where, we recall, $\mathbf{c}_0 \in C^N$ is the vector of the multipath channel coefficients that are assumed to remain constant during the transmission of the data packet (quasi-static fading), and $\mathbf{z}_j$, $j = 1, \ldots, J$, is a sequence of independent identically distributed zero mean Gaussian vectors with *unknown* covariance matrix $\mathbf{R}_z$, that represent comprehensively channel interference and noise that is independent of the data sequence of user of interest, b_j, $j = 1, \ldots, J$. We note that the Gaussian distribution assumption for the general interference plus noise component is only introduced here to facilitate the development of the packet-data receiver. In the simulation studies, the performance of the packet-data receiver will be examined under a realistic communication system setup with non-Gaussian distributed MAI.

The probability density function (pdf) of the observations, $\mathbf{X} \triangleq [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J]$, conditioned on the transmitted bits, $\mathbf{b} = [b_1, b_2, \ldots, b_J]^T$, can be expressed in the following compact form

$$f(\mathbf{X}|\mathbf{b}; E_0, \mathbf{S}_0, \mathbf{c}_0, \mathbf{R}_z) = \frac{1}{\pi^{PJ}|\mathbf{R}_z|^J} e^{\text{trace}\left[-\mathbf{R}_z^{-1}(\mathbf{X}-\sqrt{E_0}\,\mathbf{S}_0\mathbf{c}_0\mathbf{b}^T)(\mathbf{X}-\sqrt{E_0}\,\mathbf{S}_0\mathbf{c}_0\mathbf{b}^T)^H\right]} \tag{9.35}$$

GLRT Detection: Known Channel

In this section, we treat receiver design as a joint optimization problem where receiver parameter estimation is coupled with packet-data detection. In particular, for each hypothesis (information bit combination) we maximize the conditional likelihood with respect to the unknown receiver parameters and then choose the most likely hypothesis. The solution to this optimization problem is a generalized likelihood ratio test (GLRT)-type receiver (it is a likelihood ratio scheme that utilizes maximum likelihood (ML) estimates of the unknown parameters). Formal derivation of the GLRT packet-data detector is provided by the next proposition [20].

Proposition 9.1 *The GLRT test for the detection of the data packet* $\mathbf{b}$ *of size* J *in the presence of complex Gaussian disturbance with unknown covariance matrix* $\mathbf{R}_z$ *is*

$$\hat{\mathbf{b}}_{\text{GLRT}} = \arg\max_{\mathbf{b}}\left\{\max_{\mathbf{R}_z} f(\mathbf{X}|\mathbf{b}, \mathbf{v}_0, \mathbf{R}_z)\right\} = \arg\max_{\mathbf{b}} l_1(\mathbf{b}) \tag{9.36}$$

where

$$
\begin{aligned}
l_1(\mathbf{b}) &\triangleq J\mathbf{b}^T\mathbf{X}^H\hat{\mathbf{R}}^{-1}(J)\mathbf{v}_0 + J\mathbf{v}_0^H\hat{\mathbf{R}}^{-1}(J)\mathbf{X}\mathbf{b} + (\mathbf{b}^T\mathbf{X}^H\hat{\mathbf{R}}^{-1}(J)\mathbf{X}\mathbf{b})(\mathbf{v}_0^H\hat{\mathbf{R}}^{-1}(J)\mathbf{v}_0) \\
&\quad - (\mathbf{b}^T\mathbf{X}^H\hat{\mathbf{R}}^{-1}(J)\mathbf{v}_0)(\mathbf{v}_0^H\hat{\mathbf{R}}(J)\mathbf{X}\mathbf{b})
\end{aligned}
\tag{9.37}
$$

$\mathbf{v}_0 = \sqrt{E_0}\mathbf{S}_0\mathbf{c}_0$ *and* $\hat{\mathbf{R}}(J) \triangleq \frac{1}{J}\mathbf{X}\mathbf{X}^H$ *is the sample average received data correlation matrix.*

We note that direct implementation of test in Equation (9.36) has complexity exponential in the packet size J.

The GLRT test in Equation (9.36) can be contrasted with the standard *desired signal absent* SMI detection scheme summarized below

$$\hat{b}_{j_{\text{SMI-dsa}}} = \text{sgn}[\text{Re}\{\mathbf{v}_0^{\text{H}} \hat{\mathbf{R}}_z^{-1}(K)\mathbf{x}_j\}], \quad j = 1, \ldots, J \tag{9.38}$$

where $\hat{\mathbf{R}}_z(K) \overset{\Delta}{=} \frac{1}{K}\sum_{k=1}^{K} \mathbf{z}_k \mathbf{z}_k^{\text{H}}$ based on pure disturbance observations, $\mathbf{z}_k$, $k = 1, \ldots, K$, that are independent from $\mathbf{x}_j$, $j = 1, \ldots, J$, and the subscript part "dsa" is used to emphasize that the desired-signal-absent SMI detector utilizes additional pure disturbance observations. When pure disturbance observations (secondary data), $\mathbf{z}_k$, $k = 1, \ldots, K$ are not available, a popular version of the test in Equation (9.38) utilizes directly the sample-average correlation matrix of the (desired-signal-present) received data, $\hat{\mathbf{R}}(J)$, evaluated using the *same* received data $\mathbf{x}_j$, $j = 1, \ldots, J$, that are processed by the detector. This results in the widely known SMI receiver,

$$\hat{b}_{j_{\text{SMI}}} = \text{sgn}[\text{Re}\{\mathbf{v}_0^{\text{H}} \hat{\mathbf{R}}^{-1}(J)\mathbf{x}_j\}], \quad j = 1, \ldots, J \tag{9.39}$$

Recent analytical results on short data record adaptive filtering [14,16,19] indicate that for finite sample support of equal size $(K = J)$ the test in Equation (9.38) significantly outperforms the test in Equation (9.39) in terms of BER. Yet, as is shown [20], for sufficiently large transmitted energy per bit, the packet-data GLRT detector in Equation (9.36) can achieve approximately the same average[1] BER performance as the test in Equation (9.38) that utilizes extra $K = J - 1$ independent pure disturbance observations (secondary data), i.e.,

$$\lim_{\gamma \to \infty} \frac{BER_{\text{GLRT}}(J)}{BER_{\text{SMI-dsa}}(J - 1)} \approx 1 \tag{9.40}$$

where $\gamma \overset{\Delta}{=} \mathbf{v}_0^{\text{H}} \mathbf{R}_z^{-1} \mathbf{v}_0$.

Using the above result, we can derive an approximation of the average BER of the coherent GLRT packet-data detector that operates on a data packet of size $J \geqslant P + 2$ as follows [20]

$$BER_{\text{GLRT}}(J) \approx \frac{2}{3}Q\left(\sqrt{2\mu}\right) + \frac{1}{6}Q\left(\sqrt{2\mu + 2\sqrt{3}\sigma}\right) + \frac{1}{6}Q\left(\sqrt{2\mu - 2\sqrt{3}\sigma}\right) \tag{9.41}$$

where $\mu \overset{\Delta}{=} \frac{J - P + 1}{J}\gamma$, and $\sigma^2 \overset{\Delta}{=} \frac{(J - P + 1)(P - 1)}{J^2(J + 1)}\gamma^2$.

GLRT Detection: Unknown Channel

When the channel is unknown, i.e., when E_0 and $\mathbf{c}_0$ in Equation (9.34) are unknown, the GLRT receiver is given by the following proposition [20].

Proposition 9.2 *The GLRT test for the detection of the data packet,* **b**, *of size J transmitted over an unknown linear channel in the presence of complex Gaussian disturbance of unknown covariance matrix* $\mathbf{R}_z$ *is given by*

$$\hat{\mathbf{b}}_{\text{GLRT}} = \arg\max_{\mathbf{b}}\left\{\max_{\mathbf{c}_0, \mathbf{R}_z} f(\mathbf{X}|\mathbf{b}, \mathbf{S}_0, \mathbf{c}_0, \mathbf{R}_z)\right\} = \arg\max_{\mathbf{b}} l_2(\mathbf{b}) \tag{9.42}$$

where

$$l_2(b) = \frac{J\mathbf{b}^{\text{T}}\mathbf{X}^{\text{H}}[\mathbf{X}\mathbf{X}^{\text{H}}]^{-1}\mathbf{S}_0(\mathbf{S}_0^{\text{H}}[\mathbf{X}\mathbf{X}^{\text{H}}]^{-1}\mathbf{S}_0)^{-1}\mathbf{S}_0^{\text{H}}[\mathbf{X}\mathbf{X}^{\text{H}}]^{-1}\mathbf{X}\mathbf{b}}{J^2 - J\mathbf{b}^{\text{T}}\mathbf{X}^{\text{H}}[\mathbf{X}\mathbf{X}^{\text{H}}]^{-1}\mathbf{X}\mathbf{b}} \tag{9.43}$$

---

[1]The average BER of a packet-data detector is defined as the expected number of bits in error divided by the packet size.

We note that the function $l_2(\mathbf{b})$ in Equation (9.43) is ambiguous with respect to the phase of $\mathbf{b}$. In practice, phase ambiguity is resolved either by using a pilot sequence or by employing differential modulation at the transmitter; the rest of this section deals exactly with these two approaches.

Pilot Assisted GLRT Detection

Proposition 9.3 *Let $\{b_j\}_{j=1}^{J_p}$ and $\{b_i\}_{i=J_p+1}^{J}$ denote, respectively, J_p known pilot bits and $J - J_p$ unknown information bits within the data packet, $\mathbf{b}$, of size J that is transmitted over an unknown linear channel in the presence of complex Gaussian disturbance of unknown covariance. Then, the pilot assisted GLRT detector of $\{b_i\}_{j=J_p+1}^{J}$ is given by*

$$\{\hat{b}_{i_{\mathrm{GLRT}}}\}_{i=J_p+1}^{J} = arg \max_{b_i, i \geq J_p+1} l_2(\mathbf{b}) \tag{9.44}$$

It is interesting to note that in Equation (9.44) the pilot sequence $\{b_i\}_{i=1}^{J_p}$ is not used to directly estimate the phase in an explicit manner, but is incorporated implicitly in the GLRT rule. It is also interesting to observe that the GLRT test expression in Equation (9.44) maintains the same structure as in Equation (9.43). We conclude the treatment of the pilot assisted GLRT detection by providing a closed form approximation of the BER performance of the above pilot assisted GLRT detector for a data packet of size $J \geqslant P + 3$ [20]

$$BER_{\mathrm{GLRT-pilot}}(J) \approx BER_{\mathrm{SMI-dsa}}(J-2) \tag{9.45}$$

$$\approx \frac{2}{3} Q\left(\sqrt{2\mu}\right) + \frac{1}{6} Q\left(\sqrt{2\mu + 2\sqrt{3}\sigma}\right) + \frac{1}{6} Q\left(\sqrt{2\mu - 2\sqrt{3}\sigma}\right) \tag{9.46}$$

where $\mu \overset{\Delta}{=} \frac{J-P}{J}\gamma$, $\sigma^2 \overset{\Delta}{=} \frac{(J-P)(J-1)}{J^3}\gamma^2$, and $BER_{\mathrm{SMI-dsa}}(J-2)$ is the BER of the *coherent* desired-signal-absent SMI detector in Equation (9.38) that would require perfect knowledge of $\mathbf{c}_0$ and utilize $K = J - 2$ independent pure disturbance observations.

DPSK GLRT Detection

As an alternative to pilot signaling, phase ambiguity of the GLRT detector in Equation (9.43) can be resolved by employing differential encoding at the transmitter. To avoid redundancy in our presentation, in this section we keep the size of the transmitted packet equal to, J while the number of the information bits embedded in the differentially encoded packet is $J - 1$, $\{b_j\}_{j=1}^{J-1}$. The differentially encoded bits themselves are $e_0 = +1$ and $e_j = e_{j-1}b_j$, $j = 1, 2, \ldots, J-1$. The jth received vector, $\mathbf{x}_j$, is still of the form of Equation (9.34) with e_j in place of b_j. Given the transmitted bits, e_j, $j = 0, 1 \ldots J - 1$, the information bits can be uniquely determined by $b_j = e_{j-1}e_j$, $j = 1, \ldots, J-1$.

Under unknown input statistics and channel coefficients, the common approach has been to produce estimates of the unknown quantities and insert the estimates in the J-symbol (or 2-symbol) block differential detector. Instead, what we propose in this section is a GLRT-type scheme that combines into a single optimization effort estimation of interference-plus-noise covariance matrix and channel coefficients and detection of packet data. The following proposition identifies the DPSK GLRT packet-data detector [20].

Proposition 9.4 *The DPSK GLRT detector of differentially encoded packet data, $\{b_j\}_{j=1}^{J-1}$, transmitted over an unknown linear channel in the presence of complex Gaussian disturbance of unknown covariance is given by*

$$\{\hat{e}_{j_{\mathrm{GLRT}}}\}_{j=1}^{J-1} = \arg \max_{e_j, j \geqslant 1} l_2(\mathrm{e}) \tag{9.47}$$

$$\hat{b}_{j_{\mathrm{GLRT}}} = \hat{e}_{j-1_{\mathrm{GLRT}}}\hat{e}_{j_{\mathrm{GLRT}}}, \quad j = 1, 2, \ldots, J - 1 \tag{9.48}$$

where $\mathbf{e} \overset{\Delta}{=} [e_0, \ldots e_{J-1}]^T$.

Finally, we provide the following closed form approximation of the BER performance of the above DPSK GLRT detector for a data packet of size $J \geqslant P + 3$ derived in Ref. [20]

$$BER_{\text{GLRT-DPSK}}(J) \approx BER_{\text{SMI-dsa,DPSK}}(J-2) \tag{9.49}$$

$$\approx \frac{4}{3}Q\left(\sqrt{2\mu}\right) + \frac{1}{3}Q\left(\sqrt{2\mu + 2\sqrt{3}\sigma}\right) + \frac{1}{3}Q\left(\sqrt{2\mu - 2\sqrt{3}\sigma}\right) \tag{9.50}$$

where $\mu \triangleq \frac{J-P}{J}\gamma$, $\sigma^2 \triangleq \frac{(J-P)(J-1)}{J^3}\gamma^2$, and $BER_{\text{SMI-dsa,DPSK}}(J-2)$ is the BER of a detection scheme that utilizes the *coherent* desired-signal-absent SMI detector of $\{e_j\}_{j=1}^J$ based on $K = J - 2$ independent pure disturbance observations under perfect knowledge of $\mathbf{c}_0$, followed by the differential decoder $\hat{b}_j = \hat{e}_{j-1}\hat{e}_j$, $j = 1, \dots, J-1$.

Implementation Issues

Direct implementation of the GLRT detectors has complexity exponential in the number of bits. Below, we consider a procedure to obtain effective suboptimum implementations with linear complexity. The procedure is basically a parallel implementation of T sequential bit updates with D steps for each sequential bit update. The procedure is summarized below: We start with an initial estimate of the packet bits, $\hat{\mathbf{b}}(0) = [\hat{b}_1(0), \dots, \hat{b}_J(0)]$, replicated T times for T distinct sequential bit updates, $\hat{\mathbf{b}}^{(t)}(d)$, $t = 1, 2, \dots, T$, $d = 0, 1, \dots, D$. With each sequential bit update, t, we associate a bit update index order, $\{\pi_t(j)\}_{j=1}^J$, that is a distinct permutation of $\{1, 2, \dots, J\}$. At each step, d, we check one bit in each of T sequential bit updates, namely $\hat{b}_{\pi_t(d \bmod J)}^{(t)}$. Upon completion, step, D, we declare as our approximate GLRT decision the most likely among the T bit vectors $\hat{b}^{(t)}(D)$, $t = 1, \dots, T$.

Suboptimum GLRT Algorithm

Initialization: Number of parallel sequential bit updates, T; search depth D;
                 Initial decision vector $\hat{\mathbf{b}}^{(t)}(0) := [\hat{b}_1(0), \hat{b}_2(0), \dots, \hat{b}_J(0)]^T$, $t = 1, 2, \dots, T$;
                 Bit update index orders, $\{\pi_t(j)\}_{j=1}^J$, $t = 1, 2, \dots, T$.
For step $d = 1, 2, \dots D$
       For path $t = 1, 2, \dots, T$
            $i := \pi_t(d \bmod J)$
            $\hat{b}_i^{(t)}(d) := \arg\max_{b_i^{(t)}}\left\{\max_{\mathbf{R}_z} f(\mathbf{X}|\{\hat{b}_j^{(t)}(d-1)\}_{j\neq i}, b_i^{(t)}, \mathbf{v}_0, \mathbf{R}_z)\right\}$
            $\hat{b}_j^{(t)}(d) := \hat{b}_j^{(t)}(d-1), \quad j \neq i$
       end
end
$\hat{\mathbf{b}}_{\overline{\text{GLRT}}} := \arg\max_{\mathbf{b}\in\{\hat{b}^{(1)}(D),\dots,\hat{b}^{(T)}(D)\}} l_{1,2}(\mathbf{b}).$

The complexity of one bit update is of order, $O(P)$. Hence, the overall complexity of the above algorithm for the detection of a data packet of size J is of order $O(JP^2) + O(JP) + O(P^3) + O(DTP)$, which includes the cost of the initial evaluation of $\hat{\mathbf{R}}(J)$ and $\hat{\mathbf{R}}^{-1}(J)$. We note that a good initial estimate may allow relatively small values for T and D. In this context, we can regularly reinitialize the parallel search algorithm using the best sequence estimate among the current T alternatives.

Simulation Studies

In this section, we examined the performance of the proposed GLRT schemes for a packet-data DS-CDMA communication system.[1] At all times, the GLRT detectors are implemented via the linear complexity

---

[1]The combined effect of DS-CDMA multiple access interference (MAI) and AWGN is Gaussian-mixture distributed and not plain Gaussian. It is interesting to examine how the proposed GLRT detectors perform in such an environment.

algorithm with $T = 16$ and $D = 6J$. Initial bit estimates are either taken by conventional matched-filter (MF) outputs (Case Study 9.1) or set arbitrarily (Case Studies 9.2 and 9.3).

DS-CDMA Case Study 9.1 Synchronous Multiuser System and Single-Path Channel

We consider a system with 10 synchronous users with Gold signatures of length, $L = 31$. We select a "user of interest" and have the SNR's of the interfering users fixed in the range, [6dB,11dB]. In this study, we assume exact knowledge of the channel of the user of interest. We compare the BER of the GLRT detector with the BER of the MF, SMI, LMS (step size 10^{-4}), RLS (initialization parameter 100), and the desired-signal-absent SMI detector in Equation (9.38) that assumes availability of $J - 1$ additional pure disturbance observations. In Figure 9.15 and Figure 9.16, we plot the BER as a function of the SNR of the user of interest and the packet size J, respectively. In view of the nearly overlapping analytical and simulated GLRT BER curves, we may claim that our linear cost GLRT implementation performs very close to full GLRT and Equation (9.41) provides an accurate approximation of the BER of the GLRT packet-data detector. Furthermore, the GLRT packet-data
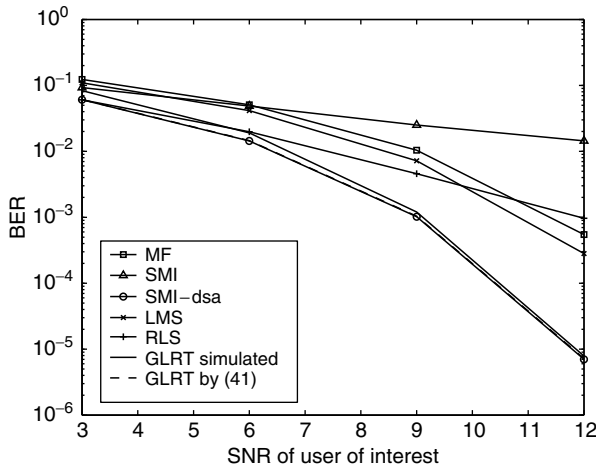
FIGURE 9.15 Case Study 9.1: BER as function of the SNR of the user of interest ($J = 127$).
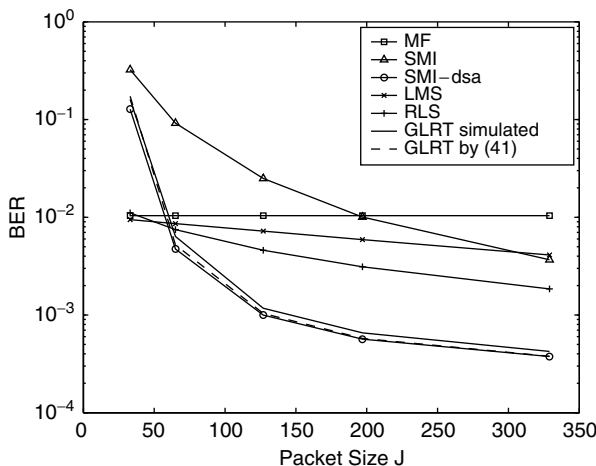
FIGURE 9.16 Case Study 9.1: BER as function of the packet size J. The SNR of the user of interest is fixed at 9 dB.

detector outperforms significantly the SMI, LMS, and RLS detectors and performs nearly the same as the desired-signal-absent SMI detector in Equation (9.38) that requires $J - 1$ *additional* pure disturbance observations.

DS-CDMA Case Study 9.2   Asynchronous Multipath Fading Channel: Pilot-Assisted Signaling

We consider the same setup as in Case Study 9.1, except that now the users transmit asynchronously and each user channel has 3 resolvable paths. The path coefficients are modeled as independent complex Gaussian random variables all of unit variance. The length of the pilot sequence is fixed at $J_p = 10$. We compare the BER of the GLRT detector with the BER of the RAKE-MF, the desired-signal-absent SMI and SMI detectors in Equation (9.38) and Equation (9.39), and the LMS and RLS detectors. We note that in this study the GLRT detector assumes no knowledge of the channel *while all other detectors assume exact knowledge of the channel*. In addition, the desired-signal-absent SMI detector in Equation (9.38) uses $J - 2$ *extra* pure disturbance observations that are assumed to be available. It is worth noting that the pilot sequence is incorporated and processed internally and elegantly by the GLRT algorithm without the need for a separate phase estimation stage.

DS-CDMA Case Study 9.3   Asynchronous Multipath Fading Channel: DPSK Signaling

We consider the same setup as in Case Study 9.2, except that the transmitter now uses DPSK encoding instead of pilot signaling; hence, at the receiver end a differential decoder is needed to recover the information bits. We compare the BER of our DPSK GLRT detector with the BER of the DPSK version of the following detectors: RAKE-MF, SMI, LMS, RLS, and ideal MMSE. A 2-symbol differential decoder is used in all cases. The *coherent* desired-signal-absent SMI detector is also included as a reference. We note that the DPSK GLRT detector assumes no knowledge of the channel, while the RAKE-MF, LMS, RLS, SMI, and ideal MMSE detectors assume perfectly known path coefficients *up to an unknown phase* (phase ambiguity is resolved by differential encoding/decoding). In addition, the ideal MMSE detector assumes perfectly known interference-plus-noise covariance matrix, and the coherent desired-signal-absent SMI detector requires perfect knowledge of the path coefficients (including the phase) and $J - 2$ additional pure disturbance observations. In Figure 9.18, we repeat the studies of Figure 9.17. We note that for data packets of size $J = 250$ and higher the DPSK GLRT detector outperforms even the ideal MMSE (2-symbol decoder) detector.
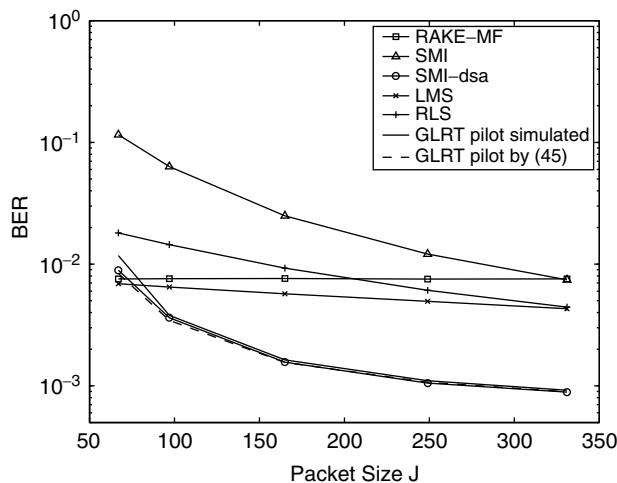
FIGURE 9.17   Case Study 9.2: BER as function of the packet size J. The SNR of the user of interest is fixed at 9 dB.
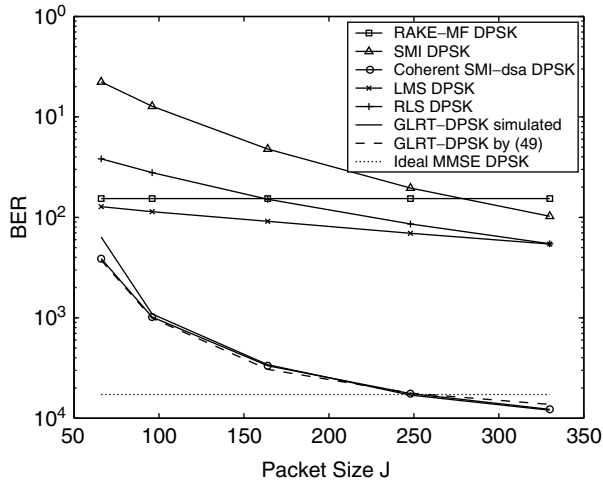
FIGURE 9.18 Case Study 9.3: BER as function of the packet size *J*. The SNR of the user of interest is fixed at 9 dB.

9.7   Concluding Remarks

Wireless cellular and personal communications services (PCS) networks experienced significant growth in the past few years, 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 (for example, code-division-multiple-access with adaptive antenna arrays). 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).

   In this chapter, we first focused on packet data receivers that are implemented by using sample average estimates in place of ideal statistics in the optimum receiver formula and perform disjoint parameter estimation and packet-data detection. In particular, we examined linear MMSE/MVDR-type receivers. We presented two alternative methods that approximate the optimum solution under perfectly known input statistics (input autocorrelation matrix and input/desired-output cross-correlation vector): The generalized sidelobe canceller, and the auxiliary-vector filters. When the input statistics are unknown and estimated, these 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.

   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 quasi-static 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, and SMI filter estimators. In addition, the troublesome, data-dependent tuning of the real-valued LMS learning gain parameter, or the RLS initialization parameter is replaced by an integer

choice among the first several members of the estimator sequence. Thus, we presented two data-driven criteria for the selection of the best AV filter estimator in the sequence.

Next, we considered the GLRT-type detection schemes that perform joint estimation of the unknown system parameters and detection of the packet-data. In particular, for the known channel case, we developed a coherent GLRT detector, while for the unknown channel case we derived a pilot assisted GLRT detector (the pilot signal is implicitly used to resolve phase ambiguity) and a DPSK GLRT detector. We also derived analytical expressions for the BER performance of each proposed GLRT-type scheme, and compared it to the corresponding conventional estimate-and-plug-in detector that replaces unknown parameters in its ideal formula by estimates obtained through a separate estimation process. In all cases, the GLRT schemes maintained the same elegant core structure, regardless of known or unknown channels and pilot or DPSK signaling. Finally, we developed suboptimum implementations of the GLRT packet-data detectors that exhibit linear complexity in the packet size.

Acknowledgments

Defining Terms

Adaptive filter: A self-designing filter that recalculates itself from data to follow changes in the statistics of the environment.

Antenna array: An electromagnetic wave receiving system composed of several antennae that is capable of directional reception/transmission.

Auxiliary-vector filters: A sequence of linear filters that approximate the optimum MMSE/MVDR linear filter.

Auxiliary-vector estimates: A sequence of estimates of the optimum MMSE/MVDR linear filter that are generated by a conditional statistical optimization procedure and are built as a linear combination of properly selected nonorthogonal auxiliary vectors. Early, nonasymptotic elements of this sequence exhibit superior estimation performance under low sample support conditions.

Biased estimator: An estimator whose mean value is not equal to the true parameter value.

Code-division-multiple-access: A physical layer multiple access technique that allows many user signals to coexist in time and frequency and, yet, be recovered/separated reliably as necessary. This is achieved by assigning a unique code sequence to each user signal, known as signature.

Cross-validation: A statistical technique that validates a model on a data set different from the one used for estimation of model parameters.

Generalized likelihood ratio test: A likelihood ratio test that utilizes maximum-likelihood estimates (MLEs) of the unknown parameters.

***J*-divergence:**   A distance measure between two probability distributions.

Linear MMSE filter: A linear filter that minimizes the mean-square-error between the desired response and the actual filter output.

Linear MVDR filter: A linear filter that minimizes the mean filter output power subject to the constraint that the filter maintains a certain response with respect to a given input.

Small-sample-support estimate: A parameter estimate that is calculated based on a data record of small (nonasymptotic) size.

Smart antenna: An antenna array with an advanced post-reception signal processing circuitry that intelligently suppresses or cancels unwanted signals.

References

1.   S.N. Batalama, "Packet-rate adaptive receivers for mobile communications," in *Wiley Encyclopedia of Telecommunications*, vol. 4, J. Proakis, Ed., New York: Wiley, 2003, pp. 1886–1905.

2. J. Proakis, *Digital Communications*, 3rd ed., New York: McGraw-Hill, 1995.

3. E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, "UMTS/IMT-2000 based on wideband CDMA," *IEEE Commun. Mag.*, vol. 36, pp. 70–80, 1998.

4. D.A. Pados and S.N. Batalama, "Joint space-time auxiliary-vector filtering for DS/CDMA systems with antenna arrays," *IEEE Trans. Commun.*, vol. 47, pp. 1406–1415, 1999.

5. S. Haykin, *Adaptive Filter Theory*, 2nd ed., Englewood Cliffs, N.J.: Prentice Hall, 1991.

6. G.H. Golub and C.F. Van Loan, *Matrix Computations*, Baltimore, MD: The Johns Hopkins University Press, 1990.

7. K.A. Byerly and R.A. Roberts, "Output power based partially adaptive array design," in *Proc. Asilomar Conf. Signal. Syst. Comput.*, CA: Pacific Grove, 1989, pp. 576–580.

8. S.N. Batalama, M.J. Medley, and D.A. Pados, "Robust adaptive recovery of spread-spectrum signals with short data records," *IEEE Trans. Commun.*, vol. 48, pp. 1725–1731, 2000.

9. D.A. Pados and G.N. Karystinos, "An iterative algorithm for the computation of the MVDR filter," *IEEE Trans. Signal Process.*, vol. 49, pp. 290–300, 2001.

10. D.A. Pados and S.N. Batalama, "Low-complexity blind detection of DS/CDMA signals: auxiliary-vector receivers," *IEEE Trans. Commun.*, vol. 45, pp. 1586–1594, 1997.

11. D.A. Pados, T. Tsao, J.H. Michels, and M.C. Wicks, "Joint domain space-time adaptive processing with small training data sets," in *Proc. IEEE Radar Conf.*, Dallas, TX, May 1998, pp. 99–104.

12. 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.*, vol. 51, pp. 1700–1708, 2003.

13. H.V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Trans. Info. Theory*, vol. 43, pp. 858–871, 1997.

14. I.N. Psaromiligkos and S.N. Batalama, "Recursive short-data-record estimation of AV and MMSE/MVDR linear filters for DS-CDMA antenna array systems," *IEEE Trans. Commun.*, vol. 52, pp. 136–148, 2004.

15. 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.*, vol. 51, pp. 983–994, 2003.

16. 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.*, vol. 52, pp. 1538–1546, 2004.

17. P. Chaudhury, W. Mohr, and S. Onoe, "The 3GPP proposal for IMT-2000," *IEEE Commun. Mag.*, vol. 37, pp. 72–81, 1999.

18. I.N. Psaromiligkos and S.N. Batalama, "Blind self-synchronized demodulation of DS-CDMA communications," in *Proc. IEEE ICC 2000 — Int. Conf. Commun.*, New Orleans, LA, June 2000, pp. 2557–2560.

19. J.M. Farrell, I.N. Psaromiligkos, and S.N. Batalama, "Design and analysis of supervised and decision-directed estimators of the MMSE/LCMV filter in data-limited environments," in *Proc. SPIE, Dig. Wireless Commun. Conf.*, vol. 5100, Orlando, FL, 2003, pp. 227–237.

20. H. Qian, S.N. Batalama, and B. Suter, "Novel GLRT packet-data detectors," in *Proc. IEEE ICASSP 2004 — Int. Conf. Acoust. Speech Signal Process.*, Montreal, Canada, May 2004.

21. I.S. Reed, J.D. Mallet, and L.E. Brennan, "Rapid convergence rate in adaptive arrays," *IEEE Trans. Aerospace Electron. Syst.*, vol. 10, pp. 853–863, 1974.

22. J.S. Goldstein and I.S. Reed, "Reduced-rank adaptive filtering," *IEEE Trans. Signal Process.*, vol. 45, pp. 492–496, 1997.

Further Information

IEEE Transactions on Information Theory is a bimonthly journal that publishes papers on theoretical aspects of estimation theory and in particular on transmission, processing, and utilization of information.

IEEE Transactions on Signal Processing is a monthly journal which presents applications of estimation theory to speech recognition and processing, acoustical signal processing, and communication.

IEEE Transactions on Communications is a monthly journal presenting applications of estimation theory to data communication problems, synchronization of communication systems, and channel equalization.

IEEE Transactions on Aerospace and Electronic Systems is a quarterly journal presenting developments in sensor systems, communications systems, command and control centers, avionics, space systems, military systems, and digital signal processing simulators.

10

Bandwidth Efficient Modulation in Optical Communications

Moncef B. Tayahi
University of Nevada, Reno

Banmali S. Rawat
University of Nevada, Reno

10.1 Introduction

The explosive growth of data, particularly Internet traffic, has led to a dramatic increase in demand for transmission bandwidth imposing an immediate requirement for expanding the capacity current networks. An additional driving force for higher capacity, enhanced functionality, and flexibility in optical networks is the increased trend in interactive exchange of data and multimedia files. Due to the unpredictable and ever growing size of data files and messages exchanged over global distances, the future communication grid must be agile in time and able to react rapidly to support end-to-end bandwidth requirements for transmission of files of any conceivable size.

Telecommunication networks currently widely deploy wavelength division multiplexing (WDM) in single mode optical fibers to interconnect discrete network locations, and offer high capacity and long range transmission capabilities. The presence of dark fiber in existing networks can accommodate a large percentage of the capacity requirements, however novel solutions are needed. Equipping the already installed equipment with conventional technologies is not the most cost-effective solution [1].

The recent advances in optical technologies have had a significant impact on optical networking solutions deployed world wide. These interconnections lead to a web of optical fibers that connect the globe, offering various services to the end users. However, further optimization of the existing solutions involves not only the physical implementation, but also software control and network management. Network providers have to address the continuous evolution of services and applications that are becoming available to the users in a

resilient and secure manner. Technical breakthroughs are expected to further accelerate the realization of transparent optical networks to offer increased transmission bandwidth, switching capabilities, and optical signal processing functionalities. This progress is not only expected in the core networks, but also in the metropolitan area and the access networks to provide scaleable, transparent, and flexible end-to-end solutions [2]. These features are offering full access to the global information network for all with reliable system performance and at reduced cost.

Because of the rapid growth of capacity requirement on current transmission systems, fiber-optic technology is advancing into high data rate per channel, and multiplexing various channels in one single fiber. The capacity of a single optical fiber has increased from a single OC-3 signal, transmitting at rate of 155 Mb/s, to terabits capacity. In order to maximize the system capacity and minimize performance degradations caused by transmission impairments, careful engineering rules have to be crafted before networks deployment. Signal modulation format is a key factor to be considered in the initial system design. Adopted modulation format determines the signal quality, signal tolerance to transmission impairments, system capacity, total system spectral efficiency, and total cost. Until not long ago, non-return-to-zero (NRZ), also known as On and Off Keying (OOK), had been the dominant modulation format of choice in intensity-modulation and direct-detection (IM/DD) fiber-optic systems. Recent needs for advanced modulation formats are motivated by the demand of systems with high capacity, better overall reliability, optimum operation conditions, and low operating cost. From an information theory point of view, a variety of signal modulation formats had been studied extensively in communications. In contrast to microwave transmission and wireless communication systems known by low data rates, fiber-optic systems have their unique properties for supporting large capacity and long range transmission systems. Although there is no magic modulation format immune to performance degradations, the choice of a suitable modulation format in fiber optic systems depends on many factors such as fiber types, per-channel data rate and aggregate link rate, wavelength spacing, system reach, and so on.

10.2   Bandwidth Efficient Modulation (BEM) Concept

Multiplexing Limitations

In order to take advantage of the finite bandwidth of state of the art optical fiber more efficiently, new multiplexing techniques have been adopted from other applications; such techniques are Time Division Multiplexing (TDM), Wavelength Division Multiplexing (WDM), Frequency Division Multiplexing (FDM), and their combinations. Apart from noise accumulation, high-speed multiplexed signals suffer from chromatic dispersion, nonlinear crosstalk, and polarization mode dispersion (PMD). In multi-wavelength DWDM optical systems with relatively high data rate per wavelength, inter-channel crosstalk originated by fiber nonlinearity such as cross-phase modulation (XPM) and four-wave mixing (FWM), can be limiting factors. Optical systems with data rates of 10 Gb/s and higher require precise dispersion compensation and careful engineering of a suitable dispersion map. However, due to the limitations in the wavelength stability of semiconductor lasers and the limited selectivity of optical filters, the minimum channel spacing is currently limited to ∼25 GHz in commercial WDM systems [3]. Although, optical filter designs have improved over the years, we are reaching physical limits that can only be overcome by new physics and novel material breakthroughs. Further improvement in the optical spectrum density can be achieved with bandwidth efficiency modulation techniques [4].

Nyquist Theoretical Minimum Bandwidth Requirement

The Nyquist minimum bandwidth limit requires specifying the transmitted pulse shape so that no Intersymbol Interference (ISI) will take place at the receiver. The requirement states that the theoretical minimum system bandwidth needed to avoid ISI is half the signal frequency [5]. Therefore the Nyquist channel for zero ISI is rectangular in shape and its impulse response is a sinusoidal function. Nyquist established that each received signal that has sinusoidal pulse shape is ISI free and can be called an ideal Nyquist signal. Therefore, when the sampling time is perfect, there will be no ISI penalty induced. For baseband transmission, the required system bandwidth is half the pulse period without ISI. From this

assumption, the transmission rate per hertz is two symbols per second per hertz (Sym/s/Hz) that can be obtained at best. It is clear from this requirement that achieving a matching filter with a perfect rectangular shape is not realizable, and only a quasi shape can be targeted.

The benefit of multilevel signaling can be seen in the following example. A fundamental figure of merit for communication systems is the bandwidth efficiency whose unit is bits/s/Hz. The Nyquist limit is set at 2 Sym/s/Hz, although a symbol may have more than one bit per symbol. For example, consider M-ary Quadrature Amplitude Modulation (QAM-XX) signal, the value for XX is equal to 2^k where k is the number of bits symbol. A 10 Gb/s data stream occupies a minimum of 20 GHz of the optical band when NRZ modulation is used, however, when QAM-256 is used the optical signal will only occupy 2.5 GHz; the 17.5 GHz spared from the spectrum can be populated by other channels. Theoretically, k can be as high as 100, however practical value for k in optical transmission systems has not exceeded 10 yet. Higher-order modulation for bandwidth efficiency is becoming more attractive today because of available hardware for pulse shaping to achieve optimum spectral efficiency, the availability of low cost forward-error-correction (FEC) with low overheads, and the advances in RF signal multiplexing and demultiplexing techniques.

Sampling Amplitude Modulated Signals

Sampling is a process that allows modern communication systems to carry thousands of simultaneous signals within a limited bandwidth. Sampling theory is fundamental to the understanding of signals in analog and digital communications; it describes the conversion of signals from the continuous-time (analog) form to a discrete-time form (digital). In practice, discrete-time signals are coded into a numeric form and transmitted as a digital signal or stored digitally in a computer memory. The sampling theorem dictates that the sampling rate must be at least twice the frequency of the highest frequency component of the signal being sampled.

Digital systems must break up the signal in the time domain, a process known as time sampling or time discretization. This is a separate process from quantization, which is the breaking of signals in various levels. The discrete information on a compact disc, for example, is a series of samples of continuous audio signals that may have been generated by conventional instrument or a singer's voice. The sampling process is used to cut down excess information storage achieved by a process called time discretization.

There are several features to note: digital signals are only non-zero at certain times, they actually spend most of the time being zero. Analog signals would normally be denoted $x(t)$, the (t) denoting that the signal x varies continuously with time t. The digital signal would normally be denoted x_n, the subscript n being the sample number. As the subscript n can only be an integer (0, 1, 2, 3...), this means that x_n only has values at discrete times. This nomenclature is important to note as it is used throughout the industry [6]. Time samples are evenly spaced for the digital signal (the time space between adjacent samples is constant). This time spacing is called the sampling interval and will be denoted Δt. The sampling frequency f_s is related to the sampling interval Δt as $f_s = 1/\Delta t$. Although the digital waveform should be drawn as a histogram for most real signals, a continuous line would be drawn through the top of the histogram to produce a continuous waveform. The digitization can be viewed as a pulse train amplitude modulating an all positive version of the continuous signal, hence the term Pulse Code Modulation (PCM).

Shannon–Hartley Capacity Theorem

The Shannon–Hartley capacity theorem describes the maximum possible efficiency of error correcting methods versus levels of noise interference and data corruption [7]. The theory does not describe how to construct the error-correcting method; it only tells us how good the best possible method can be. Shannon's theorem has wide-ranging applications in both communications and data storage applications. This theorem is the foundation of the modern field of information theory.

In the communication domain, Shannon's theorem is also known as the Shannon limit or Shannon capacity. The maximum rate of clean data C that can be sent through an analog communication channel subject to Gaussian-distribution noise interference is given by

$$C \leq W \log_2 (1 + S/N) \tag{10.1}$$

where C is the post-correction effective channel capacity in bits per second, W is raw channel capacity in hertz, and S/N is the signal-to-noise ratio of the communication signal to the Gaussian noise interference expressed as a straight power ratio (not as decibels).

Simple schemes such as "send the message 3 times and use a best 2 out of 3 voting scheme if the copies differ" are inefficient users of bandwidth and thus are far from the Shannon limit. Advanced techniques such as Reed–Solomon codes and, more recently, Turbo codes come much closer to reaching the theoretical Shannon limit, but at a cost of high computational complexity.

The approach of using bandwidth efficient modulation is not a new concept even in modern fiber optics transmission networks, however it has been reconsidered recently due to the availability of low overhead forward error correction modules (FEC) that are capable of reducing the required signal to noise ratio for a given link. With an overhead of only 7%, FEC can provide more then 9 dB coding gain [8]. This coding gain can be used to offset the OSNR penalty caused by multilevel signaling.

Shannon's Limit

The importance of Shannon's limit can be explained using the sphere-packing bound technique [9]. The sphere-packing bound technique can be used to calculate the $\dfrac{Eb}{N_o}$ needed to achieve a given word error probability P_w for a code having information length k bits, coded block length n symbols, and rate $R = k/n$. The above quantities are expressed by the following formula

$$P_w \geqslant Q_n(\theta s_n, A) \tag{10.2}$$

where $A = \sqrt{2RE_b/N_o}$ and θs_n is the solution of the following (solid angle) equation:

$$\Omega_n(\theta s_n, A) = \int_0^{\theta s_n} \frac{n-1}{n} \frac{\Gamma\left(\dfrac{n}{2}+1\right)}{\Gamma\left(\dfrac{n+1}{2}\right)} \sin(\phi)^{n-2} \mathrm{d}\phi = \frac{1}{2^{Rn}} \tag{10.3}$$

Given R, n, and P_w, the $\dfrac{Eb}{N_o}$ is obtained by solving the following equation for A

$$Q_n(\theta s_n, A) = \int_{\theta s_n}^{\pi} \frac{(n-1)\sin(\phi)^{n-2}}{2^{n/2}\sqrt{\pi}\Gamma\left(\dfrac{n+1}{2}\right)} \int_0^{\infty} s^{n-1} e^{-(s^2+nA^2-2s\sqrt{n}A\cos(\phi))/2} \mathrm{d}s \mathrm{d}\phi \tag{10.4}$$

It is possible to directly obtain numerical solutions for the above equations provided that n is small, say less than 100. Larger values of n produce numerical overflow and underflow. In order to minimize these problems, a recursive relationship has been derived for the inner integral in Equation (10.4). Let's define

$$J(n, x, \phi) = \int_0^{\infty} \frac{n-1}{2^{n/2}\sqrt{\pi}\Gamma\left(\dfrac{n+dx}{2}\right)} s^{n-1} e^{-(s^2-2x\cos(\phi)s+x^2)/2} \tag{10.5}$$

so that Equation (10.4) becomes

$$Q_n(\theta s_n, A) = \int_{\theta s_n}^{\pi} \sin(\phi)^{n-2} J(n, \sqrt{n}A, \phi) \mathrm{d}\phi = P_w \tag{10.6}$$

A recursive relationship can be obtained for J as shown below

$$J(n+2, x, \phi) = \int_0^\infty \frac{n+1}{2^{(n+2)/2}\sqrt{\pi}\Gamma\left(\frac{n+dx}{2}\right)} s^n(s - x\cos(\phi))e^{-(s^2-2x\cos(\phi)s+x^2)/2}$$
$$+ \int_0^\infty \frac{n+1}{2^{(n+2)/2}\sqrt{\pi}\Gamma\left(\frac{n+dx}{2}\right)} s^n x \cos(\phi)e^{-(s^2-2x\cos(\phi)s+x^2)/2} \qquad (10.7)$$

Upon integrating the first term by parts it is readily shown that

$$J(n+2, x, \phi) = \int_0^\infty \frac{n(n+1)}{2^{(n+2)/2}\sqrt{\pi}\Gamma\left(\frac{n+dx}{2}\right)} s^{n-1} e^{-(s^2-2x\cos(\phi)s+x^2)/2}$$
$$+ x\cos(\phi)\int_0^\infty \frac{n+1}{2^{(n+2)/2}\sqrt{\pi}\Gamma\left(\frac{n+dx}{2}\right)} s^n e^{-(s^2-2x\cos(\phi)s+x^2)/2} \qquad (10.8)$$

Hence

$$J(n+2, x, \phi) = \frac{\sqrt{2}x\cos(\phi)}{n} \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \quad J(n+1, x, \phi) + \frac{n}{n-1}J(n, x, \phi) \qquad (10.9)$$

This equation has been used to calculate the sphere packing bound exactly for an information block size k up to 1000, for different code rates, and P_w. For values of k higher than 1000 the bound has been evaluated by using Shannon's approximation as presented in Ref. [10]. A comparison between the exact solution and the approximation indicates $\frac{Eb}{N_o}$ errors of less than 0.01 dB for block length codes greater than 50.

Forward Error Correction (FEC)

One of the main reasons that BEM in optical transmission systems is getting serious attention is the availability of Forward Error Correction (FEC). The limitation factors for closely spaced levels in multilevel coding can be overcome with FEC coding gain to enable a 70% increase in the number of channels or a 60% increase in the transmission distance [11]. Additionally, FEC allows an improvement in the Quality of Service (QoS) by guaranteeing a received bit-error rate (BER) better than 10^{-15}. FEC coding relies on Reed-Solomon algorithms to add redundancy bits to the data stream, enabling the identification and correction of corrupted bits [12]. These added redundant bits take the optical carrier (OC)–192 data rate from 9.953 Gb/s to 10.709 Gb/s that is an overhead of only 7% to yield a minimum of 9 dB improvement in optical signal to noise ratio (OSNR) margin. This FEC related OSNR improvement allows for an increase in channel capacity and/or transmission distance.

Super FEC is an enhancement of Reed-Solomon FEC and adds additional gain to an optical link, typically 8 dB of gain rather than 5 or 6, extending the reach of 40 Gb/s systems.

FEC has some interesting associated benefits as well, which may play out as the market evolves to accommodate non-standard transmission formats such as 10 Gb/s Ethernet and video. Since FEC is a kind of "wrapper" around any optical signal, it can be used as a link performance monitor for any optically transmitted signal.

To further enhance performance, FEC supported transponders utilize optimized threshold crossing control in the receiver to set the decision circuit threshold to the optimum decision level in the received data; when

multiple traces of a random data stream are superimposed on top of each other, the 0s and 1s form an eye diagram of the modulated waveform. The more open the eye, the more reliably the 0s and 1s will be detected and the better the BER. However, amplitude noise from optical amplifiers, phase noise, dispersion effects, and interference resulting from conversion of phase into amplitude modulation start to close the eye. As the eye closes, the decision circuit that determines if a bit is a 0 or a 1 gives fewer errors if the decision threshold level can be adaptively adjusted to an optimum threshold decision level [13,14]. The optical receiver with FEC line extender modules and features such as adaptive threshold crossing control result in an improved received optical signal to noise ratio error-free transmission.

Bandwidth Efficiency Consideration in Optical Systems

Among many factors to be considered when evaluating the overall bandwidth efficiency of a modulation format and the most obvious one is the channel spectrum efficiency or channel loading. Channel loading is usually expressed as function of the number of bits-per-symbol transmitted and the amount of frequency guard band required to keep one channel from interfering with another. In the literature the notations used for bit rate and symbol rate sometimes have different meanings [15]. The information bit rate R_b and the symbol rate R_s refer to the channel bit and symbol rates before and after the modulator. The notation used to denote the coded symbol rate is measured at the input of the modulator. If no error correction coding or formatting is used, then R_s is equal to the information bit rate R_b. Likewise, T_b is the bit period and T_s is the symbol period. In the absence of error correction coding or formatting, $T_s = T_b$.

For bandwidth efficiency illustration, M-ary Phase Shift Keying (MPSK) has a bit loading of k bits-per-symbol; it requires a frequency guard band of 25% of its real band. As a result, the bandwidth efficiency of a MPSK modulated channel is Log k bits/sec/Hz [16]. This technique increases channel loading by increasing the number of bits per symbol transmitted in a given symbol period. Other factors besides channel loading impact the bandwidth efficiency in the optical spectrum are immunity to noise impairments, restricted guard bands requirements, and close levels in multilevel signaling. Bandwidth efficiency modulations open new regions in the frequency spectrum that were occupied by redundant information. In order to maintain quality signaling in any assigned channel, the reserved channels are no longer required to be sitting idle waiting for traffic from other channels to be switched into them.

Power Efficiency Consideration in Optical Systems

Since power is scarce in all transmission systems, techniques that use less power for the same transmitted bandwidth are always sought. In power limited systems, the following tradeoffs are to be considered [17]:

i) Improve the power at the expense the usable bandwidth for a fixed $\frac{Eb}{N_o}$

ii) Reduce $\frac{Eb}{N_o}$ at the expense of the useable bandwidth for a fixed power.

iii) Use practical modulations for power efficient systems are M-ary Frequency Shift Keying (MFSK). If the IF minimum Nyquist bandwidth is given by: $B_N = M.R_s$.

where R_s is the symbol rate, the required transmission bandwidth is expanded M-fold, which is why MFSK is called bandwidth expansive technique and can be used to reduce the required $\frac{Eb}{N_o}$ at the expense of increased bandwidth.

Two modulations techniques (MPSK and MFSK) have been used as examples to provide some insight into the design issues and trade-offs between bandwidth and power efficiencies. As M increase, MPSK signaling provides more bandwidth efficiency at the cost of $\frac{Eb}{N_o}$, while MFSK signaling allows a reduced $\frac{Eb}{N_o}$ at the cost of increased bandwidth.

Four general rules can be followed in the design stage:

1) The modulation must not expand the required transmission bandwidth beyond the available bandwidth.
2) BEM based systems must be interoperable with other technologies.

3) The required optical signal to noise ratio must be met even under worst-case scenario.
4) The hardware and software needed for BEM implementation must be simple and inexpensive.

10.3   Transmission Impairments and Technology Limitations

The finite bandwidth of state of the art optical fibers has pushed many researchers to look for alternatives to increase the spectral efficiency. Time Division Multiplexing (TDM), WDM, and their combinations were intended for this purpose, but these current techniques use binary On-and-Off Keying (OOK) with direct detection, which is neither power efficient nor bandwidth efficient. The following is a list of additional implementation imperfections whose effect on the performance of various modulations should be considered; most of them have a greater impact on phase modulations [18]:

1. Phase and amplitude imbalance in MPSK, MFSK, QAM.
2. Imperfect or noisy reference at a coherent demodulator.
3. Power loss due to filtering of the modulated signals.
4. Degradation due to non-ideal detection filters.
5. Degradation due to pre-detection filtering.
6. Degradation due to bit synchronization timing errors.
7. Local oscillator phase noise and spurious.
8. Envelope amplitude variations.
9. Stability and slope of Quadrature circuits in discriminator detectors.

Signal distortions arising from transmission impairments must be controlled so that the associated penalties are accounted for during the initial design stage when the power and OSNR budgets are allocated. These penalties are typically taken into consideration once the systems specifications are known (capacity, reach, cost target, etc.). A critical performance edge is found in the area of dispersion tolerance [19]. Chromatic dispersion impairs performance of all high-rate signals, and therefore must be compensated for every link and channel. While chromatic dispersion is a linear process and it is straight forward to mitigate it, Polarization Mode Dispersion (PMD) becomes one of the most challenging factors in optical transmission systems.

PMD Effects in BEM Transmission Systems

It is well known that the fundamental mode of a circularly symmetric dielectric waveguide is degenerate in two dimensions. In real fiber the degeneracy is split by the birefringence properties of the fiber. The birefringence may be introduced deliberately, as in polarization-maintaining fiber for example, or it may be a by-product of fiber geometry [20]. In this case the birefringence is introduced randomly by geometrical or stress-induced perturbations.

The propagation constants, $\beta_i(\omega)$, of the two orthogonal modes can be expanded in a Taylor series around the center frequency, ω_o

$$\beta_i(\omega_o) = \beta_i(\omega_o) + \frac{\partial \beta_i}{\partial \omega}\bigg|_{\omega=\omega_o} (\omega - \omega_o) + \frac{1}{2}\frac{\partial^2 \beta_i}{\partial \omega^2}\bigg|_{\omega=\omega_o} (\omega - \omega_o)^2 + \dots \qquad (10.10)$$

where the $\beta_i(\omega_o)$ is the phase velocity v_p, $\frac{\partial \beta_i}{\partial \omega}$ is related to the group velocity v_g, $\frac{\partial^2 \beta_i}{\partial \omega^2}$ and is related to the dispersion of the group velocity.

With the development of dispersion-shifted fibers and the deployment of systems operating near the zero dispersion wavelength, the contribution to dispersion from the second order term that accounts for the chromatic dispersion component can be eliminated, and the first order term now becomes significant. For the case of birefringent fibers, the first order term leads to a group delay called polarization mode dispersion (PMD). This polarization dispersion introduces a differential group delay between orthogonal states of polarization. Although the effect of PMD is to change randomly the polarization state of a pulse propagating in a fiber, it is possible to define a pair of orthogonal states or "principal states" at the input whose output

states are orthogonal, and show no dependence on wavelength to first order. (In some situations, however, this approximation falls apart and the principal states can show wavelength dependency, leading to further system degradation through coupling to chromatic dispersion [22].)

The birefringence in the fiber can also be caused by local random and asymmetric mechanisms such as stress, bending, and twisting. These random birefringence mechanisms redefine the local birefringence axes along the length of the fiber, thus causing random coupling between the polarization modes along the length of the fiber. The cabling process also introduces a certain amount of random birefringence and random mode coupling. The fiber length between such changes is usually referred to as the coupling length, which for a fiber is usually quoted as the ensemble average of all of the local coupling lengths. Furthermore, changes in local environmental conditions such as temperature, for example, cause fluctuations in the local birefringence axes, thus causing random polarization coupling. As a result of the randomly changing polarization coupling, the magnitude of the Differential Group Delay (DGD) becomes a statistically varying function [22]. It can be shown that distribution of differential group delays is described by a Maxwellian distribution function, defined by

$$P(\Delta\tau) = \frac{32\Delta\tau^2}{\pi^2\langle\Delta\tau\rangle^3}\exp\left[-\frac{4\Delta\tau^2}{\pi\langle\Delta\tau\rangle^2}\right] \tag{10.11}$$

where $\Delta\tau$ is the differential group delay between the two principal states, and $\langle\Delta\tau\rangle$ is the mean differential group delay also known as PMD value. As a consequence of the statistical nature of polarization mode dispersion, the magnitude of $\langle\Delta\tau\rangle$ increases with the square root of the fiber or cable length, for lengths much longer than the coupling length. Polarization mode dispersion is usually quoted in units of ps or ps/$\sqrt{\text{km}}$. The unit of ps is usually reserved for single optical elements that have a fixed dispersion (e.g., a coupler or isolator) or short fiber sections that do not exhibit mode coupling [23].

PMD Induced Limitations in Binary Modulate Signals

In a digital transmission system the principal effect of polarization mode dispersion is to cause intersymbol interference. As a rule of thumb, a 1-dB penalty in the optical signal to noise ratio (OSNR) occurs when the total instantaneous differential group delay equals $0.3T$, where T is the bit period. Both principal states of polarization are excited equally (gamma = 1/2). This value is commonly used to account for the maximum tolerable system power penalty. The related Differential Group Delay (DGD) can also be approximated in the same way, as the maximum tolerable system value: $\text{DGD}_{\text{max}} \leq 0.3T$. A relationship between DGD_{max} and PMD values can be set on the bases of the Maxwellian probability distribution choosing a proper adjustment factor that is related to the desired maximum outage probability. Higher bit-rate systems have shorter bit periods so they tolerate less differential group delay. Current studies indicate that optical fibers will be specified according either to the DGD_{max}, defined above or to the mean level of Differential Group Delay (PMD value).

In optical transmission systems operating at 10 Gb/s, a statistical specification of 0.5 ps/sqrt(km) has been proposed for concatenated links of optical fiber cable. From the Maxwell statistics, the probability that the 1 dB in OSNR penalty at 10 Gb/s is exceeded for a 400 km span is less than 4×10^{-5}. Here, the contributions of other components to PMD are not taken into consideration. The PMD impairment can therefore be seen as a system power penalty for a given bit rate and a targeted bit error rate.

For systems operating at 40 Gb/s the mean differential group delay equal to one-tenth of a bit period, $0.1T$, corresponds to 2.5 ps. As a general assumption, part of this tolerated value could be allocated to the cable and part to optical repeaters, depending on the link characteristics. The total PMD of a link encompassing optical fibers and optical sub-systems is the quadratic sum of the fiber and sub-systems PMD

$$\text{PMD}_{\text{TOT}} = \left[\text{PMD}_{\text{F}}^2 + \sum_i \text{PMD}_{Ci}^2\right]^{1/2} \tag{10.12}$$

where PMD_{TOT} is Total PMD link (ps), PMD_{F} is PMD of concatenated optical fiber cables (ps), and PMD_{Ci} is PMD value of the ith sub-system (ps).

For example, considering a PMD value of the concatenated optical fiber cables of the link of 0.1 ps/sqrt (km) (that can be considered advisable at 40 Gbit/s), 2.0 ps is the PMD cable contribution on 400 km long links. According to the previous formula, this still leaves a 1.5 ps PMD margin for optical sub-systems. Assuming the use of 4 optical sub-systems with a PMD value of 0.6 ps, the total PMD will be below 2.5 ps limit stated before for 40 Gb/s systems.

$$\text{PMD}_{\text{TOT}} = \sqrt{(0.1 \cdot 20)^2 + 4 \cdot (0.6)^2} = 2.33 \, \text{ps} < 2.5 \, \text{ps} \qquad (10.13)$$

Furthermore, in long-haul amplified systems employing polarization scramblers (devices that deliberately modulate the polarization state of a signal laser so that it appears to be unpolarized), the polarization mode dispersion causes an increase in the degree of polarization of the signal. This degrades system performance through interactions with polarization-dependent loss and polarization hole burning.

In directly modulated bandwidth efficient modulation systems and due to the presence of analog signals, the interaction of polarization mode dispersion with laser chirp leads to a second order distortion, proportional to the modulation frequency. A further second order penalty, independent of modulation frequency, incurs when additional polarization-dependent loss is present in the system. The second order effect can cause a coupling between polarization mode dispersion and chromatic dispersion. This is caused by the wavelength dependence of the differential group delay, and more importantly, the wavelength dependence of the principal states of polarization. This leads to a statistical contribution to the chromatic dispersion. This is an area that is not well understood, and is under study. The use of chromatic dispersion compensating devices also has an unclear impact on the PMD penalty.

Methods to Minimize PMD Effect in BEM Systems

Given that the problem arises from birefringence, much of the effort in reducing the effects of polarization mode dispersion have been concerned with minimizing the birefringence introduced by fiber or cable manufacturing. Care is taken to optimize fiber production to ensure geometrical and optical circular symmetry, and to induce polarization mode coupling. Optical cables are manufactured using materials and processes that minimize the residual strain in the cable structure across the fiber. Elaborate cable structures can also be used, which introduce a circular component to the induced birefringence. By careful design, such an effect can counteract linear birefringence to produce a cable with a resultant zero polarization mode dispersion. Typically, the mean polarization mode dispersion of fibers and cables lie in the range:

$$0 < \langle \Delta \tau \rangle < 0.5 \, \text{ps}/\sqrt{\text{km}} \qquad (10.14)$$

Furthermore, advanced fiber designs show less PMD value, e.g., 0.1 ps/sqrt (km), as mentioned earlier. Another method to reduce the effect of PMD is PMD compensation. A PMD compensator can accept at its input a signal affected by PMD, and can mitigate to give out a signal properly restored. It generally consists of a PMD equalizer, a PMD monitor, and a feed-back controller. The equalizer and the monitor can be implemented in the optical or in the electrical domain. There is also the possibility for mixed or hybrid solutions. A feed-back controller makes decisions on the basis of the monitored information according to a predefined algorithm, and drives the equalizer according to the decisions.

Optical Components Limitations

Optical Filters

Optical filters that isolate a narrow range of wavelengths from a broader spectrum with high resolution are not yet well developed, especially for wavelength separation below 25 GHz. Acceptable performance optical filters should have Flat In-Band Spectral Response (FBSR), low insertion loss, center wavelength stability under various temperatures, and be tunable over a wide band [24]. Optical channel spacing is ultimately limited by

the bandwidth of the information carried by the wavelengths. Optical channels in one fiber can not be placed infinitely close together, even if the optical sources have very pure spectrum; since the modulating signal caused the spectrum to broaden, this is an ultimate (theoretical) limit. One should also consider technical limits such as the characteristics of the optical filters used to separate the optical channels before the optical receiver. These filters have to be able to sufficiently suppress the signal of adjacent channels, even in the event of very close channel spacing. In present-day networks, the suppression is supported by the electronic bandwidth of the receiver, which is usually much smaller than the 100 or 50 GHz optical channels spacing.

In an Optical Add Drop Multiplexing (OADM), where a signal in one particular wavelength is dropped and another signal having almost the same wavelength is added, the dropped signal has to be completely blocked; otherwise it will interfere with the newly added signal and can seriously degrade the BER performance. The optical receiver is a non-linear element, and as such, signals with closely spaced wavelengths will create beating products. In systems where the channels are very closely spaced the wavelength stability of the deployed optical sources and the filtering elements are also very important. Concatenation of filters leads to spectral narrowing [25]. This poses stringent requirements on the wavelength stability of the laser sources and the absolute wavelength position of the wavelength selective elements, which may make necessary the use of an absolute wavelength reference and an active feedback control. Experiments on cascadability showed a reduction of the bandwidth of more than factor of two after ten elements, and a factor of three after 20 elements. By using BEM, which will compress the signals and allow the transmission of the same capacity in smaller spectral bandwidth, the use of narrow band filters is avoided.

A further question is whether, by increasing the number of channels, the wavelength spacing has to be reduced. The bandwidth of the single-window Erbium Doped Fiber Amplifiers (EDFAs) limits the maximum number of channels that a system can support with a given channel spacing. Although double-window EDFAs have been proposed, widening the amplified spectral range from 30 to 60 nm, most of the suppliers are opting for the use of single-window EDFAs, at least for the time being. This is because double-window amplifiers are still new and very expensive. Double-window EDFAs contain two EDFAs with different operating wavelength ranges connected in parallel. One of the EDFAs has an operating wavelength range shifted towards the higher wavelengths, in contrast to traditional EDFAs. The incoming optical signal is wavelength demultiplexed and each amplifier is amplifying the channels falling into its operating wavelength range. The outputs of the EDFAs are recombined. There is also the question how the growth in number of channels will be implemented. Some manufacturers might consider upgrading by inserting new channels between the existing ones, effectively halving the channel spacing. Such a solution would not only require the replacement of the WDM elements, but would give rise to nonlinear effects, as discussed below.

Amplifier Spontaneous Emission

ASE degrades the optical signal to noise ratio (OSNR). The required OSNR level (ROSNR) depends on the bit rate and transmitter-receiver technology (e.g., FEC). The OSNR margins allocated for the impairments needs to be maintained at the receiver. In order to satisfy these requirements, vendors often provide some general engineering rule in terms of maximum length of the transparent segment and number of spans. For example, current transmission systems are often limited to up to six spans each 80 km long before deploying a dynamic-gain-flattening device. For larger transparent domains, more detailed OSNR computations will be needed to determine whether the OSNR level through a domain of transparency is acceptable. This would provide flexibility in provisioning or restoring a link through a transparent link. Assume that the average optical power launched at the transmitter is uniform and constant. The signal from the transmitter to the receiver goes through various optical amplifiers, with each introducing some noise power. Unity gain can be used at all amplifier sites to maintain constant signal power at the input of each span to minimize noise power and nonlinearity. A constraint on the maximum number of spans can be obtained that is proportional to P and inversely proportional to ROSNR, optical bandwidth, amplifier gain, and spontaneous emission factor of the optical amplifier, assuming all spans have identical gain and noise figure.

10.4   Practical Bandwidth Efficient Modulations

Recently, there has been a great deal of interest in the use of multilevel signaling in order to increase the transmission capacity in bandwidth-limited systems. Pulse amplitude modulation is used to increase the fiber capacity. This modulation format helps in increasing the bandwidth efficiency compared to the basic binary modulation. Multilevel signaling such as *M*-ary (where *M* is the number of signal levels) provides better bandwidth utilization by allowing a factor of $Log_2 (M)$ reduction in frequency over conventional binary formats. In a 4-ary PAM system, for example, there are two bits in every baud (or symbol), therefore a 10 Gbaud/s can transmit 20 Gb/s signal (native data rate). It is important to note that the transmission bandwidth of *M*-ary is scaled down by a factor of $1/Log_2 (M)$ compared to the usual on and off keying (OOK) at the same rate. This is a very strong argument if one is looking to upgrade a network capacity. No major changes in hardware are needed to double a capacity of network operating at 10 Gb/s, for instance. Previous work in multilevel signaling has been concentrated on wireless and point-to-point links [26–27]. Multilevel signaling has also been used for very short reach (VSR) links where high data rates are transmitted over a relatively short reach, and where high speed serial links using copper cables become more attractive for a reach less then 15 m, this limit in reach and data rates are dictated by the cable skin-effect loss.

PAM Modulation

Here, the focus is on 4-ary PAM signaling that can double the capacity of a DWDM metro network operating at 10 Gb/s or higher. This upgrade in capacity will not require an upgrade of the system hardware. The proposed PAM modulation uses a simple scheme to realize quadratic signal spacing to obtain optimum receiver sensitivity in a system dominated by amplified spontaneous emission (ASE) noise. The quadratic spacing allows an improvement of the receiver sensitivity by a factor of ~6 dB compared with equal signal spacing. The quadratic level spacing can be realized in three different ways: (a) by electronic means that consists of squaring the output voltage levels by an off the shelf voltage squarer. If the four levels of a PAM-4 are defined as (0, 1, 2, 3), the output of the squarer will be (0, 1, 4, 9), which is a quadratic spacing that closely matched the response of MZ-modulator, (b) by optical means, which consists of adding two streams of optical data interferomerecally with one signal attenuated by 6 dB in power with respect to the other; a phase shifter is used to ensure in-phase addition of the two modulated optical signals into the optical coupler, and (c) optoelectronic means, which uses a MZ modulator with suitable driving voltage and bias position.

PAM System Design

Implementing multilevel coding at high bit rates is rather straightforward. The 4-ary PAM is simply generated from two non-correlated data sources where one of data sources has half amplitude of the other. Both signals are power combined to form a 4-ary PAM composite signal. If the input bit streams are correlated, a delay line may be used to decorrelate the two input signals. A 6-dB RF attenuator can be used in one of the data sources to generate four levels with equal spacing. In the non-return to zero (NRZ) setup, the 4-ary PAM signal drives a single arm MZ modulator biased at $V_\pi/2$. This shows one of the advantages of multilevel coding that does not require high RF voltage drivers. Figure 10.1 shows the transmitter (encoder) used to generate the 4-ary PAM modulation. At the receiver, direct detection is used to convert the optical 4-ary PAM signal into two separate electrical binary tributaries. Also shown in Figure 10.1 is the decoder, which uses three independent decision circuits; each circuit is preset to a threshold level corresponding to respective levels. The threshold levels can be dynamically adjusted for optimal bit error rate.

Equal level spacing (0, 1, 2, 3) can easily be obtained by applying an equally spaced electrical 4-ary PAM to a MZ modulator driven at its linear regime. In order to achieve optimal receiver sensitivity in ASE noise dominated systems [28], quadratic level spacing is preferred for both direct modulation and external modulation cases. This is because in an optically preamplified receiver, the signal-ASE beating is stronger at high power levels than at low power levels. Simple calculations show that quadratic level spacing, such as (0, 1, 4, 9), can improve the receiver sensitivity by over 5 dB compared to equal level spacing. Quadratic spacing can be realized by driving the MZ modulator in such a way that the lowest level is at V_π and the highest
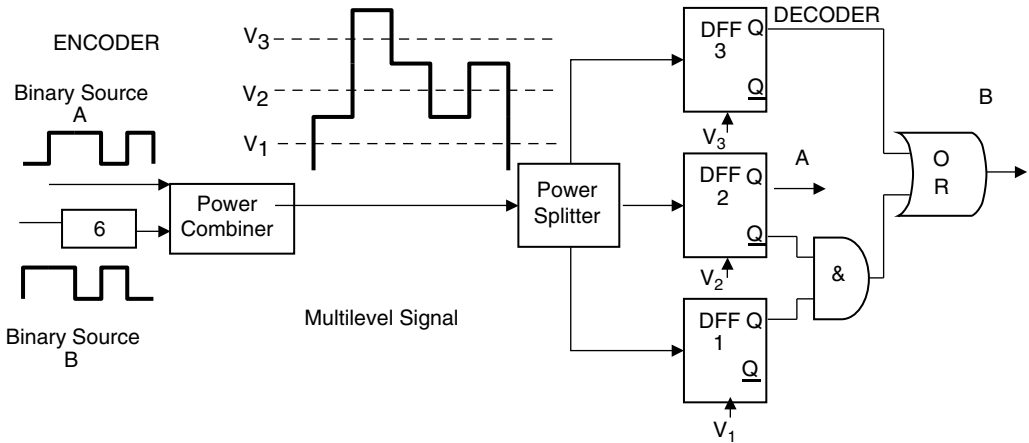
FIGURE 10.1   PAM-4 encoder decoder setup.
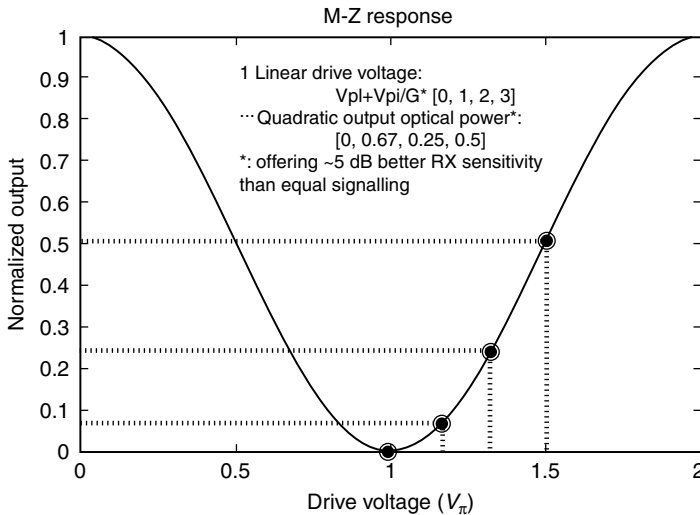
FIGURE 10.2   Schematic of the driving conditions to achieve $\sim$ quadratic signaling of 4-ary PAM.

level is at $\sim V_{\pi}/2$, as shown in Figure 10.2. It is worthy noting that subquadratic signal leveling may be needed in practical systems where other degradations have to be considered. For example, thermal noise in the receiver, finite extinction ratio (ER) of the transmitter, and inter-symbol-interference (ISI) favor a wider spacing between the lowest two levels. We point out that the proposed scheme can flexibly accomplish sub-quadratic leveling by simply increasing the driving voltage.

PAM-4 Experimental Results

Back-to-back and transmission through a 75 km standard single mode fiber (SSMF) experiment on 20 Gb/s 4-ary PAM are shown in Figure 10.3, which shows the measured back-to-back NRZ and RZ optical waveforms. Subquadratic signal leveling is obtained by suitably driving the M–Z modulator in the RZ modulation case. Figure 10.4 shows the measured NRZ and RZ eyes after transmission. No obvious degradation was noticed. The required OSNR for back to back transmission was 26 dB for a bit error rate of 10^{-10}. By using efficient forward error correction (EFEC), more then 9 dB coding can be achieved.
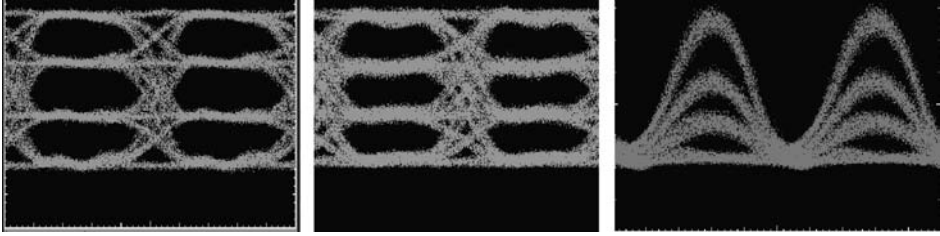
FIGURE 10.3   Electrical back to back, NRZ back to back, and RZ back to back.
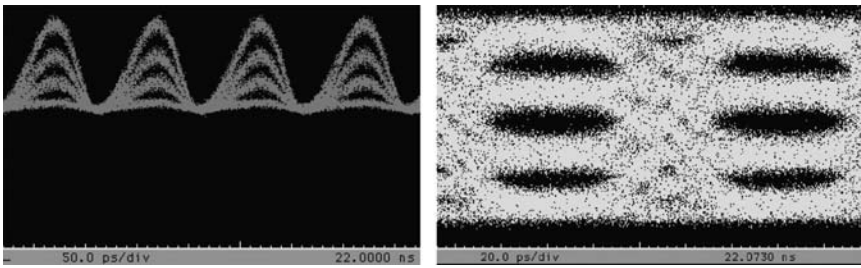
FIGURE 10.4   RZ and NRZ output eyes after 75 km of SSMF.
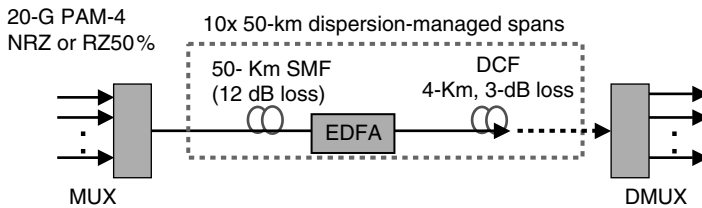
FIGURE 10.5   Schematic of the simulated transmission link using 4-ary PAM format.

PAM-4 Simulation and Modeling

To assess the performance of 20 Gb/s 4-ary PAM over longer transmission distance, a numerical simulation, using commercial software was performed. The schematic of the simulated RZ-DPSK transmission is shown in Figure 10.5. Several WDM channels are then multiplexed before being launched into a dispersion-managed link, which consists of 10 spans of 50-km SSMF (D = 17 ps/km/nm), following each of which is a dispersion compensation fiber (DCF) to bring the accumulated dispersion to zero. The fiber non-linear coefficient is 1.5/W/km. Fiber loss of each span (12 dB) is compensated by an EDFA. At the receiver site the WDM channels are demultiplexed and decoded using a three-level decision circuit as described previously. The 4-ary PAM transmissions in both single-channel and DWDM configurations were modeled. For the DWDM transmission we assume five 50-GHz spaced 20-Gb/s, 4-ary PAM channels, each of which is modulated by a 2^7-1 pseudo-random binary sequence (PRBS). Modulator bandwidth is assumed to be 20 GHz. The WDM channels are multiplexed and de-multiplexed with 50-GHz third-order super-Gaussian and 20-GHz fourth-order Gaussian filters, respectively. A fifth-order Bessel electrical filter with 7-GHz bandwidth is used at the receiver. We adjust the ASE noise level at the receiver to find the required OSNR (defined as the ratio between the per-channel signal power and the ASE noise power in 0.1-nm bandwidth) for achieving BER=10^{-3}, the threshold for a corrected BER of 10^{-16} by ultra-FEC.

NRZ 4-ary PAM Transmission

Figure 10.6 shows the optical spectrum of the simulated NRZ 4-ary PAM signals and back-to-back eye diagram. The required OSNR for BER = 10^{-3} (uncorrected) is ~19.5 dB. The optimum decision levels are (0.23, 0.8, 1.7). Figure 10.7 shows the eye diagrams after 200 km and 400 km single-channel transmissions with 0 dBm per channel launch power. The nonlinear penalties (in terms of the increase of required OSNR) are ~1.2 dB and ~4 dB after 200 km and 400 km transmission, respectively. Figure 10.8 shows the eye diagrams after 200 km and 400 km 50-GHz spaced DWDM transmissions with 0 dBm per channel launch power. The nonlinear penalties are ~1.4 dB and >5 dB after 200 km and 400 km transmission, respectively. Evidently, there exists a large nonlinear penalty due to inter-channel cross-phase modulation (XPM) induced timing jitter at 400 km.

RZ 4-ary PAM Transmission

Figure 10.8 shows the optical spectrum of the simulated 50%-RZ 4-ary PAM signals, and back-to-back eye diagram. The required OSNR for BER = 10^{-3} is ~17.5 dB, ~2dB which is better than that of NRZ.
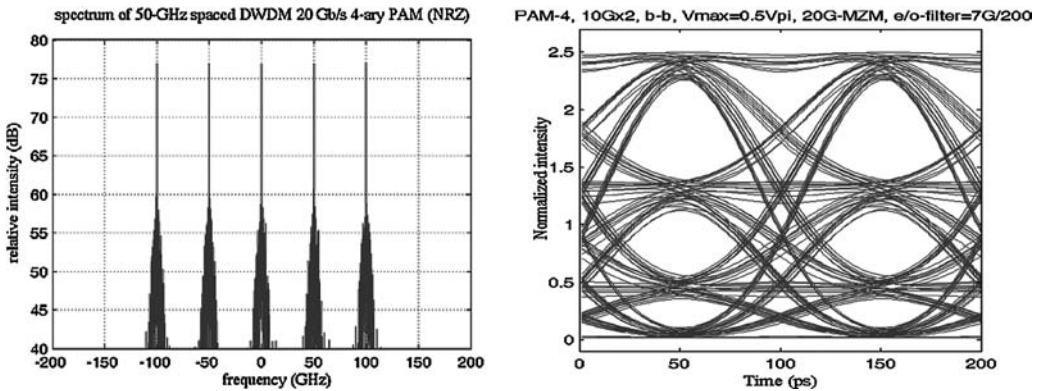
FIGURE 10.6   Optical spectrum of the simulated NRZ 4-ary PAM signals (left), and back-to-back eye diagram (right).
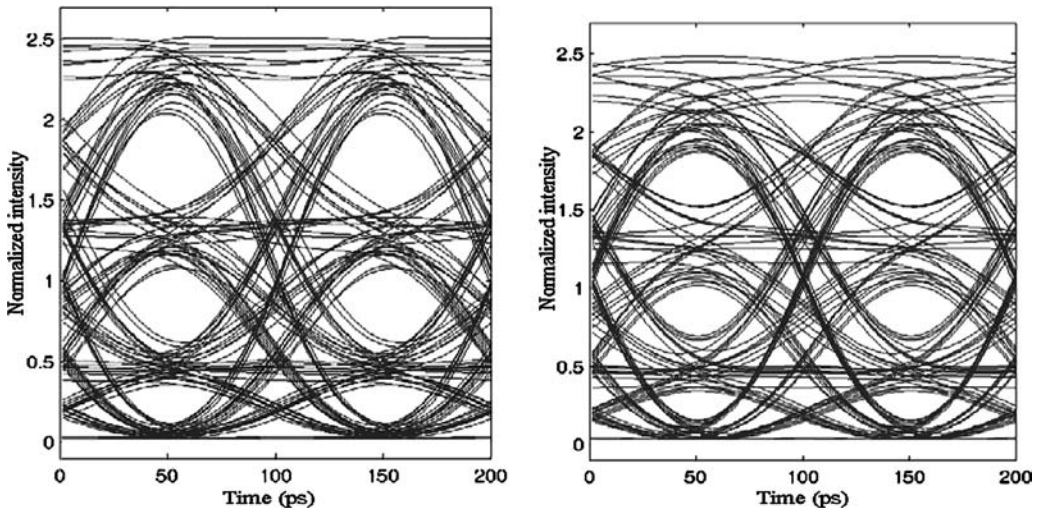
FIGURE 10.7   Eye diagrams after 200 km (left) and 400 Km (right) single-channel NRZ 4-ary PAM transmissions with 0 dBm per channel launch power.
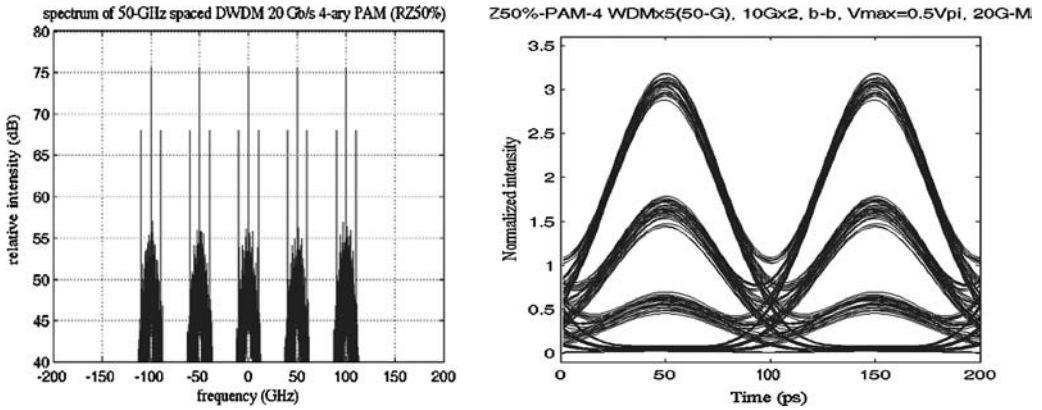
FIGURE 10.8 Optical spectrum of the simulated 50%-RZ 4-ary PAM signals (left), and back-to-back eye diagram (right).
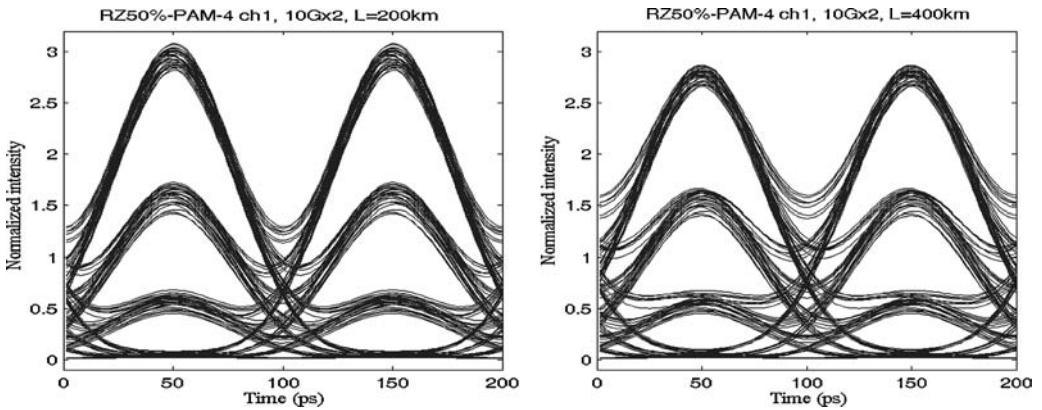
FIGURE 10.9 Eye diagrams after 200 km (left) and 400 km (right) single-channel 50%-RZ 4-ary PAM transmissions with 0 dBm per channel launch power.

The optimum decision levels are (0.25, 1, 2.25). The nonlinear penalties (in terms of the increase of required OSNR) are <0.2 dB and ~0.5 dB after 200 km and 400 km transmission, respectively. Figure 10.9 shows the eye diagrams after 200 km and 400 km 50-GHz spaced DWDM RZ 4-ary PAM transmissions with 0 dBm per channel launch power. The nonlinear penalties are <0.5 dB and ~1 dB after 200 km and 400 km transmission, respectively. As expected, RZ is much more robust against nonlinear distortions than NRZ.

For short-reach (<200 km) Metro optical networks, the required OSNR for both 20 Gb/s NRZ and RZ 4-ary PAM formats is less than 22 dB. Assuming 8 dB NF of EDFA, the received OSNR with 0-dBm/ch launch power after a 200-km transmission link (consisting of four 50-km SSMF spans of 12-dB loss each) is ~32 dB, indicating a ~10 dB OSNR margin. It is thus reasonable to expect 200-km reach with 0.4 spectral efficiency. With it superior robustness against nonlinear distortions, RZ 4-ary PAMF may allow a system reach of around 400 km with reasonably large OSNR margin (~10 dB).

Figure 10.10 shows the eye diagrams after 200 km (left) and 400 km (right) single-channel 50%-RZ 4-ary PAM transmissions with 0 dBm per channel launch power. The 4-ary PAM coding can be used for doubling the capacity of current 10 Gb/s Metro optical networks. A 20 Gb/s 4-ary PAM transmitter is constructed by a novel scheme using a single M–Z modulator. Multilevel PAM coding may find applications in future Metro optical communications systems in which both capacity and robustness against non-perfect dispersion compensation and PMD are important.
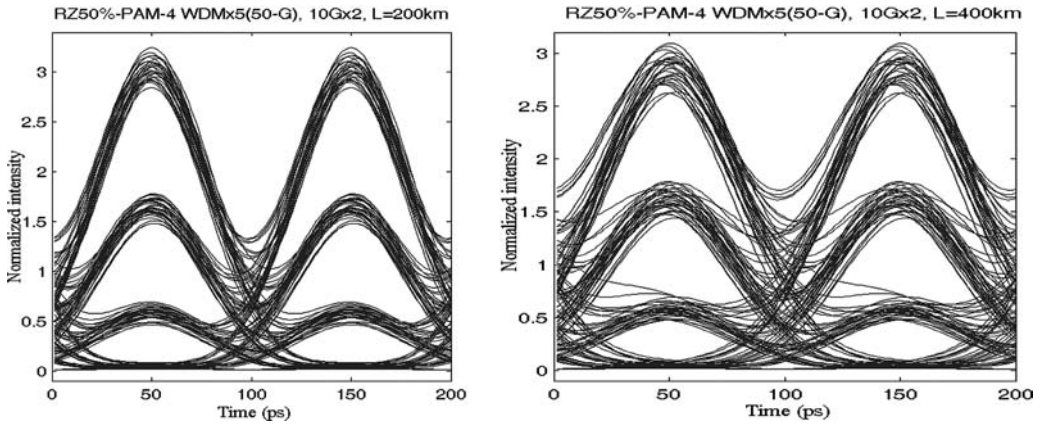
FIGURE 10.10   Eye diagrams after 200 km (left) and 400 km (right) DWDM (with 50-GHz spacing) 50%-RZ 4-ary PAM transmissions with 0 dBm per channel launch power.

SCM-WDM Optical Systems

Optical subcarrier multiplexing (SCM) is a scheme where multiple signals are multiplexed in the RF domain and transmitted by a single wavelength. A significant advantage of SCM is that microwave devices are more mature than optical devices; the stability of a microwave oscillator and the frequency selectivity of a microwave filter are much better than their optical counterparts. In addition, the low phase noise of RF oscillators makes coherent detection in the RF domain easier than optical coherent detection, and advanced modulation formats can be applied easily. A popular application of SCM technology in fiber optic systems is analog CATV distribution [29]. Because of its simple and low-cost implementation, SCM has also been proposed to transmit multichannel digital optical signals using direct detection [30] for local area optical networks. The performance of high-speed digital fiber-optic transmission using SCM both analytically and numerically were investigated by various research groups. Fiber nonlinearities such as cross phase modulation (XPM) and four-wave mixing (FWM) may generate significant amounts of nonlinear crosstalk between adjacent SCM channels because they are very closely spaced. Chromatic dispersion when properly compensated for is not a limiting factor in modulated SCM systems because the data rate at each subcarrier is relatively low, but carrier fading due to PMD is significant because of high subcarrier frequencies. In order to optimize the system performance, tradeoffs have to be made between data-rate per subcarrier, levels of modulation, channel spacing between subcarriers, optical power, and modulation indices. An experiment of 10 Gb/s SCM fiber-optical system was performed in which 4×10 Gb/s data streams were combined into one wavelength, which occupied an approximately 80 GHz optical bandwidth. The combination of SCM and WDM may provide a more flexible platform for high-speed optical transport networks with high optical bandwidth efficiency and high dispersion tolerance. For example, n independent high-speed digital signals are mixed by N different microwave carrier frequencies f_i. They are combined and optically modulated on to an optical carrier. M wavelengths are then multiplexed together in an optical WDM configuration. At the receiver an optical demultiplexer separates the wavelengths for individual optical detectors. Then RF coherent detection is used at the SCM level to separate the digital signal channels. Channels add/drop are also possible at both the wavelength and SCM levels. While this SCM-WDM is in fact an ultra-dense WDM system, sophisticated microwave and RF technology enables the channel spacing to be comparable to the spectral width of the baseband, which is otherwise not feasible by using optical technology. Compared to conventional high-speed TDM systems, SCM is less sensitive to fiber dispersion because the dispersion penalty is determined by the width of the baseband of each individual signal channel. Compared to conventional WDM systems, on the other hand, it has better optical spectral efficiency because much narrower channel spacing is allowed.

SCM-BEM Systems

SCM is a complementary technology to traditional TDM and WDM systems. It provides an additional dimension of multiplexing to increase the efficiency and flexibility of an optical transmission link. It has largely been used in wireless and CATV applications, but it can be adopted for increasing the capacity and bandwidth on a given optical transmission link. SCM, when combined with a bandwidth efficient modulation (i.e., QAM, QPSK, DPSK...), permits many additional capabilities where designers can take advantages of two proven techniques (i.e., SCM and QAM). Optical SCM-QAM approach is a valuable technology due to several key improvements in pulse shaping for spectral efficiency, the availability of forward-error-correction (FEC) for OSNR improvement and the advances in RF signal pre- and post processing [31].

SCM-QAM Technique

The most commonly used modulation with SCM for capacity improvement is Quadrature Amplitude Modulation (QAM), where multiple bits are mapped in amplitude and phase plane. An example is 256 QAM, which transmits eight information bits per symbol by transmitting one of 256 symbols in the amplitude phase plane. The digital clock speed used in timing these symbols is much lower than the pure bit clock speed used in an on/off-keyed (OOK) modulation. This translates to lower cost electronics and to an ability to transmit more information in a given bandwidth, referred to as bandwidth-efficient modulation (BEM).

Quadrature amplitude modulation (QAM) is a modulation scheme in which two sinusoidal carriers, one exactly 90° out of phase with respect to the other, are used to transmit data over a given physical channel. Because the orthogonal carriers occupy the same frequency band and differ by a 90° phase shift, each can be modulated independently, transmitted over the same frequency band, and separated by demodulation at the receiver. For a given available bandwidth, QAM enables data transmission at twice the rate of standard pulse amplitude modulation (PAM) without any degradation in the bit error rate (BER) [32]. QAM and its derivatives are used in both mobile radio and satellite communication systems, and are starting to appear in fiber optics networks. Numerically controlled oscillator (NCO) Compiler can be used to design a dual-output oscillator that accurately generates the in-phase and quadrature carriers used by a QAM modulator. The carrier frequency of each sinusoid can be set to any precision by defining the phase increment input to the NCO. A raised cosine finite impulse response filter is used to filter the data streams before modulation onto the quadrature carriers. When passed through a band-limited channel, rectangular pulses suffer from the effects of time dispersion and tend to smear into one another. This pulse shaping filter eliminates inter-symbol interference by ensuring that at a given sampling instance, the contribution to the response from all other symbols is zero.

In addition to transmitting several information bits in each symbol period, digital signal processing is used to shape the spectrum of the transmitted pulses so that most of the modulated signal energy is retained in a minimum frequency band, which significantly reduces the required spacing between carriers, enabling SCM systems to provide very high bandwidth efficiency since having 256 states (symbols) to choose from each period would naturally increase the likelihood of error. Suitable techniques to enhance performance are also used with these modulation approaches so that link budgets comparable to or better than alternative approaches can be achieved with similar power. These include linear equalization of the received signal within the demodulator and use of multiple advanced FEC coding algorithms.

Key to the performance benefits in SCM-QAM modulation is the availability of RF signal processing components, which can be adopted from other areas such as satellite communications. The individual channels are frequency converted and filtered to a series of equally spaced subcarrier frequencies ($f_1, f_2...f_k$). The spaced subcarriers are combined together creating a single composite RF signal consisting of k independent subcarriers. This RF signal is then used to directly modulate laser or an optical Mack Zehnder modulator in external modulation or DWDM networks. Systems have been developed with these technologies that provide up to 40 Gb/s of data transmission on a single wavelength in as little as 20 GHz of spectrum where the same data rate in NRZ format will occupy a minimum of 80 GHz. These individual wavelength signals can be fed into WDM systems to provide even greater capacity per fiber. SCM provides superior bandwidth

efficiency and overall capacity to the much simpler NRZ modulation. Some additional inherent features that provide a new level of flexibility are:

- Straightforward network design
- Easy and quick system upgrades
- Improved system robustness with reduced adjacent channel interference and link performance enhancements due to equalization and FEC
- Lower susceptibility to polarization-mode dispersion and chromatic dispersion than 10-Gb/s
- Easy performance monitoring using the SCM technique itself

In addition to supporting a variety of applications with these configurations, the digital and RF processing inherent in SCM provide a level of visibility together with data transparency that is unique. The digital processing performed yields a great deal of insight into the performance of the link on a given channel with metrics ranging from encoded channel BER to received signal-to-noise ratio. That enables SONET-level protection to be provided on any traffic type carried by an SCM system. Since all traffic is digitally modulated, all traffic can be protected with less than 50-msec switching to protect a channel, including native IP data traffic without SONET framing or digital wrappers. The RF processing performed in the SCM system provides the desired transparency that network operators seek. Each channel can independently carry traffic operating at different rates. Any traffic type can be carried transparently, and systems have been designed that carry SONET, ATM, and IP traffic seamlessly on the different subcarriers. This transparency provides the network operator with the ability to reconfigure the system easily and quickly as the demand in new services changes without swapping out a lot of equipment. The rise of Gigabit Ethernet as a service provided by carriers is pushing demand for flexible systems that enable better transitions. Also SCM eliminates the need for equipment to convert non-SONET signals into SONET signals.

Because SCM is effectively orthogonal to TDM and WDM, the technologies are inherently interoperable. SCM adds another dimension to the space network operators can choose from to implement capacity. On the tributary side, SCM systems can and indeed must be designed to provide complete compatibility to SONET and TDM signals. The use of standard lasers supporting the established ITU-grid wavelengths enables SCM systems to be combined with WDM for very-high-capacity systems with superior flexibility, providing sub-wavelength granularity while supporting multiple services in their native format. With so many technologies being applied to optical networking, traditional notions of performance standards are being challenged and adapted. Network design is more challenging at higher rates, over more wavelengths, for longer distances. Any edge in simplifying network design through enhanced or more robust performance can make a real difference to designers and to cost. SCM with digital signal processing for bandwidth-efficient modulation provides some significant performance benefits to optical networks. Optical-link budgets can be difficult to maintain through a network that, despite standards used for initial design, has a range of channel conditions across its spans; each change can mean having to take a new look at the link budgets. However, the optical power for an SCM signal remains constant regardless of how many subcarrier channels are in service. By modulating a common light source with the aggregate signal, the optical power is a function of the laser, not the number of channels in the signal. Any power not used by a channel is returned to the unmodulated portion of the SCM waveform, which enables the add/ drop of subcarriers without reengineering the link.

The SCM and BEM signal processing combination also provides a high level of robustness in system link performance over a wide range of channel conditions. Because of the pulse shaping and filtering of the subcarriers, SCM signals have very low adjacent-channel interference (see Figure 10.4a). Unlike an OOK signal, the spectral side lobes of an SCM signal are negligible; an ideal SCM signal will have no sidelobes at all, whereas nonlinearities in any practical SCM modulator will produce small side lobes well below those of an OOK signal. The lack of out-of-band power results in very low adjacent-channel interference in DWDM systems. The digital signal processing in the SCM modems provides additional robustness with linear equalization to improve signals received through varied channels and with FEC more powerful than that used in other optical systems. The relatively low subcarrier clock rate allows the implementation of multitap, fractionally spaced linear equalizers at each receiver to combat linear distortion in the channel (see Figure 10.4(b)). The improved signal quality results in lower BER in highly distorted channels.

Many optical systems are now employing linear block-code FEC, typically Reed–Solomon block codes. However, with binary signaling the full power of FEC cannot be tapped. The use of higher-order modulation with SCM allows the traditional Reed–Solomon FEC to be paired with trellis coded modulation (TCM), a form of FEC that takes advantage of the modulation format to maximize the coding gains. Together these codes provide coding gains of 7 dB or more. Single-sideband (SSB) SCM eliminates the need for chromatic-dispersion compensation in all but the longest optical networks, even at very high aggregate bit rates. By eliminating the energy on one side of the optical carrier, the chromatic-dispersion tolerance is limited only by the symbol rate of the individual subcarriers, not by the aggregate signal bandwidth. Thus, a 20-Gbit/sec SCM signal, operating with OC-48 rate subcarriers, will exhibit chromatic-dispersion tolerance far exceeding that of a 10-Gbit/sec OC-192 system, and even that of an OC-48 system (see Figure 10.5).

Polarization-mode dispersion (PMD) is another impairment that gets worse with increased bandwidth. Unfortunately, SSB does not help in this case. However, the inherent bandwidth efficiency of an SCM signal limits the impact of PMD. A 40-Gb/s SCM signal exhibits greater PMD tolerance than an OC-192 signal. At 10 Gb/s, the PMD tolerance of SCM approaches that of an OC-48 signal. In networks where PMD is the primary limitation, an SCM signal can be designed to maximize the PMD tolerance.

Extensive performance metrics provided by the digital signal processing together with a decentralized architecture provide SCM systems with a very high level of performance monitoring. The underlying digital signal processing and RF processing at low clock rate provide BER and SNR measurement in real time. The channel BER is much higher in the operating region, providing BER degradation information to the operator as gradual or slight degradations occur, well before they accumulate to result in an outage. That provides a valuable maintenance corrective-action tool to avoid costly downtime. While SCM clearly provides the capability to increase optical-network capacity and improve fiber bandwidth efficiency, it also provides many other potential benefits to future optical networks. Flexibility in configuration and traffic mix combined with true transparency and carrier-class protection of all traffic reduces operations cost and improves provisioning. Highly robust optical links with constant optical power together with highly informative performance metrics and lower susceptibility to dispersions make SCM systems easier and less costly to design and maintain. Sub carrier multiplexing is clearly part of the future of optical networks, providing more than just capacity.

10.5 Conclusion

BEM importance in optical networks is evident with its efficient bandwidth use, enabling more users within a limited channel bandwidth, and the flexibility in upgrading existent networks as needed. PAM, SCM, and QAM are making important contributions to optical communications from local area to long haul. The results above indicate that there is no single, prominent modulation scheme. Bandwidth efficient modulation and DWDM have strong features that provide a desirable optical communication capacity and performance. When it comes to any one particular application it is important to look at the tradeoffs involved. While DWDM effectively utilizes bandwidth, SCM when combined with QAM requires less bandwidth. Furthermore, due to its frequency modulating characteristic, BEM shows a greater immunity to signal fluctuations. DWDM and BEM each provide beneficial features although neither dominates the other; both contribute to the advancement of optical communication systems.

Nomenclature

Symbol	Quantity	Symbol	Quantity
ASE	Amplified Spontaneous Emission	CSRZ	Carrier Suppressed RZ
ASON	Automatically Switched Optical Networks	DPSK	Differential Phase Shift Keying
		FEC	Forward Error Correction
ATM	Asynchronous Transfer Mode	GMPLS	Generalized Multi-Protocol
BER	Bit-Error Rate		Label Switching

Symbol	Quantity	Symbol	Quantity
IETF	Internet Engineering Task Force	PAM	Pulse Amplitude Modulation
IP	Internet Protocol	QAM	Quadrature Amplitude Modulation
ITU	International Telecommunications Union	QoS	Quality of Service
		RWA	Routing and Wavelength Assignment
MPLS	Multi-Protocol Label Switching		
OBS	Optical Burst Switching	RZ	Return to Zero
OADM	Optical Add Drop Multiplexer	SDH	Synchronous Digital Hierarchy
O-PNNI	Optical Private Network to Network Interface	SONNET	Synchronous Optical Network
		TCP/IP	Transmission Control Protocol/ Internet Protocol
OPS	Optical Packet Switching		
OSNR	Optical Signal-to-Noise Ratio	VPN	Virtual Private Network
OTDM	Optical Time Division Multiplexing	VSRZ	Vestigial Sideband RZ
OTN	Optical Transport Network	WDM	Wavelength Division Multiplexing
OXC	Optical Cross-Connect	WSS	Wavelength Selective Switch

References

1. A.R. Pratt, B. Charbonnier, P. Harper, D. Nesset, B.K. Nayar, and N.J. Doran, "40 × 10.7 Gb/s DWDM transmission over meshed ULH network with dynamically re-configurable optical cross connects," PD09, *Post deadline OFC*, Atlanta, March 2003.
2. D. Banerje and B. Mukerjee, "Wavelength-routed optical networks: linear Formulation, resource budgeting tradeoffs, and reconfiguration study," *Proc. IEEE INFOCOM*, 1997, pp. 269–210.
3. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*, Los Altos, CA: Morgan Kaufmann Publishers, 1998.
4. S. Waklin and J. Conradi, "Multilevel signaling for increasing the reach of 10 Gb/s lightwave systems," *J. Lightwave Technol.*, vol. 17, pp. 2235–2248, 1999.
5. H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. AIEE*, vol. 47, pp. 617–644, 1928.
6. W.H. Joseph, *Applications of Discrete and Continuous Fourier Analysis*, New York: Wiley, 1983.
7. C.E. Shannon, "Communication in the presence of noise," *Proc. Inst. Radio Eng.*, vol. 37, no.1, pp. 10–21, 1949.
8. C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding: turbo-codes (1)," *IEEE Int. Conf. Commun.*, May, 1064–1070, 1993.
9. W. Weaver and C.E. Shannon, *The Mathematical Theory of Communication*, Illinois, IL: Urbana, 1963, republished in paperback.
10. J.R. Marks II, *Introduction to Shannon Sampling and Interpolation Theory*, Berlin: Springer-Verlag, 1991.
11. W.W. Peterson, *Error Correcting Codes*, New York: Wiley, 1961.
12. S. Lin and D.J. Costello, *Error Control Coding: Fundamentals and Applications*, Englewood Cliffs, NJ: Prentice Hall, 1983.
13. O.E. Agazzi, T. Koh, S.S. Haider, R.W. Walden, D.R. Cassiday, G.A. Wilson, T.M. Lalumia, C.M. Gerveshi, J. Kumar, R.E. Crochiere, R.F. Shaw, R.A. Wilson III, W.R. McDonald, N.L. Gottfried, N.S. Ramesh, and R.B. Blake Jr., "A digital signal processor for an ANSI standard ISDN transceiver," *IEEE J. Solid-State Circ.*, vol. 24, no. 6, pp. 1605–1613, 1989.
14. A.M. Gottlieb, P.M. Crespo, J.L. Dixon, and T.R. Hsing, "The DSP implementation of a new timing recovery technique for high speed digital data transmission," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Albuquerque, New Mexico, 1990, pp. 1679–1682.
15. J.B. Anderson and D.P. Taylor, "A bandwidth-efficient class of signal-space codes," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 6, pp. 703–712, 1978.
16. J.B. Anderson, C-E.W. Sundberg, T. Aulin, and N. Rydbeck, "Power-bandwidth performance of smoothed phase modulation codes," *IEEE Trans. Commun.*, vol. COM-29, no. 3, pp. 187–195, 1981.

17. M. Shtaif and A. Mecozzi *Proceedings of Optical Fiber Communication Conference, Paper MM1*, Washington DC: Optical Society of America, 2001.
18. D. Dahan and G. Eisenstein, "Numerical comparison between distributed and discrete amplification in a point-to-point 40 Gb/s 40 WDM-based transmission system with three different modulation formats," *J. Lightwave Technol.*, vol. 20, no. 3, pp. 371–378, March 2002.
19. N. Antoniades, M. Lee, J.K. Rhee, M. Sharma, and A. Boskovic, "Extending the reach of WDM networks by combating high channel resolution power divergence using dynamic power equalization," LEOS, 14th Annual Meeting of the IEEE, vol. 1.1, pp. 354–355, 2001.
20. S. Lanne, D. Pennincks, J-P. Thiéry, and J-P. Hamaide, "Extension of polarization-mode dispersion limit using optical mitigation and phase-shaped binary transmission," *Proc. Opt. Fiber Commun. Conf.*, OFC, paper ThH3, 2000.
21. L.T Lima, R. Khosravani, O.H. Adamczyd, P. Ibrahimi, E. Ibragimov, A.E. Willner, and C.R. Menyuk, "Enhanced PMD mitigation using forward error correction coding and a first order compensator," Conference on Optical Communications, ThB4, 2000.
22. M. Tomizawa, Y. Kisaka, A. Hirano, and Y. Miyamoto, "PMD mitigation by frequency diverse detection receiver employing error correction function," *Proc. ECOC2002*, Copenhagen, Denmark.
23. F.Q. Zhou, M. Zhou, and J.J. Pan, "Optical coating computer simulation of narrow bandpass filters for dense wavelength division multiplexing," *Optical Interference Coating*, OSA Technical Digest Series 9, pp. 223–224, 1988.
24. J.D. Downie, I. Tomkos, N. Antoniades, and A. Boskovic, "Effects of filter concatenation for directly modulated transmission lasers at 2.5 and 10 Gb/s," *J. Lightwave Technol.*, vol. 20, no. 2, pp. 218–228, 2002.
25. R. Farjad-Rad, C.K. Yang, M. Horowitz, and T. Lee, "A 0.3-um CMOS 8 Gb/s 4-PAM serial link transceivers," *IEEE J. Solid State Circ.*, vol. 35, pp. 757–764, 2000.
26. R.A. Griffin, "Multi-level signaling with quadratic level spacing," UK Patent Application GB 2 366106A, 2001.
27. J. Poultron and W.J. Dally, "A tracking clock recovery receiver for 4 Gs/s signaling," *Hot Interconnects Symposium*, August 1997.
28. T.E. Darcie and G.E. Bodeep, "Lightwave subcarrier CATV transmission systems," *IEEE Trans. Microwave Theory Tech.*, vol. 38, no. 5, pp. 524–533, 1990.
29. L. Pophillat, "Optical modulation depth improvement in SCM lightwave systems using a dissymmetrization scheme," *IEEE Photon. Technol. Lett.*, vol. 6, no. 6, pp. 750–753, 1994.
30. J.H. Angenent, "Simple model for calculation of distortion in an optical analogue subcarrier multiplexed CATV system," *Electron. Lett.*, vol. 26, no. 24, pp. 2049–2050, 1990.
31. A.A.M. Saleh, "Fundamental limit on number of channels in subcarrier-multiplexed lightwave CATV system," *Electron. Lett.*, vol. 25, no. 12, pp. 710–777, 1989.

This page intentionally left blank

11

Phase-Locked Loop

Steven L. Maddy
SpectraLink Corporation

11.1   Introduction

A *phase-locked loop* (PLL) is a system that uses feedback to maintai.n an output signal in a specific phase relationship with a reference signal. PLLs are used in many areas of electronics to control the frequency and/or phase of a signal. These applications include frequency synthesizers, analog and digital modulators and demodulators, and clock recovery circuits. Figure 11.1 shows the block diagram of a basic PLL system. The *phase detector* consists of a device that produces an output voltage proportional to the phase difference of the two input signals. The VCO (voltage-controlled oscillator) is a circuit that produces an ac output signal whose frequency is proportional to the input control voltage. The *divide by N* is a device that produces an output signal whose frequency is an integer (denoted by *N*) division of the input signal frequency. The **loop filter** is a circuit that is used to control the PLL dynamics and therefore the performance of the system. The *F(s)* term is used to denote the Laplace transfer function of this filter.

Servo theory can now be used to derive the equations for the output signal phase relative to the reference input signal phase. Because the VCO control voltage sets the frequency of the oscillation (rather than the phase), this will produce a pure integration when writing this expression. Several of the components of the PLL have a fixed gain associated with them. These are the **VCO** control voltage to output frequency conversion gain (K_v), the **phase detector** input signal phase difference to output voltage conversion gain ($K\phi$), and the feedback division ratio (N). These gains can be combined into a single factor called the loop gain (K). This loop gain is calculated using Equation (11.1) and is then used in the following equations to calculate the loop transfer function.

$$K = \frac{K_\phi \times K_V}{N} \tag{11.1}$$

The closed-loop transfer function [$H(s)$] can now be written and is shown in Equation (11.2). This function is typically used to examine the frequency or time-domain response of a PLL and defines the relationship of the phase of the VCO output signal (θ_o) to the phase of the reference input (θ_i). It also describes the relationship of a change in the output frequency to a change in the input frequency. This function is low-pass in nature.

$$H(s) = \frac{\theta_o(s)}{\theta_i(s)} = \frac{KF(s)}{s + KF(s)} \tag{11.2}$$
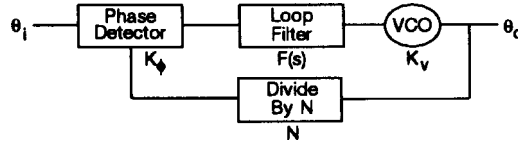
FIGURE 11.1 PLL block diagram.

The loop error function, shown in Equation (11.3), describes the difference between the VCO phase and the reference phase and is typically used to examine the performance of PLLs that are modulated. This function is high-pass in nature.

$$\frac{\theta_i(s) - \theta_o(s)}{\theta_i(s)} = \frac{\theta_e(s)}{\theta_i(s)} = \frac{s}{s + KF(s)} \tag{11.3}$$

The open-loop transfer function $[G(s)]$ is shown in Equation (11.4). This function describes the operation of the loop before the feedback path is completed. It is useful during the design of the system in determining the gain and phase margin of the PLL. These are indications of the stability of a PLL when the feedback loop is connected.

$$G(s) = \frac{KF(s)}{s} \tag{11.4}$$

These functions describe the performance of the basic PLL and can now be used to derive synthesis equations. The synthesis equations will be used to calculate circuit components that will give a desired performance characteristic. These characteristics usually involve the low-pass corner frequency and shape of the closed-loop response characteristic (Equation (11.2)) and determine such things as the loop lock-up time, the ability to track the input signal, and the output signal noise characteristics.

11.2 Loop Filter

The loop filter is used to shape the overall response of the PLL to meet the design goals of the system. There are two implementations of the loop filter that are used in the vast majority of PLLs: the passive lag circuit shown in Figure 11.2 and the active circuit shown in Figure 11.3. These two circuits both produce a PLL with a second-order response characteristic.

The transfer functions of these loop filter circuits may now be derived and are shown in Equation (11.5) for the passive circuit (Figure 11.2) and (11.6) for the active circuit (Figure 11.3).

$$F_p(s) = \frac{sC_1R_2 + 1}{s(R_1 + R_2)C_1 + 1} \tag{11.5}$$
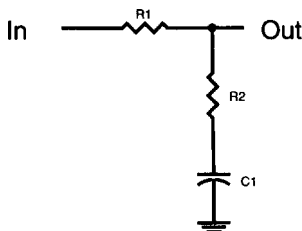
FIGURE 11.2 Passive loop filter.
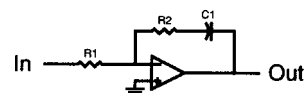
FIGURE 11.3 Active loop filter.

$$F_a(s) = \frac{sR_2G_1 + 1}{sR_1C_1} \tag{11.6}$$

These loop filter equations may now be substituted into Equation (11.2) to form the closed-loop transfer functions of the PLL. These are shown as Equation (11.7) for the case of the passive filter and 11.8 for the active.

$$H_p(s) = \frac{s\dfrac{KR_2}{R_1 + R_2} + \dfrac{K}{(R_1 + R_2)C_1}}{s^2 + s\left[\dfrac{1}{(R_1 + R_2)C_1} + \dfrac{KR_2}{R_1 + R_2}\right] + \dfrac{K}{(R_1 + R_2)C_1}} \tag{11.7}$$

$$H_a(s) = \frac{s\dfrac{KR_2}{R_1} + \dfrac{K}{R_1C_1}}{s^2 + s\dfrac{KR_2}{R_1} + \dfrac{K}{R_1C_1}} \tag{11.8}$$

These closed-loop equations can also be written in the forms shown below to place the function in terms of the **damping factor** (ζ) and the loop natural frequency (ω_n). It will be shown later that these are very useful parameters in specifying PLL performance. Equation (11.9) is the form used for the PLL with a passive loop filter, and Equation (11.10) is used for the active loop filter case.

$$H_p(s) = \frac{s[2\zeta\omega_n - (\omega_n^2/K)] + \omega_n^2}{s^2 + s2\zeta\omega_n + \omega_n^2} \tag{11.9}$$

$$H_a(s) = \frac{s2\zeta\omega_n + \omega_n^2}{s^2 + s2\zeta\omega_n + \omega_n^2} \tag{11.10}$$

Solving Equation (11.7) and Equation (11.9) for R_1 and R_2 in terms of the loop parameters ζ and ω_n, we now obtain the synthesis equations for a PLL with a passive loop filter. These are shown as Equation (11.11) and Equation (11.12).

$$R_2 = \frac{2\zeta}{\omega_n C} - \frac{1}{KC} \tag{11.11}$$

$$R_1 = \frac{K}{\omega_n^2 C} - R_2 \tag{11.12}$$

To maintain resistor values that are positive the passive loop filter PLL must meet the constraint shown in Equation (11.13).

$$\zeta > \frac{\omega_n}{2K} \tag{11.13}$$

For the active loop filter case Equation (11.8) and Equation (11.10) are solved and yield the synthesis equations shown in Equation (11.14) and Equation (11.15). It can be seen that no constraints on the loop damping factor exist in this case.
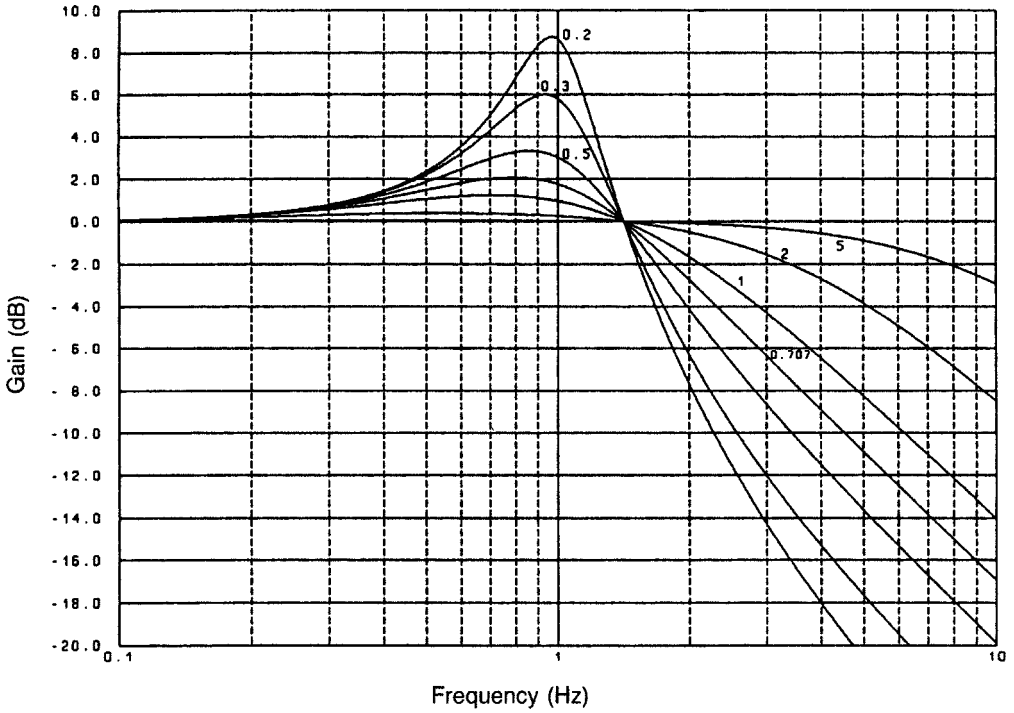
FIGURE 11.4   Closed-loop second-order type-2 PLL error response for various damping factors.
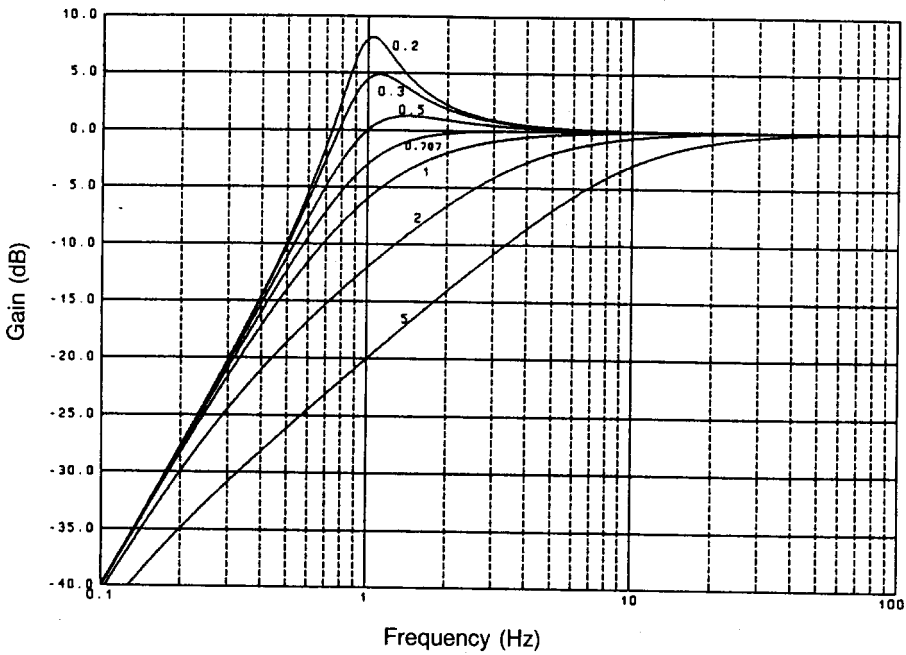
FIGURE 11.5   Closed-loop second-order type-2 PLL step response for various damping factors.
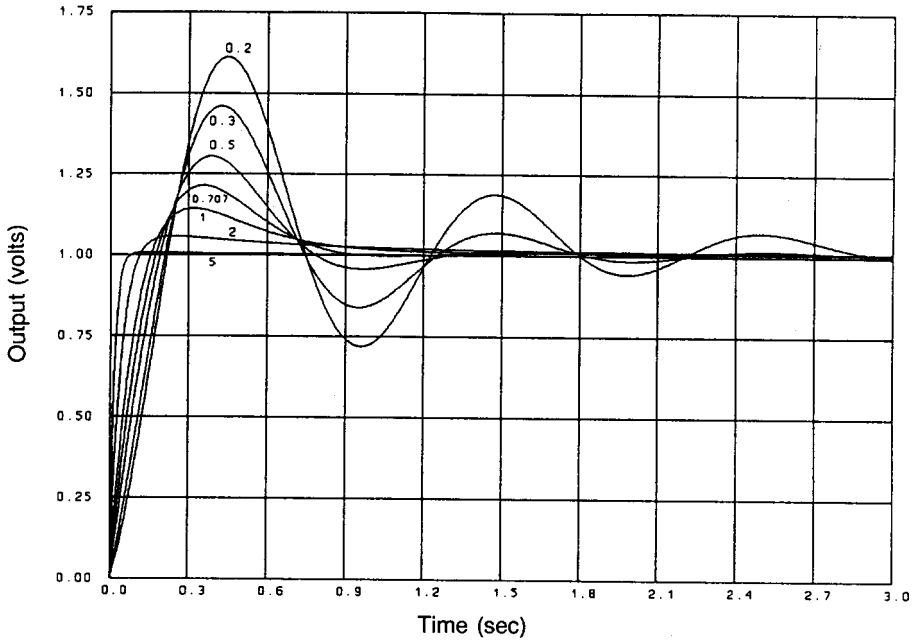
FIGURE 11.6   Closed-loop PLL response for various damping factors.

$$R_1 = \frac{K}{\omega_n^2 C} \tag{11.14}$$

$$R_2 = \frac{2\zeta}{\omega_n C} \tag{11.15}$$

A typical design procedure for these loop filters would be, first, to select the loop damping factor and natural frequency based on the system requirements. Next, all the loop gain parameters are determined. A convenient capacitor value may then be selected. The remaining resistors can now be computed from the synthesis equations presented above.

Figure 11.4 shows the closed-loop frequency response of a PLL with an active loop filter (Equation (11.10)) for various values of damping factor. The loop natural frequency has been normalized to 1 Hz for all cases.

Substituting Equation (11.6) into Equation (11.3) will give the loop error response in terms of damping factor. This function is shown plotted in Figure 11.5. These plots may be used to select the PLL performance parameters that will give a desired frequency response shape.

The time response of a PLL with an active loop filter to a step in input phase was also computed and is shown plotted in Figure 11.6.

11.3   Noise

An important design aspect of a PLL is the noise content of the output. The dominant resultant noise will appear as phase noise (jitter) on the output signal from the VCO. Due to the dynamics of the PLL some of these noise sources will be filtered by the loop transfer function (Equation (11.2)) that is a low-pass characteristic. Others will be processed by the loop error function (Equation (11.3)) that is a high-pass characteristic. Table 11.1 shows the major sources of noise in a PLL and the effect of the loop dynamics on this noise. All these factors must be combined to evaluate the complete noise performance of a PLL. Often it will be

TABLE 11.1   PLL Noise Sources

Noise Source	Filter Function
Reference oscillator phase noise	Low pass
Phase detector noise	Low pass
Active loop filter input noise	Low pass
Digital divider noise	Low pass
Active loop filter output noise	High pass
VCO free-running phase noise	High pass

found that one particular noise source will be dominant and the PLL performance can then be adjusted to minimize the output noise.

A PLL is frequently used to enhance the noise performance of an oscillator by taking advantage of these noise-filtering properties. For example, a crystal oscillator typically has very good low-frequency noise characteristics, and a free-running LC oscillator can be designed with very good high-frequency noise performance but will exhibit poor low-frequency noise characteristics. By phase-locking an LC oscillator to a crystal oscillator and setting the loop response corner frequency to the noise crossover point between the two oscillators, the desirable characteristics of both oscillators are realized.

When designing frequency synthesizers using PLLs, care must be taken to prevent noise from the PLL components from introducing excessive noise. The divider ratio (N) used in the feedback of the loop has the effect of multiplying any noise that appears at the input or output of the phase detector by this factor. Frequently, a large value of N is required to achieve the desired output frequencies. This can cause excessive output noise. All these effects must be taken into account to achieve a PLL design with optimum noise performance.

11.4   PLL Design Procedures

The specific steps used to design a PLL depend on the intended application. Typically the architecture of the loop will be determined by the output frequency agility required (frequency synthesizer) and the reference sources available. Other requirements such as size and cost play important factors, as well as available standard components. Once the topology has been determined, then the desired loop transfer function must be synthesized. This may be dictated by noise requirements as discussed above or other factors such as loop lock-up time or input signal tracking ability. The design Equation (11.11) through Equation (11.15) may then be used to determine the component values required in the loop filter.

Frequently several of these factors must be balanced or traded off to obtain an acceptable design. A design that requires high performance in several of these areas usually can be realized at the expense of design complexity or increased component cost.

11.5   Components

The development of large-scale integrated circuits over the past several years has made the design and implementation of PLLs and frequency synthesizers much cheaper and easier. Several major manufacturers (Motorola, Signetics, National, Plessey, etc.) currently supply a wide range of components for PLL implementation. The most complex of these are the synthesizer circuits that provide a programmable reference divider, programmable divide by N, and a phase detector. Several configurations of these circuits are available to suit most applications. Integrated circuits are also available to implement most of the individual blocks shown in Figure 11.1.

A wide variety of phase detector circuits are available, and the optimum type will depend on the circuit requirements. An analog multiplier (or mixer) may be used and is most common in applications where the

comparison frequency must be very high. This type of phase detector produces an output that is the multiplication of the two input signals. If the inputs are sine waves, the output will consist of a double-frequency component as well as a dc component that is proportional to the cosine of the input phase difference. The double-frequency component can be removed with a **low-pass filter**, leaving only the dc component. The analog multiplier has a somewhat limited phase range of ± 90 degrees. The remainder of the phase detector types discussed here are digital in nature and operate using digital edges or transitions of the signals to be compared.

The sample-and-hold phase detector is widely used where optimum noise performance is required. This circuit operates by using one of the phase detector inputs to sample the voltage on the other input. This latter input is usually converted to a triangle wave to give a linear phase detector characteristic. Once the input is sampled, its voltage is held using a capacitor. The good noise performance is achieved since most of the time the phase detector output is simply a stored charge on this capacitor. The phase range of the sample-and-hold phase detector depends on the type of waveform shaping used and can range from ± 90 to ± 180 degrees.

One of the simplest types of phase detectors to implement uses an exclusive OR gate to digitally multiply the two signals together. The output must then be low-pass filtered to extract only the dc component. The main drawback to this circuit is the large component that exists in the output at twice the input frequency. This requires a large amount of low-pass filtering and may restrict the PLL design. The phase range of this type of circuit is ± 90 degrees.

One of the main drawbacks of all the above types of phase detectors is that they only provide an output that is proportional to phase and not to a frequency difference in the input signals. For many applications the PLL input signals are initially not on the same frequency. Several techniques have been used in the past to resolve this such as sweeping the VCO or using separate circuitry to first acquire the input frequency. The sequential (sometimes called phase/frequency) phase detector has become the most commonly used solution due to its wide availability in integrated form. This type of phase detector produces pulses with the width of the pulses indicating the phase difference of the inputs. It also has the characteristic of providing the correct output to steer the VCO to the correct frequency. The noise characteristic of this type of phase detector is also quite good since either no or very narrow pulses are produced when the inputs are in phase with each other. The phase range of this type of circuit is ± 360 degrees.

Digital dividers are widely available and may either have programmable or fixed division ratios depending on the application. For optimum noise performance a synchronous type of divider should be used. When a programmable divider is required to operate at a high frequency (>50 MHz), a dual modulus circuit is normally used. This circuit uses a technique called *pulse swallowing* to extend the range of normal programmable divider integrated circuits by using a dual modulus prescaler (usually ECL). The dual modulus prescaler is a high-frequency divider that can be programmed to divide by only two sequential values. A second programmable divider section is then used to control the prescaler. Further details of this type of divider are available from component manufacturers' data sheets as well as in the references.

The voltage-controlled oscillator is typically the most critical circuit in determining the overall noise performance of a PLL. For this reason it is often implemented using discrete components, especially at the higher frequencies. Some digital integrated circuits exist for lower-frequency VCOs, and microwave integrated circuit VCOs are now available for use to several gigahertz. The major design parameters for a VCO include the operating frequency, tuning range, tuning linearity, and phase noise performance. Further information on the design of VCOs is contained in the references.

Loop filters used in PLLs may be either active or passive depending on the specific application. Active filters are normally used in more critical applications when superior control of loop parameters and reference frequency suppression is required. The loop filter is typically followed by a low-pass filter to remove any residual reference frequency component from the phase detector. This low-pass filter will affect the calculated loop response and will typically appear to reduce the loop damping factor as its corner frequency is brought closer to the loop natural frequency. To avoid this degradation the corner frequency of this filter should be approximately one order of magnitude greater than the loop natural

frequency. In some cases a notch filter may be used to reduce the reference frequency when it is close to the reference frequency.

11.6   Applications

Phase-locked loops are used in many applications including frequency synthesis, modulation, demodulation, and clock recovery. A frequency synthesizer is a PLL that uses a programmable divider in the feedback. By selecting various values of division ratio, several output frequencies may be obtained that are integer multiples of the reference frequency (Fref). Frequency synthesizers are widely used in radio communications equipment to obtain a stable frequency source that may be tuned to a desired radio channel. Since the output frequency is an integer multiple of the reference frequency, this will determine the channel spacing obtained. The main design parameters for a synthesizer are typically determined by the required channel change time and output noise.

Transmitting equipment for radio communications frequently uses PLLs to obtain frequency modulation (FM) or phase modulation (PM). A PLL is first designed to generate a radio frequency signal. The modulation signal (i.e., voice) is then applied to the loop. For FM the modulating signal is added to the output of the loop filter. The PLL will maintain the center frequency of the VCO, while the modulation will vary the VCO frequency about this center. The frequency response of the FM input will exhibit a high-pass response and is described by the error function shown in Equation (11.3). Phase modulation is obtained by adding the modulation signal to the input of the loop filter. The modulation will then vary the phase of the VCO output signal. The frequency response of the PM input will be a low-pass characteristic described by the closed-loop transfer function shown in Equation (11.2).

A communications receiver must extract the modulation from a radio frequency carrier. A PLL may be used by phase locking a VCO to the received input signal. The loop filter output will then contain the extracted FM signal, and the loop filter input will contain the PM signal. In this case the frequency response of the FM output will be a low-pass function described by the closed-loop transfer function and the PM output response will be a high-pass function described by the error function.

In digital communications (modems) it is frequently necessary to extract a coherent clock signal from an input data stream. A PLL is often used for this task by locking a VCO to the input data. Depending on the type of data encoding that is used, the data may first need to be processed before connecting the PLL. The VCO output is then used as the clock to extract the data bits from the input signal.

Defining Terms

Capture range: The range of input frequencies over which the PLL can acquire phase lock.

Damping factor: A measure of the ability of the PLL to track an input signal step. Usually used to indicate the amount of overshoot present in the output to a step perturbation in the input.

Free-run frequency: The frequency at which the VCO will oscillate when no input signal is presented to the PLL. Sometimes referred to as the rest frequency.

Lock range: The range of input frequencies over which the PLL will remain in phase lock once acquisition has occurred.

Loop filter: The filter function that follows the phase detector and determines the system dynamic performance.

Loop gain: The combination of all dc gains in the PLL.

Low-pass filter: A filter that usually follows the loop filter and is used to remove the reference frequency components generated by the phase detector.

Natural frequency: The characteristic frequency of the PLL dynamic performance. The frequency of the closed-loop transfer function dominant pole.

Phase detector gain: The ratio of the dc output voltage of the phase detector to the input phase difference. This is usually expressed in units of volts/radian.

VCO gain: The ratio of the VCO output frequency to the dc control input level. This is usually expressed in units of radians/second/volt.

References

AFDPLUS Reference Manual, Boulder, Colo.: RLM Research, 1991 (software used to generate the graphs in this section).

R.G. Best, *Phase-Locked Loops—Theory, Design & Applications*, New York: McGraw-Hill, 1984.

A. Blanchard, *Phase-Locked Loops, Application to Coherent Receiver Design*, New York: Wiley Interscience, 1976.

W.F. Egan, *Frequency Synthesis by Phase Lock*, New York: Wiley Interscience, 1981.

F.M. Gardner, *Phaselock Techniques*, New York: Wiley, 1979.

J. Gorski-Popiel, *Frequency Synthesis; Techniques & Applications*, Piscataway, N.J.: IEEE Press, 1975.

W.C. Lindsey and M.K. Simon, *Phase-Locked Loops & Their Applications*, Piscataway, N.J.: IEEE Press, 1978.

V. Manassewtsch, *Frequency Synthesizers: Theory and Design*, New York: Wiley Interscience, 1980.

U.L. Rhode, *Digital PLL Frequency Synthesizers Theory and Design*, Englewood Cliffs, N.J.: Prentice-Hall, 1983.

Further Information

Recommended periodicals that cover the subject of PLLs include *IEEE Transactions on Communications*, *IEEE Transactions on Circuits and Systems*, and *IEEE Transactions on Signal Processing*. Occasionally articles dealing with PLLs may also be found in *EDN*, *Electronic Design*, *RF Design*, and *Microwaves and RF Magazine*. A four-part PLL tutorial article titled *PLL Primer*, by Andrzej B. Przedpelski, appeared in *RF Design Magazine* in the March/April 1983, May/June 1983, July/August 1983, and November 1987 issues.

Another good source of general PLL design information can be obtained from application notes available from various PLL component manufacturers. *Phase-Locked Loop Design Fundamentals*, by Garth Nash, is available from Motorola, Inc. as AN-535 and gives an excellent step-by-step synthesizer design procedure.

This page intentionally left blank

12
Telemetry

Stephen Horan
New Mexico State University

12.1   Introduction

Telemetry systems are found in a variety of applications—from automobiles, to hospitals, to interplanetary spacecraft. Although these examples represent a broad range of applications, they all have many characteristics in common: a natural parameter is measured by a sensor system, the measurement is converted to numbers or data, the data are transported to an analysis point, and an end user makes use of the data gathered. This series of actions fulfills the definition of *telemetry*, which is to "measure at a distance." It is not unusual for the data collection function to be part of an overall control system where information also flows back from the monitoring point to modify the measurement discipline or to initiate some action. For example, in controlling a manufacturing process, the telemetry data about the process will cause the process monitor to change a parameter such as a temperature or flow rate to optimize the process. The flow of information back is telecommand data from the monitoring point.

In this chapter, we will look at an overview of the system-level design issues involved in the synthesis of a telemetry and telecommand system. Further details can be found in Ref. [1].

12.2   Basic Concepts

The general purpose of a telemetry system is to gather information about a subject of interest and present that measurement to a user. The physical quantity or property that is being measured is known as the **measurand**. Each measurand is sampled by a sensor at a rate appropriate for the bandwidth of the signal. The sampled analog signals are then frequently digitized by using an analog-to-digital converter. This produces a **pulse-coded modulation** (**PCM**) representation of the measurand. This process is illustrated in Figure 12.1 where an analog signal is sampled as a function of time. The sampled signal is represented by a voltage at the sensor output. This voltage is converted to the PCM representation in the analog-to-digital conversion process to produce one digital sample for each sample. The user of the measurements will reverse this process to estimate the measurand value as a function of time. To produce an accurate estimate of the measurand, the sensor and conversion process will need to be calibrated and the calibration coefficients incorporated into the inverse transformation. Measurement systems requiring the sampling of more than a single sensor require more complicated support and coordination to allow sampling of several measurands.

Time	Volt	PCM
0.00	0.0509	257
0.05	−0.0712	254
0.10	0.0000	256
0.15	0.0734	257
0.20	−0.0554	254
0.25	−0.1558	252
0.30	0.0000	256
0.35	0.1620	259
0.40	0.0606	257

Voltage

Time

Analog Voltage Signal

Digital Representation
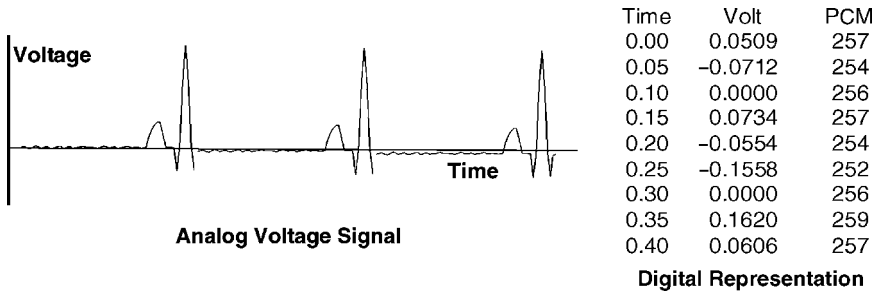
FIGURE 12.1   Sampling an analog voltage signal and converting it to a PCM digital signal.

The measurement system is typically located at a distance from the eventual data user, which can be a person or a machine. The measurements are transmitted over some type of data channel, for example, a radio link, a fiber optic cable, or a telecommunications network. This architecture produces a telemetry system having the characteristics of the standard telecommunications definition of **telemetry** that is [2]:

1. The use of telecommunication for automatically indicating or recording measurements at a distance from the measuring instrument.
2. The transmission of non-voice signals for the purpose of automatically indicating or recording measurements at a distance from the measuring instrument.

The control functions in a telemetry system are accomplished through **actuators**, which are devices that respond to transmitted control signals to affect the measurement system characteristics or interact with the measurement environment. The control is achieved through a **telecommand** link, which is the "use of telecommunication for the transmission of signals to initiate, modify or terminate functions of equipment at a distance." [2] The data flow in a telemetry/telecommand system is illustrated in Figure 12.2. The telemetry measurements from the sensors flow to the user over the data channel.

The sampling and multiplexing components format and sequence order the data for transmission over the channel. The data processing returns the measurement values to estimates of the actual measurand values for presentation to the user. In the user interface, the PCM signal will either be converted to an analog waveform or left as a series of discrete measurements for use and analysis. The user interface may be a display screen, a set of gages, or a chart recorder. The user interface often contains data-logging capabilities to provide a permanent record of the measurements.

The user also enters commands at the user interface to control the actuators. The data processing prepares the user input into telecommand data for transmission over the data channel. The control functions then interpret these commands and cause the actuators to respond. The design of the telemetry and telecommand data structures is generally individualized for each system by the system designers. The overall structure of the data structures can be classified as either frames or packets as discussed later.

Most telecommand systems transmit the command data as a small packet or series of data words. It is not unusual for the telemetry data transmission rate to exceed the telecommand transmission rate by a factor of 100 to 1000 to account for the relative differences in the real-time data volumes between the two types of data.

The data channel in Figure 12.2 can take on different configurations depending upon the system. Many telemetry systems found in aerospace, remote sensing, and even hospitals use some form of radio link to transmit the data. These channels are usually susceptible to reasonably high error rates and interference from other users. At measurement sites where networking drops are available, the network backbone can be used as the data channel. In this case, the channel is high quality coaxial cable or fiber optic cable; both having much lower channel error rates than possible with radio links. This allows the system designer to have access to the high quality of the network channel and leverage off the networking protocols found in the telecommunications infrastructure. The use of error correcting encoding of the data will need to be determined by the system designer based on the reliability of the channel and the criticality of the data.
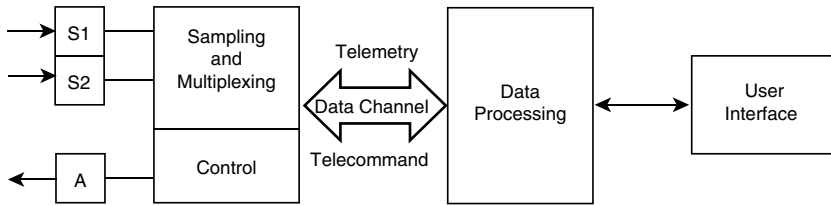
FIGURE 12.2   Overall data flow in a telemetry and telecommand system composed of sensors (S) and actuators (A).

Remote science telemetry systems measuring natural phenomena are a type of one-of-a-kind data systems that may have error-correcting coding applied since the data cannot be reacquired if the channel corrupts the transmission. However, process-control data may not have error corrections applied because the corrupted data can be replaced by a new sample almost immediately.

12.3   Data Measurements

The telemetry system will primarily produce PCM data as the result of sampling the analog sensor output. However, there are other data types that may also be present in the system. The telemetry data types include:

- PCM data—sampled analog values.
- Digital data—naturally occurring digital values such as event counts.
- Bi-level data—binary values such as switch settings.
- Timing data—to record when the measurements were taken.
- Command echoes—repeats of telecommand data for operator verification.

The sampling rate for each sensor is determined by the signal's **Nyquist sampling rate**. If W is the signal bandwidth in hertz, then the Nyquist sampling rate, f_N, in samples per second, is given by:

$$f_N = 2W \tag{12.1}$$

In practice, a minimal sampling rate of five times the signal bandwidth is typically used to accurately reconstruct the measurand.

The process of making a PCM measurement involves more than sampling the output of a sensor. As shown in Figure 12.3, the actual measurement process involves the response functions of:

- the transducer converting the input signal to be measured into its electrical analog (voltage or current),
- the signal conditioning electronics that make the signal capable of being reliably measured by providing any necessary signal gain, signal filtering, or buffering that is required by the measurement device,
- the measurement device, usually an analog-to-digital converter, to transform the electrical signal into its digital representation, and
- the data formatter to efficiently package the data measurements for transmission.

Each of these processes may impart a non-linear transformation to the measurement that will need to be removed by inverse processing before presenting the data to the user. This will need to be calibrated as discussed below.

Digital, bilevel, and command echo data types are usually read directly from a digital storage location and do not need to be further processed. Rather, they are taken directly by the data formatter to place those measurements in the transmission data stream.

As mentioned above, it is typical to time tag the data to record when it was measured. Telecommands may also be time tagged to specify the execution time. Time tags come in two typical formats: an integer count or a formatted set of values to represent time and date. The position tags can be encoded as a character string or floating point number.
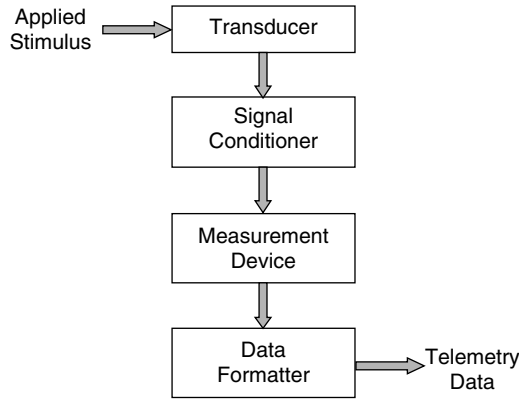
FIGURE 12.3 Measurement electronics to convert a signal to its electronic representation.
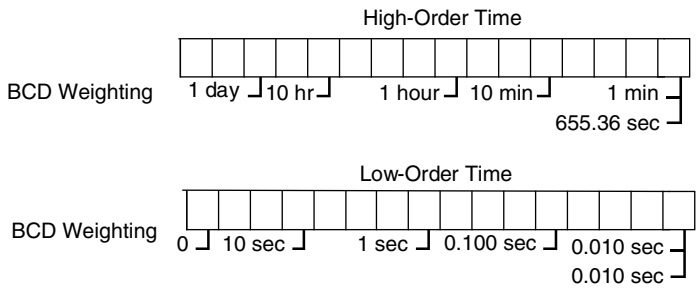
FIGURE 12.4 Formatted time tag configuration for representing time and date.

The integer time tag represents an elapsed count of seconds since a specific reference time. Many computer languages have this as a built-in system function and the time tag can be this system value. For a single day, the number of elapsed seconds since midnight can be represented as a 17-bit integer. Generally, a 32-bit integer can hold the total number of elapsed seconds for most applications and reference times. An alternative to the integer method is to use a formatted sequence such as that shown in Figure 12.4. Here, the fields are used to encode date and time. While this format requires more bits than the integer method to hold the result, it is easier for humans to understand. In both cases, the time can be referenced to one of several standards. The traditional time standard was provided by timing frames broadcast by radio station WWVB or other standard time services. In the past several years, the use of Global Positioning Satellites (GPS) with receivers having standard computer interfaces have become more popular as the means to provide a local time reference. For networked systems, time servers using the Network Time Protocol (NTP) are a convenient way to keep the system synchronized with the correct time.

In many systems, the data also needs to be position tagged to specify where the data sample was made. The position tags can be easily determined from one of the standard GPS navigation messages having position information in the message. An example of a GPS GGA-type navigation message, with both time and position information, is given in Figure 12.5 where the fields for the time, longitude, and latitude are indicated. These messages are in text format and the fields can be incorporated directly into the telemetry data stream.

The measurement system will need to be calibrated to properly convert each measurement from the received raw PCM value, sometimes called **engineering units**, to the natural set of units for the measurement. The calibration process involves supplying a known input, which should be referenced to a calibrated version of the standard value, to the measurement system and observing the resulting output. The application of the known standard input values should cover the whole expected range to be experienced during the measurement

```
$GPGGA,190719.72,3216.5110,N,10644.9837,W,1,03,3.7,00025,M,,,,*30
```

Message   Time (UTC)   Latitude   Longitude
ID

FIGURE 12.5 Example of the standard GGA type of GPS navigation message showing timing and position information.
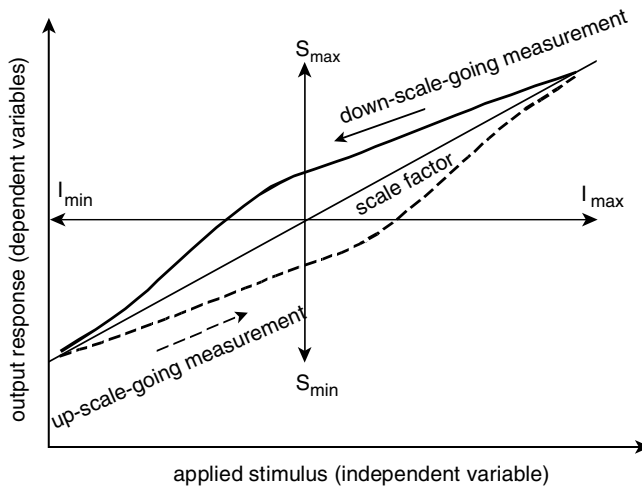
FIGURE 12.6 An example of a transducer response function showing hysteresis in the measurement profile.

process and include both increasing and decreasing ranges of the input. This will show any hysteresis that may be in the transducer response function as illustrated in Figure 12.6. The calibration process should also include any signal conditioning and signal processing electronics to ensure that the transfer functions from the devices in Figure 12.3 are also included. From this set of measurements, a mathematical model is generated to map the measured outputs onto the input values. This mathematical model is applied in the user interface to produce a processed measurement that can be related to a standard and with proper units.

12.4 Data Channels

The configuration and quality of the data channel will influence the design of the data transmission structure. Telemetry systems have two general channel configurations: either a common channel mode or a distributed channel mode, as illustrated in Figure 12.7. In common channel mode, all of the telemetry and telecommand data traverses a single, logical link. This link may be either full-duplex or parallel simplex links, depending upon the nature of the physical channel and the access protocol. In this mode, all of the sensor data is combined at a common data acquisition system prior to transmission over the channel. Typical environments for this mode include rocket payloads, laboratory data acquisition systems, and remote vehicle telemetry systems. The common channel is frequently a radio link, a wire, or a local area network.

In the distributed mode channel, each sensor connects to the channel via a transceiver. For example, this transceiver can be a simple IEEE 802.11 interface that permits the sensors to be connected as part of a wireless network with a central computer receiving the data via a wireless access point. Typical environments for this mode include building temperature and fire sensing or factory process measurement and control.

With the common channel transmission mode, the system will need to balance the data sampling requirements from the Nyquist rate with the need to share the channel between the sensors. This is done via a time division multiplexing within the frame or packet format used for the data transmission. With the time

multiplexing of the sensor data, there will need to be an accompanying synchronization algorithm to synchronize the receiver to the incoming data stream to fully recover the data.

The distributed channel can use either polling techniques or code division multiple access techniques to give each sensor shared access to the channel. Because the individual sensor is only sending its own data, synchronization is easier to achieve.

There are specific frequencies allocated for telemetry and telecommand transmission in the radio spectrum. This is in addition to other, generic frequency allocations for data transmission. Examples of telemetry bands are given in Table 12.1. These bands represent those used for specific applications such as biomedical telemetry
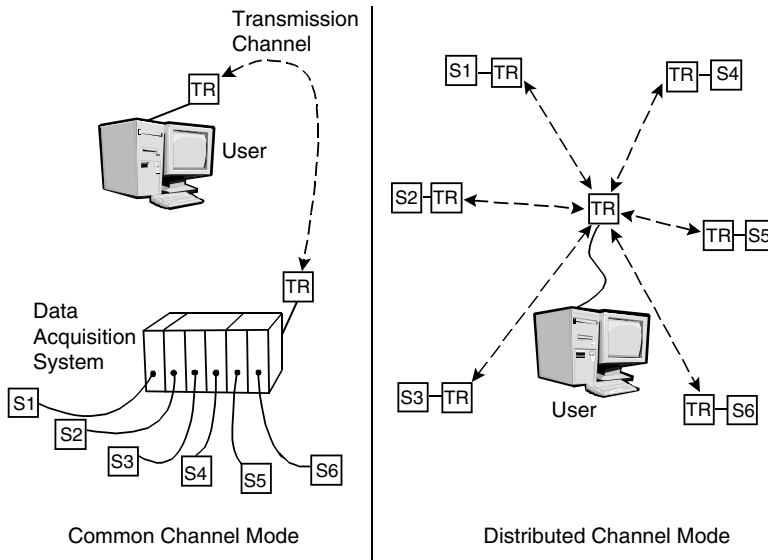
FIGURE 12.7 Common and distributed data channels for telemetry and telecommand systems. The data sources are the sensors (S) and the channel interface occurs at a transceiver (TR).

TABLE 12.1 Specific Telemetry Bands for Various Applications [3,5]

Frequency Range	Usage
38–41 MHz	Biomedical telemetry band
174–216 MHz	Biomedical telemetry band
400–406 MHz	Radiosonde; space operation; Earth exploration satellite
449.75–450.25 MHz	Space operation and space research services
460.650–460.875 MHz	Biomedical telemetry band
465.650–465.875 MHz	
902–928 MHz	Industrial, scientific, and medical (ISM) band
1400–1427 MHz	Earth exploration satellite
1427–1435 MHz	Land mobile (telemetering and telecommand); fixed (telemetering)
1435–1535 MHz	Aeronautical telemetry
1525–1530 MHz	Mobile service, including aeronautical telemetry, as a secondary allocation
1530–1535 MHz	Mobile service, including aeronautical telemetry, as a secondary allocation
1670–1700 MHz	Radiosonde
2200–2290 MHz	Government fixed, mobile, space research, space operation, and Earth-exploration services
2290–2300 MHz	Space research for deep space shared with fixed and mobile (except aeronautical mobile) services
2310–2390 MHz	Flight testing of manned or unmanned aircraft, missiles, space vehicles, or their major components
2310–2360 MHz	Aeronautical telemetry as a secondary service
2360–2390 MHz	Mobile service (will become secondary after 2005 or 2007, depending upon location)
2400–2500 MHz	ISM band
5725–5875 MHz	ISM band

or space operations as well as those commonly used in scientific and commercial applications such as the Industrial, Scientific, and Medical (ISM) bands. Some of these bands are allocated for governmental use as part of the national test ranges while others are open to nongovernment users as well. System designers need to consult the radio frequency allocation tables [5] to determine proper allocations for planned frequency use. If the allocations in Table 12.1 are being used outside of the United States, they will need to be checked in the country of use to verify that the planned usage is consistent with national allocations.

12.5   Data Transmission Formats

As mentioned above, the data channel will need to be shared among the various sensors in the telemetry system when using the common transmission channel mode. The channel bandwidth and quality are limiting factors in the system because they restrict the volume of data that can be reliably sent and they determine the data synchronization methodology. The channel bandwidth will affect the speed at which each sensor can be sampled to ensure that the Nyquist sampling rate is satisfied. Because different sensors will have different sampling rates, a time-division multiplexing technique is frequently chosen as the means to share the data channel. For channels where data loss caused by data drop outs is a significant problem or where tight synchronization is required between the sensor system and the user interface, a continuous frame transmission data structure is used. For systems where the channel provides high reliability or there are many different sampling rates possible, a packet data structure may be preferred. In this section, we will look at both methods.

Frame telemetry is the traditional method for time-multiplexing telemetry data from the source to the destination. The multiplexing structure repeats in fixed-time intervals to allow the receiving side to synchronize to the transmitter based on the contents of the data stream alone. Most frame systems use a format standard called IRIG 106 [3], developed by the military test ranges' Inter-Range Instrumentation Group (IRIG). This frame structure is used industry-wide and not just for military systems. The frame structure is arranged like a matrix, as shown in Figure 12.8, with the **major frame** being one complete cycle through the time-multiplexing sequence, during which time each sensor is sampled at least once. Each major
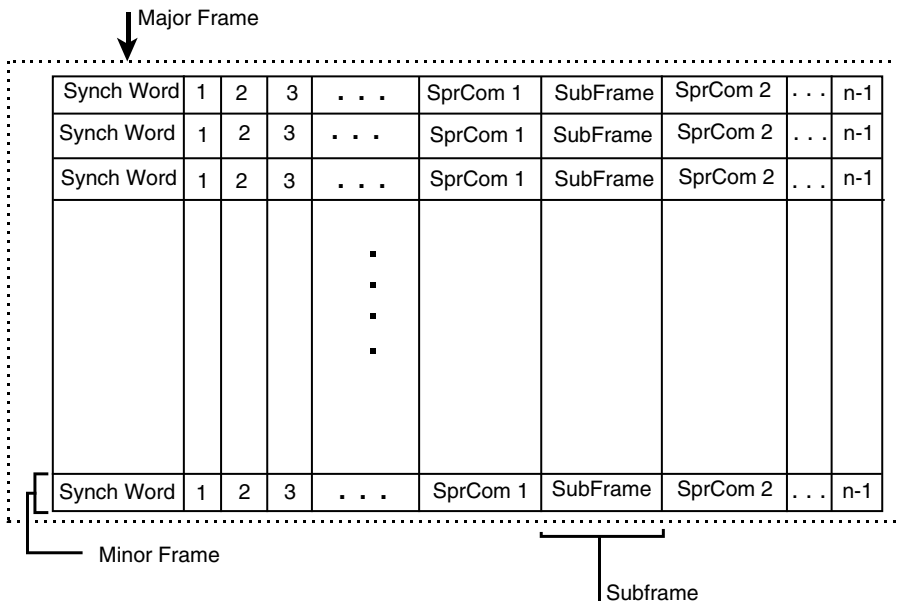
FIGURE 12.8   The standard IRIG telemetry frame structure for major frames, minor frames, and subframes.

frame is divided along rows of the matrix into **minor frames** that begin with a synchronization word. The columns of the matrix in the major frame can represent either the output from sensors or management data repeating in each minor frame. A sensor's output reading stays in its assigned column from frame to frame. If the same sensor is sampled once per minor frame—for example, columns 1, 2, and 3 in Figure 12.8—then it is sampled at the commutation rate for the system and is called **commutated data**. If a small number of sensors need to be sampled at a rate higher than the commutation rate, they are **supercommutated** sensors ("SprCom" locations in Figure 12.8) and take up more than one column in the matrix. Some columns in the matrix may represent a group of sensors rather than one sensor. As shown in the column labeled "subframe" in Figure 12.8, the sensors in that group form a **subframe** within the major frame. The sensors are sampled at a rate below the commutation rate, so the data are **subcommutated**.

Standard synchronization codes are usually 16 or 32 bits long. For example, the standard 16-bit code is EB90 in hexadecimal, whereas the standard 32-bit code is FE6BA840 in hexadecimal. These codes are optimal in the sense of having the lowest autocorrelation values when shifted by one bit. This minimizes the probability of a false lock in the synchronization-detection circuitry. All synchronization codes suffer the possibility of being mimicked by random data. Therefore, several occurrences of the code are frequently required upon initial synchronization. The condition in which the synchronizing circuitry locks onto random data and not the synchronization word is called a false lock. The probability of this false lock condition occurring, P_{FL}, is given by:

$$P_{\text{FL}} = \frac{\sum_{i=0}^{k} \binom{N}{i}}{2^N} \qquad (12.2)$$

Here N is the length of the synchronization code in bits and k is the number of differences allowed between the received code and the exact code value.

The synchronization code can also be missed if the data become corrupted in the channel. The probability of a missed synchronization code, P_{M}, caused by channel errors is given by:

$$P_{\text{M}} = \sum_{i=k+1}^{N} \binom{N}{i} p^i (1-p)^{N-i} \qquad (12.3)$$

Here p is the channel bit-error rate while N and k are as before.

If the minor frame has a length of L bits, then with a channel BER of p, the frame data has a probability of correct reception, P_0, given by:

$$P_0 = (1-p)^L \qquad (12.4)$$

This can be used by the designer to estimate necessity for any error detection and correction coding to be applied to the transmission. Equation (12.4) can also be used to estimate the probability of correct reception for command data words.

Packet telemetry systems are becoming more common, especially in systems where the data acquisition and data reception subsystems have high degrees of computational capability and the link between them can be viewed as a reliable link. Packet systems have several advantages over frame systems, with the main advantage being flexibility. With a packet system, instead of having a master commutation rate, the sampling rate for each sensor or sensor system can be individualized to the natural signal bandwidth. For example, battery voltages being measured may not change significantly over five minutes. With frame telemetry they may be sampled more frequently than once a minute because of the commutation rate specification. With packets, the voltages may be sampled only as needed and then transmitted in a data packet. Packets also have the advantage of allowing the data to be more easily routed over a computer network for analysis and distribution to end users. Examples of packet telemetry formats are given in Ref. [4].

The general packet format is composed of a header, containing accounting and addressing information, followed by the actual data similar to the format in Figure 12.9. The packet may end with a trailer composed of

Packet Header	Time Stamp	Sensor 1	Sensor 2	Sensor 3	Sensor 4
protocol specific	4 bytes	2 bytes	2 bytes	2 bytes	2 bytes

FIGURE 12.9   An example of a telemetry packet.

error-checking codes or other administrative information. The addressing information in the packet identifies the sensor system originating the packet and the destination process for analysis. Other information included in the header might be counters to identify the sequence number or a time stamp to show when the packet was created. The header will often contain a size parameter to specify the length of the data field.

For actual transmission across the data channel, a link-layer packet may be used. This link packet may multiplex the data packets from several subsystems together for efficient transport. If the data channel is synchronous, as with a radio channel, then it is common for the channel packets to be sent at regular intervals to maintain transmission synchronization. When this is done, fill packets are used to keep the channel active if there are no actual data to be sent. The packet header will then have a special code to indicate that the packet is a fill packet and should not be processed.

The packet usually begins with a synchronization marker, just as frame telemetry does. The same synchronization codes can be used in packet systems as in frame systems. After synchronization the header is analyzed to identify the type of processing to be performed based on the source of the data. The probability of missing the synchronization marker because of channel errors is computed as in Equation (12.3) and the probability of receiving the packet correctly is computed as in Equation (12.4). Many commercial protocols have error detection as part of the protocol specification so this can be used to assess data quality in the packet.

Both frame and packet telemetry systems require a synchronization phase at the receiver. Because frame telemetry systems are designed for channels with high error and drop-out rates, they use a more complicated synchronization algorithm to ensure that the receiver is properly synchronized to the transmitter. Packet systems use channels with higher inherent reliability so the probability of losing a data frame, or having a data frame arrive with errors, is smaller. In this case, the networking protocol interface does the necessary synchronization and delivers the frames to the user. If the frames are corrupted or lost, then the transmission protocol will negotiate a retransmission.

The data synchronization process begins after the receiver has successfully locked onto the data signal. This initial lock-up involves locking onto the transmission carrier and then finding the individual bits in the data stream. At this point, the data is a stream of bits without any specific context. The synchronization process places the bits into a context so that the individual measurements can be located in the data stream. The synchronization process for a telemetry frame follows these steps, as illustrated in Figure 12.10:

1. In the Search State, find the occurrences of the frame synchronization word by using a correlator comparing the data with the desired stored pattern.
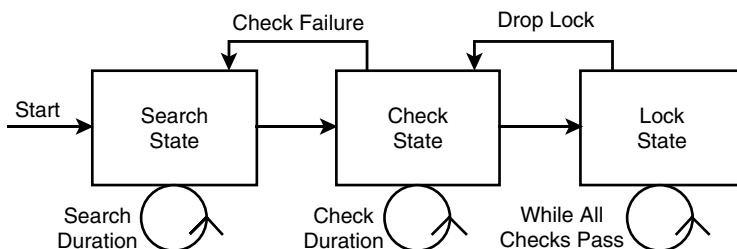
FIGURE 12.10   State diagram for the data synchronization process.

2. Once the synchronization marker is reliably found, ensure that it repeats at the minor frame rate in the Check State and that any management telemetry information in the frame is intact by using a frame synchronizer.

3. Once the frame structure is fully identified, begin processing the data in the Lock State.

At the Search State and the Check State, the hardware may stay in these states for more than one occurrence of the pattern to ensure that the correct pattern has really arrived and that it is not accidentally found in the random data.

12.6   Data Processing

The telemetry and telecommand system will require a data processing system to provide the necessary conversion function between the needs of the user and the raw data format found at the hardware level. The major real-time processing functions are those associated with maintaining the telemetry database and the user interface. Both of these are time critical functions that must keep pace with the incoming data. The processing must also validate and transmit the commands given by the user as they are entered.

The database may be configured to hold the raw data values (engineering units) or fully converted and calibrated values that are ready for display or analysis by the eventual user, or both. Additional values can be created by performing combinations of basic measurements to form a new value or a **derived parameter**. The derived parameter can then be uniquely stored in the telemetry database. This database information can be displayed in real time for the user as well as stored to a data file for more extensive later analysis. The received data is normally stored in a telemetry database as the values are received. The database is designed to allow for parameter searches of individual values to support data analysis and displaying the values for the user. The database will contain more than the received values. It will also have caution and alarm limits to warn the operator when the measurements exceed desired values as well as identifiers for the mnemonic for the value, markers for where to locate the data in the incoming data stream, and time tags for the last received value. An example of a telemetry database structure is illustrated in Figure 12.11. The first part of the figure shows how the data may be organized. The second part of the figure shows how caution and alarm limits may be configured.

No.	Type	Location	Mnemonic	Caution Low	Caution High	Alarm Low	Alarm High	Enable Warnings	Last Value	Time
1	An	MF4	XY Sp 1	5.0	8.0	3.0	10.0	Y	6.25	1058712
2	Di	MF7,20	Rx Cnt 1	16	32	8	64	Y	7	1058715
3	Bi	MF9(3)	RNG TONE				1	Y	0	1058717
4	An	SF2;4,8	Com 1 Sig	100.0	250.0	50.0	500.0	Y	550.5	1058716

(a) Telemetry database structure
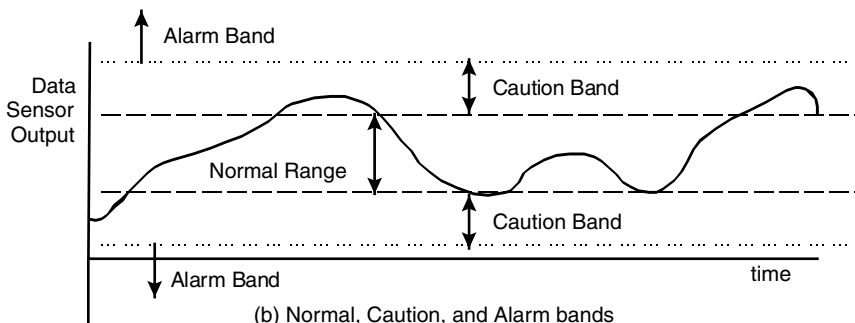
(b) Normal, Caution, and Alarm bands

FIGURE 12.11   Telemetry database structure and associated user alarm levels.
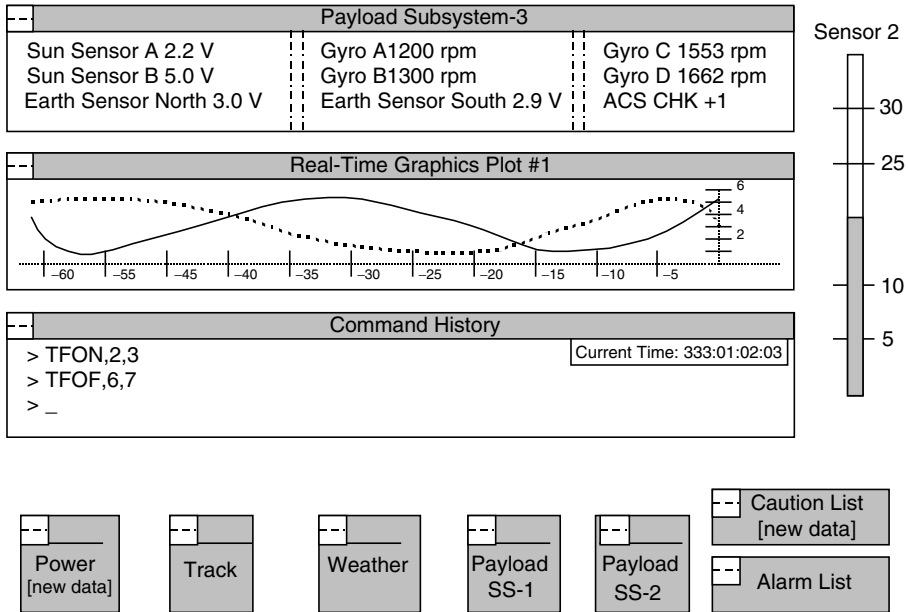
FIGURE 12.12 An example user data display for graphs, thermometer indicators, and text display windows.

The information needs to be accessible to the user who will be acting on it. The user interface may have regions for real-time plots (strip chart-type plots), numerical displays of data, "dip stick" or "thermometer" indicators for individual values, and regions for user input into the command process. This type of display is illustrated in Figure 12.12. This type of display assumes that the graphical capabilities of the display system are being used to organize the data into individual windows for easier recall by the user. The individual parameters in the display windows need to be linked to the telemetry database. This can be done via explicit coding or through structured database calls that link the display areas to the database structure.

The user will require a data input area for telecommand operations. This user input will be linked to a telecommand processor to accept the input, validate that it forms an acceptable command, and then translate the command into the basic structure understood by the actuators. The validation step will include not only syntax validation but also user authentication. Telecommand data sent over an open channel (radio or Internet) may be encrypted to prevent unauthorized listening or spoofing of the commands by an unauthorized user.

References

1. S. Horan, *Introduction to PCM Telemetering Systems*, 2nd ed., Boca Raton: CRC Press, 2002.
2. "American National Standard T1.523-2001Telecom Glossary 2000," http://www.atis.org/tg2k, February 2001.
3. Telemetry Standards, IRIG Standard 106–01, Part 1, Secretariat, Range Commanders Council, U.S. Army White Sands Missile Range, NM, February 2001.
4. Telemetry Standards, IRIG Standard 106–01, Part 2, Secretariat, Range Commanders Council, U.S. Army White Sands Missile Range, NM, February 2001.
5. *Manual of Regulations and Procedures for Federal Radio Frequency Management*, Washington, DC: National Telecommunications and Information Administration, September 2003.

This page intentionally left blank

13

Computer-Aided Design and Analysis of Communication Systems

William H. Tranter
Virginia Polytechnic Institute

Kurt L. Kosbar
University of Missouri–Rolla

13.1   Introduction

It should be clear from the preceding chapters that communication systems exist to perform a wide variety of tasks. The demands placed on today's communication systems necessitate higher data rates, greater flexibility, and increased reliability. Communication systems are therefore becoming increasingly complex, and the resulting systems cannot usually be analyzed using traditional (pencil and paper) analysis techniques. In addition, communication systems often operate in complicated environments that are not analytically tractable. Examples include channels that exhibit severe bandlimiting, multipath, fading, interference, non-Gaussian noise, and perhaps even burst noise. The combination of a complex system and a complex environment makes the design and analysis of these communication systems a formidable task. Some level of computer assistance must usually be invoked in both the design and analysis process. The appropriate level of computer assistance can range from simply using numerical techniques to solve a differential equation

defining an element or subsystem to developing a **computer simulation** of the end-to-end communication system.

There is another important reason for the current popularity of computer-aided analysis and simulation techniques. It is now practical to make extensive use of these techniques. The computing power of many personal computers and workstations available today exceeds the capabilities of many large mainframe computers of only a decade ago. The low cost of these computing resources make them widely available. As a result, significant computing resources are available to the communications engineer within the office or even the home environment. Personal computers and workstations tend to be resources dedicated to a specific individual or project. Since the communications engineer working at his or her desk has control over the computing resource, lengthy simulations can be performed without interfering with the work of others. Over the past few years a number of software packages have been developed that allow complex communication systems to be simulated with relative ease [Shanmugan, 1988]. The best of these packages contains a wide variety of subsystem models as well as integrated graphics packages that allow waveforms, spectra, histograms, and performance characteristics to be displayed without leaving the simulation environment. For those motivated to generate their own simulation code, the widespread availability of high-quality C, Pascal, and FORTRAN compilers makes it possible for large application-specific simulation programs to be developed for personal computers and workstations. When computing tools are both available and convenient to use, they will be employed in the day-to-day efforts of system analysts and designers.

The purpose of this chapter is to provide a brief introduction to the subject of **computer-aided design and analysis** of communication systems. Since computer-aided design and analysis almost always involves some level of simulation, we focus our discussion on the important subject of the simulation of communication systems.

Computer simulations can, of course, never replace a skilled engineer, although they can be a tremendous help in both the design and analysis process. The most powerful simulation program cannot solve all the problems that arise, and the process of making trade-off decisions will always be based on experience. In addition, evaluating and interpreting the results of a complex simulation require considerable skill and insight. While these remarks seem obvious, as computer-aided techniques become more powerful, one is tempted to replace experience and insight with computing power.

13.2   The Role of Simulation

The main purposes of simulation are to help us understand the operation of a complex communication system, to determine acceptable or optimum parameters for implementation of a system, and to determine the performance of a communication system. There are basically two types of systems in which communication engineers have interest: **communication links** and communication networks.

A communication link is usually a single source, a single user, and the components and channel between source and user. A typical link architecture is shown in Figure 13.1. The important performance parameter in a digital communication link is typically the reliability of the communication link as measured by the symbol or bit error rate (BER). In an analog communication link the performance parameter of interest is typically the signal-to-noise ratio (SNR) at the receiver input or the mean-square error of the receiver output. The simulation is usually performed to determine the effect of system parameters, such as filter bandwidths or code rate, or to determine the effect of environmental parameters, such as noise levels, noise statistics, or power spectral densities.
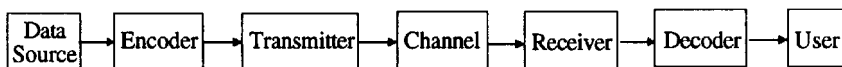
FIGURE 13.1   Basic communication link.

A communication network is a collection of communication links with many signal sources and many users. Computer simulation programs for networks often deal with problems of routing, flow and congestion control, and the network delay. While this chapter deals with the communication link, the reader is reminded that network simulation is also an important area of study. The simulation methodologies used for communication networks are different from those used on links because, in a communication link simulation, each waveform present in the system is sampled using a constant sampling frequency. In contrast, network simulations are event-driven, with the important events being such quantities as the time of arrival of a message.

Simulations can be developed to investigate either transient phenomena or steady-state properties of a system. The study of the acquisition time of a phase-lock loop receiver is an example of a transient phenomenon. Simulations that are performed to study transient behavior often focus on a single subsystem such as a receiver synchronization system. Simulations that are developed to study steady-state behavior often model the entire system. An example is a simulation to determine the BER of a system.

13.3   Motivation for the Use of Simulation

As mentioned previously, simulation is a reasonable approach to many design and analysis problems because complex problems demand that computer-based techniques be used to support traditional analytical approaches. There are many other motivations for making use of simulation.

A carefully developed simulation is much like having a breadboard implementation of the communication system available for study. Experiments can be performed using the simulation much like experiments can be performed using hardware. System parameters can be easily changed, and the impact of these changes can be evaluated. By continuing this process, parameteric studies can easily be conducted and acceptable, or perhaps even optimum, parameter values can be determined. By changing parameters, or even the system topology, one can play "what if" games much more quickly and economically using a simulation than with a system realized in hardware.

It is often overlooked that simulation can be used to support analysis. Many people incorrectly view simulation as a tool to be used only when a system becomes too complex to be analyzed using traditional analysis techniques. Used properly, simulation goes hand in hand with traditional techniques in that simulation can often be used to guide analysis. A properly developed simulation provides insight into system operation. As an example, if a system has many parameters, these can be varied in a way that allows the most important parameters, in terms of system performance, to be identified. The least important parameters can then often be discarded, with the result being a simpler system that is more tractable analytically. Analysis also aids simulation. The development of an accurate and efficient simulation is often dependent upon a careful analysis of various portions of the system.

13.4   Limitations of Simulation

Simulation, useful as it is, does have limitations. It must be remembered that a system simulation is an approximation to the actual system under study. The nature of the approximations must be understood if one is to have confidence in the simulation results. The accuracy of the simulation is limited by the accuracy to which the various components and subsystems within the system are modeled. It is often necessary to collect extensive experimental data on system components to ensure that simulation models accurately reflect the behavior of the components. Even if this step is done with care, one can only trust the simulation model over the range of values consistent with the previously collected experimental data. A main source of error in a simulation results because models are used at operating points beyond which the models are valid.

In addition to modeling difficulties, it should be realized that the digital simulation of a system can seldom be made perfectly consistent with the actual system under study. The simulation is affected by phenomena not present in the actual system. Examples are the aliasing errors resulting from the sampling operation and the finite word length (quantization) effects present in the simulation. Practical communication systems use

a number of filters, and modeling the analog filters present in the actual system by the digital filters required by the simulation involves a number of approximations. The assumptions and approximations used in modeling an analog filter using impulse-invariant digital filter synthesis techniques are quite different from the assumptions and approximations used in bilinear *z*-transform techniques. Determining the appropriate modeling technique requires careful thought.

Another limitation of simulation lies in the excessive computer run time that is often necessary for estimating performance parameters. An example is the estimation of the system BER for systems having very low nominal bit error rates. We will expand on this topic later in this chapter.

13.5   Simulation Structure

As illustrated in Figure 13.1, a communication system is a collection of subsystems such that the overall system provides a reliable path for information flow from source to user. In a computer simulation of the system, the individual subsystems must first be accurately modeled by signal processing operations. The overall simulation program is a collection of these signal processing operations and must accurately model the overall communication system. The important subject of subsystem modeling will be treated in a following section.

The first step in the development of a simulation program is to define the topology of the system, which specifies the manner in which the individual subsystems are connected. The subsystem models must then be defined by specifying the signal processing operation to be performed by each of the various subsystems. A simulation structure may be either fixed topology or free topology. In a fixed topology simulation, the basic structure shown in Figure 13.1 is modeled. Various subsystems can be bypassed if desired by setting switches, but the basic topology cannot be modified. In a free topology structure, subsystems can be interconnected in any way desired and new additional subsystems can be added at will.

A simulation program for a communication system is a collection of at least three operations, shown in Figure 13.2, although in a well-integrated simulation these operations tend to merge together. The first operation, sometimes referred to as the *preprocessor,* defines the parameters of each subsystem and the intrinsic parameters that control the operation of the simulation. The second operation is the *simulation exercisor,* which is the simulation program actually executed on the computer. The third operation performed in a simulation program is that of *postprocessing.* This is a collection of routines that format the simulation output in a way which provides insight into system operations and allows the performance of the communication system under study to be evaluated. A postprocessor usually consists of a number of graphics-based routines, allowing the user to view waveforms and other displays generated by the simulation. The postprocessor also consists of a number of routines that allow estimation of the bit error rate, signal-to-noise ratios, histograms, and power spectral densities.

When faced with the problem of developing a simulation of a communication system, the first fundamental choice is whether to develop a custom simulation using a general-purpose high-level language or to use one of the many special-purpose communication system simulation languages available. If the decision is made to develop a dedicated simulation using a general-purpose language, a number of resources are needed beyond a quality compiler and a mathematics library. Also needed are libraries for filtering routines, software models for each of the subsystems contained in the overall system, channel models, and the waveform display and data analysis routines needed for the analysis of the simulation results (postprocessing). While at least some of the required software will have to be developed at the time the simulation is being written, many of the required
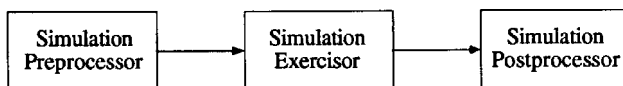
FIGURE 13.2   Typical structure of a simulation program.

routines can probably be obtained from digital signal processing (DSP) programs and other available sources. As more simulation projects are completed, the database of available routines becomes larger.

The other alternative is to use a **dedicated simulation language**, which makes it possible for one who does not have the necessary skills to create a custom simulation using a high-level language to develop a communication system simulation. Many simulation languages are available for both personal computers and work-stations [Shanmugan, 1988]. While the use of these resources can speed simulation development, the user must ensure that the assumptions used in developing the models are well understood and applicable to the problem of interest. In choosing a dedicated language from among those that are available, one should select a language that has an extensive model library, an integrated postprocessor with a wide variety of data analysis routines, on-line help and documentation capabilities, and extensive error-checking routines.

13.6   The Interdisciplinary Nature of Simulation

The subject of computer-aided design and analysis of communication systems is very much interdisciplinary in nature. The major disciplines that bear on the subject are communication theory, DSP, numerical analysis, and stochastic process theory. The roles played by these subjects is clear. The simulation user must have knowledge of the behavior of communication theory if the simulation results are to be understood. The analysis techniques of communication theory allow simulation results to be verified. Since each subsystem in the overall communication system is a signal processing operation, the tools of DSP provide the algorithms to realize filters and other subsystems. Numerical analysis techniques are used extensively in the development of signal processing algorithms. Since communication systems involve random data signals, as well as noise and other disturbances, the concepts of stochastic process theory are important in developing models of these quantities and also for determining performance estimates.

13.7   Model Design

Practicing engineers frequently use models to investigate the behavior of complex systems. Traditionally, models have been physical devices or a set of mathematical expressions. The widespread use of powerful digital computers now allows one to generate computer programs that model physical systems. Although the detailed development and use of computer models differs significantly from their physical and mathematical counterparts, the computer models share many of the same design constraints and trade-offs. For any model to be useful one must guarantee that the response of the model to stimuli will closely match the response of the target system, the model must be designed and fabricated in much less time and at significantly less
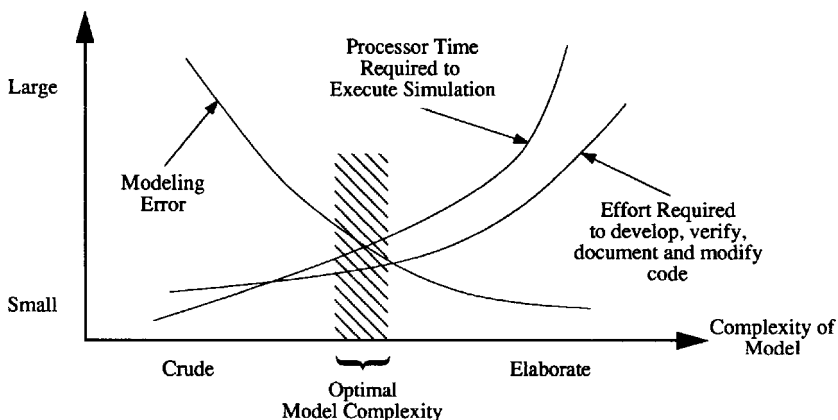
FIGURE 13.3   Design constraints and trade-offs.

expense than the target system, and the model must be reasonably easy to validate and modify. In addition to these constraints, designers of computer models must assure that the amount of processor time required to execute the model is not excessive. The optimal model is the one that appropriately balances these conflicting requirements. Figure 13.3 describes the typical design trade-off faced when developing computer models. A somewhat surprising observation is that the optimal model is often not the one that most closely approximates the target system. A highly detailed model will typically require a tremendous amount of time to develop, will be difficult to validate and modify, and may require prohibitive processor time to execute. Selecting a model that achieves a good balance between these constraints is as much an art as a science. Being ware of the trade-offs which exist, and must be addressed, is the first step toward mastering the art of modeling.

13.8   Low-Pass Models

In most cases of practical interest the physical layer of the communication system will use continuous time (CT) signals, while the simulation will operate in discrete time (DT). For the simulation to be useful, one must develop DT signals and systems that closely match their CT counterparts. This topic is discussed at length in introductory DSP texts. A prominent result in this field is the Nyquist sampling theorem, which states that if a CT signal has no energy above frequency f_h Hz, one can create a DT signal that contains *exactly* the same information by sampling the CT signal at any rate in excess of $2f_h$ samples per second. Since the execution time of the simulation is proportional to the number of samples it must process, one naturally uses the lowest sampling rate possible. While the Nyquist theorem should not be violated for arbitrary signals, when the CT signal is bandpass one can use low-pass equivalent (LPE) waveforms that contain all the information of the CT signal but can be sampled slower than $2f_h$.

Assume the energy in a bandpass signal is centered about a carrier frequency of f_c Hz and ranges from f_l to f_h Hz, resulting in a bandwidth of $f_h - f_l = W$ Hz, as in Figure 13.4. It is not unusual for W to be many orders of magnitude less than f_c. The bandpass waveform $x(t)$ can be expressed as a function of two low-pass signals. Two essentially equivalent LPE expansions are known as the envelope/phase representation [Davenport and Root, 1958],

$$x(t) = A(t)\cos[2\pi f_c t + \theta(t)] \tag{13.1}$$

and the quadrature representation,

$$x(t) = x_c(t)\cos(2\pi f_c t) - x_s(t)\sin(2\pi f_c t) \tag{13.2}$$

All four real signals $A(t)$, $\theta(t)$, $x_c(t)$, and $x_s(t)$ are low pass and have zero energy above $W/2$ Hz. A computer simulation that replaces $x(t)$ with a pair of LPE signals will require far less processor time since the LPE waveforms can be sampled at W as opposed to $2f_h$ samples per second. It is cumbersome to work with two signals rather than one signal. A more mathematically elegant LPE expansion is

$$x(t) = \text{Re}\left\{v(t)e^{j2\pi fct}\right\} \tag{13.3}$$
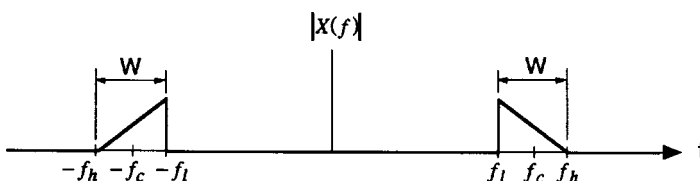
FIGURE 13.4   Amplitude spectrum of a bandpass signal.

where $v(t)$ is a low-pass, *complex-time domain signal* that has no energy above $W/2$ Hz. Signal $v(t)$ is known as the complex envelope of $x(t)$ [Haykin, 1983]. It contains all the information of $x(t)$ and can be sampled at W samples per second without aliasing. This notation is disturbing to engineers accustomed to viewing all time domain signals as real. However, a complete theory exists for complex time domain signals, and with surprisingly little effort one can define convolution, Fourier transforms, analog-to-digital and digital-to-analog conversions, and many other signal processing algorithms for complex signals. If f_c and W are known, the LPE mapping is one-to-one so that $x(t)$ can be completely recovered from $v(t)$. While it is conceptually simpler to sample the CT signals at a rate in excess of $2f_h$ and avoid the mathematical difficulties of the LPE representation, the tremendous difference between f_c and W makes the LPE far more efficient for computer simulation. This type of trade-off frequently occurs in computer simulation. A careful mathematical analysis of the modeling problem conducted *before* any computer code is generated can yield substantial performance improvements over a conceptually simpler, but numerically inefficient approach.

The fundamental reason the LPE representation outlined above is popular in simulation is that one can easily generate **LPE models** of linear time-invariant bandpass filters. The LPE of the output of a bandpass filter is merely the convolution of the LPE of the input signal and the LPE of the impulse response of the filter. It is far more difficult to determine a LPE model for nonlinear and time-varying systems. There are numerous approaches that trade off flexibility and simplicity. If the system is nonlinear and time invariant, a Volterra series can be used. While this series will exactly represent the nonlinear device, it is often analytically intractable and numerically inefficient. For nonlinear devices with a limited amount of memory the AM/AM, AM/PM [Shimbo, 1971] LPE model is useful. This model accurately describes the response of many microwave amplifiers including traveling-wave tubes, solid-state limiting amplifiers, and, under certain conditions, devices which exhibit hysteresis. The Chebyshev transform [Blachman, 1964] is useful for memoryless nonlinearities such as hard and soft limiters. If the nonlinear device is so complex that none of the conventional LPE models can be used, one may need to convert the LPE signal back to its bandpass representation, route the bandpass signal through a model of the nonlinear device, and then reconvert the output to a LPE signal for further processing. If this must be done, one has the choice of increasing the sampling rate for the entire simulation or using different sampling rates for various sections of the simulation. The second of these approaches is known as a *multirate simulation* [Cochiere and Rabiner, 1983]. The interpolation and decimation operations required to convert between sampling rates can consume significant amounts of processor time. One must carefully examine this trade-off to determine if a multirate simulation will substantially reduce the execution time over a single, high sampling rate simulation. Efficient and flexible modeling of nonlinear devices is in general a difficult task and continues to be an area of active research.

13.9   Pseudorandom Signal and Noise Generators

The preceding discussion was motivated by the desire to efficiently model filters and nonlinear amplifiers. Since these devices often consume the majority of the processor time, they are given high priority. However, there are a number of other subsystems that do not resemble filters. One example is the data source that generates the message or waveform which must be transmitted. While signal sources may be analog or digital in nature, we will focus exclusively on binary digital sources. The two basic categories of signals produced by these devices are known as *deterministic* and *random*. When performing worst-case analysis, one will typically produce known, repetitive signal patterns designed to stress a particular subsystem within the overall communication system. For example, a signal with few transitions may stress the symbol synchronization loops, while a signal with many regularly spaced transitions may generate unusually wide bandwidth signals. The generation of this type of signal is straightforward and highly application dependent. To test the nominal system performance one typically uses a random data sequence. While generation of a truly random signal is arguably impossible [Knuth, 1981], one can easily generate pseudorandom (PN) sequences. PN sequence generators have been extensively studied since they are used in Monte Carlo integration and simulation [Rubinstein, 1981] programs and in a variety of wideband and secure communication systems. The two basic
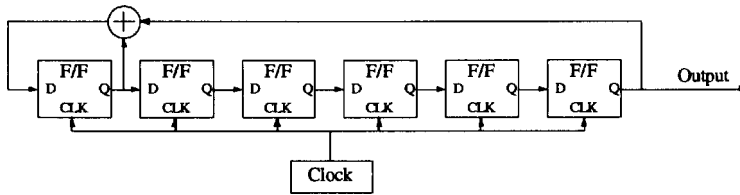
FIGURE 13.5   Six-stage binary shift register PN generator.
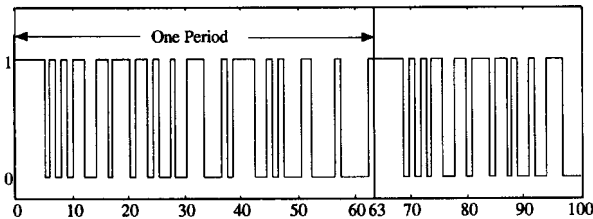
FIGURE 13.6   Output of a six-stage maximal length BSR.

structures for generating PN sequences are binary shift registers (BSRs) and linear congruential algorithms (LCAs).

Digital data sources typically use BSRs, while noise generators often use LCAs. A logic diagram for a simple BSR is shown in Figure 13.5. This BSR consists of a clock, six D-*type* flip–flops (F/F), and an exclusive OR gate denoted by a modulo-two adder. If all the F/F are initialized to 1, the output of the device is the waveform shown in Figure 13.6. Notice that the waveform is periodic with period $63 = 2^6 - 1$, but within one cycle the output has many of the properties of a random sequence. This demonstrates all the properties of the BSR, LCA, and more advanced PN sequence generators. All PN generators have memory and must therefore be initialized by the user before the first sample is generated. The initialization data is typically called the seed. One must choose this seed carefully to ensure the output will have the desired properties (in this example, one must avoid setting all F/F to zero). All PN sequence generators will produce periodic sequences. This may or may not be a problem. If it is a concern, one should ensure that one period of the PN sequence generator is longer than the total execution time of the simulation. This is usually not a significant problem, since one can easily construct BSRs that have periods greater than 10^{27} clock cycles. The final concern is how closely the behavior of the PN sequence generator matches a truly random sequence. Standard statistical analysis algorithms have been applied to many of these generators to validate their performance.

Many digital communication systems use *m* bit (*M-ary*) *sources where* $m > 1$. Figure 13.7 depicts a simple algorithm for generating a *M*-ary random sequence from a binary sequence. The clock must now cycle through *m* cycles for every generated symbol, and the period of the generator has been reduced by a factor of *m*. This may force the use of a longer-period BSR. Another common application of PN sequence generators is to produce samples of a continuous stochastic process, such as Gaussian noise. A structure for producing these samples is shown in Figure 13.8. In this case the BSR has been replaced by an LCA [Knuth, 1981]. The LCA is very similar to BSR in that it requires a seed value, is clocked once for each symbol generated, and will generate a periodic sequence. One can generate a white noise process with an arbitrary first-order probability density function (pdf) by passing the output of the LCA through an appropriately designed nonlinear, memoryless mapping. Simple and well-documented algorithms exist for the uniform to Gaussian mapping. If one wishes to generate a nonwhite process, the output can be passed through the appropriate filter. Generation of a wide-sense stationary Gaussian stochastic process with a specified power spectral density is a well-understood and -documented problem. It is also straightforward to generate a white sequence with an arbitrary first-order pdf or to generate a specified power spectral density if one does not attempt to control the pdf.
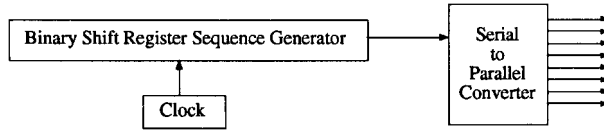
FIGURE 13.7   *M*-ary PN sequence generator.
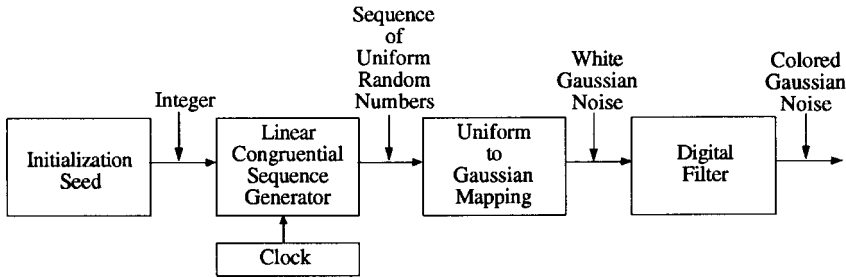
FIGURE 13.8   Generation of Gaussian noise.

However, the problem of generating a noise source with an arbitrary pdf *and* an arbitrary power spectral density is a significant challenge [Sondhi, 1983].

13.10   Transmitter, Channel, and Receiver Modeling

Most elements of transmitters, channels, and receivers are implemented using standard DSP techniques. Effects that are difficult to characterize using mathematical analysis can often be included in the simulation with little additional effort. Common examples include gain and phase imbalance in quadrature circuits, nonlinear amplifiers, oscillator instabilities, and antenna platform motion. One can typically use LPE waveforms and devices to avoid translating the modulator output to the carrier frequency. Signal levels in physical systems often vary by many orders of magnitude, with the output of the transmitters being extremely high energy signals and the input to receivers at very low energies. To reduce execution time and avoid working with extremely large and small signal level simulations, one often omits the effects of linear amplifiers and attenuators and uses normalized signals. Since the performance of most systems is a function of the signal-to-noise ratio, and not of absolute signal level, normalization will have no effect on the measured performance. One must be careful to document the normalizing constants so that the original signal levels can be reconstructed if needed. Even some rather complex functions, such as error detecting and correcting codes, can be handled in this manner. If one knows the uncoded error rate for a system, the coded error rate can often be closely approximated by applying a mathematical mapping. As will be pointed out below, the amount of processor time required to produce a meaningful error rate estimate is often inversely proportional to the error rate. While an uncoded error rate may be easy to measure, the coded error rate is usually so small that it would be impractical to execute a simulation to measure this quantity directly. The performance of a coded communication system is most often determined by first executing a simulation to establish the channel **symbol error rate**. An analytical mapping can then be used to determine the decoded BER from the channel symbol error rate.

Once the signal has passed though the channel, the original message is recovered by a receiver. This can typically be realized by a sequence of digital filters, feedback loops, and appropriately selected nonlinear devices. A receiver encounters a number of clearly identifiable problems that one may wish to address independently. For example, receivers must initially synchronize themselves to the incoming signal. This may involve detecting that an input signal is present, acquiring an estimate of the carrier amplitude, frequency,

phase, symbol synchronization, frame synchronization, and, in the case of spread spectrum systems, code synchronization. Once acquisition is complete, the receiver enters a steady-state mode of operation, where concerns such as symbol error rate, mean time to loss of lock, and reaction to fading and interference are of primary importance. To characterize the system, the user may wish to decouple the analysis of these parameters to investigate relationships that may exist.

For example, one may run a number of acquisition scenarios and gather statistics concerning the probability of acquisition within a specified time interval or the mean time to acquisition. To isolate the problems faced in synchronization from the inherent limitation of the channel, one may wish to use perfect synchronization information to determine the minimum possible BER. Then the symbol or carrier synchronization can be held at fixed errors to determine sensitivity to these parameters and to investigate worst-case performance. Noise processes can be used to vary these parameters to investigate more typical performance. The designer may also wish to investigate the performance of the synchronization system to various data patterns or the robustness of the synchronization system in the face of interference. The ability to measure the system response to one parameter while a wide range of other parameters are held fixed and the ability to quickly generate a wide variety of environments are some of the more significant advantages that simulation enjoys over more conventional hardware and analytical models.

13.11   Symbol Error Rate Estimation

One of the most fundamental parameters to measure in a digital communication system is the steady-state BER. The simplest method for estimating the BER is to perform a **Monte Carlo (MC) simulation**. The simulation conducts the same test one would perform on the physical system. All data sources and noise sources produce typical waveforms. The output of the demodulator is compared to the output of the message source, and the BER is estimated by dividing the number of observed errors by the number of bits transmitted. This is a simple technique that will work with any system that has ergodic [Papoulis, 1965] noise processes. The downside of this approach is that one must often pass a very large number of samples through the system to produce a reliable estimate of the BER. The question of how many samples must be collected can be answered using confidence intervals. The confidence interval gives a measure of how close the true BER will be to the estimate produced by the MC simulation. A typical confidence interval curve is shown in Figure 13.9. The ratio of the size of the confidence interval to the size of the estimate is a function of the number of errors observed. Convenient rules of thumb for this work are that after one error is observed the point estimate is
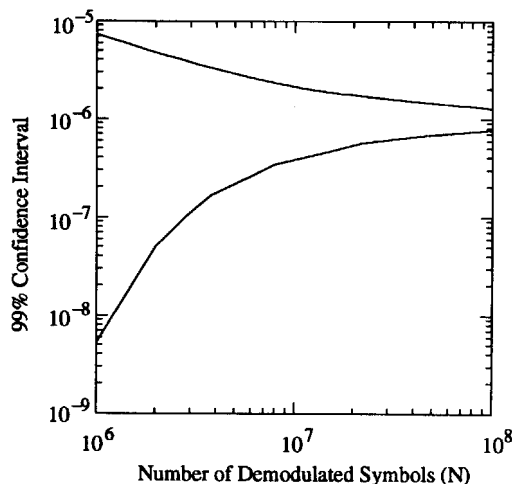
FIGURE 13.9   Typical confidence interval (BER) point estimate $= 10^{-6}$.
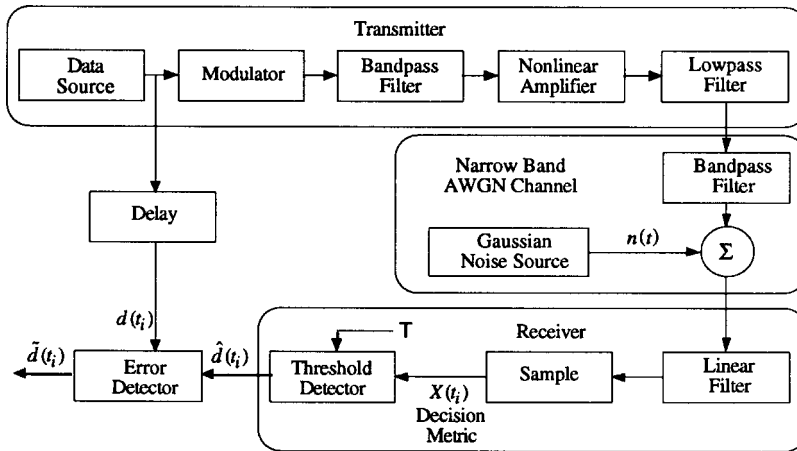
FIGURE 13.10   Typical digital communication system.

accurate to within 3 orders of magnitude, after 10 errors the estimate is accurate to within a factor of 2, and after 100 errors the point estimate will be accurate to a factor of 1.3. This requirement for tens or hundreds of errors to occur frequently limits the usefulness of MC simulations for systems that have low error rates and has motivated research into more efficient methods of estimating BER.

Perhaps the fastest method of BER estimation is the semi-analytic (SA) or quasi-analytic technique [Jeruchim, 1984]. This technique is useful for systems that resemble Figure 13.10. In this case the mean of the decision metric is a function of the transmitted data pattern and is independent of the noise. All other parameters of the pdf of the decision metric are a function of the noise and are independent of the data. This means that one can analytically determine the conditional pdf of the decision metric given the transmitted data pattern. By using total probability one can then determine the unconditional error rate. The problem with conventional mathematical analysis is that when the channel has a significant amount of memory or the nonlinearity is rather complex, one must compute a large number of conditional density functions. Simulation can easily solve this problem for most practical systems. A noise-free simulation is executed, and the value of the decision metric is recorded in a data file. Once the simulation is complete, this information can be used to reconstruct the conditional and ultimately the unconditional error rate. This method generates highly accurate estimates of the BER and makes very efficient use of computer resources, but can only be used in the special cases where one can analytically determine the conditional pdf.

The MC and SA techniques fall at the two extremes of BER estimation. MC simulations require no *a priori* information concerning the system performance or architecture but may require tremendous amounts of computer time to execute. SA techniques require an almost trivial amount of computer time for many cases but require the analyst to have a considerable amount of information concerning the system. There is a continuing search for algorithms that fall in between these extremes. These variance reduction algorithms all share the property of making a limited number of assumptions concerning system performance and architecture, then using this information to reduce the variance of the MC estimate. Popular techniques are summarized in [Jeruchim, 1984] and include importance sampling, large deviation theory, extremal statistics, and tail extrapolation. To successfully use one of these techniques one must first understand the basic concept behind the technique. Then one should carefully determine what assumptions were made concerning the system architecture to determine if the system under study satisfies the requirements. This can be a difficult task since it is not always clear what assumptions are required for a specified technique to be applicable. Finally, one should always determine the accuracy of the measurement through some technique similar to confidence interval estimation.

13.12   Validation of Simulation Results

One often constructs a simulation to determine the value of a single parameter, such as the system BER. However the estimate of this parameter has little or no value unless one can ensure that the simulation model closely resembles the physical system. A number of methods can be used to validate a simulation. Individually, none of them will guarantee that the simulation results are accurate, but taken as a group, they form a convincing argument that the results are realistic. Seven methods of validation are mathematical analysis, comparison with hardware, bounding techniques, degenerate case studies, reasonable relationship tests, subsystem tests, and redundant simulation efforts.

If one has access to mathematical analysis or hardware that predicts or approximates the performance of the system, one should obviously compare the simulation and mathematical results. Unfortunately, in most cases these results will not be available. Even though exact mathematical analysis of the system is not possible, it may be possible to develop bounds on the system performance. If these bounds are tight, they may accurately characterize the system performance, but even loose bounds will be useful since they help verify the simulation results. Most systems have parameters that can be varied. While it may be mathematically difficult to determine the performance of the system for arbitrary values, it is often possible to mathematically determine the results when parameters assume extreme or degenerate values.

Other methods of validation are decidedly less mathematical. One may wish to vary parameters and ascertain whether the performance parameter changes in a reasonable manner. For example, small changes in SNR rarely cause dramatic changes in system performance. When constructing a simulation, each subsystem, such as filters, nonlinear amplifiers, and noise and data sources, should be thoroughly tested before being included in a larger simulation. Be aware, however, that correct operation of all the various subsystems that make up a communication system does not imply that the overall system performs correctly. If one is writing his or her own code, one must verify that there are no software bugs or fundamental design errors. Even if one purchases a commercial software package, there is no guarantee that the designer of the software models made the same assumptions the user will make when using the model. In most cases it will be far easier to test a module before it is inserted into a simulation than it will be to isolate a problem in a complex piece of code. The final check one may wish to perform is a redundant simulation. There are many methods of simulating a system. One may wish to have two teams investigate a problem or have a single team implement a simulation using two different techniques to verify that the results are reasonable.

13.13   A Simple Example Illustrating Simulation Products

To illustrate the output that is typically generated by a communication system simulation, a simple example is considered. The system is that considered in Figure 13.10. An OQPSK (offset quadrature phase-shift keyed) modulation format is assumed so that one of four waveforms is transmitted during each symbol period. The data source may be viewed as a single binary source, in which the source symbols are taken two at a time when mapped onto a transmitted waveform, or as two parallel data sources, with one source providing the direct channel modulation and the second source providing the quadrature channel modulation. The signal constellation at the modulator output appears as shown in Figure 13.11(a), with the corresponding eye diagram appearing as shown in Figure 13.11(b). The eye diagram is formed by overlaying successive time intervals of a time domain waveform onto a single graph, much as would be done with a common oscilloscope. Since the simulation sampling frequency used in generating Figure 13.11(b) was 10 samples per data symbol, it is easily seen that the eye diagram was generated by retracing every 2 data symbols or 20 simulation samples. Since Figure 13.11(a) and (b) correspond to the modulator output, which has not yet been filtered, the transitions between binary states occur in one simulation step. After filtering, the eye diagram appears as shown in Figure 13.11(c). A seventh-order Butterworth bilinear *z*-transform digital filter was assumed with a 3-dB bandwidth equal to the bit rate. It should be noted that the bit transitions shown in Figure 13.11(c) do not occur at the same times as the bit transitions shown in Figure 13.11(b). The difference
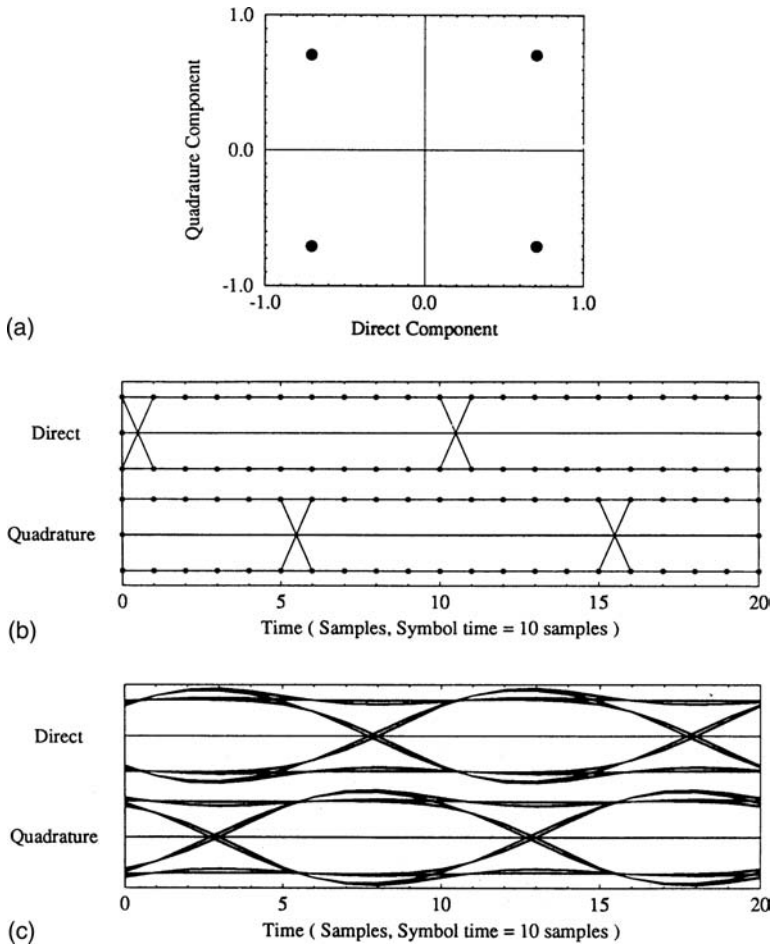
FIGURE 13.11   Transmitter signal constellation and eye diagrams: (a) OQPSK signal constellation; (b) eye diagram of modulator output; (c) eye diagram of filtered modulator output.

is due to the group delay of the filter. Note in Figure 13.10 that the transmitter also involves a nonlinear amplifier. We will see the effects of this component later in this section.

Another interesting point in the system is within the receiver. Since the communication system is being modeled as a baseband system due to the use of the complex-envelope representation of the bandpass waveforms generated in the simulation, the data detector is represented as an integrate-and-dump detector. The detector is then modeled as a sliding-average integrator, in which the width of the integration window is one bit time. The integration is therefore over a single bit period when the sliding window is synchronized with a bit period. The direct-channel and quadrature-channel waveforms at the output of the sliding-average integrator are shown in Figure 13.12(a). The corresponding eye diagrams are shown in Figure 13.12(b). In order to minimize the error probability of the system, the bit decision must be based on the integrator output at the time for which the eye opening is greatest. Thus the eye diagram provides important information concerning the sensitivity of the system to timing errors.

The signal constellation at the sliding integrator output is shown in Figure 13.12(c) and should be carefully compared to the signal constellation shown in Figure 13.11(a) for the modulator output. Three effects are apparent. First, the signal points exhibit some scatter, which, in this case, is due to intersymbol interference resulting from the transmitter filter and additive noise. It is also clear that the signal is both compressed and rotated. These effects are due to the nonlinear amplifier that was mentioned previously. For this example
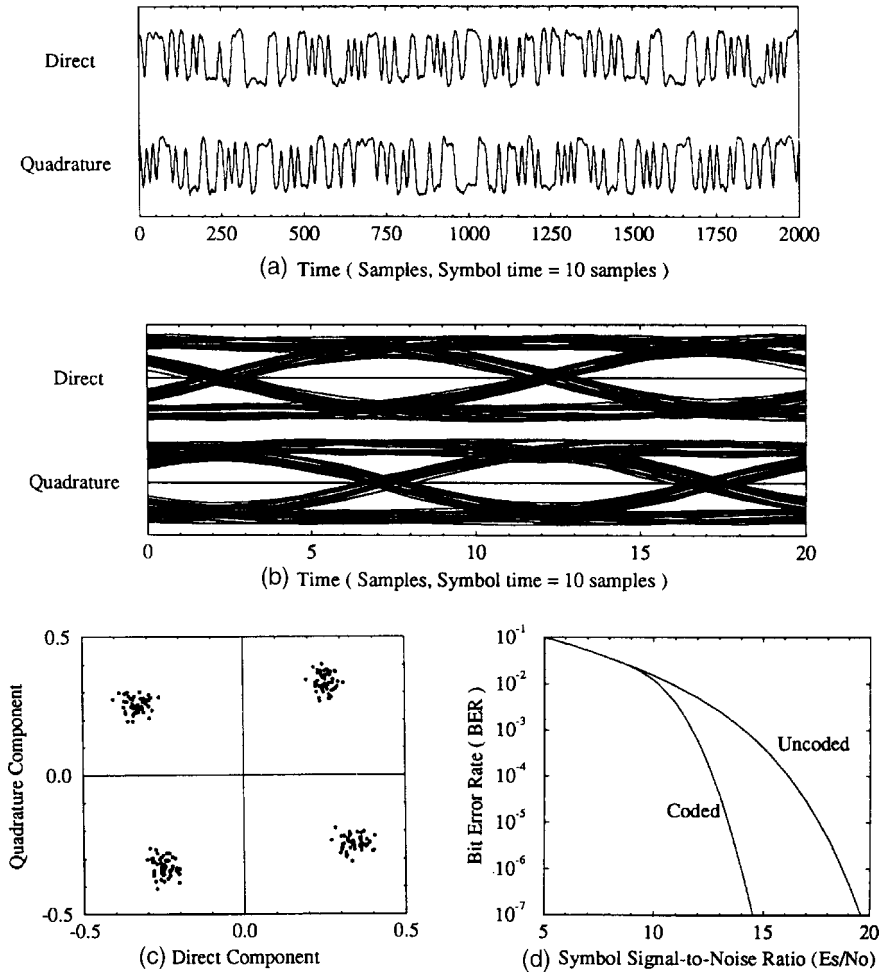
FIGURE 13.12   Integrator output signals and system error probability: (a) sliding integrator output signals; (b) sliding integrator output eye diagram; (c) sliding integrator output signal constellation; (d) error probability.

simulation the nonlinear amplifier is operating near the saturation point, and the compression of the signal constellation is due to the AM/AM characteristic of the nonlinearity and the rotation is due to the AM/PM characteristic of the nonlinearity.

The performance of the overall communication system is illustrated in Figure 13.12(d). The error probability curve is perhaps the most important simulation product. Note that both uncoded and coded results are shown. The coded results were calculated analytically from the uncoded results assuming a (63, 55) Reed–Solomon code. It should be mentioned that **semi-analytic simulation** was used in this example since, as can be seen in Figure 13.10, the noise is injected into the system on the receiver side of the nonlinearity so that linear analysis may be used to determine the effects of the noise on the system performance.

This simple example serves to illustrate only a few of the possible simulation products. There are many other possibilities including histograms, correlation functions, estimates of statistical moments, estimates of the power spectral density, and estimates of the signal-to-noise ratio at various points in the system.

A word is in order regarding spectral estimation techniques. Two basic techniques can be used for spectral estimation: Fourier techniques and model-based techniques. In most simulation problems one is blessed with a tremendous amount of data concerning sampled waveforms but does not have a simple model describing

how these waveforms are produced. For this reason model-based spectral estimation is typically not used. The most common form of spectral estimation used in simulation is the Welch periodogram. While this approach is straightforward, the effects of windowing the data sequence must be carefully considered, and tens or even hundreds of data windows must be averaged to achieve an accurate estimate of the power spectral density.

13.14   Conclusions

We have seen that the analysis and design of today's complex communication systems often requires the use of computer-aided techniques. These techniques allow the solution of problems that are otherwise not tractable and provide considerable insight into the operating characteristics of the communication system.

Defining Terms

Communication link: A point-to-point communication system that typically involves a single information source and a single user. This is in contrast to a communications network, which usually involves many sources and many users.

Computer-aided design and analysis: The process of using computer assistance in the design and analysis of complex systems. In our context, the design and analysis of communication systems, computer-aided design and analysis often involves the extensive use of simulation techniques. Computer-aided techniques often allow one to address design and analysis problems that are not tractable analytically.

Computer simulation: A set of computer programs which allows one to imitate the important aspects of the behavior of the specific system under study. Simulation can aid the design process by, for example, allowing one to determine appropriate system design parameters or aid the analysis process by, for example, allowing one to estimate the end-to-end performance of the system under study.

Dedicated simulation language: A computer language, either text based or graphics based, specifically developed to facilitate the simulation of the various systems under study, such as communication systems.

Low-pass equivalent (LPE) model: A method of representing bandpass signals and systems by low-pass signals and systems. This technique is extremely useful when developing discrete time models of bandpass continuous-time systems. It can substantially reduce the sampling rate required to prevent aliasing and does not result in any loss of information. This in turn reduces the execution time required for the simulation. This modeling technique is closely related to the quadrature representation of bandpass signals.

Monte Carlo simulation: A technique for simulating systems that contain signal sources producing stochastic or random signals. The signal sources are modeled by pseudorandom generators. Performance measures, such as the symbol error rate, are then estimated by time averages. This is a general-purpose technique that can be applied to an extremely wide range of systems. It can, however, require large amounts of computer time to generate accurate estimates.

Pseudorandom generator: An algorithm or device that generates deterministic waveforms which in many ways resemble stochastic or random waveforms. The power spectral density, autocorrelation, and other time averages of pseudorandom signals can closely match the time and ensemble averages of stochastic processes. These generators are useful in computer simulation where one may be unable to generate a truly random process, and they have the added benefit of providing reproducible signals.

Semi-analytic simulation: A numerical analysis technique that can be used to efficiently determine the symbol error rate of digital communication systems. It can be applied whenever one can analytically determine the probability of demodulation error given a particular transmitted data pattern. Although this technique can only be applied to a restricted class of systems, in these cases it is far more efficient, in terms of computer execution time, than Monte Carlo simulations.

Simulation validation: The process of certifying that simulation results are reasonable and can be used with confidence in the design or analysis process.

Symbol error rate: A fundamental performance measure for digital communication systems. The symbol error rate is estimated as the number of errors divided by the total number of demodulated symbols. When the communication system is ergodic, this is equivalent to the probability of making a demodulation error on any symbol.

References

P. Balaban, K.S. Shanmugan, and B.W. Stuck (Eds.), "Special issue on computer-aided modeling, analysis and design of communication systems," *IEEE J. Selected Areas Commun.,* no. 1, 1984.

P. Balaban, E. Biglieri, M.C. Jeruchim, H.T. Mouftah, C.H. Sauer, and K.S. Shanmugan (Eds.), "Computer-aided modeling, analysis and design of communication systems II," *IEEE J. Selected Areas Commun.,* no. 1, 1988.

N. Blachman, "Bandpass nonlinearities," *IEEE Trans. Inf. Theory*, no. 2, 1964.

P. Bratley, B.L. Fox, and L.E. Schrage, *A Guide to Simulation,* New York: Springer-Verlag, 1987.

R. Cochiere and L. Rabiner, *Multirate Digital Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.

W. Davenport and W. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.

R.L. Freeman, *Telecommunications Systems Engineering,* New York: Wiley, 1996.

J. Gagliardi, *Optical Communication,* New York: Wiley, 1995.

J. Gibson, *The Mobile Communications Handbook,* Boca Raton, Fla.: CRC Press, 1996.

S. Haykin, *Communication Systems*, New York: Wiley, 1983.

S. Haykin, *Communication Systems,* New York: Wiley, 1994.

M. Jeruchim, P. Balaban, and K. Shanmugan, *Simulation of Communication Systems*, New York: Plenum, 1992.

M. Jeruchim, "Techniques for estimating the bit error rate in the simulation of digital communication systems," *IEEE J. Selected Areas Commun.,* no. 1, January 1984.

D. Knuth, *The Art of Computer Programming,* vol. 2, *Seminumerical Algorithms,* 2nd ed., Reading, Mass.: Addison-Wesley, 1981.

H.T. Mouftah, J.F. Kurose, and M.A. Marsan (Eds.), "Computer-aided modeling, analysis and design of communication networks I," *IEEE J. Selected Areas Commun.,* no. 9, 1990.

H.T. Mouftah, J.F. Kurose, and M.A. Marsan (Eds.), "Computer-aided modeling, analysis and design of communication networks II," *IEEE J. Selected Areas Commun.,* no. 1, 1991.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965.

R. Rubinstein, *Simulation and the Monte Carlo Method,* New York: Wiley, 1981.

K. Shanmugan, "An update on software packages for simulation of communication systems (links)," *IEEE J. Selected Areas Commun.,* no. 1, 1988.

N.D. Sherali, "Optimal Location of Transmitters," *IEEE J. on Selected Areas in Communications,* pp. 662–673, May 1996.

O. Shimbo, "Effects of intermodulation, AM-PM conversion, and additive noise in multicarrier TWT systems," *Proc. IEEE*, no. 2, 1971.

M. Sondhi, "Random processes with specified spectral density and first-order probability density," *Bell Syst. Tech. J.* vol. 62, 1983.

Further Information

Until recently the subject of computer-aided analysis and simulation of communication systems was a very difficult research area. There were no textbooks devoted to the subject, and the fundamental papers were scattered over a large number of technical journals. While a number of excellent books treated the subject of simulation of systems in which random signals and noise are present [Rubinstein, 1981; Bratley et al., 1987], none of these books specifically focused on communication systems.

Starting in 1984, the *IEEE Journal on Selected Areas in Communications (JSAC)* initiated the publication of a sequence of issues devoted specifically to the subject of computer-aided design and analysis of communication systems. A brief study of the contents of these issues tells much about the rapid development of the discipline. The first issue, published in January 1984 [Balaban et al., 1984], emphasizes communication links, although there are a number of papers devoted to networks. The portion devoted to links contained a large collection of papers devoted to simulation packages.

The second issue of the series was published in 1988 and is roughly evenly split between links and networks [Balaban et al., 1988]. In this issue the emphasis is much more on techniques than on simulation packages. The third part of the series is a two-volume issue devoted exclusively to networks [Mouftah et al., 1990, 1991].

As of this writing, the book by Jeruchim et al. is the only comprehensive treatment of the simulation of communication links [Jeruchim, 1984]. It treats the component and channel modeling problem and the problems associated with using simulation techniques for estimating the performance of communication systems in considerable detail. This textbook, together with the previously cited JSAC issues, gives a good overview of the area.

This page intentionally left blank

II

Mathematics, Symbols, and Physical Constants

Ronald J. Tallarida
Temple University

THE GREAT ACHIEVEMENTS in engineering deeply affect the lives of all of us and also serve to remind us of the importance of mathematics. Interest in mathematics has grown steadily with these engineering achievements and with concomitant advances in pure physical science. Whereas scholars in nonscientific fields, and even in such fields as botany, medicine, geology, etc., can communicate most of the problems and results in nonmathematical language, this is virtually impossible in present-day engineering and physics. Yet it is interesting to note that until the beginning of the twentieth century, engineers regarded calculus as something of a mystery. Modern students of engineering now study calculus, as well as differential equations, complex variables, vector analysis, orthogonal functions, and a variety of other topics in applied analysis. The study of systems has ushered in matrix algebra and, indeed, most engineering students now take linear algebra as a core topic early in their mathematical education.

This section contains concise summaries of relevant topics in applied engineering mathematics and certain key formulas, that is, those formulas that are most often needed in the formulation and solution of engineering problems. Whereas even inexpensive electronic calculators contain tabular material (e.g., tables of trigonometric and logarithmic functions) that used to be needed in this kind of handbook, most calculators do not give symbolic results. Hence, we have included formulas along with brief summaries that guide their use. In many cases we have added numerical examples, as in the discussions of matrices, their inverses, and their use in the solutions of linear systems. A table of derivatives is included, as well as key applications of the derivative in the solution of problems in maxima and minima, related rates, analysis of curvature, and finding approximate roots by numerical methods. A list of infinite series, along with the interval of convergence of each, is also included.

Of the two branches of calculus, integral calculus is richer in its applications, as well as in its theoretical content. Though the theory is not emphasized here, important applications such as finding areas, lengths, volumes, centroids, and the work done by a nonconstant force are included. Both cylindrical and spherical polar coordinates are discussed, and a table of integrals is included. Vector analysis is summarized in a separate section and includes a summary of the algebraic formulas involving dot and cross multiplication, frequently needed in the study of fields, as well as the important theorems of Stokes and Gauss. The part on special functions includes the gamma function, hyperbolic functions, Fourier series, orthogonal functions, and both Laplace and z-transforms. The Laplace transform provides a basis for the solution of differential equations and is fundamental to all concepts and definitions underlying analytical tools for describing feedback control systems. The z-transform, not discussed in most applied mathematics books, is most useful in the analysis of discrete signals as, for example, when a computer receives data sampled at some prespecified time interval. The Bessel functions, also called cylindrical functions, arise in many physical applications, such as the heat transfer in a "long" cylinder, whereas the other orthogonal functions discussed—Legendre, Hermite, and Laguerre polynomials—are needed in quantum mechanics and many other subjects (e.g., solid-state electronics) that use concepts of modern physics.

The world of mathematics, even applied mathematics, is vast. Even the best mathematicians cannot keep up with more than a small piece of this world. The topics included in this section, however, have withstood the test of time and, thus, are truly *core* for the modern engineer.

This section also incorporates tables of physical constants and symbols widely used by engineers. While not exhaustive, the constants, conversion factors, and symbols provided will enable the reader to accommodate a majority of the needs that arise in design, test, and manufacturing functions.

Mathematics, Symbols, and Physical Constants

Greek Alphabet

Greek Letter		Greek Name	English Equivalent	Greek Letter		Greek Name	English Equivalent
A	α	Alpha	a	N	ν	Nu	n
B	β	Beta	b	Ξ	ξ	Xi	x
Γ	γ	Gamma	g	O	o	Omicron	ŏ
Δ	δ	Delta	d	Π	π	Pi	P
E	ε	Epsilon	ĕ	P	ρ	Rho	r
Z	ζ	Zeta	z	Σ	σ	Sigma	s
H	η	Eta	ē	T	τ	Tau	t
Θ	θ ϑ	Theta	th	Y	υ	Upsilon	u
I	ι	Iota	i	Φ	φ φ	Phi	ph
K	κ	Kappa	k	X	χ	Chi	ch
Λ	λ	Lambda	l	Ψ	ψ	Psi	ps
M	μ	Mu	m	Ω	ω	Omega	ō

International System of Units (SI)

The International System of units (SI) was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960. It is a coherent system of units built form seven *SI base units,* one for each of the seven dimensionally independent base quantities: they are the meter, kilogram, second, ampere, kelvin, mole, and candela, for the dimensions length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity, respectively. The definitions of the SI base units are given below. The *SI derived units* are expressed as products of powers of the base units, analogous to the corresponding relations between physical quantities but with numerical factors equal to unity.

In the International System there is only one SI unit for each physical quantity. This is either the appropriate SI base unit itself or the appropriate SI derived unit. However, any of the approved decimal prefixes, called *SI prefixes,* may be used to construct decimal multiples or submultiples of SI units.

It is recommended that only SI units be used in science and technology (with SI prefixes where appropriate). Where there are special reasons for making an exception to this rule, it is recommended always to define the units used in terms of SI units. This section is based on information supplied by IUPAC.

Definitions of SI Base Units

Meter: The meter is the length of path traveled by light in vacuum during a time interval of 1/299,792,458 of a second (17th CGPM, 1983).

Kilogram: The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).

Second: The second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967).

Ampere: The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 m apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length (9th CGPM, 1948).

Kelvin: The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (13th CGPM, 1967).

Mole: The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kg of carbon-12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, or other particles or specified groups of such particles (14th CGPM, 1971).

Examples of the use of the mole:

1 mol of H_2 contains about 6.022×10^{23} H_2 molecules, or 12.044×10^{23} H atoms.

1 mol of HgCl has a mass of 236.04 g.

1 mol of Hg_2Cl_2 has a mass of 472.08 g.

1 mol of Hg_2^{2+} has a mass of 401.18 g and a charge of 192.97 kC.

1 mol of $Fe_{0.91}S$ has a mass of 82.88 g.

1 mol of e^- has a mass of 548.60 μg and a charge of -96.49 kC.

1 mol of photons whose frequency is 10^{14} Hz has energy of about 39.90 kJ.

Candela: The candela is the luminous intensity in a given direction of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of (1/683) watt per steradian (16th CGPM, 1979).

Names and Symbols for the SI Base Units

Physical Quantity	Name of SI Unit	Symbol for SI Unit
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

SI Derived Units with Special Names and Symbols

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units	
Frequency[1]	hertz	Hz	s^{-1}	
Force	newton	N	$m\ kg\ s^{-2}$	
Pressure, stress	pascal	Pa	$N\ m^{-2}$	$= m^{-1}\ kg\ s^{-2}$
Energy, work, heat	joule	J	$N\ m$	$= m^2\ kg\ s^{-2}$
Power, radiant flux	watt	W	$J\ s^{-1}$	$= m^2\ kg\ s^{-3}$
Electric charge	coulomb	C	$A\ s$	
Electric potential, electromotive force	volt	V	$J\ C^{-1}$	$= m^2\ kg\ s^{-3}\ A^{-1}$
Electric resistance	ohm	Ω	$V\ A^{-1}$	$= m^2\ kg\ s^{-3}\ A^{-2}$
Electric conductance	siemens	S	Ω^{-1}	$= m^{-2}\ kg^{-1}\ s^3\ A^2$
Electric capacitance	farad	F	$C\ V^{-1}$	$= m^{-2}\ kg^{-1}\ s^4\ A^2$
Magnetic flux density	tesla	T	$V\ s\ m^{-2}$	$= kg\ s^{-2}\ A^{-1}$
Magnetic flux	weber	Wb	$V\ s$	$= m^2\ kg\ s^{-2}\ A^{-1}$
Inductance	henry	H	$V\ A^{-1}\ s$	$= m^2\ kg\ s^{-2}\ A^{-2}$
Celsius temperature[2]	degree Celsius	℃	K	

(continued)

SI Derived Units with Special Names and Symbols (continued)

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units	
Luminous flux	lumen	lm	cd sr	
Illuminance	lux	lx	cd sr m^{-2}	
Activity (radioactive)	becquerel	Bq	s^{-1}	
Absorbed dose (of radiation)	gray	Gy	J kg^{-1}	= m^2 s^{-2}
Dose equivalent (dose equivalent index)	sievert	Sv	J kg^{-1}	= m^2 s^{-2}
Plane angle	radian	rad	1	= m m^{-1}
Solid angle	steradian	sr	1	= m^2 m^{-2}

[1]For radial (circular) frequency and for angular velocity the unit rad s^{-1}, or simply s^{-1}, should be used, and this may not be simplified to Hz. The unit Hz should be used only for frequency in the sense of cycles per second.

[2]The Celsius temperature θ is defined by the equation:

$$\theta/°C = T/K - 273.15$$

The SI unit of Celsius temperature interval is the degree Celsius, ℃, which is equal to the kelvin, K. ℃ should be treated as a single symbol, with no space between the ° sign and the letter C. (The symbol 𝖪 and the symbol ° should no longer be used.)

Units in Use Together with the SI

These units are not part of the SI, but it is recognized that they will continue to be used in appropriate contexts. SI prefixes may be attached to some of these units, such as milliliter, ml; millibar, mbar; megaelectronvolt, MeV; kilotonne, ktonne.

Physical Quantity	Name of Unit	Symbol for Unit	Value in SI Units
Time	minute	min	60 s
Time	hour	h	3600 s
Time	day	d	86,400 s
Plane angle	degree	°	(π/180) rad
Plane angle	minute	′	(π/10,800) rad
Plane angle	second	″	(π/648,000) rad
Length	ångstrom[1]	Å	10^{-10} m
Area	barn	b	10^{-28} m^2
Volume	liter	l, L	dm^3 = 10^{-3} m^3
Mass	tonne	t	Mg = 10^3 kg
Pressure	bar[1]	bar	10^5 Pa = 10^5 N m^{-2}
Energy	electronvolt[2]	eV (= e × V)	≈1.60218 × 10^{-19} J
Mass	unified atomic mass unit[2,3]	u (= m_a(^{12}C)/12)	≈1.66054 × 10^{-27} kg

[1]The ångstrom and the bar are approved by CIPM for "temporary use with SI units," until CIPM makes a further recommendation. However, they should not be introduced where they are not used at present.

[2]The values of these units in terms of the corresponding SI units are not exact, since they depend on the values of the physical constants e (for the electronvolt) and N_a (for the unified atomic mass unit), which are determined by experiment.

[3]The unified atomic mass unit is also sometimes called the dalton, with symbol Da, although the name and symbol have not been approved by CGPM.

Conversion Constants and Multipliers

Recommended Decimal Multiples and Submultiples

Multiples and Submultiples	Prefixes	Symbols	Multiples and Submultiples	Prefixes	Symbols
10^{18}	exa	E	10^{-1}	deci	d
10^{15}	peta	P	10^{-2}	centi	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ (Greek mu)
10^{6}	mega	M	10^{-9}	nano	n
10^{3}	kilo	k	10^{-12}	pico	p
10^{2}	hecto	h	10^{-15}	femto	f
10	deca	da	10^{-18}	atto	a

Conversion Factors—Metric to English

To Obtain	Multiply	By
Inches	centimeters	0.3937007874
Feet	meters	3.280839895
Yards	meters	1.093613298
Miles	kilometers	0.6213711922
Ounces	grams	$3.527396195 \times 10^{-2}$
Pounds	kilogram	2.204622622
Gallons (U.S. liquid)	liters	0.2641720524
Fluid ounces	milliliters (cc)	$3.381402270 \times 10^{-2}$
Square inches	square centimeters	0.155003100
Square feet	square meters	10.76391042
Square yards	square meters	1.195990046
Cubic inches	milliliters (cc)	$6.102374409 \times 10^{-2}$
Cubic feet	cubic meters	35.31466672
Cubic yards	cubic meters	1.307950619

Conversion Factors—English to Metric⋆

To Obtain	Multiply	By
Microns	mils	**25.4**
Centimeters	inches	**2.54**
Meters	feet	**0.3048**
Meters	yards	**0.9144**
Kilometers	miles	**1.609344**
Grams	ounces	28.34952313
Kilograms	pounds	**0.45359237**
Liters	gallons (U.S. liquid)	**3.785411784**
Millimeters (cc)	fluid ounces	29.57352956
Square centimeters	square inches	**6.4516**
Square meters	square feet	**0.09290304**
Square meters	square yards	**0.83612736**
Milliliters (cc)	cubic inches	**16.387064**
Cubic meters	cubic feet	$2.831684659 \times 10^{-2}$
Cubic meters	cubic yards	0.764554858

⋆Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

Conversion Factors—General★

To Obtain	Multiply	By
Atmospheres	feet of water @ 4°C	2.950×10^{-2}
Atmospheres	inches of mercury @ 0°C	3.342×10^{-2}
Atmospheres	pounds per square inch	6.804×10^{-2}
BTU	foot-pounds	1.285×10^{-3}
BTU	joules	9.480×10^{-4}
Cubic feet	cords	**128**
Degree (angle)	radians	57.2958
Ergs	foot-pounds	1.356×10^{7}
Feet	miles	**5280**
Feet of water @ 4°C	atmospheres	33.90
Foot-pounds	horsepower-hours	1.98×10^{6}
Foot-pounds	kilowatt-hours	2.655×10^{6}
Foot-pounds per min	horsepower	3.3×10^{4}
Horsepower	foot-pounds per sec	1.818×10^{-3}
Inches of mercury @ 0°C	pounds per square inch	2.036
Joules	BTU	1054.8
Joules	foot-pounds	1.35582
Kilowatts	BTU per min	1.758×10^{-2}
Kilowatts	foot-pounds per min	2.26×10^{-5}
Kilowatts	horsepower	0.745712
Knots	miles per hour	0.86897624
Miles	feet	1.894×10^{-4}
Nautical miles	miles	0.86897624
Radians	degrees	1.745×10^{-2}
Square feet	acres	**43,560**
Watts	BTU per min	17.5796

★Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

Temperature Factors

$$°F = 9/5 \ (°C) + 32$$
$$\text{Fahrenheit temperature} = 1.8 \ (\text{temperature in kelvins}) - 459.67$$
$$°C = 5/9 \ [(°F) - 32)]$$
$$\text{Celsius temperature} = \text{temperature in kelvins} - 273.15$$
$$\text{Fahrenheit temperature} = 1.8 \ (\text{Celsius temperature}) + 32$$

Conversion of Temperatures

From	To	
°Celsius	°Fahrenheit	$t_F = (t_C \times 1.8) + 32$
	Kelvin	$T_K = t_C + 273.15$
	°Rankine	$T_R = (t_C + 273.15) \times 18$
°Fahrenheit	°Celsius	$t_C = \dfrac{t_F - 32}{1.8}$
	Kelvin	$T_k = \dfrac{t_F - 32}{1.8} + 273.15$
	°Rankine	$T_R = t_F + 459.67$
Kelvin	°Celsius	$t_C = T_K - 273.15$
	°Rankine	$T_R = T_K \times 1.8$
°Rankine	Kelvin	$T_K = \dfrac{T_R}{1.8}$
	°Fahrenheit	$t_F = T_R - 459.67$

Physical Constants

General

Equatorial radius of the Earth $=$ 6378.388 km $=$ 3963.34 miles (statute)
Polar radius of the Earth, 6356.912 km $=$ 3949.99 miles (statute)
1 degree of latitude at 40° $=$ 69 miles
1 international nautical mile $=$ 1.15078 miles (statute) $=$ 1852 m $=$ 6076.115 ft
Mean density of the earth $=$ 5.522 g/cm^3 $=$ 344.7 lb/ft^3
Constant of gravitation (6.673 $\pm$ 0.003) $\times$ 10^{-8} cm^3 gm^{-1} s^{-2}
Acceleration due to gravity at sea level, latitude 45° $=$ 980.6194 cm/s^2 $=$ 32.1726 ft/s^2
Length of seconds pendulum at sea level, latitude 45° $=$ 99.3575 cm $=$ 39.1171 in.
1 knot (international) $=$ 101.269 ft/min $=$ 1.6878 ft/s $=$ 1.1508 miles (statute)/h
1 micron $=$ 10^{-4} cm
1 ångstrom $=$ 10^{-8} cm
Mass of hydrogen atom $=$ (1.67339$\pm$0.0031) $\times$ 10^{-24} g
Density of mercury at 0°C $=$ 13.5955 g/ml
Density of water at 3.98°C $=$ 1.000000 g/ml
Density, maximum, of water, at 3.98°C $=$ 0.999973 g/cm^3
Density of dry air at 0°C, 760 mm $=$ 1.2929 g/l
Velocity of sound in dry air at 0°C $=$ 331.36 m/s $-$ 1087.1 ft/s
Velocity of light in vacuum $=$ (2.997925$\pm$0.000002) $\times$ 10^{10} cm/s
Heat of fusion of water 0°C $=$ 79.71 cal/g
Heat of vaporization of water 100°C $=$ 539.55 cal/g
Electrochemical equivalent of silver 0.001118 g/s international amp
Absolute wavelength of red cadmium light in air at 15°C, 760 mm pressure $=$ 6438.4696 Å
Wavelength of orange-red line of krypton 86 $=$ 6057.802 Å

π Constants

π = 3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37511
$1/\pi$ = 0.31830 98861 83790 67153 77675 26745 02872 40689 19291 48091
π^2 = 9.8690 44010 89358 61883 44909 99876 15113 53136 99407 24079
$\log_e\pi$ = 1.14472 98858 49400 17414 34273 51353 05871 16472 94812 91531
$\log_{10}\pi$ = 0.49714 98726 94133 85435 12682 88290 89887 36516 78324 38044
$\log_{10}\sqrt{2\pi}$ = 0.39908 99341 79057 52478 25035 91507 69595 02099 34102 92128

Constants Involving e

e = 2.71828 18284 59045 23536 02874 71352 66249 77572 47093 69996
$1/e$ = 0.36787 94411 71442 32159 55237 70161 46086 74458 11131 03177
e^2 = 7.38905 60989 30650 22723 04274 60575 00781 31803 15570 55185
$M = \log_{10}e$ = 0.43429 44819 03251 82765 11289 18916 60508 22943 97005 80367
$1/M\cdot = \log_e10$ = 2.30258 50929 94045 68401 79914 54684 36420 67011 01488 62877
$\log_{10}M$ = 9.63778 43113 00536 78912 29674 98645 $-$10

Numerical Constants

$\sqrt{2}$ = 1.41421 35623 73095 04880 16887 24209 69807 85696 71875 37695
$3\sqrt{2}$ = 1.25992 10498 94873 16476 72106 07278 22835 05702 51464 70151
$\log_e2$ = 0.69314 71805 59945 30941 72321 21458 17656 80755 00134 36026
$\log_{10}2$ = 0.30102 99956 63981 19521 37388 94724 49302 67881 89881 46211

$$\sqrt{3} = 1.73205\ 08075\ 68877\ 29352\ 74463\ 41505\ 87236\ 69428\ 05253\ 81039$$
$$\sqrt[3]{3} = 1.44224\ 95703\ 07408\ 38232\ 16383\ 10780\ 10958\ 83918\ 69253\ 49935$$
$$\log_e 3 = 1.09861\ 22886\ 68109\ 69139\ 52452\ 36922\ 52570\ 46474\ 90557\ 82275$$
$$\log_{10} 3 = 0.47712\ 12547\ 19662\ 43729\ 50279\ 03255\ 11530\ 92001\ 28864\ 19070$$

Symbols and Terminology for Physical and Chemical Quantities

Name	Symbol	Definition	SI Unit
Classical Mechanics			
Mass	m		kg
Reduced mass	μ	$\mu = m_1 m_2/(m_1 + m_2)$	kg
Density, mass density	ρ	$\rho = M/V$	kg m^{-3}
Relative density	d	$d = \rho/\rho^\theta$	l
Surface density	ρ_A, ρ_S	$\rho_A = m/A$	kg m^{-2}
Momentum	p	$p = mv$	kg m s^{-1}
Angular momentum, action	L	$l = r ¥ p$	J s
Moment of inertia	I, J	$I = \Sigma m_i r_i^2$	kg m^2
Force	F	$F = \mathrm{d}\mathbf{p}/\mathrm{d}t = ma$	N
Torque, moment of a force	$T, (M)$	$T = r \times \mathbf{F}$	N m
Energy	E		J
Potential energy	E_p, V, Φ	$E_p = Fds$	J
Kinetic energy	E_k, T, K	$e_k = (1/2)mv^2$	J
Work	W, w	$w = Fds$	J
Hamilton function	H	$H(q, p) = T(q, p) + V(q)$	J
Lagrange function	L	$L(q, \dot{q})T(q, \dot{q}) - V(q)$	J
Pressure	p, P	$p = F/A$	Pa, N m^{-2}
Surface tension	γ, σ	$\gamma = \mathrm{d}W/\mathrm{d}A$	N m^{-1}, J m^{-2}
Weight	$G, (W, P)$	$G = mg$	N
Gravitational constant	G	$F = Gm_1 m_2/r^2$	N m^2 kg^{-2}
Normal stress	σ	$\sigma = F/A$	Pa
Shear stress	τ	$\tau = F/A$	Pa
Linear strain, relative elongation	ε, e	$\varepsilon = \Delta l/l$	l
Modulus of elasticity, Young's modulus	E	$E = \sigma/\varepsilon$	Pa
Shear strain	γ	$\gamma = \Delta x/d$	l
Shear modulus	G	$G = \tau/\gamma$	Pa
Volume strain, bulk strain	θ	$\theta = \Delta V/V_0$	l
Bulk modulus, compression modulus	K	$K = -V_0(\mathrm{d}p/\mathrm{d}V)$	Pa
Viscosity, dynamic viscosity	η, μ	$\tau_{x,z} = \eta(\mathrm{d}v_x/\mathrm{d}z)$	Pa s
Fluidity	ϕ	$\phi = 1/\eta$	m kg^{-1} s
Kinematic viscosity	v	$v = \eta/\rho$	m^2 s^{-1}
Friction coefficient	$\mu, (f)$	$F_{frict} = \mu F_{norm}$	l
Power	P	$P = \mathrm{d}W/\mathrm{d}t$	W
Sound energy flux	P, P_a	$P = \mathrm{d}E/\mathrm{d}t$	W
Acoustic factors			
Reflection factor	ρ	$\rho = P_t/P_0$	l
Acoustic absorption factor	$\alpha_a, (\alpha)$	$\alpha_a = 1 - \rho$	l
Transmission factor	τ	$\tau = P_{tr}/P_0$	l
Dissipation factor	δ	$\delta = \alpha_a - \tau$	l

(continued)

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	Symbol	Definition	SI Unit
Electricity and Magnetism			
Quantity of electricity, electric charge	Q		C
Charge density	ρ	$\rho = Q/V$	$C\ m^{-3}$
Surface charge density	σ	$\sigma = Q/A$	$C\ m^{-2}$
Electric potential	V, ϕ	$V = dW/dQ$	$V, J\ C^{-1}$
Electric potential difference	$U, \Delta V, \Delta\phi$	$U = V_2 - V_1$	V
Electromotive force	E	$E = (F/Q)ds$	V
Electric field strength	$\mathbf{E}$	$\mathbf{E} = \mathbf{F}/Q = -\text{grad}\ V$	$V\ m^{-1}$
Electric flux	Ψ	$\Psi = \mathbf{D}dA$	C
Electric displacement	$\mathbf{D}$	$\mathbf{D} = \varepsilon\mathbf{E}$	$C\ m^{-2}$
Capacitance	C	$C = Q/U$	$F, C\ V^{-1}$
Permittivity	ε	$D = \varepsilon E$	$F\ m^{-1}$
Permittivity of vacuum	ε_0	$\varepsilon_0 = \mu_0^{-1}\ c_0^{-2}$	$F\ m^{-1}$
Relative permittivity	ε_r	$\varepsilon_r = \varepsilon/\varepsilon_0$	1
Dielectric polarization (dipole moment per volume)	$\mathbf{P}$	$\mathbf{P} = \mathbf{D} - \varepsilon_0\mathbf{E}$	$C\ m^{-2}$
Electric susceptibility	χ_e	$\chi_e = \varepsilon_r - 1$	1
Electric dipole moment	$\mathbf{p}, \mu$	$\mathbf{p} = Q\mathbf{r}$	C m
Electric current	I	$I = dQ/dt$	A
Electric current density	$\mathbf{j}, \mathbf{J}$	$I = \mathbf{j}dx\mathbf{A}$	$A\ m^{-2}$
Magnetic flux density, magnetic induction	$\mathbf{B}$	$\mathbf{F} = Qv \times \mathbf{B}$	T
Magnetic flux	Φ	$\Phi = \mathbf{B}dA$	Wb
Magnetic field strength	$\mathbf{H}$	$\mathbf{B} = \mu\mathbf{H}$	$A\ M^{-1}$
Permeability	μ	$\mathbf{B} = \mu\mathbf{H}$	$N\ A^{-2}, H\ m^{-1}$
Permeability of vacuum	μ_0		$H\ m^{-1}$
Relative permeability	μ_r	$\mu_r = \mu/\mu_0$	1
Magnetization (magnetic dipole moment per volume)	$\mathbf{M}$	$\mathbf{M} = \mathbf{B}/\mu_0 - \mathbf{H}$	$A\ m^{-1}$
Magnetic susceptibility	$\chi, \kappa, (\chi_m)$	$\chi = \mu_r - 1$	1
Molar magnetic susceptibility	χ_m	$\chi_m = V_m\chi$	$m^3\ mol^{-1}$
Magnetic dipole moment	$\mathbf{m}, \mu$	$E_p = -\mathbf{m} \cdot \mathbf{B}$	$A\ m^2, J\ T^{-1}$
Electrical resistance	R	$\mathbf{P} = \mathbf{Y}/\mathbf{I}$	Ω
Conductance	G	$G = 1/R$	S
Loss angle	δ	$\delta = (\pi/2) + \phi_I - \phi_U$	1, rad
Reactance	X	$X = (U/I)\sin\delta$	Ω
Impedance (complex impedance)	Z	$Z = R + iX$	Ω
Admittance (complex admittance)	Y	$Y = 1/Z$	S
Susceptance	B	$Y = G + iB$	S
Resistivity	ρ	$\rho = E/j$	Ω m
Conductivity	κ, γ, σ	$\kappa = 1/\rho$	$S\ m^{-1}$
Self-inductance	L	$E = -L(dI/dt)$	H
Mutual inductance	M, L_{12}	$E_1 = L_{12}(Di_2/dt)$	H
Magnetic vector potential	$\mathbf{A}$	$\mathbf{B} = \nabla \times \mathbf{A}$	$Wb\ m^{-1}$
Poynting vector	$\mathbf{S}$	$\mathbf{S} = \mathbf{E} \times \mathbf{H}$	$W\ m^{-2}$
Electromagnetic Radiation			
Wavelength	λ		m
Speed of light			$m\ s^{-1}$
in vacuum	c_0		
in a medium	c	$c = c_0/n$	

(continued)

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	Symbol	Definition	SI Unit
Electromagnetic Radiation			
Wavenumber in vacuum	V	$V = V/c_0 = 1/n\lambda$	m^{-1}
Wavenumber (in a medium)	σ	$\sigma = 1/\lambda$	m^{-1}
Frequency	v	$v = c/\lambda$	Hz
Circular frequency, pulsatance	ω	$\omega = 2\pi v$	s^{-1}, rad s^{-1}
Refractive index	n	$n = c_0/c$	l
Planck constant	h		J s
Planck constant/2π	$\hbar$	$\hbar = h/2\pi$	J s
Radiant energy	Q, W		J
Radiant energy density	ρ, w	$\rho = Q/V$	J m^{-3}
Spectral radiant energy density			
in terms of frequency	ρ_v, w_v	$\rho_v = \delta\rho/dv$	J m^{-3} Hz^{-1}
in terms of wavenumber	$\rho_{\bar{v}}, w_{\bar{v}}$	$\rho_{\bar{v}} = d\rho/d\bar{v}$	J m^{-2}
in terms of wavelength	ρ_λ, w_λ	$\rho_\lambda = \delta\rho/d\lambda$	J m^{-4}
Einstein transition probabilities			
Spontaneous emission	A_{nm}	$dN_n/dt = -A_{nm}N_n$	s^{-1}
Stimulated emission	B_{nm}	$dn_n/dt = -\rho\bar{v}(\bar{V}_{nm}) \times B_{nm}N_n$	s kg^{-1}
Radiant power, radiant energy per time	Φ, P	$\Phi = dQ/dt$	W
Radiant intensity	I	$I = d\Phi/d\Omega$	W sr^{-1}
Radiant exitance (emitted radiant flux)	M	$M = d\Phi/dA_{source}$	W m^{-2}
Irradiance (radiant flux received)	$E, (I)$	$E = d\Phi/\delta A$	W m^{-2}
Emittance	ε	$\varepsilon = M/M_{bb}$	l
Stefan–Boltzmann constant	σ	$M_{bb} = \sigma T^4$	W m^{-2} K^{-4}
First radiation constant	c_1	$c_1 = 2\pi hc_0^2$	W m^2
Second radiation constant	c_2	$c_2 = hc_0/k$	K m
Transmittance, transmission factor	τ, T	$\tau = \Phi_{tr}/\Phi_0$	l
Absorptance, absorption factor	α	$\alpha = \phi_{abs}/\phi_0$	l
Reflectance, reflection factor	ρ	$\rho = \phi_{refl}/\Phi_0$	l
(Decadic) absorbance	A	$A = \lg(1 - \alpha_i)$	l
Napierian absorbance	B	$B = \ln(1 - \alpha_i)$	l
Absorption coefficient			
(Linear) decadic	a, K	$a = A/l$	m^{-1}
(Linear) napierian	α	$\alpha = B/l$	m^{-1}
Molar (decadic)	ε	$\varepsilon = a/c = A/cl$	m^2 mol^{-1}
Molar napierian	κ	$\kappa = \alpha/c = B/cl$	m^2 mol^{-1}
Absorption index	k	$k = \alpha/4\pi\bar{v}$	l
Complex refractive index	$\hat{n}$	$\hat{n} = n + ik$	l
Molar refraction	R, R_m	$R = \frac{(n^2-1)}{(n^2+2)} V_m$	m^3 mol^{-1}
Angle of optical rotation	α		l, rad
Solid State			
Lattice vector	$\mathbf{R}, \mathbf{R}_0$		m
Fundamental translation vectors for the crystal lattice	$\mathbf{a}_1; \mathbf{a}_2; \mathbf{a}_3, \mathbf{a}; \mathbf{b}; \mathbf{c}$	$R = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$	m
(Circular) reciprocal lattice vector	$\mathbf{G}$	$G \cdot R = 2\pi m$	m^{-1}

(continued)

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	Symbol	Definition	SI Unit
Solid State			
(Circular) fundamental translation vectors for the reciprocal lattice	$\mathbf{b}_1$; $\mathbf{b}_2$; $\mathbf{b}_3$, $\mathbf{a}^\star$; $\mathbf{b}^\star$; $\mathbf{c}^\star$	$\mathbf{a}_i \cdot \mathbf{b}_k = 2\pi\delta_{ik}$	m^{-1}
Lattice plane spacing	d		m
Bragg angle	θ	$n\lambda = 2d \sin\theta$	l, rad
Order of reflection	n		l
Order parameters			
Short range	σ		l
Long range	s		l
Burgers vector	b		m
Particle position vector	r, R_j		m
Equilibrium position vector of an ion	R_o		m
Displacement vector of an ion	$\mathbf{u}$	$\mathbf{u} = \mathbf{R} - \mathbf{R}_0$	m
Debye–Waller factor	B, D		l
Debye circular wavenumber	q_D		m^{-1}
Debye circular frequency	ω_D		s^{-1}
Grüneisen parameter	γ, Γ	$\gamma = \alpha V/\kappa C_V$	l
Madelung constant	α, $\mathscr{M}$	$E_{coul} = \frac{\alpha N_A z_+ z_- e^2}{4\pi\varepsilon_0 R_0}$	l
Density of states	N_E	$N_E = dN(E)/dE$	$J^{-1}\ m^{-3}$
(Spectral) density of vibrational modes	N_ω, g	$N_\omega = dN(\omega)/d\omega$	$s\ m^{-3}$
Resistivity tensor	ρ_{ik}	$E = \rho \cdot j$	$\Omega\ m$
Conductivity tensor	σ_{ik}	$\sigma = \rho^{-1}$	$S\ m^{-1}$
Thermal conductivity tensor	λ_{ik}	$J_q = -\lambda \cdot \mathrm{grad}\ T$	$W\ m^{-1}\ K^{-1}$
Residual resistivity	ρ_R		$\Omega\ m$
Relaxation time	τ	$\tau = l/v_F$	s
Lorenz coefficient	L	$L = \lambda/\sigma T$	$V^2\ K^{-2}$
Hall coefficient	A_H, R_H	$\mathbf{E} = \rho \cdot \mathbf{j} + R_H(\mathbf{B} \times \mathbf{j})$	$m^3\ C^{-1}$
Thermoelectric force	E		V
Peltier coefficient	Π		V
Thomson coefficient	μ,(τ)		$V\ K^{-1}$
Work function	Φ	$\Phi = E_\infty - E_F$	J
Number density, number concentration	n, (p)		m^{-3}
Gap energy	E_γ		J
Donor ionization energy	E_δ		J
Acceptor ionization energy	E_α		J
Fermi energy	E_Φ, ε_F		J
Circular wave vector, propagation vector	$\boldsymbol{k}$, $\boldsymbol{q}$	$k = 2\pi/\lambda$	m^{-1}
Bloch function	$u_k(\boldsymbol{r})$	$\psi(\boldsymbol{r}) = u_k(\boldsymbol{r})\exp(i\mathbf{k} \cdot \mathbf{r})$	$m^{-3/2}$
Charge density of electrons	ρ	$\rho(\boldsymbol{r}) = -e\psi^\star(\mathbf{r})\psi(\mathbf{r})$	$C\ m^{-3}$
Effective mass	$m^\star$		kg
Mobility	μ	$\mu = v_{drift}/E$	$m^2\ V^{-1}\ s^{-1}$
Mobility ratio	b	$b = \mu_n/\mu_p$	l
Diffusion coefficient	D	$dN/dt = -DA(dn/dx)$	$m^2\ s^{-1}$
Diffusion length	L	$L = \sqrt{D\tau}$	m
Characteristic (Weiss) temperature	ϕ, ϕ_W		K
Curie temperature	T_C		K
Néel temperature	T_N		K

Credits

Material in Section II was reprinted from the following sources:

D. R. Lide, Ed., *CRC Handbook of Chemistry and Physics,* 76th ed., Boca Raton, FL: CRC Press, 1992: International System of Units (SI), conversion constants and multipliers (conversion of temperatures), symbols and terminology for physical and chemical quantities, fundamental physical constants, classification of electromagnetic radiation.

D. Zwillinger, Ed., *CRC Standard Mathematical Tables and Formulae,* 30th ed., Boca Raton, FL: CRC Press, 1996: Greek alphabet, conversion constants and multipliers (recommended decimal multiples and submultiples, metric to English, English to metric, general, temperature factors), physical constants, series expansion.

Probability for Electrical and Computer Engineers

Charles W. Therrien

The Algebra of Events

The study of probability is based upon experiments that have uncertain outcomes. Collections of these outcomes comprise *events* and the collection of all possible outcomes of the experiment comprise what is called the *sample space*, denoted by S. Outcomes are members of the sample space and events of interest are represented as *sets* of outcomes (see Figure II.1).

The algebra $\mathcal{A}$ that deals with representing events is the usual set algebra. If A is an event, then A^c (the *complement* of A) represents the event that "A did not occur." The complement of the sample space is the *null event*, $\varnothing = S^c$. The event that *both* event A_1 and event A_2 have occurred is the intersection, written as "$A_1 \cdot A_2$" or "$A_1 A_2$" while the event that *either* A_1 or A_2 *or both* have occurred is the union, written as "$A_1 + A_2$."[1]

Table II.1 lists the two postulates that define the algebra $\mathcal{A}$, while Table II.2 lists seven axioms that define properties of its operations. Together these tables can be used to show all of the properties of the algebra of events. Table II.3 lists some additional useful relations that can be derived from the axioms and the postulates.

Since the events "$A_1 + A_2$" and "$A_1 A_2$" are included in the algebra, it follows by induction that for any finite number of events $A_1 + A_2 + \cdots + A_N$ and $A_1 \cdot A_2 \cdots \cdot A_N$ are also included in the algebra. Since problems often involve the union or intersection of an *infinite* number of events, however, the algebra of events must be defined to include these infinite intersections and unions. This extension to infinite unions and intersections is known as a sigma algebra.

A set of events that satisfies the two conditions:

1. $A_i A_j = \varnothing \neq$ for $\neq i \neq j$
2. $A_1 + A_2 + A_3 + \cdots = S$

is known as a *partition* and is important for the solution of problems in probability. The events of a partition are said to be *mutually exclusive* and *collectively exhaustive*. The most fundamental partition is the set outcomes defining the random experiment, which comprise the sample space by definition.

Probability

Probability measures the likelihood of occurrence of events represented on a scale of 0 to 1. We often estimate probability by measuring the *relative frequency* of an event, which is defined as

$$\text{relative frequency} = \frac{\text{number of occurrences of the event}}{\text{number of repetitions of the experiment}}$$

(for a large number of repetitions). Probability can be defined formally by the following axioms:

(I) The probability of any event is nonnegative:

$$\Pr[A] \geqslant 0 \tag{II.1}$$

(II) The probability of the universal event (i.e., the entire sample space) is 1:

$$\Pr[S] = 1 \tag{II.2}$$

---

[1]Some authors use $\cap$ and $\cup$ rather than $\cdot$ and $+$, respectively.
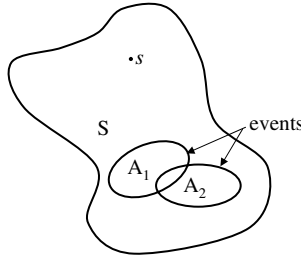
FIGURE II.1 Abstract representation of the sample space S with outcome s and sets A_1 and A_2 representing events.

(III) If A_1 and A_2 are mutually exclusive, i.e., $A_1 A_2 = \varnothing$, then

$$\Pr[A_1 + A_2] = \Pr[A_1] + \Pr[A_2] \tag{II.3}$$

(IV) If $\{A_i\}$ represent a countably infinite set of mutually exclusive events, then

$$\Pr[A_1 + A_2 + A_3 + \cdots] = \sum_{i=1}^{\infty} \Pr[A_i] \quad (\text{ if } A_i A_j = \varnothing \quad i \neq j) \tag{II.4}$$

Note that although the additivity of probability for any finite set of disjoint events follows from (III), the property has to be stated explicitly for an infinite set in (IV). These axioms and the algebra of events can be used to show a number of other important properties which are summarized in Table II.4. The last item in the table is an especially important formula since it uses probabilistic information about

TABLE II.1 Postulates for an Algebra of Events

1.	If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
2.	If $A_1 \in \mathcal{A}$ and $A_2 \in \mathcal{A}$ then $A_1 + A_2 \in \mathcal{A}$

TABLE II.2 Axioms of Operations on Events

$A_1 A_1^c = \varnothing$	Mutual exclusion
$A_1 S = A_1$	Inclusion
$(A_1^c)^c = A_1$	Double complement
$A_1 + A_2 = A_2 + A_1$	Commutative law
$A_1 + (A_2 + A_3) = (A_1 + A_2) + A_3$	Associative law
$A_1(A_2 + A_3) = A_1 A_2 + A_1 A_3$	Distributive law
$(A_1 A_2)^c = A_1^c + A_2^c$	DeMorgan's law

TABLE II.3 Additional Identities in the Algebra of Events

$S^c = \varnothing$	
$A_1 + \varnothing = A_1$	Inclusion
$A_1 A_2 = A_2 A_1$	Commutative law
$A_1(A_2 A_3) = (A_1 A_2)A_3$	Associative law
$A_1 + (A_2 A_3) = (A_1 + A_2)(A_1 + A_3)$	Distributive law
$(A_1 + A_2)^c = A_1{}^c A_2{}^c$	DeMorgan's law

TABLE II.4   Some Corollaries Derived from the Axioms of Probability

$\Pr[A^c] = 1 - \Pr[A]$
$0 \leq \Pr[A] \leq 1$
If $A_1 \subseteq A_2$ then $\Pr[A_1] \leq \Pr[A_2]$
$\Pr[\varnothing] = 0$
If $A_1 A_2 = \varnothing -$ then $= \Pr[A_1 A_2] = 0$
$\Pr[A_1 + A_2] = \Pr[A_1] + \Pr[A_2] - \Pr[A_1 A_2]$

individual events to compute the probability of the union of two events. The term $\Pr[A_1 A_2]$ is referred to as the *joint probability* of the two events. This last equation shows that the probabilities of two events add as in Equation (II.3) only if their joint probability is 0. The joint probability is 0 when the two events have no intersection ($A_1 A_2 = \varnothing$).

Two events are said to be statistically *independent* if and only if

$$\Pr[A_1 A_2] = \Pr[A_1] \cdot \Pr[A_2] \quad \text{(independent events)} \tag{II.5}$$

This definition is not derived from the earlier properties of probability. An argument to give this definition intuitive meaning can be found in Ref. [1]. Independence occurs in problems where two events are not influenced by one another and Equation (II.5) simplifies such problems considerably.

A final important result deals with partitions. *A partition* is a finite or countably infinite set of events $A_1, A_2, A_3, \ldots$ that satisfy the two conditions:

$$A_i A_j = \varnothing \text{ for } i \neq j$$

$$A_1 + A_2 + A_3 + \cdots = S$$

The events in a partition satisfy the relation:

$$\sum_i \Pr[A_i] = 1 \tag{II.6}$$

Further, if B is *any* other event, then

$$\Pr[B] = \sum_i \Pr[A_i B] \tag{II.7}$$

The latter result is referred to as the *principle of total probability* and is frequently used in solving problems. The principle is illustrated by a Venn diagram in Figure II.2. The rectangle represents the sample space and other events are defined therein. The event B is seen to be comprised of all of the pieces
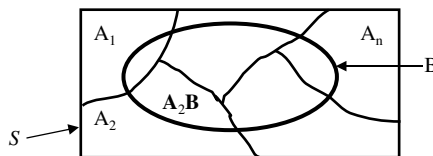
FIGURE II.2   Venn diagram illustrating the principle of total probability.

that represent intersections or overlap of event B with the events A_i. This is the graphical interpretation of Equation (II.7).

An Example

Simon's Surplus Warehouse has large barrels of mixed electronic components (parts) that you can buy by the handful or by the pound. You are not allowed to select parts individually. Based on your previous experience, you have determined that in one barrel, 29% of the parts are bad (faulted), 3% are bad resistors, 12% are good resistors, 5% are bad capacitors, and 32% are diodes. You decide to assign probabilities based on these percentages. Let us define the following events:

Event	Symbol
Bad (faulted) component	B
Good component	G
Resistor	R
Capacitor	C
Diode	D

A Venn diagram representing this situation is shown below along with probabilities of various events as given:

B Pr[B] = 0.29
Pr[BR] = 0.03
Pr[GR] = 0.12
Pr[BC] = 0.05
Pr[D] = 0.32
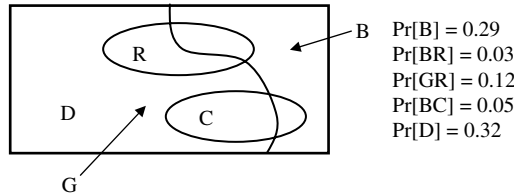
Note that since any component must be a resistor, capacitor, or diode, the region labeled D in the diagram represents everything in the sample space which is not included in R or C.

We can answer a number of questions.

1. What is the probability that a component is a resistor (either good *or* bad)?
   Since the events B and G form a partition of the sample space, we can use the principle of total probability Equation (II.7) to write:

$$\Pr[R] = \Pr[GR] + \Pr[BR] = 0.12 + 0.03 = 0.15$$

2. Are bad parts and resistors independent?
   We know that $\Pr[BR] = 0.03$ and we can compute:

$$\Pr[B] \cdot \Pr[R] = (0.29)(0.15) = 0.0435$$

   Since $\Pr[BR] \neq \Pr[B] \cdot \Pr[R]$, the events are *not* independent.
3. You have no use for either bad parts or resistors. What is the probability that a part is either bad and/or a resistor?

Using the formula from Table II.4 and the previous result we can write:

$$\Pr[B + R] = \Pr[B] + \Pr[R] - \Pr[BR] = 0.29 + 0.15 - 0.03 = 0.41$$

4. What is the probability that a part is useful to you?
   Let U represent the event that the part is useful. Then (see Table II.4):

$$\Pr[U] = 1 - \Pr[U^c] = 1 - 0.41 = 0.59$$

5. What is the probability of a bad diode?
   Observe that the events *R*, *C*, and *D* form a partition, since a component has to be one and only one type of part. Then using Equation (II.7) we write:

$$\Pr[B] = \Pr[BR] + \Pr[BC] + \Pr[BD]$$

   Substituting the known numerical values and solving yields

$$0.29 = 0.03 + 0.05 + \Pr[BD] \text{ or } \Pr[BD] = 0.21$$

Conditional Probability and Bayes' Rule

The *conditional* probability of an event A_1 given that an event A_2 has occurred is defined by

$$\Pr[A_1|A_2] = \frac{\Pr[A_1 A_2]}{\Pr[A_2]} \tag{II.8}$$

($\Pr[A_1|A_2]$ is read "probability of A_1 *given* A_2.") As an illustration, let us compute the probability that a component in the previous example is bad given that it is a resistor:

$$\Pr[B|R] = \frac{\Pr[BR]}{\Pr[R]} = \frac{0.03}{0.15} = 0.2$$

(The value for $\Pr[R]$ was computed in question 1 of the example.) Frequently the statement of a problem is in terms of conditional probability rather than joint probability, so Equation (II.8) is used in the form:

$$\Pr[A_1 A_2] = \Pr[A_1|A_2] \cdot \Pr[A_2] = \Pr[A_2|A_1] \cdot \Pr[A_1] \tag{II.9}$$

(The last expression follows because $\Pr[A_1 A_2]$ and $\Pr[A_2 A_1]$ are the same thing.) Using this result, the principle of total probability Equation (II.7) can be rewritten as

$$\Pr[B] = \sum_j \Pr[B|A_j] \Pr[A_j] \tag{II.10}$$

where *B* is any event and $\{A_j\}$ is a set of events that forms a partition.
   Now, consider any one of the events A_i in the partition. It follows from Equation (II.9) that

$$\Pr[A_i|B] = \frac{\Pr[B|A_i] \cdot \Pr[A_i]}{\Pr[B]}$$

Then substituting in Equation (II.10) yields:

$$Pr[A_i|B] = \frac{Pr[B|A_i] \cdot Pr[A_i]}{\sum_j Pr[B|A_j] Pr[A_j]} \tag{II.11}$$

This result is known as *Bayes' theorem* or *Bayes' rule*. It is used in a number of problems that commonly arise in electrical engineering. We illustrate and end this section with an example from the field of communications.

Communication Example

The transmission of bits over a binary communication channel is represented in the drawing below:
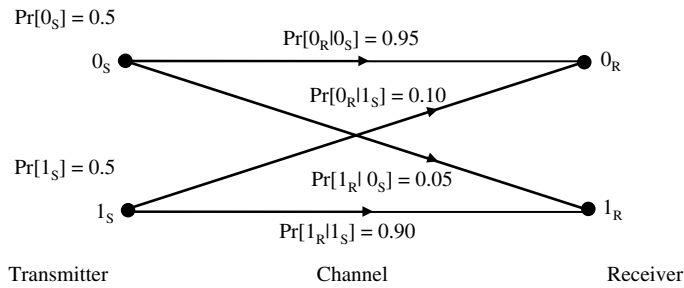
where we use notation like 0_S, 0_R ... to denote events "0 sent," "0 received," etc. When a 0 is transmitted, it is correctly received with probability 0.95 or incorrectly received with probability 0.05. That is, $Pr[0_R|0_S] = 0.95$ and $Pr[1_R|0_S] = 0.05$. When a 1 is transmitted, it is correctly received with probability 0.90 and incorrectly received with probability 0.10. The probabilities of sending a 0 or a 1 are denoted by $Pr[0_S]$ and $Pr[1_S]$. It is desired to compute the *probability of error* for the system.

This is an application of the principle of total probability. The two events 0_S and 1_S are mutually exclusive and collectively exhaustive and thus form a partition. Take the event B to be the event that an error occurs. It follows from Equation (II.10) that

$$Pr[error] = Pr[error|0_S] Pr[0_S] + Pr[error|1_S] Pr[1_S]$$

$$= Pr[1_R|0_S]Pr[0_S] + Pr[0_R|1_S] Pr[1_S]$$

$$= (0.05)(0.5) + (0.10)(0.5) = 0.075$$

Next, given that an error has occurred, let us compute the probability that a 1 was sent or a 0 was sent. This is an application of Bayes' rule. For a 1, Equation (II.11) becomes

$$Pr[1_S|error] = \frac{Pr[error|1_S] Pr[1_S]}{Pr[error|1_S] Pr[1_S] + Pr[error|0_S] Pr[0_S]}$$

Substituting the numerical values then yields:

$$Pr[1_S|error] = \frac{(0.10)(0.5)}{(0.10)(0.5) + (0.05)(0.5)} \approx 0.667$$

For a 0, a similar analysis applies:

$$\Pr[0_S|\text{error}] = \frac{\Pr[\text{error}|0_S]\,\Pr[0_S]}{\Pr[\text{error}|1_S]\,\Pr[1_S] + \Pr[\text{error}|0_S]\,\Pr[0_S]}$$

$$= \frac{(0.05)(0.5)}{(0.10)(0.5) + (0.05)(0.5)} \approx 0.333$$

The two resulting probabilities sum to 1 because 0_S and 1_S form a partition for the experiment.

Reference

1.   C. W. Therrien and M. Tummala, *Probability for Electrical and Computer Engineers*. Boca Raton, FL: CRC Press, 2004.

Indexes

This page intentionally left blank

Author Index

Subject Index

Page on which a term is defined is indicated in bold.