Digital Telephony

WILEY SERIES IN TELECOMMUNICATIONS AND SIGNAL PROCESSING

John G. Proakis, Editor
Northeastern University

Introduction to Digital Mobil Communications
Yoshihiko Akaiwa
Digital Telephony, 3rd Edition
John Bellamy
Elements of Information Theory
Thomas M. Cover and Joy A. Thomas
Fundamentals of Telecommunications
Roger L. Freeman
Practical Data Communications
Roger L. Freeman
Radio System Design for Telecommunications, 2nd Edition
Roger L. Freeman
Telecommunication System Engineering, 3rd Edition
Roger L. Freeman
Telecommunications Transmission Handbook, 4th Edition
Roger L. Freeman
Introduction to Communications Engineering, 2nd Edition
Robert M. Gagliardi
Optical Communications, 2nd Edition
Robert M. Gagliardi and Sherman Karp
Active Noise Control Systems: Algorithms and DSP Implementations
Sen M. Kuo and Dennis R. Morgan
Mobile Communications Design Fundamentals, 2nd Edition
William C. Y. Lee
Expert System Applications for Telecommunications
Jay Liebowitz
Digital Signal Estimation
Robert J. Mammone, Editor
Digital Communication Receivers: Synchronization, Channel Estimation, and Signal Processing
Heinrich Meyr, Marc Moeneclaey, and Stefan A. Fechtel
Synchronization in Digital Communications, Volume I
Heinrich Meyr and Gerd Ascheid
Business Earth Stations for Telecommunications
Walter L. Morgan and Denis Rouffet
Wireless Information Networks
Kaveh Pahlavan and Allen H. Levesque
Satellite Communications: The First Quarter Century of Service
David W. E. Rees
Fundamentals of Telecommunication Networks
Tarek N. Saadawi, Mostafa Ammar, with Ahmed El Hakeem
Meteor Burst Communications: Theory and Practice
Donald L. Schilling, Editor
Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics
Henry Stark and Yongyi Yang
Signaling in Telecommunication Networks
John G. van Bosse
Telecommunication Circuit Design
Patrick D. van der Puije
Worldwide Telecommunications Guide for the Business Manager
Walter H. Vignault
Polynomial Signal Processing
V. John Mathews and Giovanni L. Sicuranza
ADSL, VDSL, and Multicarrier Modulation
John A. C. Bingham

Digital Telephony
Third Edition

John C. Bellamy
Coppell, TX

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper. ♾

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging-in-Publication Data:

Bellamy, John, 1941–
Digital telephony / John Bellamy.—3rd ed.
p. cm. — (Wiley series in telecommunications and signal processing)
"A Wiley-Interscience publication."
Includes index.
ISBN 0-471-34571-7
1. Digital telephone systems. I. Title. II. Series.
TK5103.7.B44 2000
621.385–dc21 99-34015

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my father for passing on the enjoyment of being an engineer

CONTENTS

PREFACE

As mentioned in the preface of the first two editions, the term *digital telephony* refers to the use of digital technology in the message path of voice communications networks. In this case the term *digital* refers to a method of encoding the signal—that is, a form of modulation. Hence digital telephony implies voice transmission and switching applications, not data communications. Although the primary focus of this book is not data communications, this edition contains an expanded treatment of data communications networks, particularly as they relate to providing voice communications services in addition to data communications.

This book covers all aspects of digital voice communications technology and networks. It is not a technical book in the traditional, analytical sense of communications theory. Since numerous books covering communications theory are already available, this book stresses the application and operational aspects of communications system design. Some basic theory is presented in both qualitative and, when appropriate, quantitative terms. The main purpose, however, is to introduce concepts, terminology, and how applications influence implementations. In most cases the concepts are supported by citing example implementations in the U.S. telephone network, although examples from other (ITU) public telephone networks are also provided.

The primary audience for this book are graduate electrical engineers. The electrical engineering student is most capable of appreciating occasional references to communications theory and its influence on the practice. However, because analytical rigor is waived in favor of operational descriptions, less analytically oriented readers should have no difficulty understanding the principles. Chapter 6 (covering digital radio and modulation) is the most analytical but is easily skipped without losing continuity for the other chapters. Similarly, Chapter 12 (covering traffic analysis) contains numerous equations that are unnecessary for understanding the material in other chapters.

When the first edition of *Digital Telephony* was written (1980), public telephone networks around the world were primarily implemented with analog technology, but it was clear that digital technology was rapidly taking over. When the second edition was written (1990), the internal portions of the network had, for the most part, been converted to an all-digital network. Then and today (1999) the main remnants of the

original analog telephone networks are analog subscriber loops and analog telephones connected to them.

Although Integrated Services Digital Network (ISDN) technology was developed as a means of replacing analog loops to complete the transformation of the network to support end-to-end digital connections, ISDN deployment is below expectations for several reasons. One of these reasons is a growing need for more bandwidth than what is available from a basic rate ISDN subscriber loop (128 kbps). There is currently much activity within the industry to develop new technologies for medium- and high-bandwidth digital subscriber access. A new chapter (Chapter 11) has been added to this edition to specifically address alternative technologies for digital subscriber access.

Another relatively recent application of digital technology added to this edition involves digital cellular telephones, which first appeared in the marketplace in the mid-1990s. Digital mobile radio is enabled by the emergence of low-cost, high-performance digital signal processing (DSP) technology for compressing speech signals to low bit rates and for providing sophisticated coding, modulation, and equalization required for digital radios in a bandwidth-constrained mobile application.

A complete list of chapter topics is:

Chapter 1: Overview of analog telephone technology followed by an introduction of how digital technology is used to fulfill the same functions

Chapter 2: Discussion of advantages and disadvantages of digital technology for voice communications

Chapter 3: Descriptions of the most common voice digitization algorithms

Chapter 4: Fundamentals of digital wire-line transmission and multiplexing

Chapter 5: Basic concepts and operations of digital switching machines

Chapter 6: Digital modulation and radio fundamentals

Chapter 7: Network synchronization, control, and management requirements

Chapter 8: Fiber optic transmission systems and SONET

Chapter 9: Digital cellular telephone systems

Chapter 10: Data networks

Chapter 11: Digital subscriber access technology

Chapter 12: Fundamentals of traffic analysis for designing networks

The appendices cover the derivation of equations, PCM voice coding relationships, fundamentals of digital communications theory, and traffic tables.

JOHN C. BELLAMY

Coppell, Texas
October 1999
johncbellamy@ieee.org

ACKNOWLEDGMENT

Once again I am indebted to Wanda Fox and Alcatel USA for allowing me access to the corporate library for research materials for this edition. I also owe a great deal of gratitude to Gerald Mitchell of the University of Colorado for thoroughly reviewing and enhancing the last chapter on traffic theory.

J.B.

ACRONYMS

AAL	ATM adaptation layer
ABR	available bit rate
ACD	automatic call distributor
ACK	acknowledgment (positive)
ADM	adaptive delta modulation; add-drop multiplexer
ADPCM	adaptive differential PCM
AIN	advanced intelligent network
AMPS	advanced mobile phone system
ANI	automatic number identification
APC	adaptive predictive coding
APON	ATM based passive optical network
ARPANET	advanced research projects agency network
ARQ	automatic repeat request
ATM	asynchronous transfer mode
ATPC	adaptive transmit power control
B8ZS	binary 8 zero substitution
BER	bit-error rate
BISDN	broadband integrated services digital network
BLSR	bidirectional line switched ring
CAC	connection admission control
CAP	competitive access provider; carrierless amplitude/phase modulation
CAS	channel associated signaling
CBR	constant bit rate (ATM)
CCC	clear channel capability
CCIS	common channel interoffice signaling
CCITT	Consultative Committee for International Telephony and Telegraphy (now ITU)
CCS	common channel signaling; hundred call seconds
CDMA	code division multiple access
CDO	community dial office
CDPD	cellular digital packet data (for AMPS networks)

CELP	code excited linear prediction
CES	circuit emulation service (ATM)
CGSA	Cellular Geographic Service Area
CLASS	custom local area signaling services
CLEC	competitive local exchange carrier
CLP	cell loss priority (ATM)
CMI	coded mark inversion
CODEC	CODer/DECoder
CPFSK	continuous phase frequency shift keying
CRC	cyclic redundancy check
CSMA/CD	carrier sense multiple access/collision detection
CSU	channel service unit
CTD	cell transfer delay
CTI	computer telephony integration
D-AMPS	digital advanced mobile phone service
DAVIC	digital audio video council
DBS	direct broadcast satellite
DCM	digital circuit multiplication
DCME	digital circuit multiplication equipment
DECT	digital enhanced cordless telephony
DFE	decision feedback equalization
DID	direct inward dialing
DLC	digital loop carrier
DM	delta modulation; degraded minute
DMT	discrete multitone
DNIS	dialed number identification service
DPCM	differential pulse code modulation
DQDB	distributed queue dual bus
DS1	digital signal level 1 at 1.544 Mbps
DS3	digital signal level 3 at 44.736 Mbps
DSI	digital speech interpolation
DSS	digital satellite system
DTE	data terminal equipment
DTMF	dual tone multifrequency (signaling tones)
DVB	digital video broadcasting group
DVD	digital video disc
E1	European digital signal level 1 at 2.048 Mbps
E3	European digital signal level 3 at 34.368 Mbps
ECMA	European Computer Manufacturers Association
EMI	electromagnetic interference
ERMES	enhanced radio message system
ESF	extended super frame
ETSI	European Telecommunications Standards Institute
FDDI	fiber distributed data interchange

FDM	frequency division multiplexing
FEC	forward error correction
FEXT	far end crosstalk
FIFO	first in-first out
FRAD	frame relay access device
FSK	frequency shift keying
FTTC	fiber to the curb
FTTH	fiber to the home
GPS	global positioning system
GSM	global system for mobile communications
HDB3	high density bipolar of order 3
HDLC	high-level data link control
HIPPI	high performance parallel interface
HTTP	hypertext transport protocol
IDLC	integrated digital loop carrier
IEC	International Electrotechnical Commission
IETF	internet engineering task force
ILEC	incumbent local exchange carrier
IMT	international mobile telecommunications
IP	internet protocol
ISDN	integrated services digital network
ISI	intersymbol interference
ISO	International Standards Organization
ITU	International Telecommunications Union
IVR	interactive voice response
JPEG	Joint Photographic Experts Group
LAN	local area network
LATA	local access transport area
LD-CELP	low-delay CELP
LEC	local exchange carrier
LMDS	local microwave distribution service
MAN	metropolitan area network
MCM	multicarrier modulation
MLCM	multilevel coded modulation
MMDS	multichannel multipoint distribution service
MPEG	Motion Pictures Experts Group
MPLS	multiprotocol label switching
MSK	minimum shift keying
MTIE	maximum time interval error
MTSO	mobile telephone switching office
MULDEM	multiplexer-demultiplexer
NAK	acknowledgment (negative)
NCP	network control point; network control protocol (ARPANET)
NEXT	near end crosstalk

NMT	Nordic mobile telephone system
NNI	network-to-network interface
NRZ	nonreturn to zero
OFDM	orthogonal frequency division multiplexing
OSI	open systems interconnection
PABX	private automatic branch exchange (also PBX)
PAM	pulse amplitude modulation
PBX	private branch exchange
PCM	pulse code modulation
PCME	packet circuit multiplication equipment
PCR	peak cell rate (ATM)
PCS	personal communication system (or service)
PDC	personal digital cellular (Japan)
PDH	plesiochronous digital hierarchy
PHS	personal handyphone system (Japan)
PLL	phase locked loop
PON	passive optical network
POTS	plain old telephone service
PRC	primary reference clock
PRK	phase reversal keying
PRS	partial response signaling; primary reference source
PSK	phase shift keying
PSTN	public switched telephone network
PVC	permanent virtual circuit
QoS	quality of service
QPRS	quadrature partial response signaling
QPSK	quaternary phase shift keying (4-PSK)
RADSL	rate adaptive digital subscriber loop
RCC	radio common carrier
SDH	synchronous digital hierarchy
SDLC	synchronous data link control
SES	severely errored seconds
SF	super frame
SIM	Subscriber Identification Module (GSM)
SLIC	subscriber loop interface circuit
SMDS	switched multimegabit data service
SMR	specialized mobile radio
SMS	short message service
SNMP	simple network management protocol
SOHO	small office/home office
SONET	synchronous optical network
SRTS	synchronous residual time stamp
SS7	signaling system version 7
STM	synchronous transfer mode

STS	space-time-space digital switching structure
STS-n	synchronous transport signal-n
SVC	switched virtual circuit
T1	TDM transmission system at 1.544 Mbps
T3	TDM transmission system at 44.736 Mbps
TACS	total access communications system
TASI	time assignment speech interpolation
TCM	trellis coded modulation
TCP/IP	transmission control protocol/internet protocol
TDM	time division multiplexing
TDMA	time division multiple access
TETRA	Trans-European trunked radio
TMN	Telecommunications Management Network
TST	time-space-time digital switching structure
UBR	unspecified bit rate
UDP	User datagram protocol
UI	unit interval
UMTS	universal mobile telephone service
UNI	user-to-network interface
UPSR	unidirectional path switched ring
URL	universal resource locator
UTP	unshielded twisted pair
VBR	variable bit rate
VCC	virtual channel connection
VCI	virtual channel identifier
VPC	virtual path connection
VPCI	virtual path connection identifier
WAN	wide area network
WATS	wide area telecommunications services

1

BACKGROUND AND TERMINOLOGY

Beginning in the 1960s, telecommunications in the United States and around the world began undergoing radical changes in several different areas. First, the conventional analog telephone network was being called upon to provide many new and different services, most of which emanated from the data processing industry. Second, the marketplace and the regulatory agencies in the United States stimulated competition in both old and new areas of traditionally monopolistic services. Third, digital technology emerged to implement many of the fundamental transmission and switching functions within the U.S. telephone network and other networks around the world. The main purpose of this book is to describe the design, application, and operational aspects of this new digital equipment. As background, the technology of the analog telephone network is reviewed to provide a framework for the introduction of digital equipment.

It must be emphasized that the introduction of digital technology into the telephone network was motivated by desires to improve the quality, add new features, and reduce the costs of conventional voice services. Digitization of the network did not arise from the needs of the data processing industry for better data transmission services. Indeed, most of the digital technology introduced into the network was initially inaccessible to data traffic, except through analog channels. Of course, a digital network is a natural environment for data communications services. As more of the network became digitized, more support for direct use of the facilities became available for data applications. Initially, direct digital access existed only for relatively high-end business applications. It was not until facilities of the Integrated Services Digital Network (ISDN) became available that end-to-end switched digital channels could be used by individual subscribers for both voice and data. By the late 1990s numerous other approaches to providing digital access to digital facilities became available, primarily for Internet access. These various digital access technologies are described in Chapter 11.

As a point of historical reference, Figure 1.1 has been included to show that the idea of integrated voice and data is not new. This figure depicts a concept of a German inventor named Phillip Reis [1] to add voice communications to the prevailing means of electrical communications at the time—the telegraph. Reis developed the equipment in the 1860s and died in 1874—two years before Alexander Graham Bell re-

ceived his patent for the telephone. As indicated, the figure implies alternate use of the wires for voice or data communications (i.e., integrated transmission). Reis actually used the telegraph attachment to signal information pertaining to voice tests, an indication of inadequate voice quality.

To implement simultaneous voice and telegraph communications, the telephone in Figure 1.1 would have to have been digital. Because of technology limitations at the time, such an implementation was impossible and telephone systems necessarily evolved with analog technology. One hundred years later the situation changed significantly. Telephone equipment developers and service providers had an abundance of new technology, and they were challenged with how to make effective use of it.

This book describes digital telephone technology from two perspectives. The first perspective describes individual equipments or subsystems and technical reasons for transitions from conventional analog equipment to seemingly less natural digital counterparts. Thus, one purpose of this book is to describe how digital technology improves and expands the capabilities of various subsystems within voice telephone networks. Another purpose of the book is to describe the ultimate benefits derived when an entire network is implemented with digital techniques. A great degree of synergism exists when individual systems are designed into one cohesive network utilizing digital implementations throughout. The synergistic effect benefits conventional voice services and newer services such as the Internet.

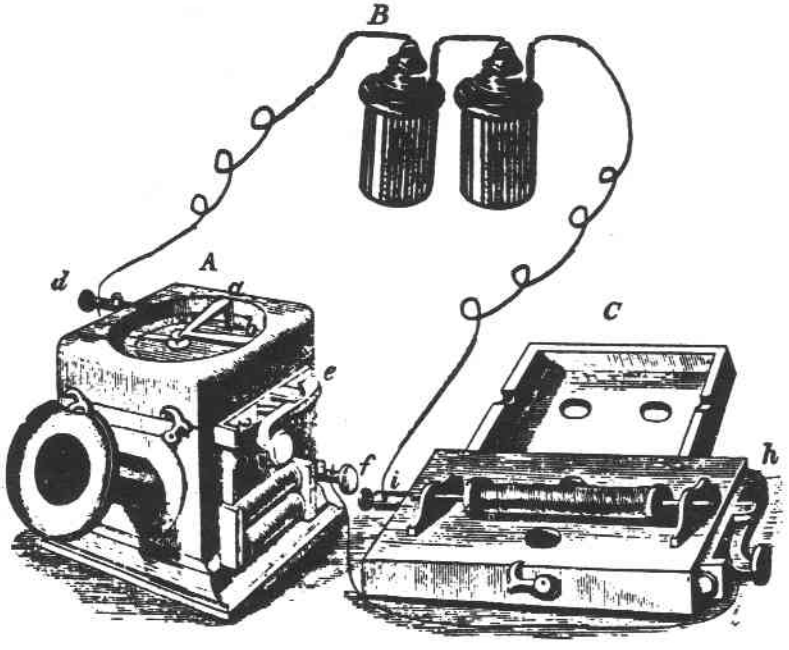


Figure 1.1 Back to the future: the first integrated voice/data communication system.

Most of the equipment descriptions and design examples presented in this book come from material authored by engineers at AT&T Laboratories (now Lucent Technologies) and other suppliers for the public telephone network. The basic principles, however, are by no means unique to the public telephone network. The concepts and implementation examples are applicable to any communications network: public or private, voice or data. An inherent attribute of a digital network is that it can, to a large extent, be designed independently of its application.

Terminals, Transmission, and Switching

The three basic elements of a communications network are terminals, transmission systems, and switches. The first part of this chapter provides an overview of these elements as implemented in analog telephone networks. Then, the last part of this chapter provides a brief overview of digital implementations *within* the analog network. Following a detailed discussion of the motivation for digital implementations in Chapter 2, the next four chapters describe the operation and design of the basic elements of a digital voice network. Chapter 3 discusses digital voice terminals and the most common algorithms used to convert analog voice signals into digital bit streams. Chapter 4 presents the basics of digital transmission systems. Fundamentals of digital switching follow in Chapter 5. Basic digital modulation techniques and their application to point-to-point digital microwave and digital cellular systems are described in Chapter 6. A discussion of various synchronization and control considerations for digital networks is provided in Chapter 7. Chapter 8 describes fiber optic transmission systems and the synchronous multiplexing standard (SONET). Chapter 9 discusses the basic architecture and operation of prevailing digital cellular systems in use in the United States and around the world.

The main emphasis of the first nine chapters involves circuit switching as traditionally implemented for voice telephone networks. A circuit-switched network is one that assigns a complete end-to-end connection in response to each request for service. Each connection, with its associated network facilities, is held for the duration of the call. Chapter 10 describes a different type of network, generically referred to as a packet-switched network, that is particularly suited to servicing data traffic. Included in Chapter 10 is a discussion of Asynchronous Transfer Mode (ATM), a form of a packet-switched network. Chapter 11 discusses various technologies and systems for achieving direct digital access to a digital network (voice or data). The last chapter presents the basics of traffic theory: the fundamental mathematics for analyzing and predicting telecommunications network performance.

1.1 TELECOMMUNICATIONS STANDARD ORGANIZATIONS

Prior to the breakup of the Bell System on January 1, 1984, telecommunications standards in North America were essentially established by the dominant equipment designer and supplier to the Bell System: Bell Telephone Laboratories and Western Electric. Independent telephone companies that provided local service to the 20% of the country not covered by the Bell System relied on the U.S. Independent Telephone

Association [USITA; later referred to as the U.S. Telephone Association (USTA)] to formulate and disseminate standards, particularly for interconnecting with the Bell System.

In anticipation of the divestiture of the Regional Bell Operating Companies (RBOCs) from AT&T, the Exchange Carriers Standards Association (ECSA) was formed in 1983 as a nonprofit trade association to represent the interests of all exchange carriers (RBOCs and independents). In February 1984 the ECSA sponsored the establishment of the T1 standards committee to formulate new interconnection standards for the U.S. national network. The T1 committee is accredited by the American National Standards Institute (ANSI) to ensure that standards approvals follow principles of openness. Thus T1 committee standards are designated as ANSI T1.nnn-date (T1 stands for Telecommunications standards entity number 1). Table 1.1 lists the major subcommittees within T1 and the respective responsibilities.

Other organizations in North America that establish standards related to telecommunications are the Electronic Industries Association (EIA), the Institute of Electrical and Electronic Engineers (IEEE), and Bell Communications Research (Bellcore). Bellcore was an organization chartered to establish standards and qualify equipment for the RBOCs. Bellcore has since been reorganized as Telcordia Technologies. The IEEE is most known for its data communications standards listed in Table 1.2 but has also established numerous standards for measuring and characterizing telecommunications equipment.

Most of the world outside of North America relies on international telecommunications standards committees established under the auspices of the International Telecommunication Union (ITU). In the past, two major entities within the ITU were established: the International Telegraph and Telephone Consultative Committee (CCITT) and the International Radio Consultative Committee (CCIR). CCITT established recommendations for telephone, telegraph, and data transmission circuits and equipment. CCIR was concerned with coordinating the use of the radio spectrum. CCITT and CCIR activities are no longer identified as being distinct from the ITU. CCITT has become ITU-T and CCIR is now ITU-R. In the United States, use of the radio spectrum is controlled by the Federal Communications Commission (FCC).

North American standards and ITU standards have often been incompatible in the past. North American standards established by the Bell System were therefore incor-

TABLE 1.1 T1 Standards Subcommittees

Committee	Responsibility
T1A1	Performance and signal processing
T1E1	Interfaces, power and protection for networks
T1M1	Internetwork operations, administration, maintenance, and provisioning (IOAM&P)
T1P1	Wireless/mobile services and systems
T1S1	Services, architectures, and signaling
T1X1	Digital hierarchy and synchronization

TABLE 1.2 IEEE Local Area Network/Metropolitan Area Network (LAN/MAN) Data Communications Standards

802.1	Overview and Architecture, Bridging, Virtual bridged LAN (VLAN)
802.2	Logical Link Control (LLC)
802.3	Carrier Sense Multiple Access (CSMA) with Collision Detection (CD) (Ethernet)
802.4	Token Bus (Arcnet)
802.5	Token Ring (IBM Ring)
802.6	Queued Packet Synchronous Exchange (QPSX)
802.7	Broadband
802.8	Optical Fiber Technologies
802.9	Integrated Services
802.10	Security
802.11	Wireless
802.12	Demand Priority
802.14	Cable TV

porated into CCITT recommendations as a subset. Due to the need for more international compatibility, the T1 subcommittees and ITU-T committees now work closely together to establish joint standards. A first major example of the joint effort is the standard for synchronous fiber transmission systems referred to as SONET in the United States and Synchronous Digital Hierarchy (SDH) in ITU-T standards.

The International Standards Organization (ISO) is an organization with standards activities in a wide range of subject matters, some of which involve telecommunications. Technical subcommittees within ISO work closely with ITU study groups in formulating ITU recommendations, particularly those related to ISDN protocols that adhere, as much as possible, to the ISO data communications standard for Open Systems Interconnection (OSI) Reference Model.

1.2 THE ANALOG NETWORK HIERARCHY

Because the analog telephone networks of the world evolved over a period of almost 100 years, a great amount of diversity in equipment implementations also developed. It is a remarkable achievement that vast networks, like the U.S. network, could accommodate the myriad of equipment types and function properly. In 1980* in the United States alone, there were 181 million telephones [2], almost all of which could directly dial any public telephone number and have a good quality connection. This achievement was made possible by standardized interfaces and well-defined functional hierarchies. As newer digital equipment was installed, it necessarily adhered to the

*Although there is no specific date at which digital technology took over from analog technology, 1980 is significant in that it represents the time frame in which fiber optics emerged to begin displacing analog radios for intercity transmission, the last stronghold of analog technology in the internal portions of the public network.

standardized practices of the analog network. The fact that the equipment was implemented with digital technology was transparent to the rest of the network.

1.2.1 Bell System Hierarchy

Alexander Graham Bell invented the first practical telephone in 1876. It soon became apparent, however, that the telephone was of little use without some means of changing connections on an "as-needed" basis. Thus the first switching office was established in New Haven, Connecticut, only two years later. This switching office, and others following, was located at a central point in a service area and provided switched connections for all subscribers in the area. Because of their locations in the service areas, these switching offices are often referred to as central offices.

As telephone usage grew and subscribers desired longer distance connections, it became necessary to interconnect the individual service areas with trunks between the central offices. Again, switches were needed to interconnect these offices, and a second level of switching evolved. Continued demand for even longer distance connections, along with improved long-distance transmission facilities, stimulated even more levels of switching. In this manner the analog public telephone network in the United States evolved to a total of five levels. These levels are listed in Table 1.3.

At the lowest level of the network are class 5 switching offices, also called central offices (COs) or end offices (EOs). The next level of the network was composed of class 4 toll offices. The toll network of the Bell System contained three more levels of switching: primary centers, sectional centers, and regional centers.

To illustrate the structure and motivation for hierarchical networks, a symbolic, three-level example is shown in Figure 1.2. In contrast, Figure 1.3 depicts a different network structure for interconnecting all of the first-level switches: a fully connected mesh structure. Obviously, the hierarchical network requires more switching nodes but achieves significant savings in the number of trunks: the transmission links between switching offices. Determination of the total number of trunk circuits in either network is necessarily a function of the amount of traffic between each pair of switching nodes. (Chapter 12 provides the mathematics for determining the number of trunk circuits.) As a first approximation, the trunk costs of a mesh can be determined as the total number of connections (trunk groups) N_c between switching offices:

TABLE 1.3 Public Network Hierarchy of the Bell System (1982) [3]

Switch Class	Functional Designation	No. in Bell System	No. in Independents	Total
1	Regional center	10	0	10
2	Sectional center	52	0	67
3	Primary center	148	20	168
4	Toll center	508	425	933
5	End office	9803	9000	18,803

$$N_c = \frac{1}{2} N(N-1) \tag{1.1}$$

where N is the number of nodes.

Thus the mesh network of Figure 1.3 has 36 connections, as compared to 12 connections in Figure 1.2. In the case of fiber optic transmission the cost comparison of

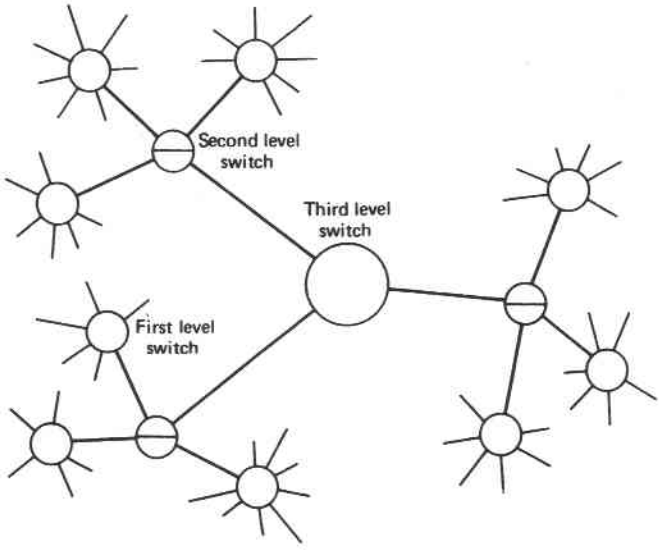


Figure 1.2 Three-level switching hierarchy.

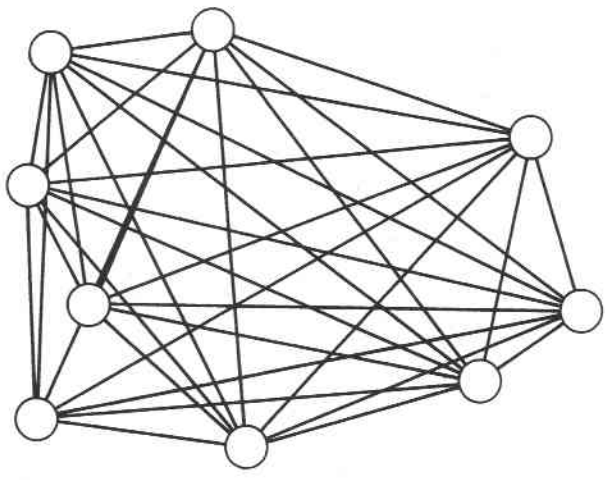


Figure 1.3 Mesh-connected network.

the trunks is almost exactly 3 : 1 because a single fiber system can provide more voice capacity than is needed between any two switches.

A less obvious difference between the networks of Figures 1.2 and 1.3 involves the method of establishing connections between two offices. In the hierarchical network there is one and only one route between any two switching nodes. In the mesh network most connections would be established on the direct route between the two offices. However, if the direct route is unavailable (because of a traffic overload or an equipment failure) and the first-level switches can provide trunk-to-trunk connections (called tandem switching functions), the mesh network provides many alternatives for establishing connections between any two nodes. Hence the reliability of a network architecture must be considered in addition to just the costs. In general, neither a pure mesh nor a purely hierarchical network is desirable.

Taking these factors into account, Figure 1.4 depicts alternate routing as implemented in the former Bell System. As indicated, the basic backbone hierarchical network was augmented with high-usage trunks. High-usage trunks are used for direct connections between switching offices with high volumes of interoffice traffic. Normally, traffic between two such offices is routed through the direct trunks. If the direct trunks are busy (which may happen frequently if they are highly utilized), the backbone hierarchical network is still available for alternate routing.

Traffic was always routed through the lowest available level of the network. This procedure not only used fewer network facilities but also implied better circuit quality because of shorter paths and fewer switching points. Figure 1.4 shows the basic order of selection for alternate routes. The direct interoffice trunks are depicted as dashed lines, while the backbone, hierarchical network is shown with solid lines.

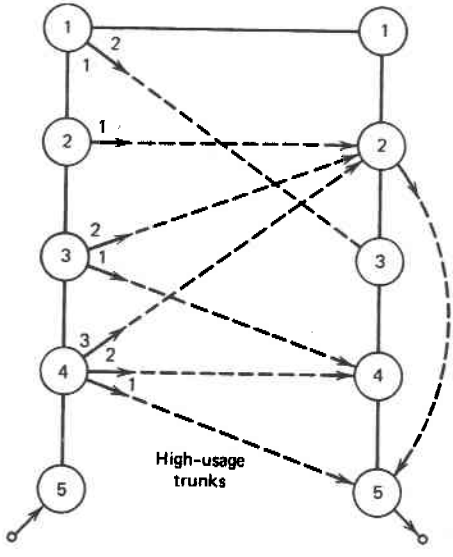


Figure 1.4 Alternate routing in North American network.

In addition to the high-usage trunks, the backbone network was also augmented with additional switching facilities called tandem switches. These switches were employed at the lowest levels of the network and provided switching between end offices. Tandem switches were not part of the toll network, as indicated in Figure 1.5, but were (and are) part of what is referred to as an exchange area. Generally speaking, an exchange area is an area within which all calls are considered to be local calls (i.e., toll free).

In general terms, any switching machine in a path between two end offices provides a tandem switching function. Thus toll switches also provide tandem switching functions. Within the public telephone network, however, the term *tandem* refers specifically to intermediate switching within the exchange area.

The basic function of a tandem office is to interconnect those central offices within an exchange area having insufficient interoffice traffic volumes to justify direct trunks. Tandem offices also provide alternate routes for exchange area calls that get blocked on direct routes between end offices. Although Figure 1.5 depicts tandem offices as being physically distinct from end offices and toll offices, tandem switches were often colocated with either or both types. Operationally, exchange area switching and toll network switching in the Bell system were always separated. The primary reason for the separation was to simplify tandem switching by avoiding billing and network routing. A toll switch had to measure call duration for billing purposes but a tandem switch did not. Prior to the introduction of computer-controlled switching, billing functions were a significant consideration. The operational separation also implied that toll-connecting trunk groups were separate from tandem trunk groups. The flexibility of computer-controlled switching has eliminated the need for the separation.

The separation of exchange facilities from toll facilities had an important effect on the transmission and switching equipment utilized in the respective applications. Exchange area connections were usually short and only involved a few switching offices. Toll connections, on the other hand, could involve numerous switching offices with

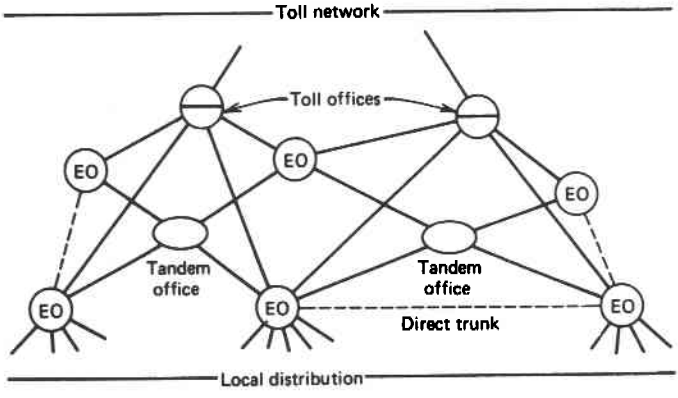


Figure 1.5 Exchange area network.

relatively long transmission links between them. Thus, for comparable end-to-end quality, individual analog exchange area equipment did not have to provide as much quality as did toll network counterparts.

1.2.2 Postdivestiture U.S. Network

In the decade of the 1980s the structure of the public telephone network in the United States changed significantly as a result of changes in the technology and the regulatory environment. The main technological changes were (1) extensive deployment of very large digital switching machines, (2) the adaptation of computer-controlled switches to provide multiple switching functions in one machine (e.g., the integration of end-office, tandem, and toll switching functions), and (3) the deployment of fiber optic transmission systems that could carry very large cross sections of traffic. Notice that all three of these technological developments suggest a network with fewer and larger switching offices. These technological influences on the network topology are discussed more fully in Chapters 8–10.

The most dramatic and immediate effect on the network occurred on January 1, 1984, when the breakup of AT&T officially took effect. Because the breakup involved divestiture of Bell Operating Companies (BOCs) from AT&T, the network itself became partitioned at a new level. The new partitioning is shown in Figure 1.6, which depicts AT&T as one of several competing long-distance carriers referred to as interexchange carriers (IXCs) and local access and transport areas (LATAs), which were originally the exclusive domain of local exchange carriers (LECs). In addition to AT&T, the other two main IXCs are MCI and U.S. Sprint. The LECs originally included 23 BOCs (organized into 7 RBOCs), former independent telephone companies like GTE, Contel, and United Telecommunications, and some 1500 mostly small-town telephone companies. Mergers within the industry have subsequently reduced the number of LECs and RBOCs.

The number of LATAs in the United States was initially 164, but the number has changed as adjustments in service boundaries are sometimes made. Because a LATA entails an area that includes many exchange areas, LECs complete toll calls that traverse different exchange areas within one LATA. The IXCs were not allowed to carry intra-LATA traffic. Similarly, an LEC was not allowed to carry traffic between two LATAs even when both LATAs might be service areas of a single BOC. Only an IXC was allowed to carry inter-LATA traffic. To ensure that these service partitions were adhered to, each IXC interfaced with a LATA at a single point in the LATA, referred to as a point of presence (POP). IXC equipment at a POP could be a switching office or merely a junction for collecting traffic carried elsewhere to be switched.

A major aspect of the modified final judgment (MFJ) that specified the divestiture was the condition of equal access, which meant that an LEC (specifically a BOC) was to treat all IXCs equally in regards to exchange access. The conditions of equal access meant that access to all end offices in a LATA would be equal in type, quality, and price for all IXCs. The LATA network structure [4] established to accomplish equal access is shown in Figure 1.7.

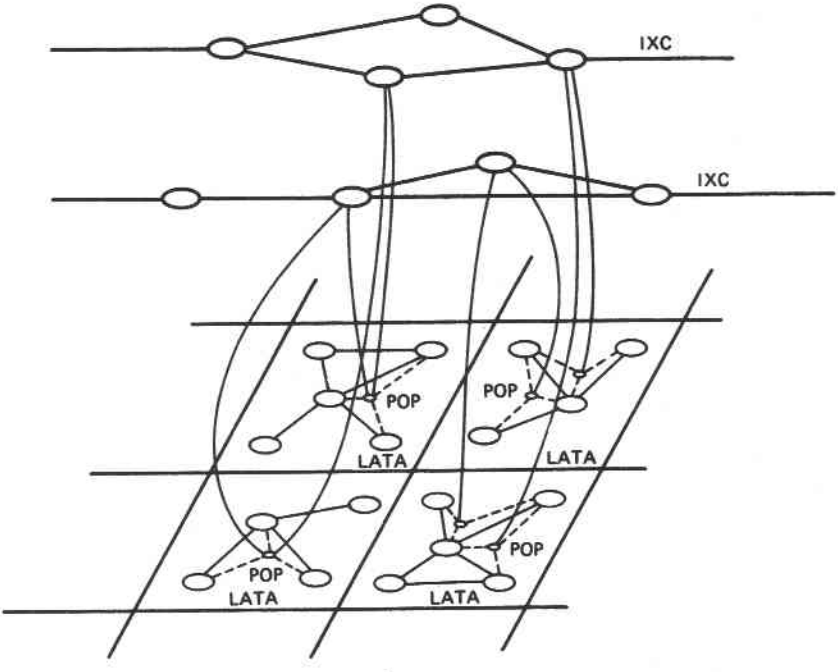


Figure 1.6 U.S. network partitioning.

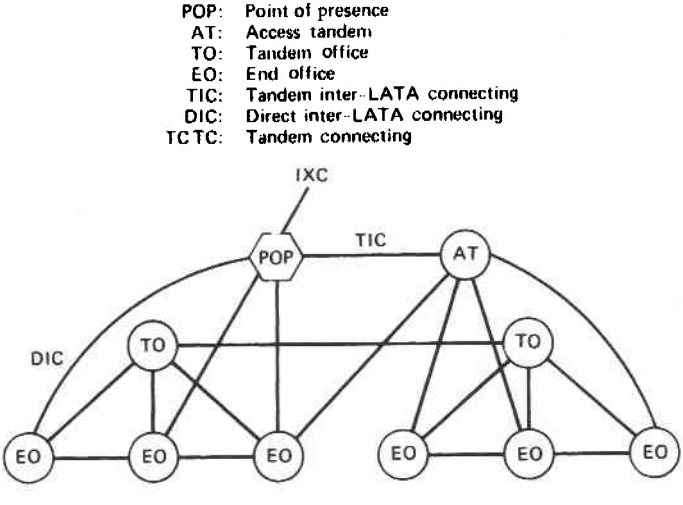


Figure 1.7 LATA hierarchy and access architecture.

The design of the LATA network for intra-LATA traffic was left to the discretion of the LECs. Thus intra-LATA connections can involve multiple switching offices between end offices. However, connections between an EO and a POP could involve at most one intermediate switching office referred to as an access tandem (AT). With respect to the previous Bell System hierarchy, an AT takes the place of a class 4 toll switch. However, long-distance billing functions, which were formerly performed in class 4 switches, are now performed within the IXC network. Although Figure 1.7 shows access tandem and basic tandem switching functions as being distinct, access tandem functions can be integrated into regular tandem switches if the tandem switch provides AT features. Foremost among these features are the ability to forward automatic number identification (ANI) information to an IXC for billing and the ability to route calls to different IXC POPs depending on presubscription or per-call three-digit carrier designations.

In 1997 the FCC issued some rulings with the intent of stimulating competition in both the local exchange and long-distance networks. Under this ruling, LECs that want to enter the long-distance market can do so if they open their local exchange facilities to long-distance carriers or other competitive access providers. A key aspect of making the local facilities available to competition is the establishment of unbundled pricing for local services: the separation of the cost of the local loop, the local switching equipment, maintenance, and ancillary services such as 911 emergency calling. Another key requirement is number portability, which allows a subscriber to change local service providers without having to change telephone numbers. The introduction of competition for local distribution instigated the use of two terms: competitive local exchange carrier (CLEC) for the competition and incumbent local exchange carrier (ILEC) for the established carrier.

1.2.3 Switching Systems

Manual Switchboards

The first telephone switching equipment utilized operators at manual switchboards. The operators asked a caller for the number they wanted to call and then established the connection by plugging in a cord between terminal jacks. Although switchboards are no longer used, a legacy of their existence lives on: the use of the terms "tip and ring." As shown in Figure 1.8, one wire of a wire pair was connected to the tip of a plug connector and the other wire was connected to the ring. Ever since, one wire of a wire pair is com-

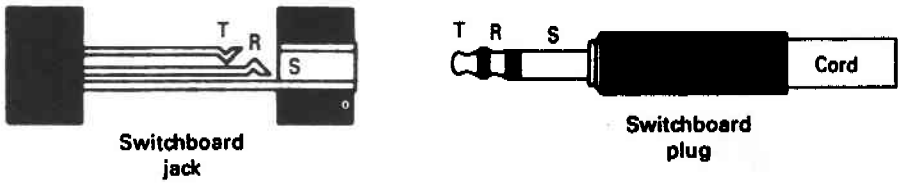


Figure 1.8 Switchboard plug with corresponding jack (R, S, and T are ring, sleeve, and tip, respectively). (From Freeman, *Fundamentals of Telecommunications*, Wiley, New York.)

monly referred to as the tip and the other is referred to as the ring, even on digital wire pairs, which have never used plugs in a switchboard. On some of the original switchboards a third connection would be provided by the sleeve conductor shown in Figure 1.8.

Automated Switching

In general terms the equipment associated with any particular switching machine can be categorized as providing one of the following functions:

1. Signaling
2. Control
3. Switching

The basic function of the signaling equipment is to monitor the activity of the incoming lines and forward appropriate status or control information to the control element of the switch. Signaling equipment is also used to place control signals onto outgoing lines under direction of the switch control element.

The control element processes incoming signaling information and sets up connections accordingly. The switching function itself is provided by a switching matrix: an array of selectable crosspoints used to complete connections between input lines and output lines. These basic constituents of a switching machine are shown in Figure 1.9.

Electromechanical Switching. Prior to the introduction of digital electronic switching machines in the late 1970s, switching offices in North America and around the world were equipped with one of two basic types of electromechanical switches: step-by-step* and crossbar. As shown in Figure 1.10, crosspoints of a step-by-step switch are wiper contacts that move in direct response to dial pulses. As the pulses of the first digit enter the switch, they immediately "step" the vertical wiper to a horizontal row corresponding to the first digit. After the proper row is selected, the wiper is rotated across another set of contacts until an idle line to the next stage of switching is located. The next set of dial pulses, representing the second digit, then steps the second stage in like manner. The process continues through however many stages are needed for a particular switch size.

As the name implies, a step-by-step switch uses *direct progressive control*: Successive segments of a path through the switch are established as each digit is dialed. With progressive control, the control elements of the switch are integrated into the switching matrix. This feature is very useful for implementing a variety of switch sizes and allowing relatively easy expansion. A progressive control switch, however, has a number of significant limitations:

1. A call may be blocked even though an appropriate path through the switch exists but is not attempted because an unfortunate path gets selected in an early stage.
2. Alternate routing for outgoing trunks is not possible. That is, the outgoing line is directly selected by incoming dial pulses and cannot be substituted.

*A step-by-step switch is also referred to as a *Strowger switch*, in honor of its inventor Almon B. Strowger.

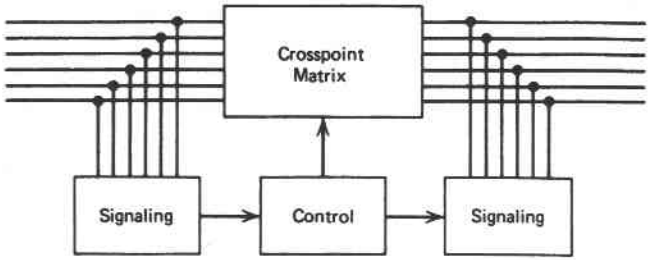


Figure 1.9 Switching system components.

3. Signaling schemes other than dial pulses (e.g., tone signaling) are not directly usable.
4. Number translation is impossible.

In contrast to a step-by-step switch, a crossbar switch is one that used centralized, common control for switch path selection. As digits were dialed, the control element of the switch received the entire address before processing it. When an appropriate path through the switch was determined (which may have involved number translation or alternate routing), the control element transferred the necessary information in the form of control signals to the switching matrix to establish the connection. The fundamental feature, and advantage, of a common control switch is that control function implementation is separate from the switch implementation. Common control cross-

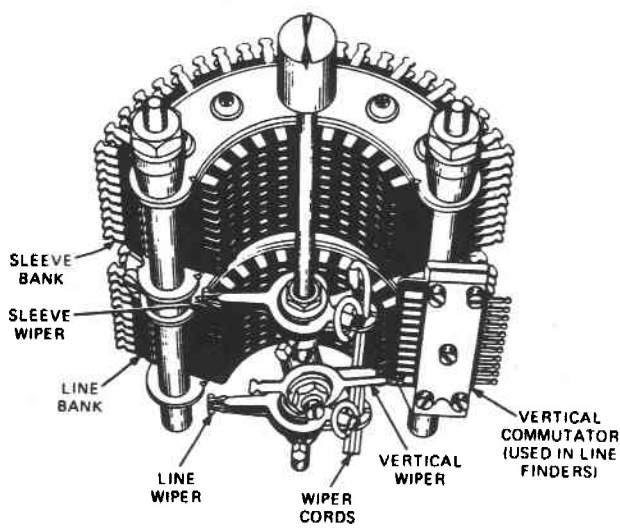


Figure 1.10 Step-by-step switching element. (Copyright 1977 by Bell Telephone Laboratories. Reprinted by permission.)

bar systems introduced the ability to assign logical addresses (telephone numbers) independently of physical line numbers.

The crosspoints of a crossbar switch (Figure 1.11) are mechanical contacts with magnets to set up and hold a connection. The term *crossbar* arises from the use of crossing horizontal and vertical bars to initially select the contacts. Once established, the switching contacts are held by electromagnets energized with direct current passing through the established circuit. When the circuit is opened, the loss of current causes the crosspoints to be released automatically.

Because of the operational limitations of progressive control, step-by-step switches were used primarily in smaller class 5 switching offices. Crossbar switches, on the other hand, were used predominantly in metropolitan areas and within the toll network. In some cases step-by-step switches were augmented with common control by receiving the digits into special control equipment. After processing the request, the control equipment generated pulses that set up a connection as if the switch was receiving dial pulses directly.

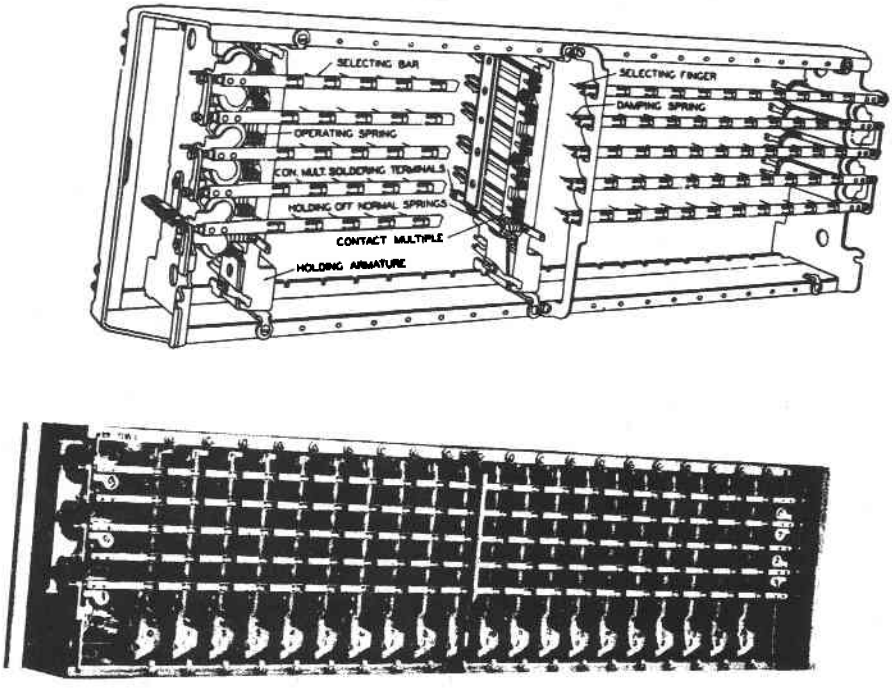


Figure 1.11 Crossbar switching element. (Copyright 1977 by Bell Telephone Laboratories. Reprinted by permission.)

Stored Program Control. Step-by-step and crossbar switching systems used electromechanical components for both the switching matrix and the control elements. In some cases the electromechanical control elements in these switches represented rudimentary forms of special-purpose digital computers. The hardwired electromechanical logic, however, had limited capabilities and was virtually impossible to modify.

A major milestone for telephony was established in 1965 when the Bell System installed its first computer-controlled switching system: the No. 1 Electronic Switching System (ESS).* This switching system uses a stored-program digital computer for its control functions. The stored-program control (SPC) feature of the No. 1 ESS allowed the introduction of new features such as abbreviated dialing, call forwarding, call waiting, and three-way calling.

The introduction of SPC not only provided significant advantages to end users but also simplified many administrative and maintenance tasks for the operating companies. A large part of line administration that formerly required many manual modifications (mainframe cross-connects) could instead be accomplished with changes in computer data tables of an SPC switch. Furthermore, physical line numbers were independent of the logical (directory) line numbers, thus making number changes easy. Other benefits enabled by SPC are automated record keeping, lower blocking probabilities, generation of traffic statistics, automated call tracing, and message unit accounting (per-call charges as opposed to flat-rate billing for unlimited local calling).

The switching matrix of the No. 1 ESS (and also the No. 2 ESS, No. 3 ESS, and No. 1A ESS) is implemented with electromechanical reed relays. Thus the term *ESS* refers in general to computer-controlled switching and not to the nature of the switching matrix itself. However, AT&T's No. 4 ESS, which was first installed in 1976, is a high-capacity toll switch using computer control and digital electronics for its switching matrix. Thus the No. 4 ESS is "electronic" in its control and its switching matrix. Furthermore, the digital multiplex system (DMS) switches of Northern Telecom, the No. 5 electronic automatic exchange (EAX) of GTE, and the No. 5 ESS of AT&T also utilize digital logic circuits for the crosspoint matrix.

Private Branch Exchanges. In the United States the term *private branch exchange* (PBX) refers generically to any switching system owned or leased by a business or organization to provide both internal switching functions and access to the public network. Thus a PBX in the United States may use either manual or automatic control. The term *PABX* is also used in the United States, and particularly in other countries, to refer specifically to automatically controlled PBXs.

The historical development of PBX systems has followed closely that of switches in the public network. PBXs with computerized control became available in 1963 (before the No. 1 ESS) when AT&T's No. 101 ESS was first installed. Since that time a

*Computer-controlled PBXs were available before 1965. The No. 1 ESS represents the first instance of computer control in the public network hierarchy.

WATS ?

large number of independent manufacturers have developed computer-controlled PBXs. In fact, the PBX market has long been one of the most competitive businesses in telecommunications.

The use of computer control for PBXs introduced numerous new features for users. Not only were customized calling features (e.g., abbreviated dialing) provided, but numerous facilities for cost management also became available. Some of the more useful, commonplace features in a PBX are the following:

1. Accounting summaries by individual employee or department
2. Multiple classes of service with priorities and access restrictions to area codes, WATS lines, and so on
3. Least-cost routing to automatically select tie lines, foreign exchange circuits, WATS, DDD, and so forth
4. Automatic callback when circuits are available
5. Traffic monitoring and analysis to determine the utilization of existing circuits or to ascertain blocking probabilities and network cost effectiveness

Centrex. Many of the features initially provided by PBXs are also offered by operating companies as Centrex features. As indicated in Figure 1.12, Centrex is a business customer service offering supported by switching equipment in the central office. Every telephone or data device at the customer premises has a dedicated channel to the switch in the central office. Originally, each channel implied a dedicated pair of wires. It is now more common to use multiplexing techniques (described in Section 1.2.5) to reduce the transmission costs. Nevertheless, from the central office point of view each Centrex extension has a dedicated appearance at the CO with a unique public network telephone number. A software partition in the central office treats the Centrex lines as a closed user group to provide the following basic features:

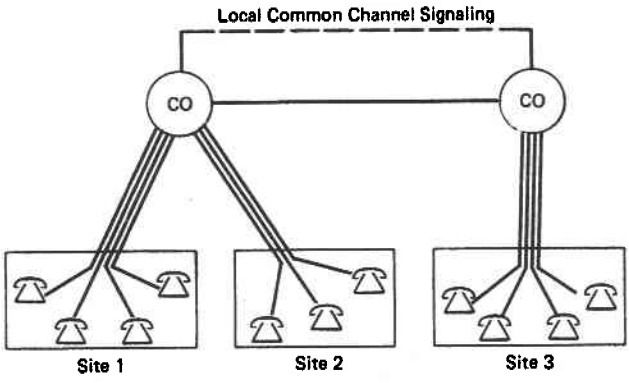


Figure 1.12 Centrex service to multiple sites.

1. Direct dialing to Centrex extensions from the public network.
2. Station-to-station calling using extension numbers as opposed to full, 7- (or 10-) digit public numbers.
3. Common voice user features such as call forwarding, call transfer, call waiting, call pick up, and three-way calling.
4. Multiple sites with transparent numbering plans and features. With citywide Centrex the sites can be supported from multiple offices interconnected by common-channel signaling (CCS) described in Section 1.2.10.
5. Centralized attendant/message desk with call origination information for informed processing of forwarded calls.
6. High availability because CO equipment and directly connected analog phones are powered at the CO with backup power sources.
7. Virtually unlimited growth.

1.2.4 Transmission Systems

Functionally, the communications channels between switching systems are referred to as trunks. In the past, these channels were implemented with a variety of facilities, including pairs of wires, coaxial cable, and point-to-point microwave radio links. Except for special situations, trunk facilities now utilize optical fibers.

Open Wire

A classical picture of the telephone network in the past consisted of telephone poles with crossarms and glass insulators used to support uninsulated open-wire pairs. Except in rural environments, the open wire has been replaced with multipair cable systems or fiber. The main advantage of an open-wire pair is its relatively low attenuation (a few hundredths of a decibel per mile at voice frequencies). Hence, open wire is particularly useful for long, rural customer loops. The main disadvantages are having to separate the wires with crossarms to prevent shorting and the need for large amounts of copper. (A single open-wire strand has a diameter that is five times the diameter of a typical strand in a multipair cable. Thus open wire uses roughly 25 times as much copper as does cable.) As a result of copper costs and the emergence of low electronics costs, open wire in rural environments has been mostly replaced with cable systems using (digital) amplifiers to offset attenuation on long loops.

Paired Cable

In response to overcrowded crossarms and high maintenance costs, multipair cable systems were introduced as far back as 1883. Today a single cable may contain anywhere from 6 to 2700 wire pairs. Figure 1.13 shows the structure of a typical cable. When telephone poles are used, a single cable can provide all the circuits required on the route, thereby eliminating the need for crossarms. More recently the preferred means of cable distribution is to bury it directly in the ground (buried cable) or use underground conduit (underground cable).

Table 1.4 lists the most common wire sizes to be found within paired-cable systems. The lower gauge (higher diameter) systems are used for longer distances where signal attenuation and direct-current (dc) resistance can become limiting factors. Figure 1.14 shows attenuation curves [5] for the common gauges of paired cable as a function of frequency. An important point to notice in Figure 1.14 is that the cable pairs are capable of carrying much higher frequencies than required by a telephone quality voice signal (approximately 3.4 kHz).

In the past, the exchange areas of the telephone network used paired cable almost exclusively for short-haul interoffice transmission. Up until the introduction of mul-

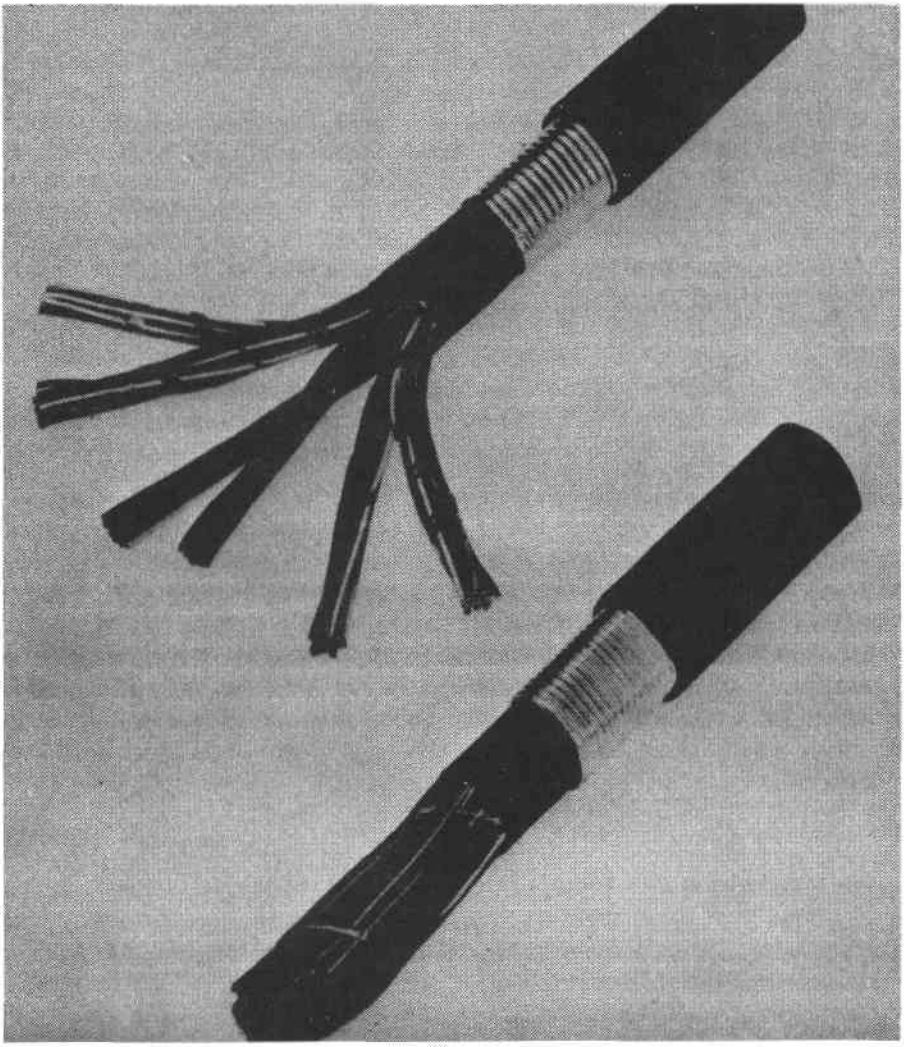


Figure 1.13 Multipair cable. (Copyright 1977 by Bell Telephone Laboratories. Reprinted by permission.)

TABLE 1.4 Wire Gauge and Resistance of Common Paired Cable

Gauge	Diameter (in.)	Direct-Current Resistance (Ω/1000 ft)[a]
30	0.010	104
28	0.013	66
26	0.016	41
24	0.020	26
22	0.025	16
20	0.032	10
19	0.036	8

[a]Note that the loop resistance of a pair is twice the resistance of a single wire given in the table.

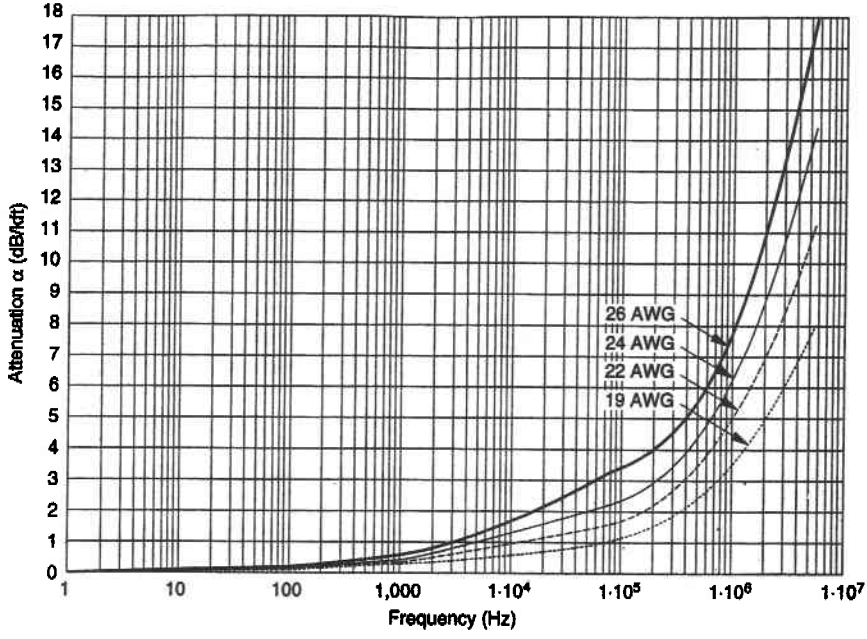


Figure 1.14 Attenuation versus frequency of common gauges of paired cable. (From W. D. Reeve, *Subscriber Loop Signaling Transmission Handbook*, IEEE Press, New York, Fig. 7-16a.)

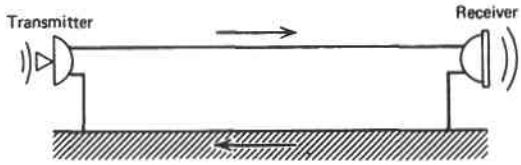


Figure 1.15 Single-wire transmission with ground return.

Figure 1.16 Two-wire transmission.

tiplexing techniques, described later in this chapter, each voice circuit (trunk) was carried on a separate pair of wires. Beginning in the early 1960s electronics began to be used for short-haul interoffice transmission systems using multiplexing techniques to carry multiple channels on a single pair of wires.

Two-Wire Versus Four-Wire

All wire-line transmission in the telephone network is based on transmission through pairs of wires. As shown in Figure 1.15, transmission through a single wire (with a ground return) is possible and has been used in the past. However, the resulting circuit is too noisy for customer acceptance. Instead, balanced pairs of wires as shown in Figure 1.16 are used with signals propagating as a voltage difference between the two wires. The electrical current produced by the difference signal flowing through the wires in opposite directions is called a "metallic current." In contrast, current propagating in the same direction in both wires is referred to as common-mode or longitudinal current. Longitudinal currents are not coupled into a circuit output unless there is an imbalance in the wires that converts some of the longitudinal signal (noise or interference) into a difference signal. Thus the use of a pair of wires for each circuit provides much better circuit quality than does single-wire transmission. Some older switching systems used single-wire (unbalanced) transmission to minimize the number of contacts. Unbalanced circuits were only feasible in small switches where noise and crosstalk could be controlled.

Virtually all subscriber loops in the telephone network are implemented with a single pair of wires.* The single pair provides for both directions of transmission. If users on both ends of a connection talk simultaneously, their conversations are superimposed on the wire pair and can be heard at the opposite ends. In contrast, wire-line (and fiber) transmission over longer distances, as between switching offices, is best implemented if the two directions of transmission are separated onto separate wire pairs.

*It is now commonplace to use fiber for the feeder portion of a subscriber loop, but the drop to a re sidence is a single pair per telephone.

Longer distance transmission requires amplification and most often involves multiplexing. These operations are implemented most easily if the two directions of transmission are isolated from each other. Thus interoffice trunks typically use two pairs of wires or two fibers and are referred to as four-wire systems. The use of two pairs of wires did not necessarily imply the use of twice as much copper as a two-wire circuit. After 1960, four-wire systems increasingly used some form of multiplexing to provide multiple channels in one direction on one wire pair. Thus, a net savings in copper could result.

Sometimes the bandwidth of a single pair of wires was separated into two subbands that were used for the two directions of travel. These systems were referred to as *derived four-wire systems*. Hence, the term *four-wire* has evolved to imply separate channels for each direction of transmission, even when wires may not be involved. For example, fiber optic and radio systems that use separate channels for each direction are also referred to as *four-wire systems*.

The use of four-wire transmission had a direct impact on the switching systems of the toll network. Since toll network circuits were four-wire, the switches were designed to separately connect both directions of transmission. Hence, two paths through the switch were needed for each connection. A two-wire switch, as used in older analog end offices, required only one path through the switch for each connection.

Two-Wire-to-Four-Wire Conversion

At some point in a long-distance connection it is necessary to convert from two-wire transmission of local loops to four-wire transmission on long-distance trunks. In the past, the conversion usually occurred at the trunk interface of the (two-wire) end office switch. Newer digital end office switches are inherently "four-wire," which means the two-wire-to-four-wire conversion point is on the subscriber (line) side of the switch as opposed to the trunk side. A generalized interconnection of two-wire and four-wire facilities for a connection is shown in Figure 1.17. The basic conversion function is provided by hybrid circuits that couple the two directions of transmission as shown. Hybrid circuits have been traditionally implemented with specially interconnected transformers. More recently, however, electronic hybrids have been developed. Ideally a hybrid should couple all energy on the incoming branch of the four-wire circuit into the two-wire circuit, and none of the incoming four-wire signal should be transferred to the outgoing four-wire branch.

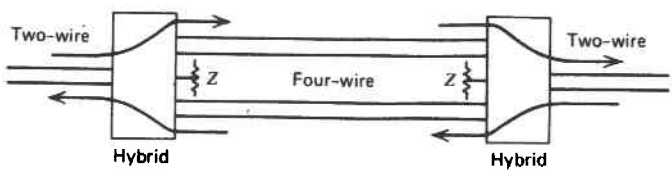


Figure 1.17 Interconnection of two-wire and four-wire circuits.

When the impedance matching network Z exactly matches the impedance of the two-wire circuit, near-perfect isolation of the two four-wire branches can be realized. Impedance matching used to be a time-consuming, manual process and was therefore not commonly used. Furthermore, the two-wire circuits were usually switched connections so the impedance that had to be matched would change with each connection. For these reasons the impedances of two-wire lines connected to hybrids were rarely matched. The effect of an impedance mismatch is to cause an echo, the power level of which is related to the degree of mismatch. The effect of echoes on voice quality and the means of controlling them is discussed later in this chapter.

Loading Coils

The attenuation curves shown in Figure 1.14 indicate that the higher frequencies of the voice spectrum (up to 3.4 kHz) experience more attenuation than the lower frequencies. This frequency-dependent attenuation distorts the voice signal and is referred to as amplitude distortion. Amplitude distortion becomes most significant on long cable pairs, where the attenuation difference is greatest.

The usual method of combating amplitude distortion on intermediate-length (3–15-mile) wire pairs is to insert artificial inductance into the lines. The extra inductance comes from loading coils that are inserted at 3000-, 4500-, or 6000-ft intervals. Figure 1.18 shows the effect of loading coils on a 24-gauge loop. Notice that the voiceband response up to 3 kHz is greatly improved, but the effect on higher frequencies is devastating.

Prior to the introduction of wire-line and fiber carrier systems, loading coils were used extensively on exchange area interoffice trunks. Loading coils are also used on the longer, typically rural, subscriber loops. Here, too, carrier systems have displaced most of the single pairs of wires being used on long routes.

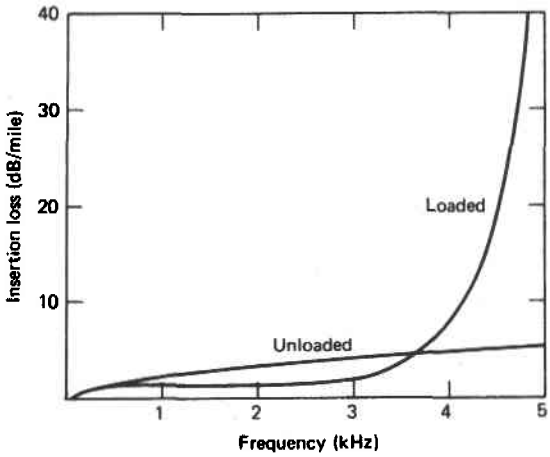


Figure 1.18 Effect of loading on 24-gauge cable pair.

1.2.5 Pair-Gain Systems

Providing service to rural subscribers has always been an expensive proposition because of the length of the routes involved and the small number of households to support the cost of both the initial installation and the maintenance. In the past, a common means of reducing the costs was to use party lines, which involved sharing of a wire pair among multiple households. A party line is particularly useful for satisfying new service requests on routes with no spare pairs but is obviously objectionable due to the lack of privacy and the lack of availability of the line.

A pair-gain system is an alternate approach to sharing pairs of wires that is much more acceptable to the users. This section describes two basic types of pair-gain systems: concentrators (remote switches) and multiplexers (carrier systems).

Concentration

The first form of a pair-gain system in Figure 1.19 depicts a basic line concentration system. When viewed from the station set end of the system, a pair-gain system provides concentration by switching some number of active stations to a smaller number of shared output lines. At the other end of the system, deconcentration (expansion) occurs by switching from the shared lines to individual inputs of the switching office corresponding to the active stations. Expanding the traffic back to the original number of stations ensures that the system is operationally transparent to both the switch and the user. Notice that a definition of which end provides concentration and which end provides expansion is dependent on the point of view.

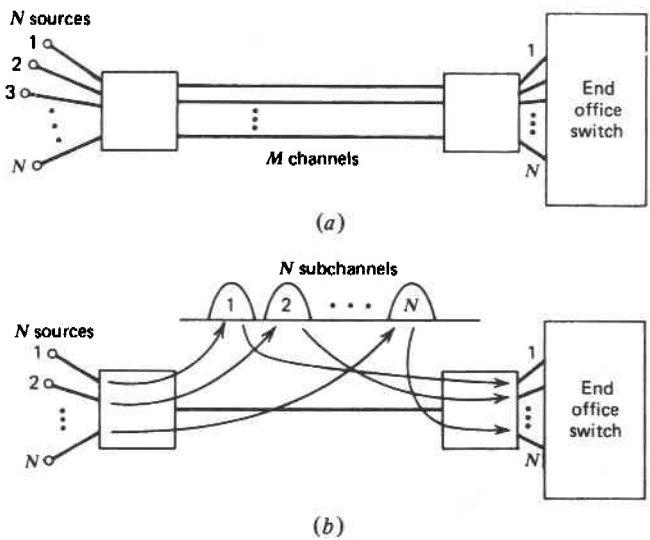


Figure 1.19. Pair-gain systems: concentration and multiplexing: (*a*) concentration ($n > M$); (*b*) multiplexing.

Since a concentrator is incapable of simultaneously connecting all stations it services, a certain amount of blocking is necessarily introduced by concentration. When the activity of individual stations is low enough, significant amounts of concentration can be achieved with acceptable blocking probabilities. For example, 40 stations that are each active only 7.5% of the time can be concentrated onto 10 lines with a blocking probability of 0.001.* This is an acceptable degradation in service since an equally active called station is busy 75 times as often.

Notice that a concentration system requires the transfer of control information between the concentrator/expander switch terminals. When one end of the system establishes a new connection to one of the shared lines, the other end must be informed to set up the appropriate reverse connection.

Multiplexing

As shown in Figure 1.14, the inherent bandwidth of a typical wire pair is considerably greater than that needed for a single voice signal. Thus, multiplexing can be used to carry multiple voice channels on a single pair of wires. The increase in attenuation implied by the higher frequencies is offset by amplifiers in the multiplex equipment and at periodic points in the transmission lines. The particular multiplexing technique shown in Figure 1.19*b* is a frequency division multiplex system. Another form of multiplexing, time division multiplexing of digital voice signals, is the preferred multiplexing approach for digital pair-gain systems discussed later.

As shown in Figure 1.19*b*, there is a one-to-one relationship between the customer lines and the subchannels of the multiplexer. Thus, unlike the concentration system, there is no possibility of blocking in a multiplexing type of pair-gain system. Also, there is no need to transfer switching information since the same one-to-one relationship defines the correspondence between customer lines at one end and switching office lines at the other end. A major drawback of multiplexing pair-gain systems is that the subchannels are highly underutilized if the sources are relatively inactive. In these situations a combination of concentration and multiplexing is normally justified.

Time Assignment Speech Interpolation

Time assignment speech interpolation (TASI) is a pair-gain system that dynamically assigns a channel to a circuit only when there is actual voice activity. Thus, a TASI system senses voice activity from a number of sources N; assigns active sources to one of M channels, where M is typically about half as large as N; and signals the far end about the connections. Normally, each participant in a conversation is active for only 40% of the time, which indicates that if $M = \frac{1}{2}N$, there is some amount of spare capacity to accommodate excess activity in one direction. If a source begins to talk when all channels are utilized, the beginning of that speech segment gets clipped until a channel becomes available. Chapter 12 provides mathematical formulations to determine the probability of clipping as a function of N, M, and the voice activity factor.

Initial TASI applications involved improving the utilization of undersea cable pairs. The obvious expense of these channels warranted the use of a rather complicated

*A discussion of traffic analysis is provided in Chapter 12, from which this result can be obtained.

multiplexing technique for the time. The same basic technique has since been used in numerous applications with digital speech for satellite and land-line applications. These systems are generally called digital speech interpolation (DSI) systems [6].

1.2.6 FDM Multiplexing and Modulation

The introduction of cable systems into the transmission plant to increase the circuit packing density of open wire is one instance of multiplexing in the telephone network. This form of multiplexing, referred to as *space division multiplexing*, involves nothing more than bundling more than one pair of wires into a single cable. The telephone network uses two other forms of multiplexing, both of which use electronics to pack more than one voice circuit into the bandwidth of a single transmission medium. Analog frequency division multiplexing (FDM) has been used extensively in point-to-point microwave radios and to a much lesser degree on some obsolete coaxial cable and wire-line systems. FDM is also utilized in fiber optic transmission systems, where it is referred to as wavelength division multiplexing (WDM).* Digital time division multiplexing (TDM) is the dominant form of multiplexing used in the telephone networks worldwide. (Even the fiber optic systems that utilize WDM commonly use digital TDM within the signals of a particular wavelength.)

Frequency Division Multiplexing

As indicated in Figure 1.19*b*, an FDM system divides the available bandwidth of the transmission medium into a number of narrower bands or subchannels. Individual voice signals are inserted into the subchannels by amplitude modulating appropriately selected carrier frequencies. As a compromise between realizing the largest number of voice channels in a multiplex system and maintaining acceptable voice fidelity, the telephone companies established 4 kHz as the standard bandwidth of a voice circuit.† If both sidebands produced by amplitude modulation are used (as in obsolete N1 or N2 carrier systems on paired cable), the subchannel bandwidth is 8 kHz, and the corresponding carrier frequencies lie in the middle of each subchannel. Since double-sideband modulation is wasteful of bandwidth, single-sideband (SSB) modulation was used whenever the extra terminal costs were justified. The carrier frequencies for single-sideband systems lie at either the upper or lower edge of the corresponding subchannel, depending on whether the lower or upper sideband is selected. The A5 channel bank multiplexer of AT&T used lower sideband modulation.

FDM Hierarchy

In order to standardize the equipment in the various broadband transmission systems of the original analog network, the Bell System established an FDM hierarchy as provided in Table 1.5. CCITT recommendations specify the same hierarchy at the lower

*Optical technology is customarily defined in terms of the wavelength of the optical signal as opposed to the corresponding frequency.

†Actually, the usable bandwidth of an FDM voice channel was closer to 3 kHz due to guard bands needed by the FDM separation filters.

TABLE 1.5 FDM Hierarchy of the Bell Network

Multiplex Level	Number of Voice Circuits	Formation	Frequency Band (kHz)
Voice channel	1		0–4
Group	12	12 voice circuits	60–108
Supergroup	60	5 groups	312–552
Mastergroup	600	10 supergroups	564–3,084
Mastergroup Mux	1,200–3,600	Various	312,564–17,548
Jumbogroup	3,600	6 master groups	564–17,548
Jumbogroup Mux	10,800	3 jumbo groups	3,000–60,000

levels. Each level of the hierarchy is implemented using a set of standard FDM modules. The multiplex equipment is independent of particular broadband transmission media.

All multiplex equipment in the FDM hierarchy used SSB modulation. Thus, every voice circuit required approximately 4 kHz of bandwidth. The lowest level building block in the hierarchy is a channel group consisting of 12 voice channels. A channel group multiplex uses a total bandwidth of 48 kHz. Figure 1.20 shows a block diagram of an A5 channel group multiplexer, the most common A-type channel bank used for first-level multiplexing. Twelve modulators using 12 separate carriers generate 12 double-sideband signals as indicated. Each channel is then bandpass filtered to select only the lower sideband of each double-sideband signal. The composite multiplex signal is produced by superposing the filter outputs. Demultiplex equipment in a receiving terminal uses the same basic processing in reverse order.

Notice that a sideband separation filter not only removes the upper sideband but also restricts the bandwidth of the retained signal: the lower sideband. These filters therefore represented a basic point in the analog telephone network that defined the bandwidth of a voice circuit. Since FDM was used on all long-haul analog circuits,

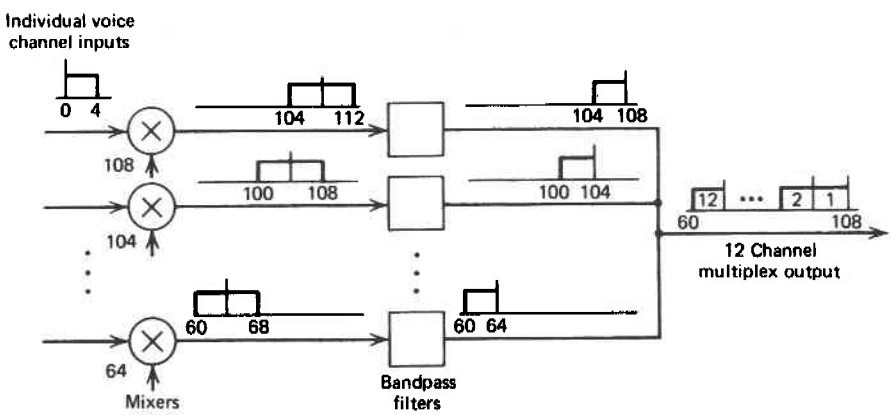


Figure 1.20 A5 channel bank multiplexer.

long-distance connections provided somewhat less than 4 kHz of bandwidth. (The loading coils discussed previously also produce similar bandwidth limitations into a voice circuit.)

As indicated in Table 1. 5, the second level of the FDM hierarchy is a 60-channel multiplex referred to as a supergroup. Figure 1.21 shows the basic implementation of an LMX group bank that multiplexes five first-level channel groups. The resulting 60-channel multiplex output is identical to that obtained when the channels are individually translated into 4-kHz bands from 312 to 552 kHz. Direct translation requires 60 separate SSB systems with 60 distinct carriers. The LMX group bank, however, uses only five SSB systems plus five lower level modules. Thus two-stage multiplexing, as implied by the LMX group bank, requires more total equipment but achieves economy through the use of common building blocks. (Engineers at AT&T Laboratories also developed a multiplexer that generated a supergroup directly [7].)

Because a second-level multiplexer packs individual first-level signals together without guard bands, the carrier frequencies and bandpass filters in the LMX group bank must be maintained with high accuracy. Higher level multiplexers do not pack the lower level signals as close together. Notice that a master group, for example, does not provide one voice channel for every 4 kHz of bandwidth. It is not practical to maintain the tight spacing between the wider bandwidth signals at higher frequencies. Furthermore, higher level multiplex signals include pilot tones to monitor transmission link quality and aid in carrier recovery.

1.2.7 Wideband Transmission Media

Wire pairs within multipair cables have usable bandwidths that range from a little under 1 MHz up to about 4 MHz depending on the length, the wire gauge, and the type of insulation used on the pairs. Multiplexed transmissions on these wire pairs consequently have capacities that range from 24 channels (on analog N3 or digital T1 carrier systems) up to 96 channels (on obsolete digital T2 carrier systems). In contrast, an ana-

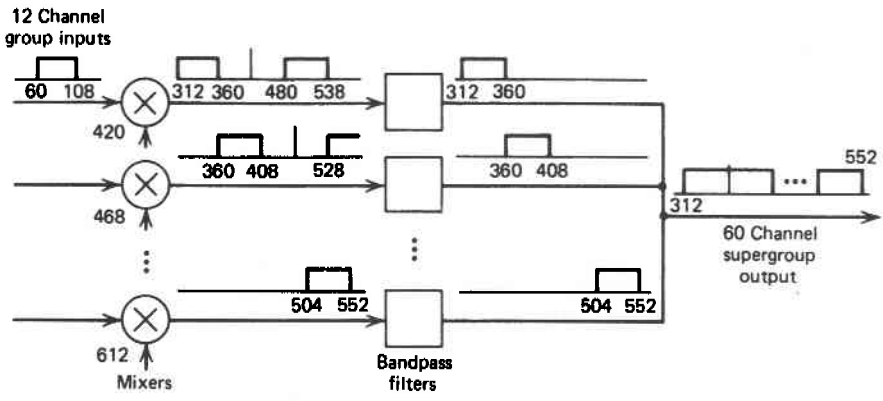


Figure 1.21 LMX group bank multiplexer.

log L5E coaxial cable system was developed to carry 13,200 voice channels. Optical fiber systems have been developed that carry in excess of 100,000 voice channels on a single wavelength. Because optical transmitters [light-emitting diodes (LEDs) and lasers] function best in a pulsed mode of operation, fiber systems are primarily digital in nature. Digital fiber optic transmission systems are described in Chapter 8. This section discusses coaxial cable and point-to-point microwave radio systems.

Coaxial Cable

Coaxial cable systems were used predominantly to satisfy long-haul requirements of the toll network. The first commercial system was installed in 1941 for transmission of 480 voice circuits over a 200-mile stretch between Minneapolis, Minnesota, and Stevens Points, Wisconsin [7]. To combat attenuation, repeater amplifiers were installed at 5.5-mile intervals. Considering the maximum capacity of 12 voice circuits on open wire or cable at the time, the introduction of "coax" was a significant development. After the first installation coaxial cable capacity was steadily increased by (1) using larger diameter cables (0.375 in.) to reduce attenuation, (2) decreasing the distance between repeaters, and (3) improving the noise figure, linearity, and bandwidth of the repeater amplifiers.

A summary of the analog coaxial cable systems used in the Bell System is provided in Table 1.6. Notice that each system reserved one pair of tubes as spares in the event of failure, a particularly important consideration since each tube carried a high volume of traffic. Because optical fibers have wider bandwidths, lower attenuation, lower maintenance, and lower cost, coaxial systems are obsolete.

Microwave Radio

Much of the impetus for terrestrial microwave radio systems came from the need to distribute television signals nationwide. As the volume of long-distance traffic increased, radio systems also became the most economical means of distributing voice circuits in the long-distance network. Beginning in 1948, when the first system was installed between New York and Boston, the number of microwave radio systems grew to supply 60% of the voice circuit miles in the U.S. toll network in 1980 [7]. It was just a few years after that that optical fibers began to take over for high-density interoffice routes and eventually for the network as a whole.

TABLE 1.6 Coaxial Cable Systems in the Bell Network

System Designation	Pairs per System[a]	Signal Designation	Repeater Spacing (miles)	Capacity per Pair	Total Capacity
L1	3/6	Mastergroup	8	600	1,800
L3	5/6	Mastergroup Mux	4	1,860	9,300
L4	9/10	Jumbogroup	2	3,600	32,400
L5	10/11	Jumbogroup Mux	1	10,800	108,000

[a]The number of pairs are shown as working/total.

Microwave radio systems require line-of-sight transmission with repeater spacings typically 26 miles apart. The major advantage of radio systems is that a continuous right-of-way is not required—only small plots of land spaced 20–30 miles apart for towers and equipment shelters. A major cost of guided transmission, for example, wire pairs, coax, or fiber, is the right-of-way costs. In many metropolitan areas, microwave routes became heavily congested and could not be expanded with allocated common-carrier frequency bands. In these situations, finding a right-of-way for a coax or eventually an optical fiber system was sometimes the only choice for high-capacity transmission.*

The frequency bands allocated by the FCC for common-carrier use in the United States are listed in Table 1.7. Of these bands, 4 and 6 GHz have been the most popular. The 2-GHz band has not been used extensively because the relatively narrow allocated channel bandwidths do not permit implementation of economical numbers of voice circuits. The basic drawback of the 11-GHz band is its vulnerability to rain attenuation. However, 11-GHz radios have been used in some short-haul applications.

The microwave radio systems of the analog Bell network are listed in Table 1.8. Notice that each radio system is designed to carry one of the multiplex hierarchies described previously. All of these radios except the AR-6A used low-index frequency modulation (FM) of the signal generated by the FDM multiplexer equipment. Thus, the FM radios transmit the SSB FDM signal as a baseband signal with a bandwidth as indicated in Table 1.5. FM modulation was chosen to permit the use of nonlinear power amplifiers in the transmitters and to take advantage of FM signal-to-noise ratio performance.

Examination of Tables 1.7 and 1.8 indicates that 13.3 kHz of bandwidth is utilized per voice circuit in TD-2 radios and 14.3 kHz in TH-3 radios. Thus, the use of FM introduced a significant increase in the 4 kHz bandwidth of the individual SSB voice circuit. In contrast, the SSB AR-6A radio introduced in 1981 provided 6000 voice circuits in the 30-MHz channels at 6 GHz.

Since a large number of voice circuits are carried by each radio channel, microwave systems usually include extra equipment and extra channels to maintain service despite outages that may result from any of the following:

1. Atmospheric-induced multipath fading
2. Equipment failures
3. Maintenance

On some routes, the most frequent source of outage in a microwave radio system arises from multipath fading. Figure 1.22 depicts a simple model of a multipath environment arising as a result of atmospheric refraction. As indicated, the model involves two rays: a primary ray and a delayed secondary ray. If the secondary ray arrives out of phase with respect to the primary ray, the primary signal is effectively canceled. The

*Microwave bands with wide-bandwidth allocations at high carrier frequencies are available for some local distribution services and are discussed in Chapter 11.

TABLE 1.7 Microwave Frequencies Allocated for Common-Carrier Use in the United States

Band (MHz)	Total Bandwidth (MHz)	Channel Bandwidths (MHz)
2110–2130	20	3.5
2160–2180	20	3.5
3700–4200	500	20
5925–6425	500	30
10,700–11,700	1000	40, 20

amount of cancelation is dependent on both the magnitude and the phase of the secondary ray. Quite often only nominal amounts of fading occur and can be accommodated by excess signal power in the transmitter, called a fade margin. In some instances, however, the received signal is effectively reduced to zero, which implies that the channel is temporarily out of service.

Frequency Diversity

Fortunately, exceptionally deep fades normally affect only one channel (carrier frequency) at a time. Thus, a backup channel including a spare transmitter and a spare receiver can be used to carry the traffic of a faded primary channel. Selection of and switching to the spare channel is performed automatically without a loss of service. This remedy for multipath fading is referred to as frequency diversity. Notice that frequency diversity also provides hardware backup for equipment failures.

A fully loaded TD-3 radio system used 12 channels: 10 main channels and 2 backup channels for protection. This is referred to variously as 2-for-10, 10-by-2, or 10×2 protection switching. Some short-haul systems used 1-for-1 protection switching because it is simpler to implement. However, since only half of the required bandwidth is actually carrying traffic, systems with 1-for-1 protection were only allowed in uncongested environments.

TABLE 1.8 Bell System Analog Microwave Radios

System	Band (GHz)	Voice Circuits	Application
TD-2	4	600–1500	Long haul
TD-3	4	1200	Long haul
TH-1	6	1800	Short/long haul
TH-3	6	2100	Short/long haul
TM-1	6	600–900	Short/long haul
TJ	11	600	Short haul
TL-1	11	240	Short haul
TL-2	11	600–900	Short haul
AR-6A	6	6000	Long haul (SSB)

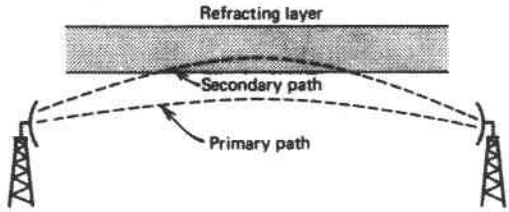


Figure 1.22 Two-ray model of multipath propagation.

Except in maintenance situations, protection switching must be automatic in order to maintain service continuity. A typical objective was to restore service within 30 msec to minimize noticeable effects in the message traffic. A more critical requirement is to restore service before the loss of signal is interpreted by some signaling schemes as a circuit disconnect. Inadvertent disconnects occur if an outage lasts for more than 1–2 sec.

Space Diversity

Since deep fades only occur when a secondary ray arrives exactly out of phase with respect to a primary ray, it is unlikely that two paths of different lengths experience fading simultaneously. Figure 1.23 depicts a technique, called space diversity, using different path lengths to provide protection against multipath fading. As indicated, a single transmitter irradiates two receive antennas separated by some distance on the tower. Although the path length difference may be less than a meter, this difference is adequate at microwave frequencies, which have wavelengths on the order of tenths of meters.

Rain is another atmospherically based source of microwave fading. As already mentioned, rain attenuation is a concern mostly in higher frequency radios (11 GHz and above). Unfortunately neither frequency diversity (at the high frequencies) nor space diversity provides any protection against rain fades.

Satellites

Following the April 1965 launch of the Soviet Union's Molniya domestic communications satellite and the first international communications satellite, INTELSAT I, the use of satellites for international telephone traffic grew phenomenally. The 1970s and early 1980s also produced significant use of satellites in the United States for television program distribution and for corporate voice and data networks. The first domes-

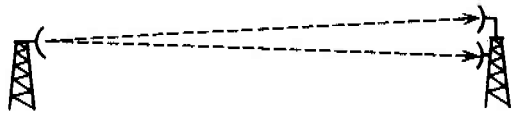


Figure 1.23 Space diversity.

tic system in North America was the Canadian Anik A in 1972 followed by Western Union's Westar system for U.S. service in 1974 [8].

In one sense a satellite system is a microwave radio system with only one repeater: the transponder in outer space. In fact, some satellite systems use the same 4- and 6-GHz frequency bands used by terrestrial microwave radios. In another sense, however, the broadcast nature of the down link offers additional opportunities for new services not available from point-to-point terrestrial systems. Distribution of network television programming is one application particularly suited to the broadcast nature of satellites. Direct broadcast satellite (DBS) transmission to receive-only home receivers is a prime example.* Two primary examples of DBS systems are the European DVB-T system [9] and the North American digital satellite system (DSS) system developed by Hughes Electronics Systems (HES).

Another application that is well suited to satellites is communications with mobile stations. The international maritime satellite (INMARSAT) system, for example, began supporting the maritime industry in 1982 and was augmented with digital Skyphone [10] for aeronautical telephone service in 1989.

One drawback to satellite communications is the inherent propagation delay of the long transmission path. For a stationary satellite, this delay (not including ground links) is 250 msec up and down. A complete circuit with satellite links for both directions of travel therefore implies greater than a one-half-second round-trip propagation time. Delays of this magnitude are noticeable in a voice conversation but not prohibitive. The effects of the propagation delays can be alleviated somewhat by pairing each satellite circuit with a ground-based circuit in the opposite direction. Thus, the round-trip delay involves only one satellite link.

As is the case with point-to-point microwave radio, fiber optic transmission has displaced the use of satellites for high-density, domestic and international telephone communications. Hence the use of satellites is primarily directed at thin-route telephone and data applications, some mobile systems, and broadcast applications.

1.2.8 Transmission Impairments

One of the most difficult aspects of designing an analog telephone network is determining how to allocate transmission impairments to individual subsystems within the network. Using subjective evaluations by listeners, certain objectives for end-to-end transmission quality were established in a relatively straightforward manner [11]. After tempering the goals with economic feasibility, the end-to-end objectives were established. However, considering the myriad of equipment types and connection combinations in the older analog network, designing the individual network elements to meet these objectives in all cases was a complex problem. A great deal of credit is due the Bell System for having developed a nationwide analog network with the level of consistent performance it had.

*A DBS system is designed to use receiver antennas that are about $\frac{1}{2}$ m in diameter. The older home satellite receiver systems that were common in the United States used 3- to 5-m antennas to receive (intercept) commercial program distribution.

The major factors to be considered in establishing transmission objectives for an analog network are signal attenuation, noise, interference, crosstalk, distortion, echoes, singing, and various modulation- and carrier-related imperfections.

Signal Attenuation

Subjective listening tests have shown that the preferred acoustic-to-acoustic loss [12] in a telephone connection should be in the neighborhood of 8 dB. A study of local telephone connections [13] demonstrates that the typical local call had only 0.6 dB more loss than ideal. Surveys of the toll network [14] indicated that the average analog toll connection had an additional 6.7 dB of loss. This same survey also showed that the standard deviation of loss in toll connections was 4 dB (most of which was attributable to the local loops).

Since trunks within the toll network used amplifiers to offset transmission losses, it would have been straightforward to design these trunks with zero-decibel nominal insertion loss. However, as discussed later, echo and singing considerations dictated a need for certain minimum levels of net loss in most analog trunk circuits.

Interference

Noise and interference are both characterized as unwanted electrical energy fluctuating in an unpredictable manner. Interference is usually more structured than noise since it arises as unwanted coupling from just a few signals in the network. If the interference is intelligible, or nearly so, it is referred to as crosstalk.* Some of the major sources of crosstalk are coupling between wire pairs in a cable, inadequate filtering or carrier offsets in older FDM equipment, and the effects of nonlinear components on FDM signals.

Crosstalk, particularly if intelligible, is one of the most disturbing and undesirable imperfections that can occur in a telephone network. Crosstalk in analog systems is particularly difficult to control since voice signal power levels vary considerably (i.e., across a dynamic range of 40 dB). The absolute level of crosstalk energy from a high-level signal must be small compared to a desired low-level signal. In fact, crosstalk is most noticeable during speech pauses, when the power level of the desired signal is zero.

Two basic forms of crosstalk of concern to telecommunications engineers are *near-end crosstalk* (NEXT) and *far-end crosstalk* (FEXT). Near-end crosstalk refers to coupling from a transmitter into a receiver at a common location. Often this form of crosstalk is most troublesome because of a large difference in power levels between the transmitted and received signals. Far-end crosstalk refers to unwanted coupling into a received signal from a transmitter at a distant location. Both forms of crosstalk are illustrated in Figure 1.24.

Noise

The most common form of noise analyzed in communications systems is white noise with a Gaussian (normal) distribution of amplitude values. This type of noise is both

*Crosstalk is also used to characterize signallike interferences in nonvoice networks. For example, crosstalk in a data circuit would refer to an interfering signal being coupled in from another similar data circuit.

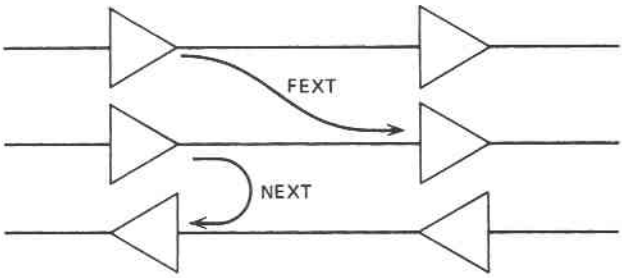


Figure 1.24 Near-end and far-end crosstalk.

easy to analyze and easy to find since it arises as thermal noise in all electrical components. Battery systems used to power customer loops are also a source of this type of noise. White noise is truly random in the sense that a sample at any instant in time is completely uncorrelated to a sample taken at any other instant in time. The other most common forms of noise in the telephone network are impulse noise and quantization noise in digital voice terminals (Chapter 3). Impulse noise can occur from switching transients in older electromechanical switching offices or from rotary dial telephones. Step-by-step switches were the most frequent culprits. More modern electromechanical switches that use glass-encapsulated reed relays for crosspoints produce much less noise. Whereas white noise is usually quantified in terms of average power, impulse noise is usually measured in terms of so many impulses per second. Impulse noise is usually of less concern to voice quality than background white noise. However, impulse noise tends to be the greatest concern in a data communications circuit.

The power level of any disturbing signal, noise or interference, is easily measured with a root-mean-square (rms) voltmeter. However, disturbances at some frequencies within the passband of a voice signal are subjectively more annoying than others. Thus, more useful measurements of noise or interference power in a speech network take into account the subjective effects of the noise as well as the power level. The two most common such measurements in telephony use a C-message weighting curve and a psophometric weighting curve, as shown in Figure 1.25. These curves essentially represent filters that weight the frequency spectrum of noise according to its annoyance effect to a listener. C-message weighting represents the response of the 500-type telephone set. As far as perceived voice quality is concerned, only the noise that gets passed by the telephone set is important. Notice that disturbances between 1 and 2 kHz are most perceptible. C-message weighting is used in North America while psophometric weighting is the European (ITU-T) standard.

A standard noise reference used by telephone engineers is 1 pW, which is 10^{-12} W, or –90 dBm (dBm is power in decibels relative to a milliwatt). Noise measured relative to this reference is expressed as so many decibels above the reference (dBrn). Thus, a noise level of 30 dBrn corresponds to –60 dBm, or 10^{-9} W of power. If the readings are made using C-message weighting, the power level is expressed by the abbreviation dBrnC. Similarly, psophometrically weighted picowatts are expressed by

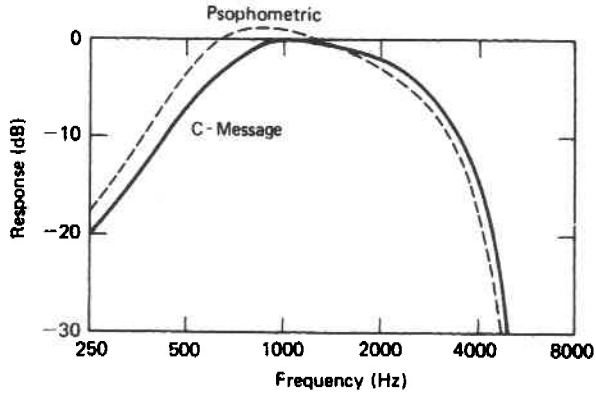


Figure 1.25 C-message and psophometic weighting.

the abbreviation pWp. The relationships between various noise power measurements are given in Table 1.9.

The quality of an analog voice circuit is usually not specified in terms of the classical signal-to-noise ratio. The reason is that relatively low levels of noise or interference are noticeable during pauses in speech, when there is no signal. On the other hand, high levels of noise can occur during speech and be unnoticeable. Thus, absolute levels of noise are more relevant than signal-to-noise ratios for specifying voice quality.* The objectives for maximum noise levels in the analog AT&T network were 28 dBrnC for connections up to 60 miles in length and 34 dBrnC for 1000-mile circuits.†

Distortion

In a previous section signal attenuations were considered with the tacit assumption that a received waveform was identical in shape to a source waveform but merely scaled down in amplitude. Actually, a received waveform generally contains certain distortions not attributable to external disturbances such as noise and interference but that can be attributed to internal characteristics of the channel itself. In contrast to noise and interference, distortion is deterministic; it is repeated every time the same signal is sent through the same path in the network. Thus distortions can be controlled or compensated for once the nature of the distortion is understood.

There are many different types and sources of distortion within the telephone network. The telephone companies minimized those types of distortion that most affected the subjective quality of speech. Later on they also became concerned with distortion effects on data transmission. Some distortions arise from nonlinearities in the network, such as carbon microphones, saturating voice-frequency amplifiers, and un-

*It is a common practice in the industry to specify the quality of a voice circuit in terms of a test-tone-to-noise ratio. However, the test tone must be at a specific power level so the ratio, in fact, specifies absolute noise power.

†These noise power values are related to a particular point in a circuit, called a *zero-transmission-level point*, discussed later.

TABLE 1.9 Relationships between Various Noise Measurements

	To Convert		
	From	To	
dBm		dBrn	Add 90 dB
dBm	3 kHz flat	dBrnC	Add 88 dB
dBm	3 kHz flat	dBp	Add 87.5 dB
dBrn	3 kHz flat	dBrnC	Subtract 2 dB
dBC		dBp	Subtract 0.5 dB
pW	3 kHz flat	pWp	Multiply by 0.562

matched compandors (Chapter 3). Other distortions are linear in nature and are usually characterized in the frequency domain as either amplitude distortion or phase distortion.

Amplitude distortion refers to attenuating some frequencies in the voice spectrum more than others. The loading coils discussed earlier represent one means of eliminating amplitude distortion on long voice-frequency wire pairs. Amplitude distortion could also be introduced by spectrum-limiting filters in FDM equipment. Ideally these filters should uniformly pass all voiceband frequencies up to 4 kHz and reject all others. Practical designs, however, imply the need for gradual attenuation "roll-offs" beginning at about 3 kHz. Figure 1.26 shows the attenuation-versus-frequency response of a typical analog toll connection of the past.

Phase distortion is related to the delay characteristics of the transmission medium. Ideally a transmission system should delay all frequency components in a signal uni-

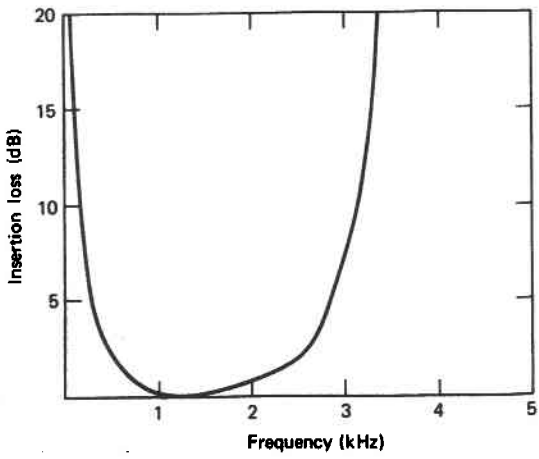


Figure 1.26 Insertion loss versus frequency of typical toll connection.

formly so the proper phase relationships exist at the receiving terminal. If individual frequency components experience differing delays, the time-domain representation at the output becomes distorted because superposition of the frequency terms is altered at the output. For reasons not discussed here the delay of an individual frequency component is usually referred to as its envelope delay. For a good explanation of envelope delay see reference [15].

Uniform envelope delay relates to a phase response that is directly proportional to frequency. Thus systems with uniform envelope delay are also referred to as linear phase systems. Any deviation from a linear phase characteristic is referred to as phase distortion. The perceptual effects of phase distortion to a voice signal are small. Thus only minimal attention need be given to the phase response of a voice network. The phase response and corresponding envelope delay provided by a typical analog toll connection is shown in Figure 1.27.

In addition to the distortions just mentioned, analog carrier systems sometimes introduced other frequency-related distortions such as frequency offsets, jitter, phase hits, and signal dropouts. The effects of these imperfections and phase distortion were adequately controlled for voice traffic but presented difficulties for high-rate voice-band data traffic.

Echoes and Singing

Echoes and singing both occur as a result of transmitted signals being coupled into a return path and fed back to the respective sources. The most common cause of the coupling is an impedance mismatch at a four-wire-to-two-wire hybrid. As shown in Figure 1.28, mismatches cause signals in the incoming branch of a four-wire circuit to get coupled into the outgoing branch and return to the source. In the older network with two-wire analog end office switches, it was impractical to provide good impedance matches at this point for all possible connections since the two-wire side of the hybrid

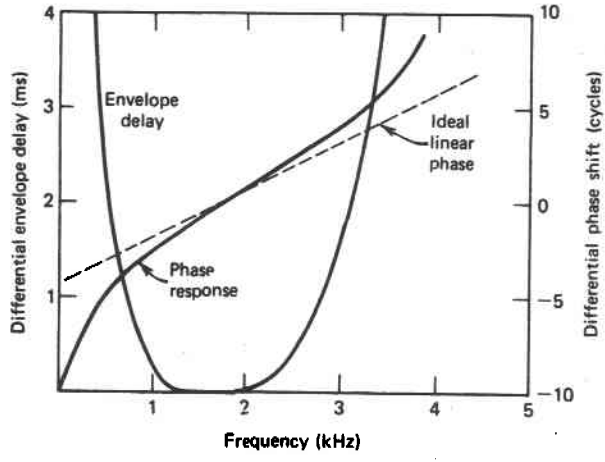


Figure 1.27 Envelope delay and phase response of typical toll connection.

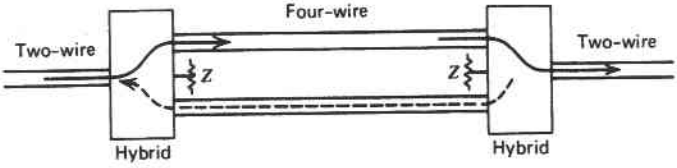


Figure 1.28 Generation of echos at two-wire-to-four-wire interface.

could be connected to many different local loops, each with its own characteristic impedance.

If only one reflection occurs, the situation is referred to as "talker echo." If a second reflection occurs, "listener echo" results. When the returning signal is repeatedly coupled back into the forward path to produce oscillations, singing occurs. Basically, singing results if the loop gain at some frequency is greater than unity. If the loop gain is only slightly less than unity, a near-singing condition causes damped oscillations. Singing and near-singing conditions have a disturbing effect on both the talker and the listener. Talker echo is usually the most noticeable and troublesome.

The degree of echo annoyance experienced by a talker is dependent on both the magnitude of the returning signal and the amount of delay involved [16, 17]. On short connections the delay is small enough that the echo merely appears to the talker as natural coupling into his ear. In fact, a telephone is purposely designed to couple some speech energy (called sidetone) into the earpiece. Otherwise, the telephone seems dead to a talker. Near-instantaneous echoes merely add to the sidetone and go unnoticed. As the roundtrip delay increases, however, it becomes necessary to increasingly attenuate the echoes to eliminate the annoyance to a talker. Hence, long-distance circuits require significant attenuation to minimize echo annoyance. Fortunately, an echo experiences twice as much attenuation as does the forward-propagating signal since it traverses twice the distance. Intermediate-length connections are typically designed with 2–6 dB of path attenuation depending on the delay. All transmission links within the Bell System were designed with specific amounts of net loss called via net loss (VNL) that depended on the length of the link and the position in the hierarchy [17]. In general, the VNL network design established end-to-end attenuation in proportion to the length of the circuit.

Connections that produce more than 45 msec of roundtrip delay (representing 1800 miles of wire) require more attenuation for echo control than can be tolerated in the forward path. In these cases one of two types of devices was used to control the echo: an echo suppressor or an echo canceller.

As shown in Figure 1.29, an echo suppressor operates on four-wire circuits by measuring the speech power in each leg and inserting a large amount of loss (35 dB typically) in the opposite leg when the power level exceeds a threshold. Thus, a returning echo is essentially blocked by the high level of attenuation. Notice that an echo suppressor converts a full-duplex circuit into a half-duplex circuit with energy sensing being the means of turning the line around.

One drawback of echo suppressors for voice circuits was that they might clip beginning portions of speech segments. If a party at one end of a connection begins talk-

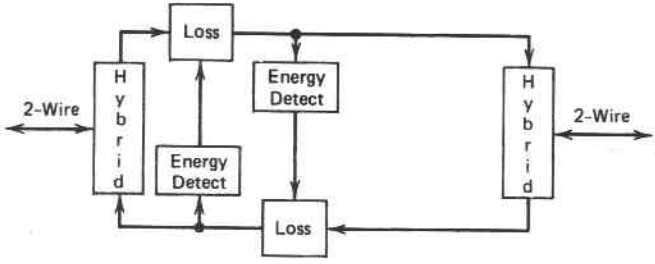


Figure 1.29 Echo suppressor.

ing at the tail end of the other party's speech, the echo suppressor does not have time to reverse directions. Echo suppressors with good performance are able to reverse directions in 2–5 msec [16]. For the fastest possible release of backward attenuation, split-echo suppressors are necessary. A split-echo suppressor is one that separates the echo control of each direction so the loss insertion of each direction is closest to the point of echo occurrence.

A second, and much preferred, form of echo control is echo cancellation [18, 18a]. As shown in Figure 1.30, an echo canceller operates by simulating the echo path to subtract a properly delayed and attenuated copy of a transmitted signal from the receive signal to remove (cancel) echo components. Thus echo cancellation requires training to determine how much delay and attenuation are needed to simulate the echo characteristics of the circuit. Notice in Figure 1.30 that echoes are canceled close to the source so that delays in the echo canceller are minimized. The important feature of an echo canceller is that it maintains a full-duplex circuit so clipping does not occur. Satellite circuits with greater than 500 msec of roundtrip delay required echo cancellers for acceptable performance. Because the cost of digital signal processing (DSP) technology has dropped so dramatically, echo cancellers are now used in any situation requiring echo control.

Full-duplex voiceband modems (V.32 and later) incorporate echo cancellers directly in their receive circuitry. Thus, network-based echo cancellers are unnecessary*—and sometimes undesirable because tandem echo cancelling may not work properly if two echo cancellers do not cooperate in the training process. Network-based echo cancellers can therefore be disabled by a modulated 2100-Hz tone (echo suppressors were also disabled with a 2100-Hz tone) transmitted at the start of a connection [19].

Another method of echo control involves impedance matching of the hybrids to reduce the magnitude of the echo. Some electronic hybrids provide dynamic balancing to automatically eliminate or reduce echoes. In fact, a common way of implementing the impedance matching is to build an echo canceller with near-zero delay. These circuits eliminate, or greatly reduce, echoes occurring at the associated hybrid but do not eliminate echoes that may occur elsewhere in the network. For a detailed discussion

*If the terminal (e.g., modem) echo canceller has insufficient delay buffering for very long echoes, the network echo canceller at the far end of a connection may be necessary.

TLP = ?

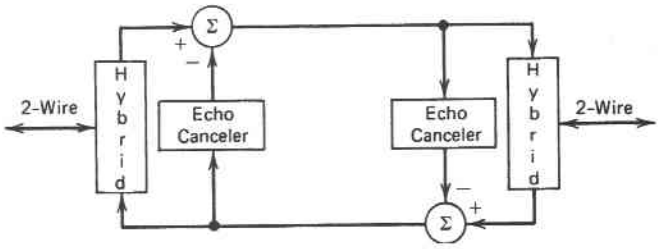


Figure 1.30 Echo canceller.

of all types of echo control, see reference [20]. In general, the procedures used to control echoes also control singing. On some fairly short connections, however, no echo control is necessary, and singing may become a problem.

1.2.9 Power Levels

As indicated in previous paragraphs, voice signal power in a long-distance connection needs to be rigidly controlled. The delivered signal power must be high enough to be clearly perceived but, at the same time, not be so strong that circuit instabilities such as echo and singing result.

To maintain rigid control on the end-to-end power level of a circuit involving a variety of transmission systems, telephone companies necessarily controlled the net attenuation and amplification of each analog transmission system. These systems were designed for a certain amount of net loss (VNL) as described previously.

To administer the net loss of transmission links, the transmission levels of various points in a transmission system are specified in terms of a reference point. ITU-T recommendations call this point the zero-relative-level point and the North American term is a zero-transmission-level point (0-TLP). The reference point may not exist as an accessible point but has long been considered to be at the sending end terminal of a two-wire switch. In North America the sending end of a four-wire switch is defined to be a –2-dB TLP. Hence, a 0-dB TLP is only a hypothetical point on a four-wire circuit. Nevertheless, it is useful in relating the signal level at one point in the circuit to the signal level at another point in the circuit.

If a 0-dBm (1-mW) test tone is applied at a 0-TLP, the power level at any other point in the circuit is determined directly (in decibels referred to 1 mW) as the TLP value at that point. It should be emphasized, however, that TLP values do not specify power levels; they specify only the gain or loss at a point relative to the reference point.

Signal power levels or noise power levels are not normally expressed in terms of local measured values. Instead, powers are expressed in terms of their value at the 0-TLP. For example, if an absolute noise power of 100 pW (20 dBrn or –70 dBm) is measured at a –6-dB TLP, it is expressed as 26 dBrn0. (The "0" indicates that the specification is relative to the 0-TLP.) If noise power is measured with C-message weighting, it is designated as so many dBrnC0. Similarly, psophometric weighted

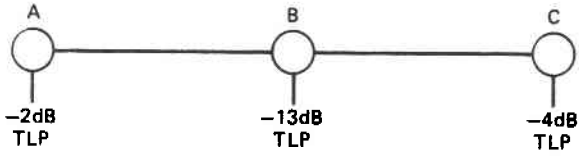


Figure 1.31 TLP values for Example 1.1.

noise is commonly expressed in units of picowatts psophometrically weighted (dBm0p or pWp0).

Example 1.1. Using Figure 1.31 for reference, determine each of the following: (a) the signal power to be applied at point B to determine if points A and C are at the proper levels; (b) the amount of gain (loss) a signal experiences when propagating from A to C; and (c) the amount of noise that would be measured at C if 27 dbrnC of absolute noise is measured at B and no additional noise occurs on the B-to-C link.

Solution. (a) Because point B is a −13-dB TLP, the proper test tone level is −13 dBm (0.05 mW). (b) Because the TLP values drop by 2 dBm, there is 2 dB net loss from A to C. (c) An absolute measurement of 27 dBrnC at B is 40 dBrnC0. This is also 40 dBrnC0 at C. The absolute noise power measured at C would be 40 − 4 = 36 dBrnC.

To put signal and noise powers in perspective, a survey of voice signals in the Bell System [21] indicated the average speech signal has −16 dBm0 of power. In comparison, the noise power objective for a 1000-mile analog circuit was 34 dBrnC0 (−56 dBmC0). Thus, the noise is 40 dB below the signal power.

1.2.10 Signaling

The signaling functions of a telephone network refer to the means for transferring network-related control information between the various terminals, switching nodes, and users of the network. There are two basic aspects of any signaling system: specially encoded electrical waveforms (signals) and how these waveforms should be interpreted. The most common control signals to a user are dial tone, ringback, and busy tone. These signals are referred to as call progress tones and may never change. The signaling procedures used internally to the network are not constrained by user convention and were often changed to suit particular characteristics of transmission and switching systems. As a result, the analog network used a wide variety of signaling schemes to transfer control information between switching offices.

Signaling Functions

Signaling functions can be broadly categorized as belonging to one of two types: supervisory or information bearing. Supervisory signals convey status or control of network elements. The most obvious examples are request for service (off-hook), ready

to receive address (dial tone), call alerting (ringing), call termination (on-hook), request for an operator or feature invocation (hook flash), called party ringing (ringback), and network or called party busy tones. Information bearing signals include called party address, calling party address, and toll charges. In addition to call-related signaling functions, switching nodes communicate between themselves and network control centers to provide certain functions related to network management. Network-related signals may convey status such as maintenance test signals, all trunks, busy, or equipment failures or they may contain information related to routing and flow control. Chapter 7 discusses some of the basic considerations of network management for routing and flow control.

In-Channel Signaling

Signals are transmitted with one of two basic techniques: in-channel signaling or common-channel signaling. In-channel signaling (sometimes referred to as "per-trunk signaling") uses the same transmission facilities or channel for signaling as for voice. Common-channel signaling, as discussed in the next section, uses one channel for all signaling functions of a group of voice channels. In the past most signaling systems in the telephone network were the in-channel variety.

In-channel signaling systems can be further subdivided into in-band and out-of-band techniques. In-band systems transmit the signaling information in the same band of frequencies used by the voice signal. The main advantage of in-band signaling is that it can be used on any transmission medium. The main disadvantage arises from a need to eliminate mutual interference between the signaling waveforms and a user's speech. The most prevalent example of in-band signaling was single-frequency (SF) signaling, which used a 2600-Hz tone as an on-hook signal for interoffice trunks. Although normal speech rarely produces a pure 2600-Hz signal, inadvertent disconnects have occurred as a result of user-generated signals. Two other common examples of in-band signaling are addressing by dual-tone multifrequency (DTMF) signals from push-button telephones or multifrequency (MF) signaling between switching offices.

In-channel but out-of-band signaling uses the same facilities as the voice channel but a different portion of the frequency band. Thus out-of-band signaling represents a form of FDM within a single voice circuit. The most common instance of out-of-band signaling is dc signaling as used on most customer loops. With this form of signaling, the central office recognizes the off-hook condition by the flow of direct current in the line. Other commonly used loop signals are dial pulses generated by a rotary dial at a rate of 10 pulses per second and a 20-Hz ringing voltage from the central office. All of these signals use lower frequencies than those generated in speech. Thus there is no possibility of one being mistaken for the other. The major disadvantage of out-of-band signaling is its dependence on the transmission system. For example, SSB carrier systems filter out the very low frequencies associated with each voice channel. Thus, the on-hook/off-hook signal must be converted to something like SF signaling for FDM transmission. Out-of-band signaling is also implemented with frequencies above the cut-off frequency of voice separation filters but below the 4-kHz limit of a channel. CCITT recommends the use of 3825 Hz for this purpose.

Common-Channel Signaling

Instead of sending signaling information over the same facilities that carry the message traffic (voice signals), common-channel signaling (CCS) uses a dedicated data link between the stored-program control elements (computers) of switching systems. Figure 1.32 depicts such a data link between two switching offices. Notice that the per-trunk signaling equipment associated with the trunks has been eliminated. The data link sends messages that identify specific trunks and events related to the trunks. Thus the type of CCS shown in Figure 1.32 is referred to as "channel-associated" common-channel signaling. The main advantages of CCS are:

1. Only one set of signaling facilities is needed for each associated trunk group instead of separate facilities for each individual circuit.
2. A single dedicated control channel allows transfer of information such as address digits directly between the control elements (computers) of switching offices. In-channel systems, on the other hand, must have the control information switched from the common control equipment of the originating office onto the outgoing channel, and then the receiving office must switch the incoming control information from the voice channel into its common control equipment. The simpler procedure for transferring information directly between switch processors is one of the main motivations for CCS.
3. Since separate channels are used for voice and control, there is no chance of mutual interference.
4. Since the control channel of a common-channel system is inaccessible to users, a major means for fraudulent use of the network is eliminated.
5. Connections involving multiple switching offices can be set up more rapidly since forwarding of control information from one office can overlap a circuit set up through the node. With in-channel systems the associated circuit must first be established before the control information can be transferred across it. The CCS data link can also operate at a much higher data rate than common analog network signaling schemes, implying even faster connection setup.
6. The channel used for CCS does not have to be associated with any particular trunk group. In fact, the control information can be routed to a centralized control facility where requests are processed and from which switching offices

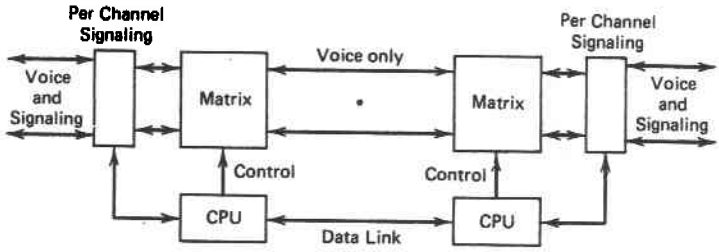


Figure 1.32 Trunk-group-associated common-channel signaling.

receive their connection control information. Figure 1.33 depicts a CCS network that is disassociated from the message network structure. One advantage of centralized control is its ability to process requests with knowledge of networkwide traffic conditions. Centralized control is also attractive for managing multiple switching offices that are too small to warrant call processing facilities of their own. The transition from in-channel signaling to disassociated CCS at the network level is analogous to the lower level transition from direct progressive control switches (step-by-step) to common control switches.

The major disadvantages of CCS are:

1. Control information pertaining to an established circuit, such as a disconnect, must be relayed from one node to the next in a store-and-forward fashion. An in-channel disconnect signal, on the other hand, automatically propagates through the network, enabling all nodes involved in the connection to simultaneously process the disconnect and release the associated facilities.
2. If one node in a common-channel system fails to relay the disconnect information properly, facilities downstream from the disconnect will not be released. Thus, a high degree of reliability is required for the common channel—both in terms of physical facilities (duplication) and in terms of error control for the data link.
3. Since the control information traverses a separate path from the voice signal, there is no automatic test of the voice circuit as when the voice channel is used to transfer control information. Initial usage of CCS systems included special provisions for testing a voice circuit when it was set up.
4. In some instances all trunks in a group do not logically terminate at the same switch. Figure 1.34 depicts a multiplexed transmission link from a PBX to the

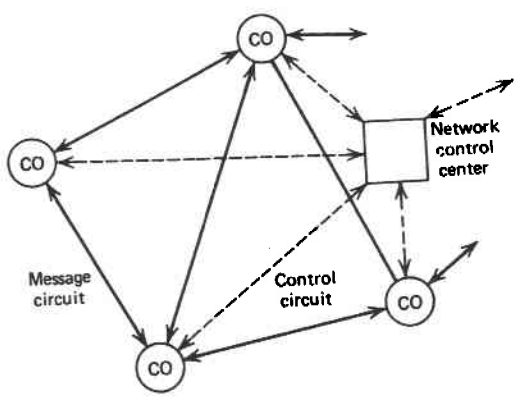


Figure 1.33 Dissociated common-channel signaling network.

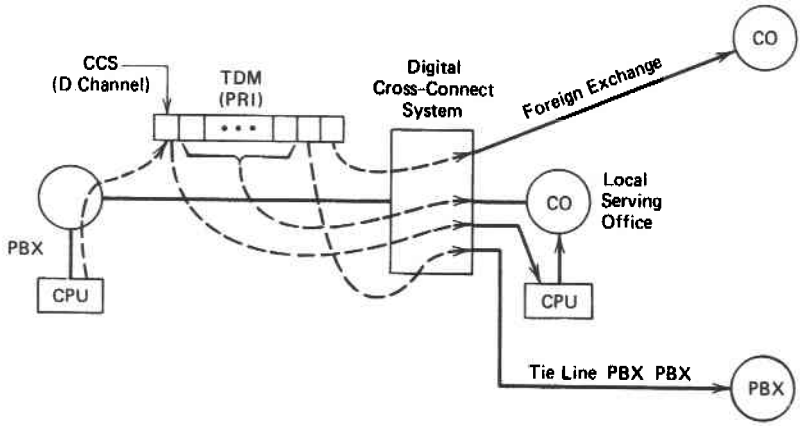


Figure 1.34 TDM link with multiple destinations for the channels.

public network. As indicated, some of the channels compose a trunk group to the local CO but other channels may represent tie lines to other PBXs or foreign exchange circuits to other COs. The digital cross-connect system (DCS) is a specialized switching system (described in Chapter 5) that routes individual circuits within a trunk group to individual locations. If CCS is implemented on the trunk group (as in an ISDN primary rate interface, described in Chapter 11), the signaling information for the latter two trunk groups terminates at the CO. The CO must then forward the information to the respective destinations (assuming those destinations are equipped with CCS). All-in-all these latter two cases are easier to handle if the signaling information accompanies the message channels as it does automatically with in-channel signaling.

As a final note, it should be pointed out that some signaling functions originating or terminating with an end user inherently require in-channel implementations. For example, data terminals wishing to disable echo suppressors or echo cancellers in dial-up connections need to send special tones that get recognized by echo control elements in the circuit. In similar fashion automatic originating facsimile equipment generate 1100-Hz tones to allow automatic recognition of the originating equipment type at the destination.

1.2.11 Analog Interfaces

The design, implementation, and maintenance of any large and complex system require partitioning of the system into subsystems. Associated with each subsystem is an interface that defines the inputs and outputs independent of internal implementations. Well-established interfaces are a fundamental requirement to maintain compatibility between old and new equipment. Within the U.S. telephone network

standardized interfaces are particularly necessary to support competition among equipment suppliers and service providers in almost all facets of the network. The principal *analog* interfaces used in the periphery of the network are subscriber loops, loop-start trunks, ground-start trunks, direct-inward-dial trunks, and E&M trunks.

Subscriber Loop Interface

The most common interface in the network involves the two-wire connection of individual telephone lines to end office switches: the subscriber loop interface. Because of the nature of industry standard telephones (2500 station sets) and the electromechanical switches to which they were originally connected, this interface has a number of characteristics that are difficult to satisfy with modern integrated circuit technology. The fundamental characteristics of this interface are:

1. *Battery*: Application of dc power to the loop (48 V normally) to enable dc signaling and provide bias current for carbon microphones.
2. *Overvoltage Protection*: Protection of equipment and personnel from lightning strikes and power line induction or shorts.
3. *Ringing*: Application of a 20-Hz signal at 86 V rms for ringer excitation. Typical cadence is 2 sec on and 4 sec off.
4. *Supervision*: Detection of off-hook/on-hook by flow/no-flow of dc current.
5. *Test*: Access to the line to test in either direction—toward the subscriber or back into the switch.

In the case of a digital end office, two more functions are necessarily required: two-wire-to-four-wire conversion (*hybrid*) and analog-to-digital *coding* (and digital-to-analog decoding). Taken together, these functions are referred to as BORSCHT. The basic implementation of a BORSCHT interface is described in the context of digital switching in Chapter 5.

Loop-Start Trunks

A loop-start (LS) trunk is a two-wire connection between switches (usually between a central office and a PBX). From an operational point of view, an LS trunk is identical to a subscriber loop. Thus an LS interface in a PBX emulates a telephone by closing the loop to draw current for call origination and by detecting ringing voltages for incoming calls. To send address information, the PBX interface generally waits for a few seconds and assumes that a dial tone is present before sending DTMF tones or generating dial pulses by interrupting loop current. Some PBXs provide dial tone detection so faulty equipment or connections are more easily recognized and addressing can be sent as soon as the other end is ready.

One significant difficulty with two-way LS trunks arises when both ends of the line seize the line at the same time (or nearly the same time). Because both ends of the line think they are originating a call, the line becomes hung. This situation is commonly referred to as glare. If the PBX detects a dial tone before sending digits, it will recognize the glare condition by timing out on the wait for dial tone and can then generate

a disconnect to release the glare condition but drop the incoming call. More commonly, the PBX blindly sends the address digits and connects the originating PBX station to the line. Generally, this means the incoming call gets connected to the wrong station. For this reason, LS trunks are normally used only as one-way trunks: either one-way incoming or one-way outgoing.

Ground-Start Trunks

The aforementioned problem with glare on two-way LS trunks can be largely resolved by augmenting the call origination process to use ground-start (GS) procedures [22]. When originating a call, the end office applies a ground potential to the tip lead of the tip and ring pair and waits for the PBX to acknowledge the seizure by drawing loop current. When the PBX originates a call, it first applies ground to the ring lead and closes the loop waiting for loop current. (The CO does not apply battery during an idle state as it does in an LS interface.) The CO acknowledges the connect request by applying battery to the tip/ring pair and momentarily applying ground to the tip. (Maintaining low-noise circuits requires removal of all paths to ground during the connected state.) A GS protocol prevents simultaneous seizures unless the originations occur within a few hundred milliseconds of each other. In contrast, an LS protocol allows multiple seizures to occur with windows up to 4 sec (the silent interval between ring bursts). Moreover, a glare condition on GS trunks can be recognized by the interface equipment so it can be resolved by redirecting the calls to different trunk circuits.

Another advantage of GS trunks is the ability of the CO to signal network disconnects to the PBX (the CO removes battery). With LS trunks the network does not generally provide disconnect signaling so the PBX must rely on the end user to hang up. (A situation that often produces hung trunks with data connections.) Furthermore, when the CO places an incoming call that eventually gets abandoned, because no one answers, a CO immediately signals the abandonment by removing ground from the tip lead. With LS trunks, abandoned calls can be recognized only by the absence of ring voltage, which can take 6 sec.

Direct-Inward-Dial Trunks

Direct-inward-dial (DID) trunks are particularly simple two-wire trunk interfaces because they are always one-way trunks: incoming only with respect to a PBX. As implied by the name, they allow a serving CO to forward the extension number of incoming calls so a PBX can immediately route the call to a destination without going through an attendant. In contrast to LS and GS trunks, the PBX end of a DID trunk provides battery voltage so the CO can signal an incoming call by merely closing the loop to draw current. After the PBX reverses battery momentarily (winks) to signify it is ready to receive digits, the CO either generates dial pulses or DTMF tones to send the extension number (two, three, or four digits). After the designated station answers, the PBX reverses battery again to signify the connected state and holds that state for the duration of the call. DID trunks are also referred to as "loop reverse-battery supervision" trunks with variations in the signaling protocol depending on the type of CO [22].

E&M Trunks

As indicated in Figure 1.35, an E&M trunk is defined as an interface to a transmission system and not a transmission system itself. The interface in Figure 1.35 has a four-wire voice path, an E lead with an associated return lead (SG), and an M lead with an associated return (SB). Thus there are eight wires in this interface (referred to as a type II E&M interface). Other types of E&M interfaces are defined with as few as four wires [23] (a two-wire voice path, an E lead, and an M lead with earth ground returns).

In any type of E&M interface, supervision signaling is always conveyed on the E and M leads and not on the voice pair (or pairs). The PBX signifies off-hook by closing the M-SB loop to draw current while the transmission equipment indicates off-hook by closing the E-SG loop to draw current. How the transmission equipment conveys the supervision is a function of the transmission link. A variety of timing protocols are defined for the start of address signaling, which can be in-channel DTMF tones or dial pulses generated by momentary opens on the respective E and M leads.

Although E&M signaling is formally defined as just an interface, they are often used (with up to four pairs of wires) as direct connections between PBXs. Because of the requirement for multiple pairs, such applications usually occur when the PBXs are located within a single building or campus complex. The availability of external control leads allows the use of E&M interfaces for special applications such as paging systems, where the M lead can be used to turn on the loudspeaker.

1.2.12 The Intelligent Network

The first common channel signaling facilities of AT&T were installed in the toll network between a No. 4A crossbar switch in Madison, Wisconsin, and a No. 4 ESS switch in Chicago, Illinois, in 1976 [24]. The 2400-bps CCS data links between SPC switching offices replaced in-channel SF/MF signaling facilities to reduce costs and significantly speed up long-distance-call connect times. As more and more of the toll switches were implemented with CCS capabilities, a CCS network evolved that not only improved the performance of the existing network but also established a platform for the introduction of new features. As indicated in Figure 1.36, the CCS network be-

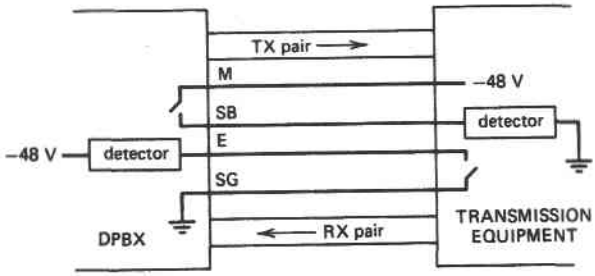


Figure 1.35 Type II E&M interface.

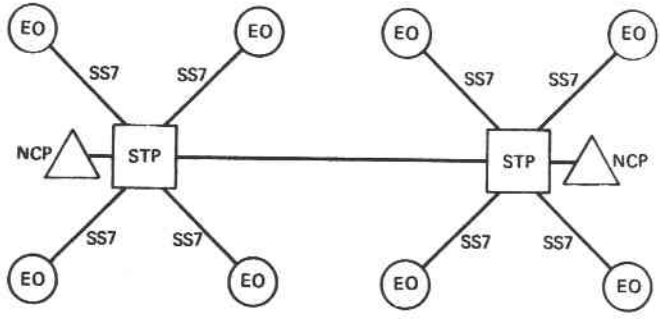


Figure 1.36 CCS network/intelligent network platform.

came a disassociated packet-switching network that separated control of the network from the switching machines themselves.

The packet-switching nodes of the CCS network are referred to as signal transfer points (STPs). The network control points (NCPs) of Figure 1.36 initially represented network database servers that defined how to route calls, verify credit cards, or process special services such as 800 numbers. The same basic structure is also installed within the LATAs to extend CCS features all the way to the end offices. The original communications protocol used between CCS entities was CCITT Signaling System No. 6 (CCS6). In the early 1980s this protocol was replaced by CCITT Signaling System No. 7 (SS7). See reference [25] for a good overview of SS7.

As the capabilities of the NCPs are expanded from being data base servers to processing service requests, the concept of an intelligent network (IN) [26] begins to take shape. In its ultimate form advanced services of the network are executed in the NCPs, or service control points (SCPs), as they are called in the IN, and not in the switching machines themselves. The switching machines provide only the switching fabric; connection commands come from an SCP. The fundamental motivations behind developing the IN are:

1. To enable the deployment of network-based features such as citywide centrex mentioned previously
2. To allow the introduction of new features without having to change the hardware or software of switching vendors
3. To speed up the design/deployment cycle of new features because only the SCP software needs upgrading
4. To allow customizing of services by the operating companies according to the needs of their customers

The obvious disadvantage of such centralized control is the vulnerability of the network to a failure in the SS7 network or the SCP. For this reason basic service is likely to remain with the switching machines. In this case, software in the switching machines recognizes special service situations as software triggers to involve an SCP.

Leaving basic service processing in the switching machines also reduces the load on the SS7 links and the processing speed in the SCPs.

1.2.13 Dynamic Nonhierarchical Routing

The alternate routing procedures shown in Figure 1.4 for the hierarchical network of the Bell System were necessarily simple and straightforward because of the original reliance on mechanical switching machines. Such simplicity, however, leads to the following limitations:

1. Even though the topology of the network would allow a large number of routing alternatives, constraints on how the routes were selected limited the actual choices to a relatively small number.
2. The routing patterns could not be changed according to the time of day or network traffic patterns.
3. The process of progressive setup precluded retracing a path to a previous switching node and trying a new, untried route when a previous choice led to a node that was blocked.
4. A completed connection may have involved a large number of intermediate facilities using a lot of network resources for one connection.

All of these deficiencies resulted from using relatively simple decision logic in each node with no knowledge of the network status as a whole. The introduction of SPC switching machines and CCS changed the situation so that more efficient routine procedures were possible. Along these lines, AT&T completed the deployment of dynamic nonhierarchical routing (DNHR) in 1987 [27, 28]. Specific features of DNHR are:

1. Routing tables in the NCPs list all possible two-link routes in the order of cost so use of network resources are minimized.
2. Routing is dynamic to take advantage of traffic noncoincidence.
3. Blocking at an intermediate node produces "crankback," which allows selection of untried routes from the originating node.
4. Routes that produce too much delay for the echo cancellers in the circuit are excluded.

Reference [28] reports that DNHR reduced blocking on Thanksgiving Day from 34% in 1986 to 3% in 1987. After installation of DNHR Mother's Day traffic saw three to four times as many first-attempt completions as before. The implementation of DNHR demonstrates that a network architecture is as much a function of the software in the network as of the physical topology. It also demonstrates that centralized control of a network is more vulnerable to failures. On January 15, 1990, an obscure bug in the

CCS software of the No. 4 ESS switches in AT&T's long-distance network prevented completion of many calls during a 9-hour period [29]. Ironically, the bug occurred in software intended to isolate the network from a node failure.

The combination of large digital switches, large-bandwidth (fiber optic) transmission links, and development of the SS7 network has stimulated a transition from hierarchical networks to "flat" networks wherein switching nodes are interconnected in a functional mesh—switches are either connected directly to each other or through cross-connect facilities (DCSs and SONET rings described later).

1.2.14 Cellular Radio Telephone System

Prior to 1983 mobile telephone users in the United States were restricted to using the services of radio common carriers (RCCs) that had limited bandwidth and limited resources available for services. The service environment was necessarily one of limited availability and extreme congestion. October 1983 marks a significant date for mobile telephony as the time when commercial cellular mobile telephone service started in Chicago, Illinois. As indicated in Figure 1.37, a cellular system consists of a number of radio base stations connected to a (digital) switching office referred to as the mobile telephone switching office (MTSO). As a mobile subscriber moves from one cell to another, the MTSO automatically switches connections with the respective base stations to maintain a continuous connection with the public switched network. The basic advantages of the cellular architecture are:

1. The ability to reuse allocated channel frequencies within a serving area. Using a combination of antenna directivity in the base stations and signal attenuation

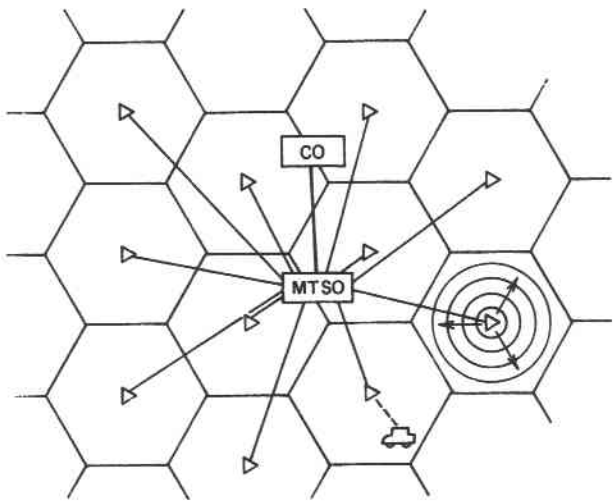


Figure 1.37 Cellular mobile telephone topology.

from distant cells, the same channel can be reused in one of every seven cells [30].

2. Reduced transmit power requirements for the mobiles. The power savings advantage for automobiles is primarily one of reducing the cost of the transmitter. For hand-held units the power savings is important for battery size and time between recharges. In fact, widespread acceptance and use of personal radio systems require much lower transmit power levels and consequently smaller cells [31].
3. Reduced occurrences of multipath propagation. Shorter distances imply less chance of signal reflections causing multipath signal degradation.
4. Expandability. A system can be installed with comparatively large cells to minimize start-up costs. After service requirements grow and revenue is being received, the capacity of the system can be expanded by subdividing the congested cells.
5. Reliability. Because the cells actually have significant overlapping coverage, if one cell fails, neighboring cells can provide service until repairs are made.

The FCC in the United States has defined 728 mobile service areas referred to as cellular geographic service areas (CGSAs). Each CGSA is allocated 832 radio channels, which are equally divided between two competing service providers: one a wireline carrier and the other a nonwireline carrier. The wireline carrier provides local telephone service in the area but competes with the nonwireline carrier (RCC) for cellular service. To ensure effective competition, the wireline carrier must not use facilities that are shared with local telephone service. Specifically, the MTSOs must be separate from the local switching offices. Thus, both types of carriers must backhaul all traffic to their respective MTSOs. The service providers can have more than one MTSO in an area but cannot interconnect them by the switched public network. They are typically interconnected by leased private lines (fiber) or digital microwave. Cellular networks have grown to cover large areas interconnected by dedicated long-distance facilities that allow some carriers to offer free long-distance calling when using a cellular phone.

When a mobile unit first activates itself, it scans the channels to determine which idle channel of a predefined set of control channels has the strongest signal. Using that channel the unit registers with the system to identify itself and place calls. After the initialization process, the network continually monitors signals from the mobile and controls it to switch channels when necessary. To complete calls to a mobile, the cellular network pages for the designated subscriber beginning with its home cell (unless the cellular control center already knows where the subscriber is located). When a subscriber crosses an MTSO boundary, in addition to a cell boundary, a common-channel signaling network is used to transfer the call to the new cell and the new MTSO. The connection to the public network is unchanged. Thus, the original MTSO then becomes a tandem node to the public network.

Examples of analog cellular systems are Advanced Mobile Phone Service (AMPS) [30, 32], developed in the United States and deployed in North America; Total Access

Communications System (TACS), developed in the United Kingdom and deployed in several European countries; and the Nordic Mobile Telephone (NMT) system, deployed in Scandinavia. As discussed in Chapter 9, digital cellular systems have been developed to either replace the analog systems or provide service alternatives.

1.2.15 Voiceband Data Transmission

The primary concern of telephone companies, voice service, is sometimes referred to as POTS: plain old telephone service. In addition to POTS, the telephone companies also provide a number of special services such as program distribution of radio and television, telephoto, teletype, facsimile, and data transmission. In most cases, these services are supported by adapting the sources to use the ubiquitous voice-grade telephone channel. An exception is television distribution, which was an initial stimulus for the nationwide microwave radio network. To a certain extent, wideband FDM voice signals were designed to conform to the network TV distribution facilities. Nationwide TV distribution is now primarily by satellites and fiber.

One significant aspect of a voiceband telephone channel in regards to nonvoice applications is the absence of dc and low-frequency transmission. Although the metallic circuit provided in customer loops passes frequencies down to and including dc, most of the rest of the network does not. The equipments that are primarily responsible for blocking the low frequencies are transformers in two-wire to four-wire hybrids, old FDM separation filters, and digital voice encoders that purposely block low frequencies to avoid 60-Hz hum (that arises from power line interference). Because baseband signals from facsimile and most data applications have frequency content down to dc, these signals must be modulated for voiceband transmission.

A second consideration when using the telephone network for data is the bandwidth restriction of approximately 3 kHz originally imposed by FDM separation filters, loading coils, and more recently band-limiting filters in digital voice terminals (Chapter 3). The main implication of a restricted bandwidth is a limitation on the signaling rate, or baud rate, which in turn directly relates to the data rate. A common signaling rate is 2400 symbols per second using carriers between 1700 and 1800 Hz. Symbol rates of 4800 symbols per second have also been used in lower sideband modems with a carrier at 2850 Hz.

As mentioned, user acceptance of voice quality does not require stringent control of the phase response (envelope delay) of the channel. High-speed data transmission, however, requires comparatively tight tolerance on the phase response to prevent intersymbol interference. Thus analog voice channels used for high-speed data transmission sometimes required special treatment. The special treatment was originally provided at additional monthly rates from the telephone companies in the form of C-type conditioning for leased lines. C-type conditioning was available in several different grades that provided various amounts of control of both phase and amplitude distortion. As digital signal processing technology became commercially viable, the need for signal conditioning became less necessary because modem equalizers accomplished the same function.

Most medium-rate (synchronous) modems include fixed equalization circuitry designed to compensate for the phase distortion in a typical connection. For higher data rates, automatically adjustable equalization is needed. An automatic equalizer first undergoes a training sequence in which the characteristics of the transmission channel are determined by measuring the response to known test signals. Then, equalizing circuitry in the receiver of each modem is adjusted (and continually readjusted during transmission) to provide compensation for the amplitude and phase distortions in the channel. High-speed transmission over dial-up lines requires automatic equalization in the modems since the channel characteristics change with each connection.

Another form of conditioning, referred to as a D-type conditioning, provided lower levels of noise and harmonic distortion than normally provided in leased lines, even with C-type conditioning. D-type conditioning usually did not involve special treatment of any particular line. Instead, the telephone company tested a number of different circuits until one with suitable quality was found. Sometimes this meant avoiding a cable that included pairs from a noisy switching office. Unlike voice, which is reasonably tolerant of impulse noise, a data circuit is more susceptible to impulse noise than to the normal background (white) noise. The elimination of older equipment (e.g., step-by-step switches) has eliminated much of the impulse noise in the network.

As more and more of the public network became digitized, fewer and fewer facilities were analog and, hence, fewer analog impairments were likely to occur in a connection. By the early 1990s virtually all of the internal portions of the public telephone network had been digitized. The only remaining segments of a connection that were analog were the subscriber loops and the interfaces in the associated end offices. The principal source of impairment for voiceband modems then became the analog-to-digital (A/D) conversion equipment in the line interfaces. Recognition of this situation completely eliminated the need for any type of conditioning on leased lines (which were often digital and not analog) and allowed for the development of V.90 modems described in Chapter 11.

An important consideration for data transmission over long-distance circuits in the analog network was the effect of echo suppressors. As mentioned in Section 1.2.8, an echo suppressor blocked the signal in a return path when the corresponding forward path of a four-wire circuit was active. Thus, operative echo suppressors effectively precluded a full-duplex operation. Even in a half-duplex mode of operation the echo suppressors might require 100 msec of deactivation time to reverse the direction of propagation. For these reasons the common carriers provided a means of disabling the echo suppressors by using an in-channel control signal from the data terminals. Echo suppressors were disabled by transmitting a pure tone between 2010 and 2240 Hz for 400 msec. The echo suppressors remained disabled as long as there was continuous energy in the channel. Thus, the modems could switch to signaling frequencies and begin full-duplex data transmission after the suppressors were first disabled. If the energy in the channel was removed for 100 msec, the echo suppressors were reactivated. Hence, rapid line turnaround was required for half-duplex modems utilizing the entire bandwidth for each direction of transmission. Contemporary modems have echo can-

cellers for full-duplex transmission and therefore automatically disable echo suppressors (which are probably nonexistent).

1.3 THE INTRODUCTION OF DIGITS

Voice digitization and transmission first became feasible in the late 1950s when solid-state electronics became available. In 1962 Bell System personnel established the first commercial use of digital transmission when they began operating a T1 carrier system for use as a trunk group in a Chicago area exchange [33]. After the T1 system a family of T-carrier systems (T1, T1C, T1D, T2, T3, T4) were developed, all of which involved time division multiplexing of digitized voice signals.

The world's first commercially designed digital microwave radio system was established in Japan by Nippon Electric Company (NEC) in 1968 [34]. In the early 1970s digital microwave systems began to appear in the United States for specialized data transmission services. The first digital microwave link in the U.S. public telephone network was supplied by NEC of Japan for a New York Telephone link between Brooklyn and North Staten Island in 1972 [34]. Digital microwave systems were subsequently developed and installed by several U.S. manufacturers for use in intermediate-length toll and exchange area circuits.

Bell System's first commercial use of digital fiber optic transmission occurred in September of 1980 on a short-haul route between Atlanta and Smyrna, Georgia [35]. Three years later the first long-haul system between New York and Washington, D.C., was put into service.

In addition to transmission systems, digital technology has proven to be equally useful for implementing switching functions. The first country to use digital switching in the public telephone network was France in 1970 [36]. The first application of digital switching in the public network of the United States occurred in early 1976 when Bell System began operating its No. 4ESS [37] in a class 3 toll office in Chicago. Two months later Continental Telephone Company began operation in Ridgecrest, California, of a digital toll switch [38]. The first digital end office switch in the United States became operational in July of 1977 in the small town of Richmond Hill, Georgia [39].

1.3.1 Voice Digitization

The basic voice-coding algorithm used in T-carrier systems and most other digital voice equipment in telephone networks around the world is shown in Figure 1.38. The first step in the digitization process is to periodically sample the waveform. As discussed at length in Chapter 3, all of the information needed to reconstruct the original waveform is contained in the samples if the samples occur at an 8-kHz rate. The second step in the digitization process involves quantization: identifying which amplitude interval of a group of adjacent intervals a sample value falls into. In essence the quantization process replaces each continuously variable amplitude sample with a dis-

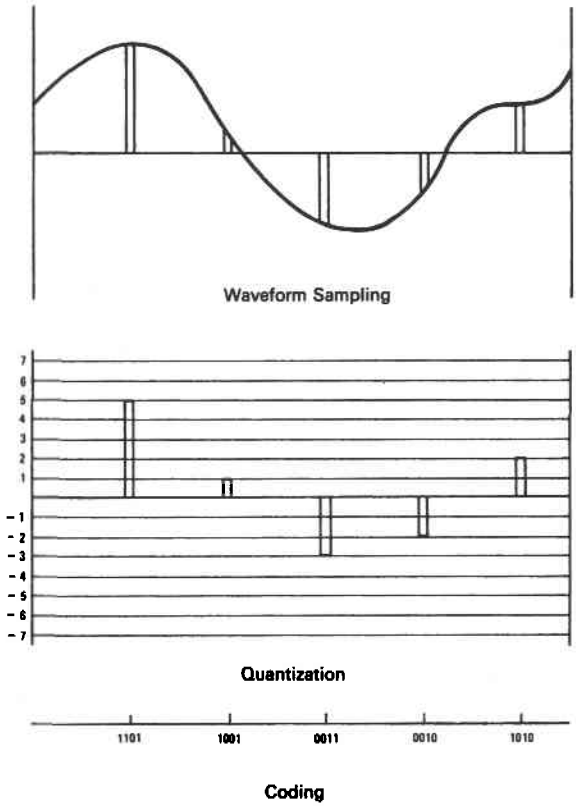


Figure 1.38 Voice digitization process.

crete value located at the middle of the appropriate quantization interval. Since the quantized samples have discrete levels, they represent a multiple-level digital signal.

For transmission purposes the discrete amplitude samples are converted to a binary codeword. (For illustrative purposes only, Figure 1.38 shows 4-bit codewords.) The binary codes are then transmitted as binary pulses. At the receiving end of a digital transmission line the binary data stream is recovered, and the discrete sample values are reconstructed. Then a low-pass filter is used to "interpolate" between sample values and re-create the original waveform. If no transmission errors have occurred, the output waveform is identical to the input waveform except for quantization distortion: the difference between a sample value and its discrete representation. By having a large number of quantization intervals (and hence enough bits in a codeword to encode them), the quantization intervals can be small enough to effectively eliminate perceptible quantization effects.

It is worth noting that the bandwidth requirements of the digital signal increase as a result of the binary encoding process. If the discrete, multiple-amplitude samples are transmitted directly, the bandwidth requirements are theoretically identical to the

bandwidth of the original signal. When each discrete sample is represented by a number of individual binary pulses, the signal bandwidth increases accordingly. The two-level pulses, however, are much less vulnerable to transmission impairments than are the multiple-amplitude pulses (or the underlying analog signal).

1.3.2 Time Division Multiplexing

Basically, time division multiplexing (TDM) involves nothing more than sharing a transmission medium by establishing a sequence of time slots during which individual sources can transmit signals. Thus the entire bandwidth of the facility is periodically available to each source for a restricted time interval. In contrast, FDM systems assign a restricted bandwidth to each source for all time. Normally, all time slots of a TDM system are of equal length. Also, each subchannel is usually assigned a time slot with a common repetition period called a frame interval. This form of TDM (as shown in Figure 1.39) is sometimes referred to as synchronous time division multiplexing to specifically imply that each subchannel is assigned a certain amount of transmission capacity determined by the time slot duration and the repetition rate. In contrast, another form of TDM (referred to as *statistical*, or *asynchronous*, *time division multiplexing*) is described in Chapter 10. With this second form of multiplexing, subchannel rates are allowed to vary according to the individual needs of the sources. The backbone digital links of the public telephone network (T-carrier, digital microwave, and fiber optics) use a synchronous variety of TDM.

Time division multiplexing is normally associated only with digital transmission links. Although analog TDM transmission can be implemented by interleaving samples from each signal, the individual samples are usually too sensitive to all varieties of transmission impairments. In contrast, time division switching of analog signals is more feasible than analog TDM transmission because noise and distortion within the switching equipment are more controllable. As discussed in Chapter 5, analog TDM techniques have been used in some PBXs before digital electronics became so inexpensive that the digitization penalty disappeared.

T-Carrier Systems

The volume of interoffice telephone traffic in the United States has traditionally grown more rapidly than local traffic. This rapid growth put severe strain on the older

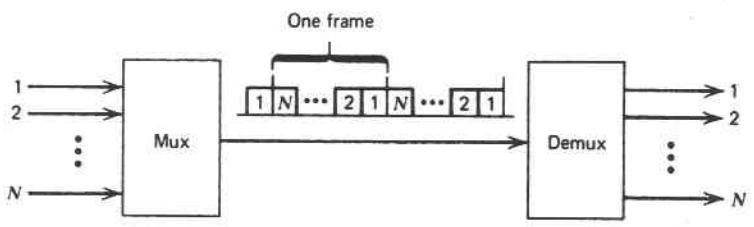


Figure 1.39 Time division multiplexing.

interoffice transmission facilities that are designed for lower traffic volumes. Telephone companies were often faced with the necessary task of expanding the number of interoffice circuits. T-carrier systems were initially developed as a cost-effective means for interoffice transmission: both for initial installations and for relief of crowded interoffice cable pairs.

Despite the need to convert the voice signals to a digital format at one end of a T1 line and back to analog at the other, the combined conversion and multiplexing cost of a digital TDM terminal was lower than the cost of a comparable analog FDM terminal. The first T-carrier systems were designed specifically for exchange area trunks at distances between 10 and 50 miles.

A T-carrier system consists of terminal equipment at each end of a line and a number of regenerative repeaters at intermediate points in the line. The function of each regenerative repeater is to restore the digital bit stream to its original form before transmission impairments obliterate the identity of the digital pulses. The line itself, including the regenerative repeaters, is referred to as a span line. The original terminal equipment was referred to as D-type (digital) channel banks, which came in numerous versions. The transmission lines were wire pairs using 16- to 26-gauge cable. A block diagram of a T-carrier system is shown in Figure 1.40.

The first T1 systems used D1A channel banks for interfacing, converting, and multiplexing 24 analog circuits. A channel bank at each end of a span line provided interfacing for both directions of transmission. Incoming analog signals were time division multiplexed and digitized for transmission. When received at the other end of the line, the incoming bit stream was decoded into analog samples, demultiplexed, and filtered to reconstruct the original signals. Each individual TDM channel was assigned 8 bits per time slot. Thus, there were (24)(8) = 192 bits of information in a frame. One additional bit was added to each frame to identify the frame boundaries, thereby producing a total of 193 bits in a frame. Since the frame interval is 125 μsec, the basic T1 line rate became 1.544 Mbps. This line rate has been established as the fundamental standard for digital transmission in North America and Japan. The standard is referred to as a DS1 signal (for digital signal 1).

A similar standard of 2.048 Mbps has been established by ITU-T for most of the rest of the world. This standard evolved from a T1-like system that provides 32 channels at the same rate as the North American channels. Only 30 of the channels in the E1 standard, however, are used for voice. The other two are used for frame synchronization and signaling. Signaling and control information for T1 systems are inserted into each voice channel (or transmitted separately by CCIS facilities). Digital signaling and control techniques for both systems are discussed in Chapter 7.

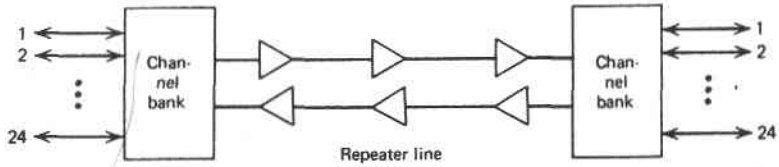


Figure 1.40 T1-carrier system.

The greatly increased attenuation of a wire pair at the frequencies of a DS1 signal (772 kHz center frequency) mandates the use of amplification at intermediate points of a T1 span line. In contrast to an analog signal, however, a digital signal can not only be amplified but also be detected and regenerated. That is, as long as a pulse can be detected, it can be restored to its original form and relayed to the next line segment. For this reason T1 repeaters are referred to as regenerative repeaters. The basic functions of these repeaters are:

1. Equalization
2. Clock recovery
3. Pulse detection
4. Transmission

Equalization is required because the wire pairs introduce certain amounts of both phase and amplitude distortion that cause intersymbol interference if uncompensated. Clock recovery is required for two basic purposes: first, to establish a timing signal to sample the incoming pulses; second, to transmit outgoing pulses at the same rate as at the input to the line.

Regenerative repeaters are normally spaced every 6000 ft in a T1 span line. This distance was chosen as a matter of convenience for converting existing voice frequency cables to T-carrier lines. Interoffice voice frequency cables typically used loading coils that were spaced at 6000-ft intervals. Since these coils were located at convenient access points (manholes) and had to be removed for high-frequency transmission, it was only natural that the 6000-ft interval be chosen. One general exception is that the first regenerative repeater is typically spaced 3000 ft from a central office. The shorter spacing of this line segment was needed to maintain a relatively strong signal in the presence of impulse noise generated by older switching machines.

The operating experience of T1 systems was so favorable that they were continually upgraded and expanded. One of the initial improvements produced T1C systems that provide higher transmission rates over 22-gauge cable. A T1C line operates at 3.152 Mbps for 48 voice channels, twice as many as a T1 system.

Another level of digital transmission became available in 1972 when the T2 system was introduced. This system was designed for toll network connections. In contrast, T1 systems were originally designed only for exchange area transmission. The T2 system provided for 96 voice channels at distances up to 500 miles. The line rate was 6.312 Mbps, which is referred to as a DS2 standard. The transmission media was special low-capacitance 22-gauge cable. By using separate cables for each direction of transmission and the specially developed cables, T2 systems could use repeater spacings up to 14,800 ft in low-noise environments. The emergence of optical fiber systems made copper-based T2 transmission systems obsolete.

TDM Hierarchy

In a manner analogous to the FDM hierarchy, AT&T established a digital TDM hierarchy that has become the standard for North America. Starting with a DS1 signal as

TABLE 1.10 Digital TDM Signals of North America and Japan

Digital Signal Number	Number of Voice Circuits	Multiplexer Designation	Bit Rate (Mbps)	Transmission Media
DS1	24	D channel bank (24 analog inputs)	1.544	T1 paired cable
DS1C	48	M1C (2 DS1 inputs)	3.152	T1C paired cable
DS2	96	M12 (4 DS1 inputs)	6.312	T2 paired cable
DS3	672	M13 (28 DS1 inputs)	44.736	Radio, Fiber
DS4	4032	M34 (6 DS3 inputs)	274.176	T4M coax, WT4 waveguide, radio

a fundamental building block, all other levels are implemented as a combination of some number of lower level signals. The designation of the higher level digital multiplexers reflects the respective input and output levels. For example, an M12 multiplexer combines four DS1 signals to form a single DS2 signal.* Table 1.10 lists the various multiplex levels, their bit rates, and the transmission media used for each. Notice that the bit rate of a high-level multiplex signal is slightly higher than the combined rates of the lower level inputs. The excess bits are included for certain control and synchronization functions discussed in Chapter 7. A similar digital hierarchy has also been established by ITU-T as an international standard. As shown in Table 1.11, this hierarchy is similar to the North American standard but involves different numbers of voice circuits at all levels.

Digital Pair-Gain Systems

Following the successful introduction of T1 systems for interoffice trunks, most major manufacturers of telephone equipment developed digital TDM systems for local distribution. These systems are most applicable to long rural loops where the cost of the electronics is offset by the savings in wire pairs. No matter what the distance is, unexpected growth can be most economically accommodated by adding electronics, instead of wire, to produce a pair-gain system. The possibility of trailer parks, apartment houses, or Internet service providers springing up almost overnight causes nightmares in the minds of cable plant forecasters. Pair-gain systems provide a networking alternative to dispel those nightmares.

Digital pair-gain systems are also useful as alternatives to switching offices in small communities. Small communities are often serviced by small automatic switching systems normally unattended and remotely controlled from a larger switching office nearby. These small community switches are referred to as community dial offices

*Because T2 transmission systems have become obsolete, the M12 function exists only in a functional sense within M13 multiplexers, which multiplex 28 DS1 signals into 1 DS3 signal.

TABLE 1.11 ITU Digital Hierarchy

Level Number	Number of Voice Circuits	Multiplexer Designation	Bit Rate (Mbps)
E1	30		2.048
E2	120	M12	8.448
E3	480	M23	34.368
E4	1920	M34	139.264
E5	7680	M45	565.148

(CDOs). A CDO typically provides only limited service features to the customers and often requires considerable maintenance. Because digital pair-gain systems lower transmission costs for moderate-sized groups of subscribers, they are a viable alternative to a CDO: Stations in the small community are serviced from the central office by way of pair-gain systems. A fundamental consideration in choosing between pair-gain systems and remote switching involves the traffic volumes and calling patterns within the small community. The basic techniques of analyzing traffic patterns and determining trunk group sizes are provided in Chapter 12.

The first two digital pair-gain systems used in the Bell System were the subscriber loop multiplex (SLM) system [40, 41] and, its successor, the subscriber loop carrier (SLC-40) system [40, 42]. Although these systems used a form of voice digitization (delta modulation) different from that used in T-carrier systems (pulse code modulation), they both used standard T1 repeaters for digital transmission at 1.544 Mbps. Both systems also converted the digitized voice signals back into individual analog interfaces at the end office switch to achieve system transparency. Figures 1.41 and 1.42 show block diagrams of these systems. Notice that the SLM system provided both concentration and multiplexing (80 subscribers for 24 channels) while the SLC-40 was strictly a multiplexer (40 subscribers assigned in a one-to-one manner to 40 channels).

The SLM and SLC-40 systems used delta modulation voice coding because it was simpler than pulse code modulation as used in T1 systems and was therefore less costly to implement on a per-channel basis—a desirable feature for modular system implementations. The original T1 systems, on the other hand, minimized electronics costs by using common encoders and decoders, which precluded implementation of

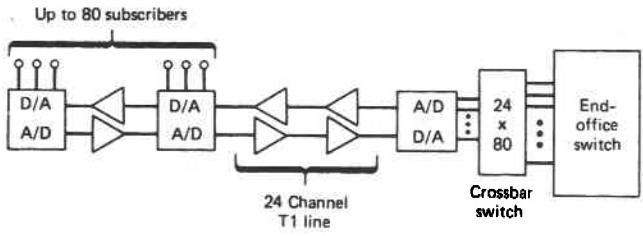


Figure 1.41 Subscriber loop multiplexer.

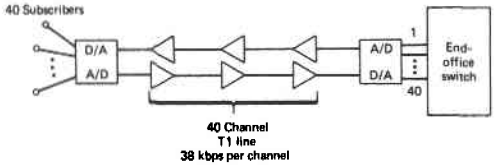


Figure 1.42 Subscriber loop carrier (SLC-40).

less than 24 channels (an unnecessary feature in an interoffice application). By the late 1970s low-cost, integrated circuit implementations of standard pulse code modulation became available that led the way to the first (1979) installation of the SLC-96, a subscriber carrier system using voice coding that was compatible with T1 systems and the emerging digital end office switching machines [43].

The SLC-96 system (which is functionally equivalent to four T1 lines) can interface directly with a digital end office and not be demultiplexed into 24 distinct analog interfaces. Thus this capability, which is referred to as integrated digital loop carrier (IDLC), greatly reduces the prove-in distance where the digital carrier becomes less expensive than separate subscriber pairs. Subsequent enhancements of the SLC-96 system include use of fiber transmission, support for local switching functions in the remote module, and configurability of analog interfaces from the central office [44, 45]. Many newer copper- and fiber-based digital loop carrier systems have been developed. These systems are discussed more thoroughly in Chapter 11 in the context of digital subscriber access.

1.3.3 Data under Voice

After the technology of T-carrier systems had been established, AT&T began offering leased digital transmission services for data communications. This service, known as Dataphone Digital Service (DDS), uses T1 transmission links with special terminals (channel banks) that provide direct access to the digital line. An initial drawback of DDS arose because T-carrier systems were originally used only for exchange area and short toll network trunks. Without some form of long-distance digital transmission, the digital circuits in separate exchange areas could not be interconnected. AT&T's original response to long-distance digital transmission was the development of a special radio terminal called the 1A radio digital terminal (1A-RDT) [46]. This terminal encoded one DS1 signal (1.544 Mbps) into less than 500 kHz of bandwidth. As shown in Figure 1.43, a signal of this bandwidth was inserted below the lowest frequency of a master group multiplex (Table 1.5). Since this frequency band is normally unused in TD or TH analog radio systems, the DS1 signal could be added to existing analog routes without displacing any voice channels. The use of frequencies below those used for voice signals leads to the designation "data under voice" (DUV).

It is important to point out that DUV represented a special development specifically intended for data transmission and not for voice services. In fact, DUV was used only to provide long-distance digital transmission facilities for DDS. The emergence of

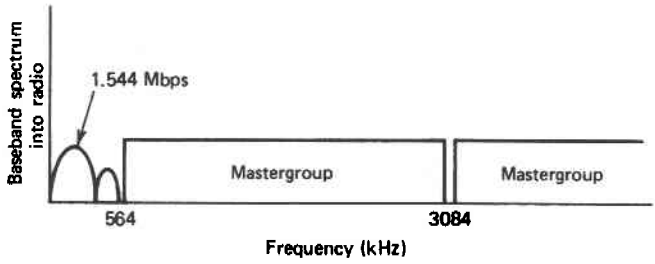


Figure 1.43 Data under voice.

long-distance fiber systems obviously eliminated the need for DUV equipment (and even the analog radios themselves).

1.3.4 Digital Microwave Radio

In contrast to DUV systems, which piggyback 1.544 Mbps onto an analog radio for data services, a common-carrier digital microwave system uses digital modulation exclusively to transmit and receive higher level digital multiplex signals for voice traffic. Digital radios use the same frequency bands allocated for analog radios as listed in Table 1.7. Thus a primary design requirement of a digital radio was that it must confine its radiated power to prevent excessive interference into adjacent, possibly analog, channels. Moreover, the FCC stipulated that the digital radios had to be capable of providing roughly the same number of voice circuits as existing analog radios. Table 1.12 lists the minimum number of voice circuits and the resulting bit rate that must be provided by digital radios in each of the common-carrier microwave bands [47].

Despite the design constraints imposed by compatibility requirements with analog radios, digital radios proved to be more economical than analog radios in several applications. Because of lower terminal costs (multiplexing and demultiplexing), digital radio systems were generally less expensive than analog radio systems for distances

TABLE 1.12 Minimum Voice Circuit Requirements of Digital Radios in the United States

Frequency Band (MHz)	Minimum Number of Circuits	Equivalent Number of DS1 Signals	Resultant Bit Rate (Mbps)[a]	Channel Bandwidth (MHz)
2110–2130	96	4	6.144	3.5
2160–2180	96	4	6.144	3.5
3700–4200	1152	48	73.7	20
5925–6425	1152	48	73.7	30
10,700–11,700	1152	48	73.7	40

[a]The actual bit rate is usually slightly greater owing to extra framing and synchronization bits. Furthermore, most radio systems provide significantly more voice circuits than the minimum.

up to about 300 miles and on longer routes requiring channel access (drop-and-insert) at intermediate points in the route [48]. The major impetus for digital radio in the United States was the emergence of digital toll switches like the No. 4ESS. The interconnection of digital radio signals to digital switches avoided costly lower level multiplex equipment.

1.3.5 Fiber Optic Transmission

Of all the new technology introduced into the telephone network, fiber has certainly had the most profound effect. Prior to its emergence, transmission engineers would have considered the combination of extremely wide bandwidth, extremely low attenuation, and immunity from interference as something akin to perpetual motion. Low attenuation allows long repeater spacings, which equates to low maintenance costs—the reason for wholesale replacement of long-haul analog radio systems. On December 5, 1986, AT&T completed the last section of a transcontinental fiber system [49]. By the end of that decade virtually all of the high-density routes were converted to fiber. Radio systems are and will continue to be used for carrying low-density traffic where right-of-way costs are a dominant consideration. However, because these routes are relatively short spurs off of digital fiber arteries, interface costs imply digital implementations, thus completing the economic foundation for the all-digital network.

As discussed in Chapter 8, an optical fiber is not an inherently digital transmission system. However, the interface electronics (drivers and receivers) and the application of connecting to digital switches have stimulated their use of digital transmission.* Furthermore, considering the amount of bandwidth available there is little incentive to use analog transmission to conserve bandwidth, as was the original situation with microwave radios.

1.3.6 Digital Switching

The original research into digital switching at Bell Laboratories was reported by Earle Vaughan in 1959 [50]. Laboratory models were developed to demonstrate the concept of integrating digital time division multiplex transmission systems with time division switching systems. Unfortunately the necessary solid-state electronics had not matured sufficiently at that time so commercial development of digital switching was not pursued, and development of the No. 1 ESS continued along the lines of space division electromechanical technology. Almost 10 years later, however, Bell Labs began development of a digital toll switch, the No. 4 ESS.

When placed in service in January 1976, the No. 4 ESS provided several new capabilities for the toll network. First, it was the first toll switch to be designed for stored-program control at the outset.† Second, its capacity was three times that of the prevailing electromechanical switch at the time: the No. 4A cross bar. The larger ca-

*Analog optical fiber systems have been widely used to carry television signals in feeders segments of CATV systems.

†Stored-program control was first implemented in the toll network beginning in 1969 by retrofitting crossbar switches [51].

pacity of the No. 4 ESS meant that many metropolitan areas could consolidate toll traffic into one switch instead of several. Third, the digital time division design of the No. 4 ESS allowed direct connection to digital T-carrier lines. This last feature illustrates the original attraction of digital switching for toll and tandem offices of the network. By 1976, when the first No. 4 ESS was installed, it was clear that digital transmission was dominating the exchange area and shorter intertoll trunks. Thus significant economies and improvements in transmission quality resulted by eliminating channel banks at the interface between digital trunks and a switching system.

The early development of digital end office switches in the United States was undertaken by independent equipment manufacturers with the first system being placed in service in 1977. These systems were primarily designed for the smaller switching offices of the independent telephone companies. Digital switches were particularly attractive to rural telephone companies because they could provide significant copper savings when directly connected to digital pair-gain transmission systems. The first large digital end office switching system to be introduced into the North American network was the DMS-100 provided by Northern Telecom. Table 1.13 lists digital switching machines developed for the North American public telephone network.

The functional essence of a digital time division switching matrix is illustrated in Figure 1.44. As indicated, all inputs are time division multiplex links. These links may represent digital pair-gain systems, T-carrier interoffice trunks, or the outputs of colocated channel banks used to interface analog lines to the digital switch. In any event the switching matrix itself is designed to service only TDM input links.

Basically, the switching matrix is required to transfer information arriving in a specified time slot (channel) on an incoming TDM link to a specified time slot on an

TABLE 1.13 Digital Central Office Switching Systems of North America

Manufacturer	Designation	Date of Introduction	Application	Line Size
AT&T	4 ESS	1976	Toll	107,000
AT&T	5 ESS	1983	Local	100,000
CIT-Alcatel	E 10-five	1982	Local	100,000
GTE	3 EAX	1978	Toll/tandem	60,000
GTE	5 EAX	1982	Local	145,000
LM Ericsson	AXE10	1978	Local/toll	200,000
NEC	NEAX-61	1979	Local/toll	80,000
ITT	System 1210	1978	Local/toll	26,000
Northern Telecom	DMS-10	1977	Local	7,000
Northern Telecom	DMS-100	1979	Local	100,000
Northern Telecom	DMS-200	1978	Toll	60,000
Siemens	EWSD	1981	Local	200,000
Stromberg Carlson	DC0	1977	Local	32,000
Vidar	ITS4	1977	Toll/tandem	7,000
Vidar	ITS4/5	1978	Local/toll	12,768

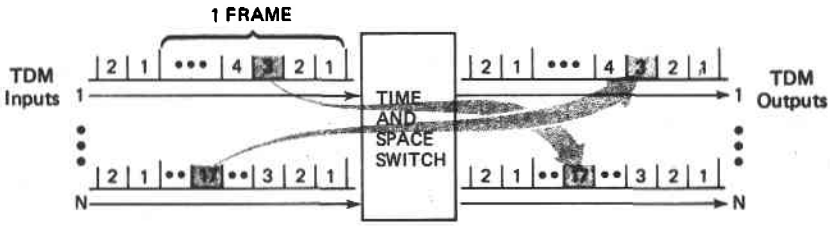


Figure 1.44 Digital time division switching operation.

outgoing TDM link. Since an arbitrary connection involves two different physical links and two different time slots, the switching process requires spatial translation (space switching) and time translation (time switching). Thus the basic operation is sometimes referred to as two-dimensional switching. Space switching is achieved with conventional digital logic selector circuits, and time switching is achieved by temporarily storing information in a digital memory or register circuit.

For convenience, the inputs to the switching matrix of Figure 1.44 are all shown on the left, and the outputs are on the right. Since the links are inherently four-wire, each input time slot is paired with the same time slot on a corresponding output link. Hence, a full-duplex circuit also requires a "connection" in the reverse direction that is achieved by transferring information from time slot 17 of link N to time slot 3 of link 1 in the example shown. Operational and implementation details of a large range of digital time division switches are provided in Chapter 5.

1.3.7 Digital Network Evolution

The evolution of the analog telephone network into one that is all digital except for the access lines is summarized in Figure 1.45. The process began in the 1960s (a) with T1 systems being installed on relatively short haul interoffice trunks within the exchange areas. Next, in the early 1970s (b), digital transmission was introduced into the toll network with T2 systems for relatively short routes between toll offices. It was in the late 1970s (c) that digitization really began to take over. T1 coverage expanded greatly, digital loop carrier (DLC) systems came into use,* and digital switches became available at all levels of the network: PBXs (DPBXs), end offices (DEOs), tandem offices, and toll offices (DTOs). Moreover, microwave digital radios (MDRs) proved to be advantageous to use in both the exchange areas and the shorter toll network routes due to low interface costs to digital switches. Thus, the late 1970s produced a number of integrated islands where digital switches within a region were interconnected by digital transmission links but there was little digital connectivity between the islands. (Data under voice was installed as an overbuild to analog routes for

*Digital subscriber carrier systems were actually introduced in the early 1970s, but these systems utilized a voice digitization technology (delta modulation) that was incompatible with the rest of the network and therefore did not figure into the integrated network. These early digital loop carrier systems are now obsolete.

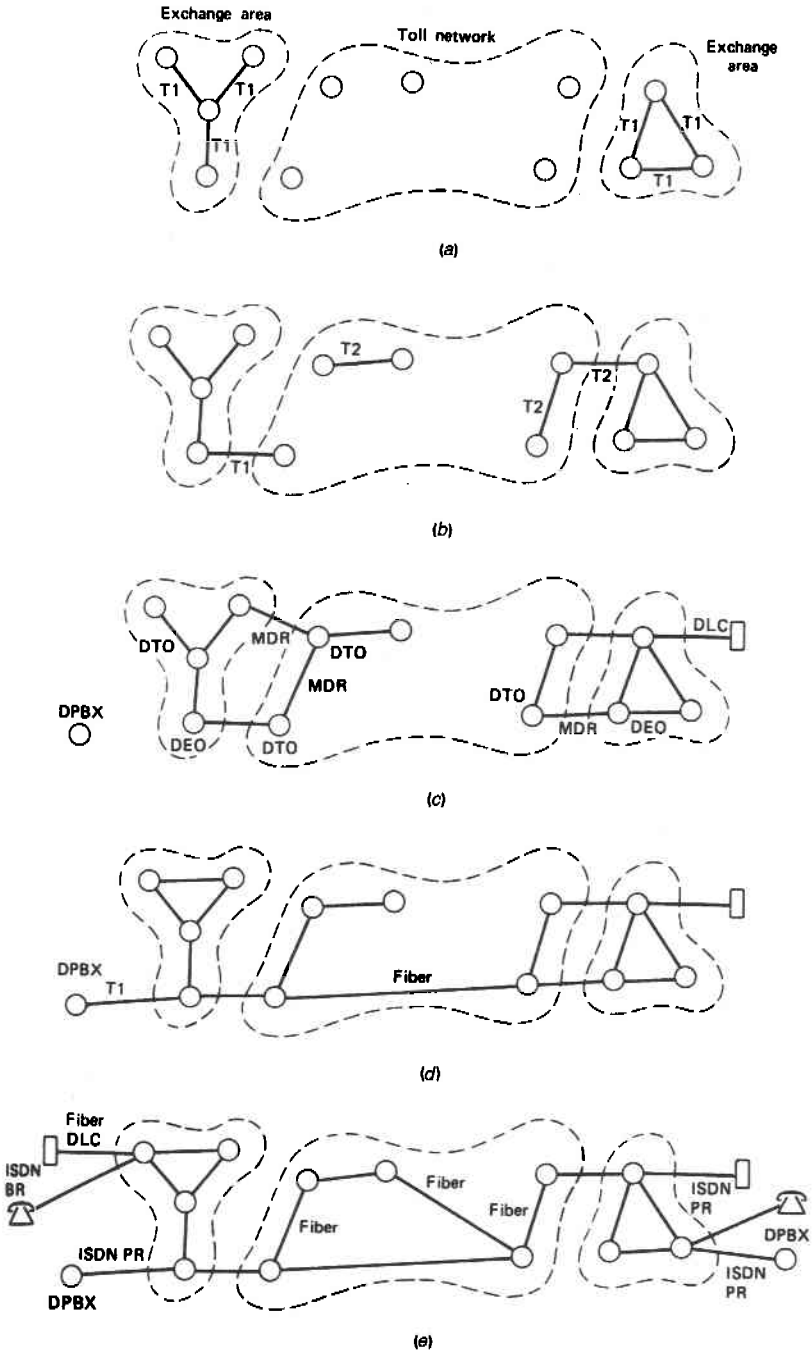


Figure 1.45 Digital network evolution.

limited-capacity data services.) A fully integrated and interconnected digital network became a reality in the early 1980s (d) when fiber optic transmission emerged as the technology of choice for high-density long-haul routes. Digital connectivity to business customer premises equipment also occurred in this time frame as T1 became the preferred voice trunk interface for large PBXs.

End-to-end digital connectivity for voice or data services became a reality in the late 1980s (e) with the introduction of ISDN basic rate (ISDN BR, 2B + D) and ISDN primary rate (ISDN PR, 23B + D) digital connections to the customer. In addition, fiber technology became much more ubiquitous as DS3 rate systems eliminated T2 systems in the toll network and fiber-based systems became the preferred technology for loop carrier and feeder systems—even at relatively short distances.

REFERENCES

1 S. P. Thompson, *Phillip Reis: Inventor of the Telephone*, Sylvanus P. Thompson, London, New York, 1883.

2 R. Smith and W. Sonnerville, "Record '81 Telco Construction Set at $23.5 Billion, Up 5.6%," *Telephone Engineer and Management*, Jan. 15, 1981, p. 55.

3 Technical Staff, Bell Laboratories, *Engineering and Operations in the Bell System*, 2nd ed., Bell Telephone Laboratories, Indianapolis, Indiana, 1983.

4 D. M. Mardon, "Plan for Access—Tandem and Trunking," *Telephone Engineer & Management*, Nov. 15, 1983, pp. 140–142.

5 W. D. Reeve, *Subscriber Loop Signaling and Transmission Handbook*, IEEE Press, New York, 1995.

6 S. J. Campanella, "Digital Speech Interpolation Systems," in *Advanced Digital Communications, Systems and Signal Processing Techniques*, K. Fehrer, Ed., Prentice-Hall, Englewood Cliffs, NJ, 1985.

7 Technical Staff, Bell Laboratories, *Engineering and Operations in the Bell System*, Bell Telephone Laboratories, Indianapolis, Indiana, 1977.

8 W. L. Pritchard, "The History and Future of Commercial Satellite Communications," *IEEE Communications Magazine*, May 1984, pp. 22–37.

9 Digital Video Broadcasting (DVB), Framing structure, channel coding and modulation for digital terrestrial television, EN 300 744, (ETSI TM 1545), 1999.

10 I. Boyd, C. B. Southcott, D. P. Crowe, and P. J. Bolingbroke, "Spech Codec for the Skyphone Aeronautical Telephone Service," *British Telecommunications Engineering*, July 1986, pp. 83–91.

11 A. J. Aikens and D. A. Lewinski, "Evaluation of Message Circuit Noise," *Bell System Technical Journal*, July 1970, pp. 879–909.

12 J. L. Sullivan, "A Laboratory System for Measuring Loudness Loss of Telephone Connections," *Bell System Technical Journal*, Oct. 1971, pp. 2663–2739.

13 W. K. Macadam, "A Basis for Transmission Performance Objective in a Telephone Communications System," *AIEE Transactions on Communications and Electronics*, May 1978, pp. 205–209.

14 F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," *Bell System Technical Journal*, Apr. 1971, pp. 1311–1347.

15 W. R. Bennett and J. R. Davey, *Data Transmission*, McGraw-Hill, New York, 1965.

16 Members of Technical Staff, Bell Telephone Laboratories, *Transmission Systems for Communications*, 4th ed., Western Electric Company, Winston-Salem, North Carolina, 1971.

17 H. R. Huntley, "Transmission Design of Intertoll Telephone Trunks," *Bell System Technical Journal*, Sept. 1953, pp. 1019–1036.

18 K. Murano, S. Unagami, and F. Amano, "Echo Cancellation and Applications," *IEEE Communications Magazine*, Jan. 1990, pp. 49–55.

18a E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publisher, Norwell, MA, 1993.

19 "A Technical Report on Echo Cancelling," Report No. 27, ANSI T1A1.6 Working Group on Specialized Signal Processing, Nov. 1993.

20 H. G. Suyderhoud, M. Onufry, and S. J. Campanella, "Echo Control in Telephone Communications," *National Telecommunications Conference Record*, 1976, pp. 8.1-1–8.1-5.

21 F. T. Andrews, Jr. and R. W. Hatch, "National Telephone Network Transmission Planning in the American Telephone and Telegraph Company," *IEEE Transactions on Communications Technology*, June 1971, pp. 302–314.

22 *Compatibility Information for Telephone Exchange Service, Bellcore Technical Advisory TA-NPL-000912*, Bell Communications Research, Issue 1, Feb. 1989.

23 *Notes on THE NETWORK*, American Telephone and Telegraph Company, Network Planning Division, AT&T, New York, 1980, pp. 44–58.

24 "Special Issue: Stored Program Control Network," *Bell System Technical Journal*, Sept. 1982.

25 P. J. Kuhn, C. D. Pack, and R. A. Skoog, "Common Channel Signaling Networks: Past, Present, Future," *IEEEE Journal on Selected Areas in Communications*, Apr. 1994, pp. 383–393.

26 J. S. Doyle and C. S. McMahon, "The Intelligent Network Concept," *IEEE Transactions on Communications*, Dec. 1988, pp. 1296–1301.

27 J. M. Akinpelu, "The Overload Performance of Engineered Networks with Nonhierarchical and Hierarchical Routing," *Bell System Technical Journal*, Sept. 1984, pp. 1261–1281.

28 G. R. Ash and E. Oberer, "Dynamic Routing in the AT&T Network—Improved Service Quality at Lower Cost," *IEEE Globecom Proceedings*, 1989, pp. 9.1.1–9.1.6.

29 C. Mason, "Software Problem Cripples AT&T Long-Distance Network," *Telephony*, Jan. 22, 1990, pp. 10–11.

30 "Special Issue on Advanced Mobile Phone Service (AMPS)," *Bell System Technical Journal*, Jan. 1979.

31 D. C. Cox, "Portable Digital Radio Communications—An Approach to Tetherless Access," *IEEE Communications Magazine*, July 1989, pp. 30–40.

32 "Mobile Station Land Station Compatibility Specification," ANSI/EIA/TIA Standard 553, Electronic Industries Association, 1989.

33 K. E. Fultz and D. B. Penick, "The T1 Carrier System," *Bell System Technical Journal*, Sept. 1965, pp. 1405–1451.

34 "Supply Record on NEC PCM Microwave Equipment," Nippon Electric Company, Tokyo, May 1974.

35 M. Schwartz, "Optical Fiber Transmission—From Conception to Prominence in 20 Years," *IEEE Communications Magazine*, May 1984, pp. 38–48.

36 J. B. Jacob, "The New Approaches and Solutions Offered by the Digital ESS.E 10 to Deal with Switching, Operation, and Maintenance Functions," *International Communications Conference Proceedings*, 1979, pp. 46.6-1–46.6-5.

37 "Special Issue: No. 4 ESS," *Bell System Technical Journal*, July/Aug. 1981.

38 J. C. McDonald and J. R. Baichtal, "A New Integrated Digital Switching System," *National Telecommunications Conference Record*, 1976, pp. 3.2-1–3.2-5.

39 "First Class 5 Digital Central Office Goes into Service in Georgia," *Telephony*, Aug. 1, 1977, pp. 24–25.

40 F. T. Andrews, Jr., "Customer Lines Go Electronic," *Bell Labs Record*, Feb. 1972, pp. 59–64.

41 M. T. Manfred, G. A. Nelson, and C. H. Sharpless, "Digital Loop Carrier Systems," *Bell System Technical Journal*, Apr. 1978, pp. 1129–1142.

42 S. J. Brolin, G. E. Harrington, and G. R. Leopold, "The Subscriber Loop Carrier System Gets Even Better," *Bell Labs Record*, May 1977, pp. 119–122.

43 S. Brolin, Y. S. Cho, W. P. Michaud, and D. H. Williamson, "Inside the New Digital Subscriber Loop System," *Bell Labs Record*, Apr. 1980, pp. 110–116.

44 G. E. Harrington and R. W. Wyndrum, "A 'Smart' SLC Carrier Meets the Changing Nature of Services in the Evolving Digital Loop," *Bell Labs Record*, Sept. 1985, pp. 13–17.

45 "Digital Interface Between the SLC-96 Digital Loop Carrier System and a Local Digital Switch," Bellcore Technical Requirements, TR-TSY-000008, Bellcore, Aug. 1987.

46 R. R. Grady and J. W. Knapp, "1A Radio Digital Terminals Put 'Data Under Voice,'" *Bell Labs Record*, May 1974, pp. 161–166.

47 FCC *Memorandum Opinion and Order*, released January 29, 1975, modifying FCC *Report and Order* of Docket 19311, released September 27, 1974.

48 W. S. Chaskin, "Digital Transmission Economics," *Digital Microwave Transmission Engineering Symposium*, Rockwell International, Sept. 1978.

49 R. C. Hansen, "AT&T's Digital Network Evolution," *IEEE International Conference on Communications*, 1987, pp. 19B.1.1–1.1.5.

50 H. E. Vaughan, "Research Model for Time-Separation Integrated Communication," *Bell System Technical Journal*, July 1959, pp. 909–932.

51 B. T. Fought and C. J. Funk, "Electronic Translator System for Toll Switching," *IEEE Transactions on Communications*, June 1970, p. 168.

PROBLEMS

1.1 Express 33 dBrnC of noise in terms of picowatts and milliwatts of absolute power.

1.2 Twenty-seven picowatts of noise with a flat spectrum between 0 and 3 kHz is how many dBrnC?

1.3 A value of 30 dBrnC0 is how many picowatts of absolute noise power at a –3-dB TLP?

1.4 An idle-channel noise power measurement of 21 dBrnC occurs at a –7-dB TLP. Express the noise power of this measurement in dBrnC0 and determine what power measurement this noise would produce at another point in the circuit that is designated as a –2-dB TLP.

1.5 A transmission link has 14 dBrn of absolute noise power at a –13-dB TLP input test point and 27 dBrn of absolute noise power at a –3-dB TLP output test point. How much absolute noise is added on the transmission link?

1.6 A transmission link with a –2-dB TLP at the transmit end and a –4-dB-TLP at the receive end is part of a voice circuit that produces the following idle-channel noise measurements: 18 dBrnC0 at the transmit end and 20 dBrnC0 at the receive end. What is the contribution (in pWp0) of the transmission link toward the total absolute power at the receive end?

1.7 Why is echo control unnecessary on short-delay voice connections? What about singing?

2

WHY DIGITAL?

The first chapter provides an overview of an analog telephone network and a brief introduction to digital transmission and switching technology introduced to replace older analog equipment. This chapter discusses the basic technical advantages of digital implementations that stimulated the replacement of the analog systems.

2.1 ADVANTAGES OF DIGITAL VOICE NETWORKS

A list of technical features of digital communications networks is provided in Table 2.1. These features are listed in the order that the author considers to be their relative importance for general telephony. In particular applications, however, certain considerations may be more or less significant. For instance, the last item, ease of encryption, is a dominant feature favoring digital networks for the military.

Most of the features of digital voice networks listed in Table 2.1 and discussed in the following paragraphs pertain to advantages of digital transmission or switching relative to analog counterparts. In some instances, however, the features pertain only to all-digital networks. Encryption, for example, is practical and generally useful only if the secure form of the message is established at the source and translated back into the clear form only at the destination. Thus an end-to-end digital system that operates with no knowledge of the nature of the traffic (i.e., provides transparent transmission) is a requirement for digital encryption applications. For similar reasons end-to-end digital transmission is needed for direct transmission of data (no modem). When a network consists of a mixture of analog and digital equipment, universal use of the network for services such as data transmission dictates conformance to the least common denominator of the network: the analog channel.

2.1.1 Ease of Multiplexing

As mentioned in Chapter 1, digital techniques were first applied to general telephony in interoffice T-carrier (time division multiplex) systems. In essence, these systems traded electronics costs at the ends of a transmission path for the cost of multiple pairs

TABLE 2.1 Technical Advantages of Digital Communications Networks

1. Ease of multiplexing
2. Ease of signaling
3. Use of modern technology
4. Integration of transmission and switching
5. Signal regeneration
6. Performance monitorability
7. Accommodation of other services
8. Operability at low signal-to-noise/interference ratios
9. Ease of encryption

of wires between them. (A trade that is more cost effective every year.) Although FDM of analog signals had also been used to reduce cable costs, FDM equipment is much more expensive than TDM equipment, even when the cost of digitization is included. After voice signals have been digitized, TDM equipment costs are quite small by comparison. Since digitization occurs only at the first level of the TDM hierarchy, high-level digital TDM is even more economical than high-level FDM counterparts.

It should be pointed out that TDM of analog signals is also very simple and does not require digitization of the sample values. The drawback of analog TDM lies in the vulnerability of narrow analog pulses to noise, distortion, crosstalk, and intersymbol interference. These degradations cannot be removed by regeneration as in a digital system. Hence, analog TDM is not feasible except for noiseless, distortion-free environments.* In essence the ability to regenerate a signal, even at the expense of greater bandwidth, is almost a requirement for TDM transmission.

2.1.2 Ease of Signaling

Control information (e.g., on-hook/off-hook, address digits, coin deposits) is inherently digital and, hence, readily incorporated into a digital transmission system. One means of incorporating control information into a digital transmission link involves time division multiplexing the control as a separate but easily identifiable control channel. Another approach involves inserting special control codes into the message channel and having digital logic in the receiving terminals decode that control information. In either case, as far as the transmission system is concerned, control information is indistinguishable from message traffic.

In contrast, analog transmission systems required special attention for control signaling. Many analog transmission systems presented unique and sometimes difficult environments for inserting control information. An unfortunate result was that many varieties of control signal formats and procedures evolved for the analog network. The control formats depend on the nature of both the transmission system and its terminal

*Analog TDM has been used in a few telephone applications. Farinon's Subscriber Radio System [1], for example, used pulse-width-modulated TDM. As discussed in Chapter 5 some older PBXs also used analog TDM.

equipment. In some interfaces between network subsystems control information had to be converted from one format to another. Signaling on analog links therefore represented a significant administrative and financial burden to the operating telephone companies.

The move to common-channel signaling removed most of the signaling costs associated with interoffice trunks but did not change the situation for individual subscriber lines, which must carry signaling on the same facility as the message channel. The use of digital subscriber lines (DSLs) reduces the signaling costs relative to analog subscriber lines, which helps offset the higher cost of a DSL and a digital telephone. DSLs are a fundamental aspect of ISDN, as described in Chapter 11.

In summary, digital systems allow control information to be inserted into and extracted from a message stream independently of the nature of the transmission medium (e.g., cable, fiber, microwave, satellite). Thus the signaling equipment can (and should) be designed separately from the transmission system. It then follows that control functions and formats can be modified independently of the transmission subsystem. Conversely, digital transmission systems can be upgraded without impacting control functions at either end of the link.

2.1.3 Use of Modern Technology

A multiplexer or switching matrix for time division digital signals is implemented with the same basic circuits used to build digital computers: logic gates and memory. The basic crosspoint of a digital switch is nothing more than an AND gate with one logic input assigned to the message signal and other inputs used for control (crosspoint selection). Thus the dramatic developments of digital integrated circuit technology for computer logic circuits and memory are applicable directly to digital transmission and switching systems. In fact, many standard circuits developed for use in computers are directly usable in a switching matrix or multiplexer. Figure 2.1 shows the basic implementation of a 16-channel, bit-interleaved, digital time division multiplexer using common digital logic circuits. As indicated, the multiplexing function involves nothing more than cyclically sampling the 16 input data streams. Such an operation assumes all of the data streams are synchronized to each other. As discussed in Chapter 7, the process of synchronizing the data streams requires logic circuitry that is much more complicated than that shown. Nevertheless, the implementation of TDM is much less expensive than analog FDM.

Even greater advantages of modern technology have been achieved by using large-scale integrated (LSI) circuits designed specifically for telecommunications functions such as voice encoding/decoding, multiplexing/demultiplexing, switching matrices, and special-purpose and general-purpose digital signal processing (DSP). Digital signal processing functions are described in Section 2.2.

The relative low cost and high performance of digital circuits allows digital implementations to be used in some applications that are prohibitively expensive when implemented with comparable analog components. Completely nonblocking switches, for example, are not practical with conventional analog implementations, except in

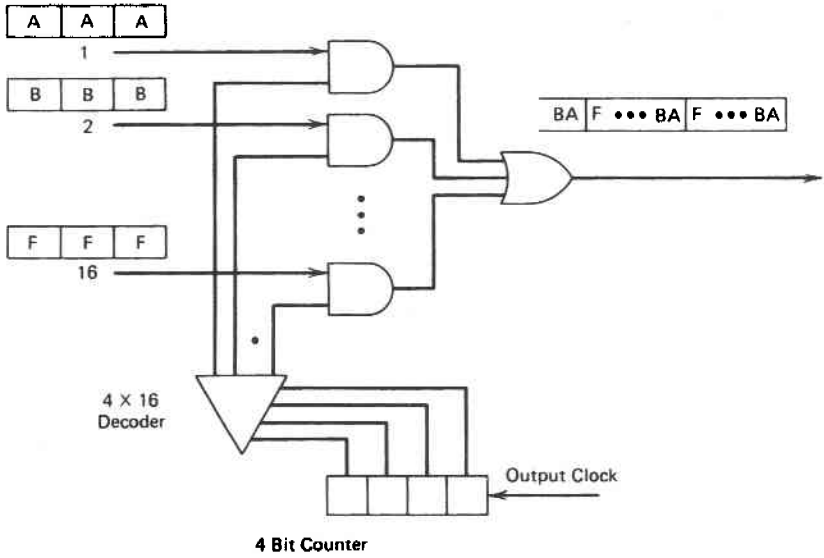


Figure 2.1 Sixteen-to-one TDM multiplexer.

small sizes. In a modern digital switch the cost of the switching matrix itself is relatively insignificant. Thus, for medium-size applications, the size of the switch matrix can be increased to provide nonblocking operations, if desired. The automatic call distributor developed by Collins-Rockwell [2] is an early example of a digital switch operating in an analog environment. A digital implementation was chosen largely because it could economically provide a nonblocking operation.

The benefits of modern device technology are not confined to digital circuits alone. Analog integrated circuits have also progressed significantly, allowing traditional analog implementations to improve appreciably. One of the primary requirements of an analog component, however, is that it be linear. It appears, if only because of research and development emphasis, that fast digital components are easier to manufacture than linear analog counterparts. In addition, digital implementations appear to have an inherent functional advantage over analog implementations. This advantage is derived from the relative ease with which digital signals can be multiplexed. A major limitation with the full use of LSI components results from limited availability of external connections to the device. With time division multiplex techniques, a single physical pin can be used for multiple-channel access into the device. Thus the same technique used to reduce costs in transmission systems can also be used within a local module to minimize the interconnections and maximize the utilization of very large scale integration. In the end, a "switch on a chip" is possible only if a great number of channels can be multiplexed onto a relatively small number of external connections.

The technological development to have the most significant impact on the telephone network is certainly fiber optic transmission. Although fibers themselves do not favor digital transmission over analog transmission, the interface electronics to a fiber

system function primarily in an on–off (nonlinear) mode of operation. Thus digital transmission dominates fiber applications, although analog optical technology is commonly used in analog video distribution.

2.1.4 Integration of Transmission and Switching

Traditionally the analog transmission and switching systems of telephone networks were designed and administered by functionally independent organizations. In the operating telephone companies, these two equipment classes are referred to as outside plant and inside plant, respectively. These equipments necessarily provide standardized interfaces, but, other than that, transmission equipment was functionally independent of switching equipment.

When TDM of digital voice signals was introduced into the exchange area and communications engineers began considering digital switching, it became apparent that TDM operations were very similar to time division switching functions. In fact, as described in later chapters, the first stages of digital switches generate first-level TDM signals by nature, even when interfaced to analog transmission links. Thus the multiplexing operations of a transmission system can be easily integrated into the switching equipment.

The basic advantage of integrating the two systems is shown in Figure 2.2. The demultiplexing equipment (channel banks) at the switching offices is unnecessary, and first-stage switching equipment is eliminated. If both ends of the digital TDM trunks are integrated into a digital switch, the channel banks at both ends of the trunk are eliminated. In a totally integrated network voice signals are digitized at or near the source and remain digitized until delivered to their destination. Furthermore, all interoffice trunks and internal links of a switching system carry TDM signals exclusively. Thus first-level multiplexing and demultiplexing are nonexistent except at the periphery of the network. Although integration of DS1 signals into switching machines is

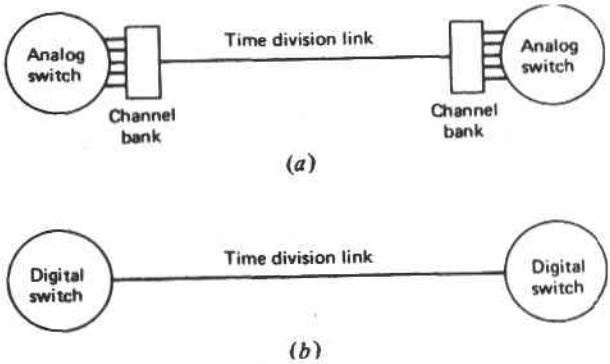


Figure 2.2 Integration of transmission and switching: (*a*) nonintegrated transmission and switching, (*b*) integrated time time division switching and transmission.

commonplace, integration of higher level signals is complicated by higher level multiplexing formats (pulse stuffing) described in Chapter 7. A newer multiplexing format (SONET) described in Chapter 8 has some operational modes that are more amenable to direct interconnection into a switching system.

Integration of transmission and switching functions not only eliminates much equipment but also greatly improves end-to-end voice quality. By eliminating multiple analog-to-digital and digital-to-analog conversions and by using low-error-rate transmission links, voice quality is determined only by the encoding/decoding processes. In summary, the implementation benefits of a fully integrated digital network are:

1. Long-distance voice quality is identical to local voice quality in all aspects of noise, signal level, and distortion.
2. Since digital circuits are inherently four-wire, network-generated echoes are eliminated, and true full-duplex, four-wire digital circuits are available.
3. Cable entrance requirements and mainframe distribution of wire pairs is greatly reduced because all trunks are implemented as subchannels of a TDM signal.

2.1.5 Signal Regeneration

As described more fully in the next chapter, the representation of voice (or any analog signal) in a digital format involves converting the continuous analog waveform into a sequence of discrete sample values. Each discrete sample value is represented by some number of binary digits of information. When transmitted, each binary digit is represented by only one of two possible signal values (e.g., a pulse versus no pulse or a positive pulse versus a negative pulse). The receiver's job is to decide which discrete values were transmitted and represent the message as a sequence of binary-encoded discrete message samples. If only small amounts of noise, interference, or distortion are impressed upon the signal during transmission, the binary data in the receiver are identical to the binary sequence generated during the digitization or encoding process. As shown in Figure 2.3, the transmission process, despite the existence of certain imperfections, does not alter the essential nature of the information. Of course, if the imperfections cause sufficient changes in the signal, detection errors occur and the binary data in the receiver does not represent the original data exactly.

A fundamental attribute of a digital system is that the probability of transmission errors can be made arbitrarily small by inserting regenerative repeaters at intermediate points in the transmission link. If spaced close enough together, these intermediate nodes detect and regenerate the digital signals before channel-induced degradations

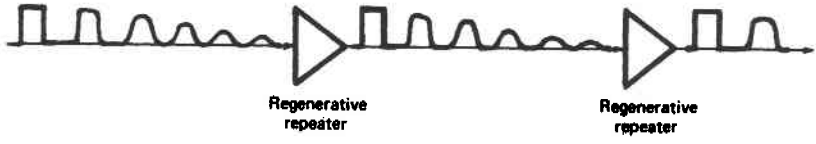


Figure 2.3 Signal regeneration in a digital repeater line.

become large enough to cause decision errors. As demonstrated in Chapter 4, the end-to-end error rate can be made arbitrarily small by inserting a sufficient number of regeneration nodes in the transmission link.

The most direct benefit of the regeneration process is the ability to localize the effects of signal degradations. As long as the degradations on any particular regenerated segment of a transmission link do not cause errors, their effects are eliminated. In contrast, signal impairments in analog transmission accumulate from one segment to the next. Individual subsystems of a large analog network must be designed with tight controls on the transmission performance to provide acceptable end-to-end quality. An individual subsystem of a digital network, on the other hand, need only be designed to ensure a certain minimum error rate—usually a readily realizable goal.

When an all-digital network is designed with enough regeneration points to effectively eliminate channel errors, the overall transmission quality of the network is determined by the digitization process and not by the transmission systems. The analog-to-digital conversion process inherently introduces a loss of signal fidelity since the continuous analog source waveform can only be represented by discrete sample values. By establishing enough discrete levels, however, the analog waveform can be represented with as little conversion error as desired. The increased resolution requires more bits and consequently more bandwidth for transmission. Hence, a digital transmission system readily provides a trade-off between transmission quality and bandwidth. (A similar trade-off exists for frequency-modulated analog signals.)

2.1.6 Performance Monitorability

An additional benefit of the source-independent signal structure in a digital transmission system is that the quality of the received signal can be ascertained with no knowledge of the nature of the traffic. The transmission link is designed to produce well-defined pulses with discrete levels. Any deviation in the receive signal, other than nominal amounts planned for in the design, represents a degradation in transmission quality. In general, analog systems cannot be monitored or tested for quality while in service since the transmitted signal structure is unknown. FDM signals typically include pilot signals to measure channel continuity and power levels. The power level of a pilot is an effective means of estimating the signal-to-noise ratio—only in a fixed-noise environment. Hence, noise and distortion are sometimes determined by measuring the energy level in an unused message slot or at the edge of the signal passband. In neither case, however, is the quality of an in-service channel being measured directly.

One common method of measuring the quality of a digital transmission link is to add parity, or cyclic redundancy check (CRC), bits to the message stream. The redundancy introduced to the data stream enables digital logic circuits in a receiver to readily ascertain channel error rates. If the error rate exceeds some nominal value, the transmission link is degraded.

Another technique for measuring in-service transmission quality is used in T-carrier lines. This technique involves monitoring certain redundancies in the signal waveform

itself. When the redundancy pattern at the receiver deviates from normal, decision errors have occurred. A complete description of the line coding format used in T-carrier systems is provided in Chapter 4. Other methods of measuring transmission quality in digital systems are discussed in Chapters 4 and 6.

2.1.7 Accommodation of Other Services

It was previously pointed out that a digital transmission system readily accommodates control (signaling) information. This fact is representative of a fundamental aspect of digital transmission: any digitally encoded message (whether inherently digital or converted from analog) presents a common signal format to the transmission system. Thus the transmission system need provide no special attention to individual services and can, in fact, be totally indifferent to the nature of the traffic it carries.

In an analog network the transmission standard is the 4-kHz voice circuit. All special services such as data or facsimile must be transformed to "look like voice." In particular, data signals must be converted to an analog format through the use of modems.

The standard analog channel was necessarily optimized for voice quality. In so doing, certain transmission characteristics (such as the phase response and impulse noise) received less attention than more noticeable voice quality impairments. Some less emphasized considerations, phase distortion in particular, are critical for high-rate data services. Use of an analog network for nonvoice services often requires special compensation for various analog transmission impairments. If the analog channel is too poor, it may be unusable for a particular application. In contrast, the main parameter of quality in a digital system is the error rate. Low-error-rate channels are readily obtainable. When desired, the effects of channel errors can be effectively eliminated with error control procedures implemented by the user.

An additional benefit of the common transmission format is that traffic from different types of sources can be intermixed in a single transmission medium without mutual interference. The use of a common transmission medium for analog signals is sometimes complicated because individual services require differing levels of quality. For example, television signals, which require greater transmission quality than voice signals, were not usually combined with FDM voice channels in a wideband analog transmission system [3].

2.1.8 Operability at Low Signal-to-Noise/Interference Ratios

Noise and interference in an analog voice network become most apparent during speech pauses when the signal amplitude is low. Relatively small amounts of noise occurring during a speech pause can be quite annoying to a listener. The same levels of noise or interference are virtually unnoticeable when speech is present. Hence it is the absolute noise level of an idle channel that determines analog speech quality. Subjective evaluations of voice quality [4, 5] led to maximum noise level standards of 28 dBrnC0 (−62 dBm0) for short-haul systems and 34 dBrnC0 (−56 dBm0) for long-haul systems. For comparison, the power level of an active talker is typically −16 dBm0.

Thus representative end-to-end signal-to-noise ratios in analog networks are 46 and 40 dB for short- and long-haul systems, respectively. Signal-to-noise ratios on individual transmission systems are necessarily higher.

In a digital system speech pauses are encoded with a particular data pattern and transmitted at the same power level as active speech. Because signal regeneration virtually eliminates all noise arising in the transmission medium, idle channel noise is determined by the encoding process and not the transmission link. Thus speech pauses do not determine maximum noise levels as they do in an analog system. As discussed in Chapter 4, digital transmission links provide virtually error-free performance at signal-to-noise ratios of 15–25 dB, depending on the type of line coding or modulation used.

The ability of a digital transmission system to reject crosstalk is sometimes more significant than its ability to operate in relatively high levels of random noise. One of the most troublesome considerations in the design and maintenance of the analog network was the need to eliminate crosstalk between conversations. The problem was most acute during pauses on one channel while an interfering channel was at maximum power. At these times relatively low level crosstalk would be noticeable. The crosstalk was particularly undesirable if it was intelligible and therefore violated someone's privacy. Again, speech pauses do not produce low-amplitude signals on digital transmission links. The transmission links maintain a constant-amplitude digital signal. Thus, low levels of crosstalk are eliminated by the regeneration process in a digital repeater or receiver. Even if the crosstalk is of sufficient amplitude to cause detection errors, the effects appear as random noise and, as such, are unintelligible.

Considering the fact that a digital system typically needs a greater bandwidth than a comparable analog system and that wider bandwidths imply greater crosstalk and noise levels, the ability to operate at lower SNRs is partly a requirement of a digital system and partly an advantage.

2.1.9 Ease of Encryption

Although most telephone users have little need for voice encryption, the ease with which a digital bit stream can be scrambled and unscrambled [6] means that a digital network (or a digital cellular system) provides an extra bonus for users with sensitive conversations. In contrast, analog voice is much more difficult to encrypt and is generally not nearly as secure as digitally encrypted voice. For a discussion of common analog voice encryption techniques, see references [7], [8], and [9]. As mentioned previously, ease of encryption stimulated early use of digital voice systems by the military.

2.2 DIGITAL SIGNAL PROCESSING

The preceding paragraphs emphasize the advantages of digital technology in implementing the transmission and switching systems of a network. Another significant application of digital technology is the area of signal processing. Basically, signal

processing refers to an operation on a signal to enhance or transform its characteristics. Signal processing can be applied to either analog or digital waveforms. Amplification, equalization, modulation, and filtering are common examples of signal processing functions.

Digital signal processing (DSP) refers to the use of digital logic and arithmetic circuits to implement signal processing functions on digitized signal waveforms. Sometimes analog signals are converted to digital representations for the express purpose of processing them digitally. Then the digital representations of the processed signals are converted back to analog. These operations are illustrated in Figure 2.4, where a sine wave corrupted by noise is digitally filtered to remove the noise. The main advantages of digitally processing signal waveforms are listed in Table 2.2.

It is important to point out that DSP in this context refers to the technology used to condition, manipulate, or otherwise transform a signal waveform (a digitized representation thereof). In another context signal processing refers to the interpretation of control signals in a network by the control processors of switching systems. In the latter case the logical interpretation of a control code is processed and not an underlying signal waveform.

2.2.1 DSP Applications

The following four sections identify applications of DSP that either represent lower cost solutions to functions that have been traditionally implemented with analog technology or are functions that have no practical implementation counterpart with analog technology.

Echo Cancellers

The cost and performance of DSP echo cancellers have improved to the point that they can be justified for any long-distance circuit, thereby providing full-duplex circuits (no echo suppression) and no artificial attenuation (no via net loss). A particularly critical need for echo cancellation occurs in high-speed, full-duplex data modems that incorporate near-end echo cancellation—an unnecessary requirement for voice circuits. Furthermore, low-cost echo canceling enables packet-switched voice applications that introduce artificial delays that are not accommodated in normal analog

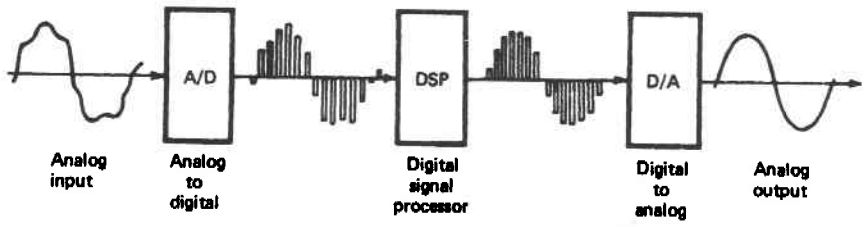


Figure 2.4 Digital signal processing of an analog signal.

TABLE 2.2 Digital Signal Processing Features

Reproducibility: The immunity of digital circuits to small imperfections and parasitic elements implies that circuits can be produced with consistent operational characteristics without fine adjustments or aging tolerances.
Programmability: A single basic structure can be used for a variety of signal types and applications by merely changing an algorithmic or parametric specification in a digital memory.
Time sharing: A single digital signal processing circuit can be used for multiple signals by storing temporary results of each process in random-access memory and processing each signal in a cyclic (time-divided) fashion.
Automatic test: Since the inputs and outputs of a digital signal processing circuit are digital data, tests can be performed routinely by comparing test responses to data patterns stored in memory.
Versatility: Because of the decision-making capabilities of digital logic, digital signal processing can perform many functions that are impossible or impractical with analog implementations.

interfaces. The adaptation logic and delay requirements of a switched network echo canceller virtually preclude any type of analog implementation.

Tone Receivers

Detection of DTMF, MF, SF, or other analog tones is easily and economically realized by converting the analog signals to digital representations for the explicit purpose of detecting the tone. Of course, a DSP implementation is even more economical when the tones are already digitized, which is the case within a digital switch. The programmability feature of a DSP circuit is particularly useful for tone receivers because one hardware implementation can be used for multiple functions by selecting different filter options (programs) depending on the application [10].

High-Speed Modems

Reliable operation (low bit error rates) of high-speed (e.g., 28.8-kbps) voiceband modems [11] over the switched telephone network requires sophisticated modulation techniques (described in Chapter 6) and sophisticated signal conditioning referred to as adaptive equalization. The only practical way to implement these functions is with DSP circuitry. Reference [12] describes an early application of DSP to a 14,400-bps modem. Reference [13] describes the use of DSP for adaptive equalization of a 400-Mbps digital radio. Previous digital radios used analog adaptive equalizers because they were cheaper. Very-high-rate digital radios require more sophisticated equalizers, which are easier to implement (perhaps only possible) with DSP.

Low-Bit-Rate Voice Encoding

The realization of low-bit-rate voice encoding algorithms described in Chapter 3 involves extensive numerical processing to remove redundancy in the digitized voice samples. DSP technology is the only economical means of implementing these algorithms on a real-time basis. References [14], [15], and [16] describe DSP implementations for 32-, 16-, and 4.8-kbps voice coders, respectively. Reference [17] describes more of the general theory and application of DSP to voice compression.

2.3 DISADVANTAGES OF DIGITAL VOICE NETWORKS

The first part of this chapter discussed the basic technical advantages of digital networks. To balance the discussion, this section reviews the basic technical disadvantages of digital implementations as listed in Table 2.3.

2.3.1 Increased Bandwidth

In the brief introduction to voice digitization presented in Chapter 1, mention is made that transmission of samples of an analog waveform requires no more bandwidth than the underlying waveform (at least in theory). The bandwidth expansion comes when the samples are encoded into binary codes and transmitted with an individual pulse for each bit in the code. Thus a T1 system requires approximately eight times as much bandwidth as do 24 analog voice channels since each sample is represented by an 8-bit codeword and each bit is transmitted as a separate discrete pulse. Although more sophisticated digitization algorithms can be used to encode voice at a lower bit rate than that used on T1 systems (64 kbps), even the most sophisticated algorithms (described in Chapter 3) cannot provide comparable voice quality without at least a two-to-one bandwidth penalty.

In some portions of the analog network, such as the local loops, the bandwidth increase did not represent much of a penalty since the inherent bandwidth was (and is) underutilized. In long-haul radio systems, however, bandwidth was at a premium, and digital systems were relatively inefficient in terms of the number of voice channels provided. One mitigating aspect of a digital radio system is its ability to overcome higher levels of noise and interference, which sometimes provides compensation for the bandwidth requirements, particularly in congested transmission environments where mutual interference can become a limiting consideration [3]. The inherent robustness of a digital system with respect to interference is one important attribute of digital cellular systems described in Chapter 9.

The bandwidth penalty imposed by voice digitization is directly dependent on the form of transmission coding or modulation used. With greater sophistication in the modulation/demodulation equipment, greater efficiency in terms of the bit rate in a given bandwidth is achievable. Basically, greater transmission efficiency is achieved by increasing the number of levels in the line code. With limited transmit power, however, the distances between discrete signal levels in the receiver are reduced dramati-

TABLE 2.3 Disadvantages of Digital Implementations

1. Increased bandwidth
2. Need for time synchronization
3. Topologically restricted multiplexing
4. Need for conference/extension bridges
5. Incompatibilities with analog facilities

cally. Thus, the transmitted signal is no longer as immune to noise and other imperfections as it is with lower information densities.

Using a combination of advanced digital modulation, lower rate digitization, and error-correcting codes, point-to-point digital radios could provide voice channel efficiencies comparable to or even better than analog microwave systems. Full development along these lines never occurred, however, because the emergence of optical fiber transmission eliminated the incentive to do so.

2.3.2 Need for Time Synchronization

Whenever digital information is transmitted from one place to another, a timing reference, or "clock," is needed to control the transfer. The clock specifies when to sample the incoming signal to decide which data value was transmitted. The optimum sample times usually correspond to the middle of the transmitted pulses. Thus, for optimum detection, the sample clock must be synchronized to the pulse arrival times. In general, the generation of a local timing reference for detecting the digital signal is not difficult. Chapter 4 discusses some of the design considerations needed to establish proper sample clocking in the receiver of a digital transmission link.

More subtle problems arise, however, when a number of digital transmission links and switches are interconnected to form a network. Not only must the individual elements of the network maintain internal synchronization, but also certain networkwide synchronization procedures must be established before the individual subsystems can interoperate properly. Chapter 7 discusses these basic network synchronization requirements and implementations.

The need for some form of synchronization is not unique to digital networks. Single-sideband FDM transmission systems present similar requirements for carrier synchronization in analog networks. In analog systems, however, the synchronization requirements are less critical by about two orders of magnitude [18].

2.3.3 Topologically Restricted Multiplexing

To the general public, the most apparent use of multiplexing is broadcast services for radio and television. In these systems the airspace is shared by using FDM of individual broadcast channels. With this system there are no operational restrictions to the geographic location of transmitters and receivers. As long as the transmitters confine their emissions to their assigned bandwidth and each receiver uses a sufficiently selective filter to pass only the desired channel, the network operates without mutual interference. On the other hand, TDM is not nearly as amenable to applications involving distributed sources and destinations. Since the time of arrival of data in a time slot is dependent on the distance of travel, distributed TDM systems require a guard time between time slots. FDM systems also require guardbands between the channels to achieve adequate channel separation. The width of the FDM guardbands, however, is not dependent on the geographic location of the transmitters. In a TDM system the guard times must be increased as the geographic separation between trans-

mitters increases. Furthermore, each time division source must duplicate the synchronization and time slot recognition logic needed to operate a TDM system. For these reasons, TDM has been used primarily in applications (e.g., interoffice trunks) where all of the information sources are centrally located and a single multiplexer controls the occurrence and assignment of time slots.

Time division multiple access (TDMA) satellites and cellular systems are examples of applications of TDM for distributed sources. These systems use sophisticated synchronization techniques so that each ground station or mobile unit times its transmission to arrive at the satellite or base station at precisely defined times, allowing the use of small guard times between time slots. Notice that these applications involve only one destination: a satellite or a base station. If an application involves multiple, distributed *sources* and *destinations* (with transmission in more than one direction), larger guard times are unavoidable. Figure 2.5 shows such an application but uses FDM instead of TDM. The main engineering consideration for this system is to ensure that the FDM channels have sufficient isolation to allow a high-powered source to be adjacent to a receiver with the worst-case receive level. Obviously, adequate FDM isolation requires a certain amount of bandwidth overhead, but it is usually fairly easy to design filters with adequate isolation for a large range of signal levels so distance considerations are minimized.

2.3.4 Need for Conference/Extension Bridges

The process of combining multiple analog signals to form a conference call or function as multiple extensions on a single telephone line can be accomplished by merely bridging the wire pairs together to superimpose all signals. Nowhere is this more convenient than when multiple extensions share a single two-wire line, as indicated in Figure 2.6. When digitized voice signals are combined to form a conference, either the signals must be converted to analog so they can be combined on two-wire analog bridges or the digital signals must be routed to a digital conference bridge, as shown in Figure 2.7. The digital bridge selectively adds the (four-wire) signals together (using digital signal processing) and routes separate sums back to the conferees as shown.

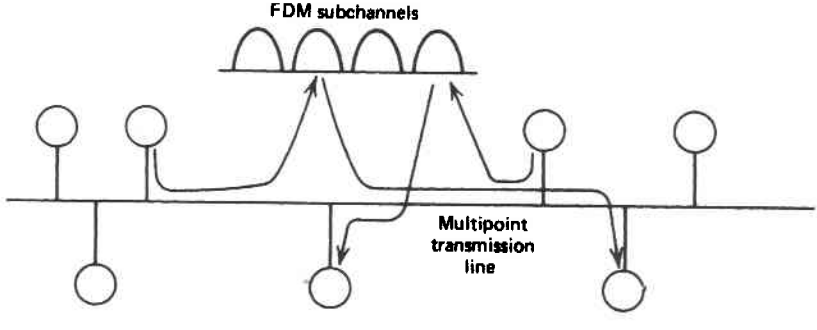


Figure 2.5 Frequency division multiplexing on distributed multipoint line.

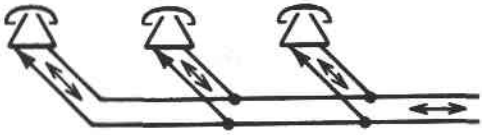


Figure 2.6 Analog telephones connected to two-wire line.

When conferencing is implemented in association with a switching system, the need for a digital conference bridge is not much of a disadvantage and in fact can significantly improve the quality of a conference by eliminating echoes and signal loss caused by power division. However, when digital extensions need to have their signals combined so multiple extensions can be active in a conversation, the need for a centralized bridge can be an onerous problem. Residential telephone wiring typically follows a daisy-chain pattern, as indicated in Figure 2.6. Thus the need to rewire all outlets and install a centralized conference box is a significant impediment to the deployment of digital station equipment in residential applications.

2.3.5 Incompatibilities with Analog Facilities

When digital equipment was first used in private and public telephone networks, it necessarily provided standard analog interfaces to the rest of the network. Sometimes these interfaces represented a major cost of the digital subsystem. The foremost example of this situation arose in digital end offices. The standard analog subscriber loop interface described in Chapter 1 is particularly incompatible with electronic switching machines (analog or digital). Another aspect of digital switching that complicates its use in analog environments is the artificial delay inserted by a typical digital matrix. Both of these aspects of digital switching are discussed in Chapter 5.

One way to eliminate the problems with the analog interface is to use digital subscriber loops and digital telephones. Unfortunately, the overwhelming investment in the loop plant for analog telephones complicates a widespread deployment of digital subscriber equipment. Most notable of the long-established practices that complicate

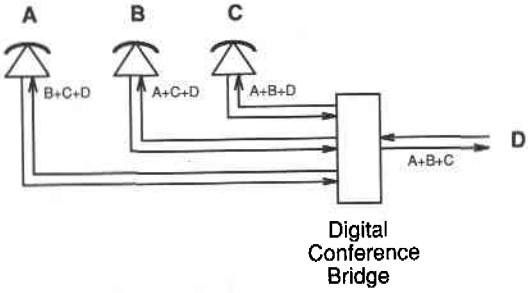


Figure 2.7 Use of conference bridge for digital telephones.

a transition to digital loops are single wire pairs, loading coils, bridged taps,* high-resistance or intermittent splices, and wiring gauge changes. The digital subscriber loop systems described in Chapter 11 accommodate most of the above impediments but do so with very sophisticated DSP circuitry.

REFERENCES

1 *SR-1536 Subscriber Radio, Technical Description*, Farinon SR Systems, Quebec, Canada, 1977.

2 R. J. Hirvela, "The Application of Computer Controlled PCM Switching to Automatic Call Distribution," *IEEE Communications Systems and Technology Conference*, Dallas, TX, May 1974.

3 M. R. Aaron, "Digital Communications—The Silent (R)evolution?" *IEEE Communications Magazine*, Jan. 1979, pp. 16–26.

4 I. Nasell, "The 1962 Survey of Noise and Loss on Toll Connections," *Bell System Technical Journal*, Mar. 1964, pp. 697–718.

5 Technical Staff, Bell Telephone Laboratories, *Transmission Systems for Communications*, Western Electric Company, Winston-Salem, North Carolina, Feb. 1970.

6 H. J. Hindin, "LSI-Based Data Encryption Discourages the Data Thief," *Electronics*, June 21, 1979, pp. 107–120.

7 N. S. Jayant, B. J. McDermott, S. W. Christensen, and A. M. Quinn, "A Comparison of Four Methods for Analog Speech Encryption," *International Communications Conference Record*, 1980, pp. 16.6.1–16.6.5.

8 A. Gersho and R. Steele, Ed., "Special Issue on Encryption of Analog Signals," *IEEE Journal on Selected Areas in Communications*, May 1984.

9 H. J. Beker and F. Piper, "Speech Scrambling," *Speech Technology*, Mar./Apr. 1987, pp. 40–45.

10 A. Fukui and Y. Fujihashi, "A Single-Chip, 4-Channel MF/MFC/PB Receiver," *IEEE Globecom Conference Record*, 1987, pp. 12.6.1–12.6.4.

11 "A Modem Operating at Data Signaling Rates Up to 28,800 bps for Use on the General Switched Telephone Network and on Leased Point-to-Point 2-Wire Telephone-Type Circuits," ITU-T Rec. V.34, Geneva, Switzerland, Sept. 1994.

12 T. Kamitake, K. Uehara, M. Abe, and S. Kawamura, "Trellis Coding 14.4 kb/s Data Modem Implemented with a Single-Chip High-Speed Digital Signal Processor," *IEEE Globecom Conference Record*, 1987, pp. 12.8.1–12.8.6.

13 H. Matsue, T. Shirato, and K. Watanabe, "256 QAM 400 Mb/s Microwave Radio System with DSP Fading Countermeasures," *IEEE International Conference on Communications*, 1988, pp. 41.5.1–41.5.6.

14 J. L. So, "Implementation on an NIC (Nearly Instantaneous Companding) 32 kbps Transcoder Using the TMS320C25 Digital Signal Processor," *IEEE Globecom Conference Record*, 1988, pp. 43.4.1–43.4.5.

*A bridged tap is an unused pair of wires connected at some point to an in-use pair as another extension or for possible future reassignment of a cable pair.

15 J. L. Dixon, V. Varma, and D. W. Lin, "Software Considerations in the Design of a 16 kbps Speed Coder for a TDMA Portable Phone," *IEEE Globecom Conference Record*, 1988, pp. 26.7.1–26.7.5.

16 K. Irie and S. Iai, "4.8 kbit/s Speech Codec Using Advanced Digital Signal Processors (DSSP)," *IEEE Globecom Conference Record*, 1987, pp. 20.4.1–20.4.5.

17 M. E. Frerking, *Digital Signal Processing in Communications Systems*, Van Nostrand Reinhold, New York, 1994.

18 J. E. Abate, E. W. Butterline, R. A. Carley, P. Greendyk, A. M. Montenegro, C. D. Near, S. H. Richman, and G. P. Zampetti, "AT&T's New Approach to the Synchronization of Telecommunication Networks," *IEEE Communications Magazine*, Apr. 1989, pp. 35–45.

3

VOICE DIGITIZATION

Because of the interesting nature of the subject and its usefulness in a variety of applications, the field of voice digitization has been and continues to be an area of intense research. This research has produced many different types of voice digitization algorithms. The choice of a particular type is primarily dependent on the implementation cost and the performance requirements implied by the application. The algorithm chosen for T1 systems (companded PCM) provides excellent quality for all types of input signals (e.g., voice or data) at a moderate data rate (64 kbps) at what was originally a moderate cost. The algorithms used in the first-generation digital PBXs used lower cost coding techniques (higher rate PCM or delta modulation) because, at the time, a switching application was more sensitive to digital conversion cost and less sensitive to quality or data rate. For example, the first digital PBX in the United States, introduced by Rolm Corporation in 1975, used uncompanded PCM at a data rate of 144 kbps because it was cheaper than companded PCM at the time [1]. Subsequent advantages derived from integrating transmission and switching and a dramatic drop in the cost of T1-compatible digital voice coders have made obsolete the use of "switching-only" voice digitization algorithms. Because of high-volume production, T1-compatible coder/decoder integrated circuits (ICs) (codecs) can be used in switching applications without a cost penalty. In fact, if the digital network were to be designed today, a more complicated but economically viable codec with a data rate significantly below 64 kbps would probably be utilized.

Transmissions applications with strict bandwidth limits such as high frequency (HF) or digital cellular radio require much more sophisticated voice digitization algorithms to achieve data rates on the order of 8–16 kbps. As a help in reducing the data rate, the performance requirements of these applications are also relaxed as much as the application allows.

Another application for digitized voice is voice storage systems—either for recorded announcements or for voice messaging. Digital storage is particularly appropriate for recorded announcements because the playback quality does not deteriorate with time and individual announcements stored in memory or on a compact disc (CD) can be randomly accessed. Speech storage with limited memory is an example of an application that can use very low rate digitization algorithms with significant quality

reductions. The original Speak-and-Spell learning machine developed by Texas Instruments, for example, stored words with a few hundred bits per word, representing a data rate of about 850 bps [2]. In a similar example requiring better quality the same encoding algorithm (LPC) was used in the Voice Alert System of Chrysler automobiles to store 20 sec of speech (40 words) in 32,000 bits of read-only memory—a data rate of 1600 bps [3].

The main reason voice messaging systems use digital storage is to have random access to the individual messages. To minimize storage space, these systems typically use 8–32-kbps data rates.

At the upper end of digital storage applications are high-fidelity recordings of voice and music. Many of the same advantages of digital transmission, as opposed to analog transmission, also apply to digital recordings. Foremost among these advantages is the ability of defining the fidelity at the time of recording and maintaining the quality indefinitely by periodically copying (regenerating) the digitally stored information before inevitable deterioration produces bit errors. Thus a high-quality (high-bit-rate) digital recording of Bing Crosby, Elvis Presley, or Luciano Pavarotti (depending on your taste in music) can be saved for posterity. This feat could not be accomplished with analog recordings no matter how well cared for or preserved. As an example of high-fidelity audio recording, compact disc players [4] record two channels of audio at 705 kbps each.

Speech analysis and synthesis make up another area of widespread research closely related to voice digitization. In fact, some of the lowest bit rate voice encoders and decoders use certain amounts of analysis and synthesis to digitally represent speech. In its ultimate form, however, analysis and synthesis have unique goals and applications fundamentally different from those of general voice digitization. Basically, goals of analysis and synthesis are to recognize words [5] or produce machine-generated speech (e.g., text-to-speech) [6].

One approach to analyzing speech is to process waveforms with the intent of recognizing speech phonemes—the basic units of speech from which spoken words are constructed. Once the phonemes have been identified, they are assigned individual codewords for storage or transmission. A synthesizer can then generate speech by recreating the combinations of phonemes. Analysis of this technique indicates that the information content of speech can be transmitted with a data rate of 50 bps [7]. It must be emphasized, however, that what is transmitted is the information content associated with the words themselves, not the more subjective qualities of speech such as naturalness, voice inflections, accents, and speaker recognizability. Thus such techniques, by themselves, are not applicable to general telephony, which customarily includes qualities other than the message content of spoken words.

Efficient encoding of facsimile images presents similar opportunities and limitations. Facsimile machines typically scan at 200 dots per inch, which implies there are 3.74 million bits of raw information on a 8.5×11-in. piece of paper. If the paper contains only recognizable text characters at 10 characters and 6 lines per inch, the same information can be encoded as 5610 ASCII characters, or 39,270 bits, a savings of almost 100 to 1. Besides being restricted to text-oriented messages, character-oriented

encoding and decoding produces the same output character font independent of the source (which could conceivably be hand written). Facsimile machines achieve one significant level of coding efficiency without sacrificing transparency by encoding white space into run length codes. Although this does not reduce the number of bits in a worst-case (random-dot) image, it greatly reduces the number of bits in the average image of interest. Similar processing is possible in voice applications by efficiently encoding silence. However, the voice problem is more complicated because real-time voice requires reconstructing the temporal aspects of the source, restricting silence encoding to relatively large intervals.

Another level of speech analysis involves the actual recognition of spoken words. High levels of success have been achieved, with the two main restrictions that the system is trained on the speakers and the speakers are trained to speak with isolated words. As an example of one implementation [8] that tries to achieve the ultimate goal of speaker independence, continuous speech, and large vocabularies, 71–96% recognition accuracy is possible depending on the level of the grammar specified. (A grammar defines allowed sequences of words.)

Voice digitization techniques can be broadly categorized into two classes: those digitally encoding analog waveforms as faithfully as possible and those processing waveforms to encode only the perceptually significant aspects of speech and hearing processes. The first category is representative of the general problem of analog-to-digital and digital-to-analog conversions and is not restricted to speech digitization. The three most common techniques used to encode a voice waveform are pulse code modulation (PCM), differential PCM (DPCM), and delta modulation (DM). Except in special cases, telephone equipment designed to transparently reproduce an analog waveform used one of these techniques. Thus, when studying these common waveform encoding techniques, we are, in fact, studying the more general realm of analog-to-digital conversion [9].

The second category of speech digitization is concerned primarily with producing very low data rate speech encoders and decoders for narrowband transmission systems or digital storage devices with limited capacity. A device from this special class of techniques is commonly referred to as a "vocoder" (voice coder). Very low data rate (e.g., 1200-bps) vocoder techniques generally produce unnatural or synthetic sounding speech. As such, low-data-rate vocoders do not provide adequate quality for general telephony.

A great deal of effort has been expended to develop medium-rate (e.g., 8-kbps) voice coders with natural speech qualities, primarily for digital cellular applications. These coders are implemented as a combination or hybrid of the low-bit-rate techniques and the waveform coders. Thus, these techniques represent a third class of voice digitization algorithm.

3.1 PULSE AMPLITUDE MODULATION

The first step in digitizing an analog waveform is to establish a set of discrete times at which the input signal waveform is sampled. Prevalent digitization techniques are

based on the use of periodic, regularly spaced sample times. If the samples occur often enough, the original waveform can be completely recovered from the sample sequence using a low-pass filter to interpolate, or "smooth out," between the sample values. These basic concepts are illustrated in Figure 3.1. A representative analog waveform is sampled at a constant sampling frequency $f_s = 1/T$ and reconstructed using a low-pass filter. Notice that the sampling process is equivalent to amplitude modulation of a constant-amplitude pulse train. Hence the technique represented in Figure 3.1 is usually referred to as a pulse amplitude modulation (PAM).

3.1.1 Nyquist Sampling Rate

A classical result in sampling systems was established in 1933 by Harry Nyquist when he derived the minimum sampling frequency required to extract all information in a continuous, time-varying waveform. This result—the Nyquist criterion—is defined by the relation

$$f_s > (2)(\mathrm{BW})$$

where f_s = sampling frequency
BW = bandwidth of input signal

The derivation of this result is indicated in Figure 3.2, which portrays the spectrum of the input signal and the resulting spectrum of the PAM pulse train. The PAM spectrum can be derived by observing that a continuous train of impulses has a frequency spectrum consisting of discrete terms at multiples of the sampling frequency. The input signal amplitude modulates these terms individually. Thus a double-sideband spectrum is produced about each of the discrete frequency terms in the spectrum of the pulse train. The original signal waveform is recovered by a low-pass filter designed to remove all but the original signal spectrum. As shown in Figure 3.2, the reconstructive low-pass filter must have a cutoff frequency that lies between BW and f_s – BW. Hence, separation is only possible if f_s – BW is greater than BW (i.e., if $f_s > 2\mathrm{BW}$).

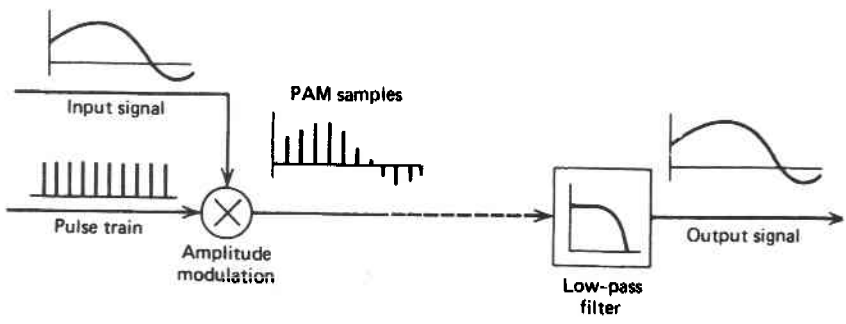


Figure 3.1 Pulse amplitude modulation.

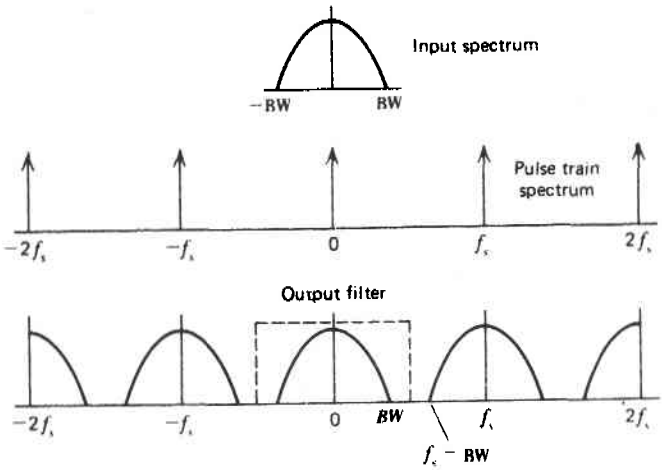


Figure 3.2 Spectrum of PAM signal.

3.1.2 Foldover Distortion

If the input waveform of a PAM system is undersampled ($f_s < 2BW$), the original waveform cannot be recovered without distortion. As indicated in Figure 3.3, this output distortion arises because the frequency spectrum centered about the sampling frequency overlaps the original spectrum and cannot be separated from the original spectrum by filtering. Since it is a duplicate of the input spectrum "folded" back on top of the desired spectrum that causes the distortion, this type of sampling impairment is often referred to as "foldover distortion."

In essence, foldover distortion produces frequency components in the desired frequency band that did not exist in the original waveform. Thus another term for this impairment is "aliasing." Aliasing problems are not confined to speech digitization processes. The potential for aliasing is present in any sample data system. Motion picture taking, for example, is another sampling system that can produce aliasing. A common example occurs when filming moving stagecoaches in old Westerns. Often the sampling process is too slow to keep up with the stagecoach wheel movements, and

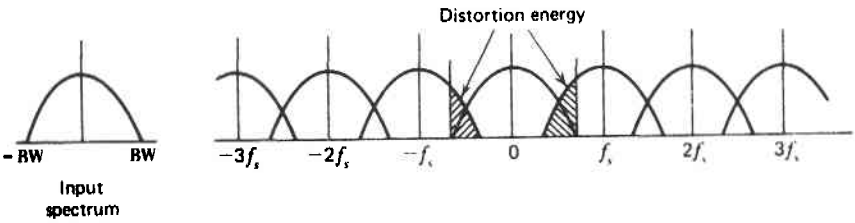


Figure 3.3 Foldover spectrum produced by undersampling an input.

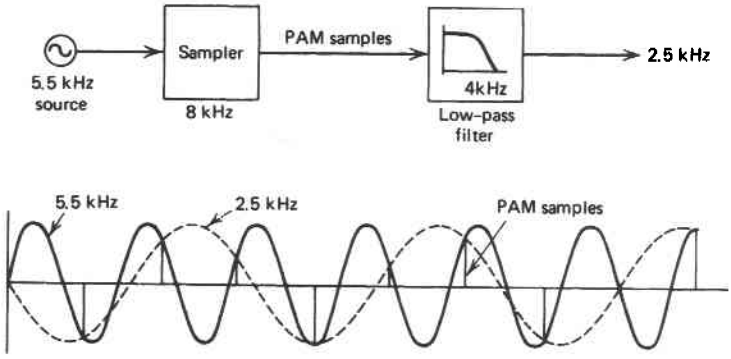


Figure 3.4 Aliasing of 5.5-kHz signal into a 2.5-kHz signal.

spurious rotation rates are produced. If the wheel rotates 355° between frames, it looks to the eye as if it has moved backward 5°.

Figure 3.4 demonstrates an aliasing process occurring in speech if a 5.5-kHz signal is sampled at an 8-kHz rate. Notice that the sample values are identical to those obtained from a 2.5-kHz input signal. Thus after the sampled signal passes through the 4-kHz output filter, a 2.5-kHz signal arises that did not come from the source. This example illustrates that the input must be bandlimited, *before sampling*, to remove frequency terms greater than $\frac{1}{2}f_s$, even if these frequency terms are ignored (i.e., are inaudible) at the destination. Thus, a complete PAM system, shown in Figure 3.5, must include a bandlimiting filter before sampling to ensure that no spurious or source-related signals get folded back into the desired signal bandwidth. The input filter of a voice codec may also be designed to cut off very low frequencies to remove 60-cycle hum from power lines.

Figure 3.5 shows the signal being recovered by a sample-and-hold circuit that produces a staircase approximation to the sampled waveform. With use of the staircase approximation, the power level of the signal coming out of the reconstructive filter is nearly the same as the level of the sampled input signal. The response of the reconstructive filter, in this case, must be modified somewhat to account for the spectrum of the wider "staircase" samples. (The modification amounts to dividing the "flat" filter spectrum by the spectrum of the finite width pulse. See Appendix C.)

The bandlimiting and reconstructive filters shown in Figure 3.5 are implied to have ideal characteristics.* Since ideal filters are physically unrealizable, a practical implementation must consider the effects of nonideal implementations. Filters with realizable attenuation slopes at the band edge can be used if the input signal is slightly oversampled.

As indicated in Figure 3.2, when the sampling frequency f_s is somewhat greater than twice the bandwidth, the spectral bands are sufficiently separated from each other

*An ideal filter is one with a frequency-independent time delay (linear phase), no attenuation in the passband (except as might be desired for pulse shaping), an arbitrarily steep cutoff, and infinite attenuation everywhere in the stopband.

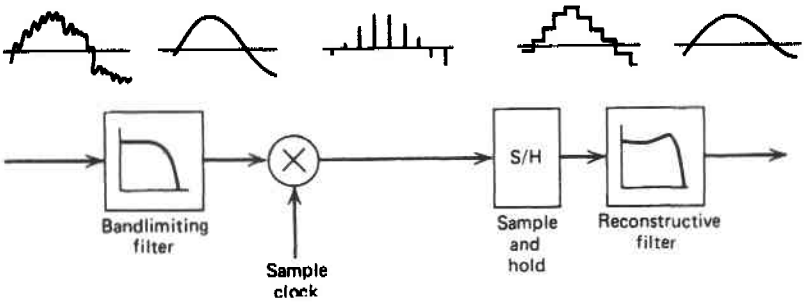


Figure 3.5 End-to-end PAM system.

that filters with gradual roll-off characteristics can be used. As an example, sampled voice systems typically use bandlimiting filters with a 3-dB cutoff around 3.4 kHz and a sampling rate of 8 kHz. Thus the sampled signal is sufficiently attenuated at the overlap frequency of 4 kHz to adequately reduce the energy level of the foldover spectrum. Figure 3.6 shows a filter template designed to meet ITU-T recommendations for out-of-band signal rejection in PCM voice coders. Notice that 14 dB of attenuation is provided at 4 kHz.

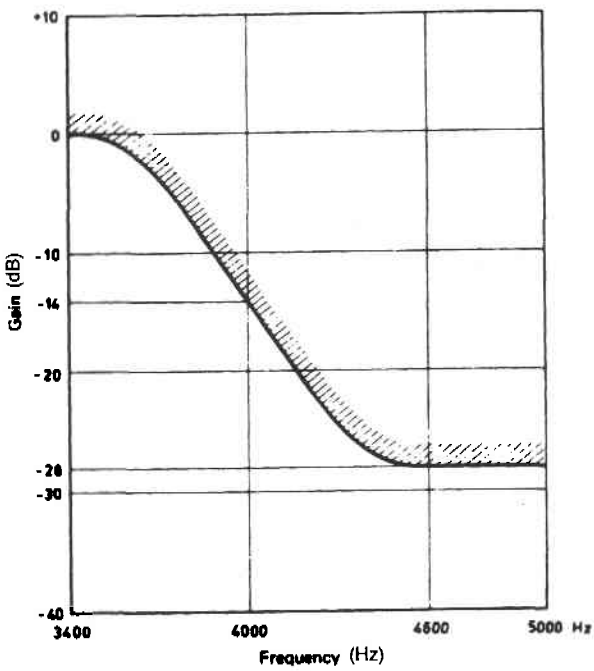


Figure 3.6 Bandlimiting filter template designed to meet ITU-T recommendations for PCM voice coders.

As mentioned in Chapter 1, the perceived quality of a voice signal is not greatly dependent upon the phase response of the channel (the relative delay of individual frequency components). For this reason the phase responses of the bandlimiting filters in the encoders and the smoothing filters in the decoders are not critical. Nonlinear phase responses in these filters, however, do impact high-rate voiceband data signals when digitized. Hence a somewhat paradoxical situation arises when voiceband data are transmitted over a T-carrier line: the process of converting the voiceband data signal (28,800 bps typical maximum) to a virtually error free 64-kbps digital signal causes distortion in the lower rate data signal. However, because of regeneration the transmission process itself does not add to the signal degradation.

By interleaving the samples from multiple sources, PAM systems can be used to share a transmission facility in a time division multiplex manner. As previously mentioned, PAM systems are not generally useful over long distances owing to the vulnerability of the individual pulses to noise, distortion, intersymbol interference, and crosstalk.* Instead, for long-distance transmission the PAM samples are converted into a digital format, thereby allowing the use of regenerative repeaters to remove transmission imperfections before errors result.

3.2 PULSE CODE MODULATION

The preceding section describes pulse amplitude modulation, which uses discrete sample times with analog sample amplitudes to extract the information in a continuously varying analog signal. Pulse code modulation (PCM) is an extension of PAM wherein each analog sample value is quantized into a discrete value for representation as a digital codeword. Thus, as shown in Figure 3.7, a PAM system can be converted into a PCM system by adding an analog-to-digital (A/D) converter at the source and a digital-to-analog (D/A) converter at the destination. Figure 3.8 depicts a typical quantization process in which a set of quantization intervals is associated in a one-to-one fashion with a binary codeword. All sample values falling in a particular quantization interval are represented by a single discrete value located at the center of the quantization interval. In this manner the quantization process introduces a certain amount of error or distortion into the signal samples. This error, known as *quantiza-*

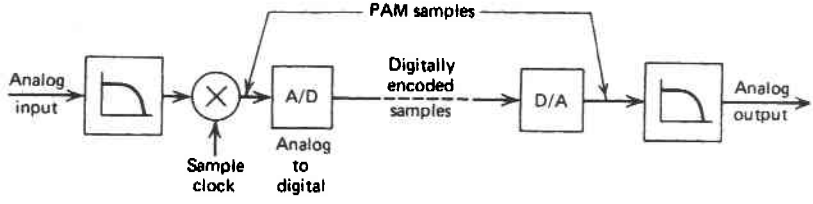


Figure 3.7 Pulse code modulation.

*As discussed in Chapter 11 the emergence of sophisticated DSP equalization algorithms in V.90 modems enables PAM transmission on analog subscriber loops.

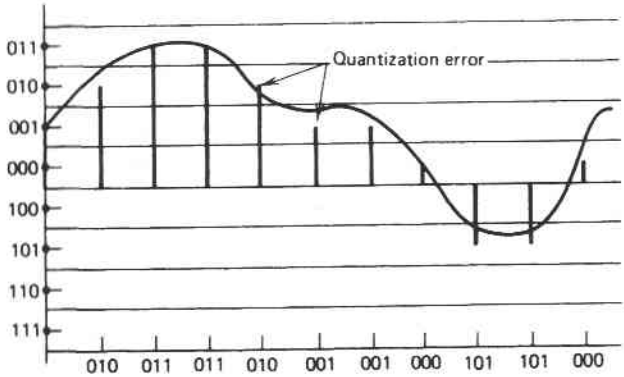


Figure 3.8 Quantization of analog samples.

tion noise, is minimized by establishing a large number of small quantization intervals. Of course, as the number of quantization intervals increases, so must the number of bits increase to uniquely identify the quantization intervals.

3.2.1 Quantization Noise

A fundamental aspect of the design and development of an engineering project is the need for analytical measures of systems performance. Only then can a system be objectively measured and its cost effectiveness compared to alternate designs. One of the measures needed by a voice communication engineer is the quality of speech delivered to the listener. Measurements of speech quality are complicated by subjective attributes of speech as perceived by a typical listener. One subjective aspect of noise or distortion on a speech signal involves the frequency content, or spectrum, of the disturbance in conjunction with the power level. These effects of noise as a function of frequency are discussed in Chapter 1 with the introduction of C-message and psophometric weighting.

Successive quantization errors of a PCM encoder are generally assumed to be distributed randomly and uncorrelated to each other. Thus the commutative effect of quantization errors in a PCM system can be treated as additive noise with a subjective effect that is similar to bandlimited white noise. Figure 3.9 shows the quantization noise as a function of signal amplitude for a coder with uniform quantization intervals. Notice that if the signal has enough time to change in amplitude by several quantization intervals, the quantization errors are independent. If the signal is oversampled (i.e., sampled much higher than the Nyquist rate), successive samples are likely to fall in the same interval, causing a loss of independence in the quantization errors.

The quantization error or distortion created by digitizing an analog signal is customarily expressed as an average noise power relative to the average signal power. Thus the signal-to-quantizing-noise ratio (SQR, also called a signal-to-distortion ratio or a signal-to-noise ratio) can be determined as

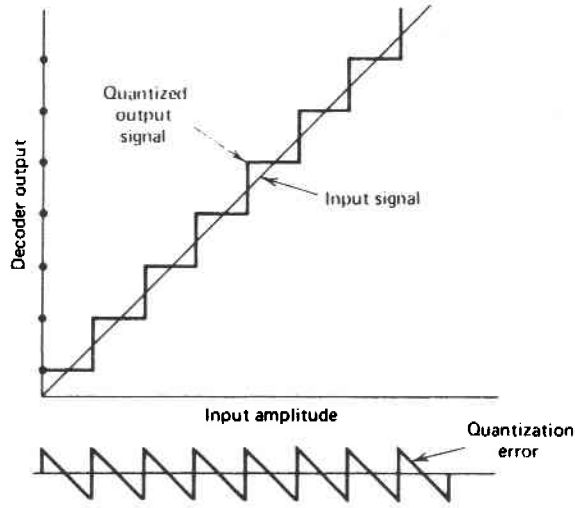


Figure 3.9 Quantization error as a function of amplitude over a range of quantization intervals.

$$\text{SQR} = \frac{E\{x^2(t)\}}{E\{[y(t) - x(t)]^2\}} \tag{3.1}$$

where $E\{\cdot\}$ = expectation or averaging
$x(t)$ = analog input signal
$y(t)$ = decoded output signal

In determining the expected value of the quantization noise, three observations are necessary:

1. The error $y(t) - x(t)$ is limited in amplitude to $q/2$, where q is the height of the quantization interval. (Decoded output samples are ideally positioned at the middle of a quantization interval.)
2. A sample value is equally likely to fall anywhere within a quantization interval, implying a uniform probability density of amplitude $1/q$.
3. Signal amplitudes are assumed to be confined to the maximum range of the coder. If a sample value exceeds the range of the highest quantization interval, overload distortion (also called peak limiting) occurs.

If we assume (for convenience) a resistance level of 1 Ω, the average quantization noise power is determined in Appendix A as

$$\text{Quantization noise power} = \frac{1}{12} q^2 \tag{3.2}$$

If all quantization intervals have equal lengths (uniform quantization), the quantization noise is independent of the sample values and the SQR is determined as*

$$\text{SQR (db)} = 10 \log_{10}\left(\frac{v^2}{q^2/12}\right)$$

$$= 10.8 + 20 \log_{10}\left(\frac{v}{q}\right) \tag{3.3}$$

where v is the rms amplitude of the input. In particular, for a sine wave input the SQR produced by uniform quantization is

$$\text{SQR (dB)} = 10 \log_{10}\left(\frac{A^2/2}{q^2/12}\right)$$

$$= 7.78 + 20 \log_{10}\left(\frac{A}{q}\right) \tag{3.4}$$

where A is the peak amplitude of the sine wave.

Example 3.1. A sine wave with a 1-V maximum amplitude is to be digitized with a minimum SQR of 30 dB. How many uniformly spaced quantization intervals are needed, and how many bits are needed to encode each sample?

Solution. Using Equation 3.4, the maximum size of a quantization interval is determined as

$$q = (1)10^{-(30-7.78)/20}$$

$$= 0.078 \text{ V}$$

Thus 13 quantization intervals are needed for each polarity for a total of 26 intervals in all. The number of bits required to encode each sample is determined as

$$N = \log_2(26) = 4.7 = 5 \text{ bits per sample}$$

When measuring quantization noise power, the spectral content is often weighted in the same manner as noise in an analog circuit. Unfortunately, spectrally weighted noise measurements do not always reflect the true perceptual quality of a voice en-

*The SQRs commonly compare unfiltered decoder outputs to unfiltered quantization errors. In actual practice, the decoder output filter reduces the power level of both the signal and the noise. The noise power experiences a greater reduction than the signal power, since the uncorrelated noise samples have a wider spectrum than the correlated voice samples. Thus filtered signal-to-noise ratios are usually higher than the values calculated here by 1–2 dB.

coder/decoder. If the spectral distribution of the quantization noise more or less follows the spectral content of the speech waveform, the noise is masked by the speech and is much less noticeable than noise uncorrelated to the speech [10]. On the other hand, if the quantization process produces energy at voiceband frequencies other than those contained in particular sounds, they are more noticeable.

High-quality PCM encoders produce quantization noise that is evenly distributed across voice frequencies and independent of the encoded waveforms. Thus quantization noise ratios defined in Equation 3.4 are good measures of PCM performance. In some of the encoders discussed later (vocoders in particular), quantization noise power is not very useful. References [9], [11], and [12] describe other measures of encoder speech quality providing better correlations to quality as perceived by a listener.

3.2.2 Idle Channel Noise

Examination of Equations 3.3 and 3.4 reveals that the SQR is small for small sample values. In fact, as shown in Figure 3.10, the noise may actually be greater than the signal when sample values are in the first quantization interval. This effect is particularly bothersome during speech pauses and is known as idle channel noise. Figure 3.11 depicts one method of minimizing idle channel noise in PCM systems by establishing a quantization interval that straddles the origin. In this case all sample values in the central quantization interval are decoded as a constant zero output. PCM systems of this type use an odd number of quantization intervals since the encoding ranges of positive and negative signals are usually equal.

The quantization characteristics required to produce the output waveforms shown in Figures 3.10 and 3.11 are shown in Figures 3.12 and 3.13, respectively. The first characteristic (midriser) cannot produce a zero output level. The second characteristic (midtread) decodes very low signals into constant, zero-level outputs. However, if the signal amplitude is comparable to the size of the quantization interval or if a dc bias exists in the encoder, midtread quantization will produce about as much idle channel noise as midriser quantization.

As mentioned in Chapter 1, noise occurring during speech pauses is more objectionable than noise with equivalent power levels during speech. Thus idle channel

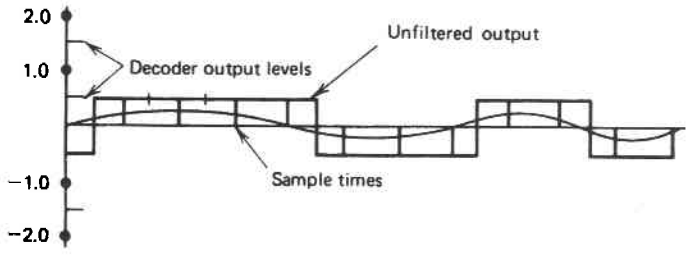


Figure 3.10 Idle channel noise produced by midriser quantization.

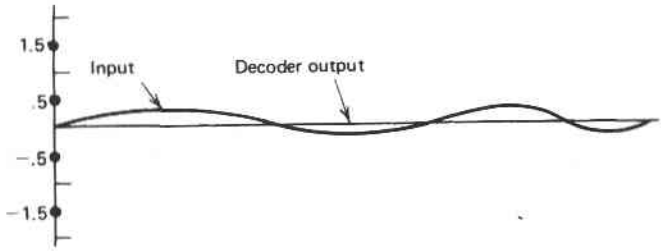


Figure 3.11 Elimination of idle channel noise by midtread quantization.

noise is specified in absolute terms separate from quantization noise, which is specified relative to the signal level. For example, Bell system D3 channel bank specifications list the maximum idle channel noise as 23 dBrnC0 [13].

3.2.3 Uniformly Encoded PCM

An encoder using equal-length quantization intervals for all samples produces codewords linearly related to the analog sample values. That is, the numerical equivalent of each codeword is proportional to the quantized sample value it represents. In this manner a uniform PCM system uses a conventional analog-to-digital converter to generate the binary sample codes. The number of bits required for each sample is determined by the maximum acceptable noise power. Minimum digitized voice quality requires a signal-to-noise ratio in excess of 26 dB [14]. For a uniform PCM system to achieve a SQR of 26 dB, Equation 3.4 indicates that $q_{max} = 0.123A$. For equal positive and negative signal excursions (encoding from $-A$ to A), this result indicates that just over 16 quantization intervals, or 4 bits per sample, are required.*

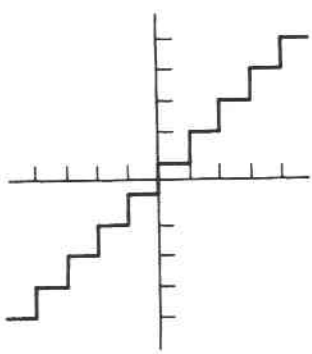


Figure 3.12 Midriser quantizer characteristic.

*This SQR objective is for minimum acceptable performance and assumes all degradations occur in a sin gle encoder. If additional signal impairments occur (such as multiple A/D conversions), the encoder must use more bits to provide noise margin for other elements in the network.

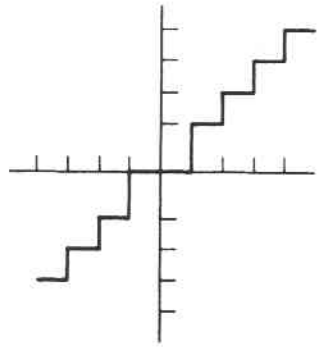


Figure 3.13 Midtread quantizer characteristic.

In addition to providing adequate quality for small signals, a telephone system must be capable of transmitting a large range of signal amplitudes, referred to as dynamic range. Dynamic range (DR) is usually expressed in decibels as the ratio of the maximum amplitude signal to the minimum amplitude signal:

$$\begin{aligned} \mathrm{DR} &= 10 \log_{10}\left(\frac{P_{\max}}{P_{\min}}\right) \\ &= 20 \log_{10}\left(\frac{V_{\max}}{V_{\min}}\right) \end{aligned} \tag{3.5}$$

A typical minimum dynamic range is 30 dB [14]. Thus signal values as large as 31 times A must be encoded without exceeding the range of quantization intervals. Assuming equally spaced quantization intervals for uniform coding, the total number of intervals is determined as 496, which requires 9-bit codewords.*

The performance of an n-bit uniform PCM system is determined by observing that

$$q = \frac{2A_{\max}}{2^n} \tag{3.6}$$

where $A_{\max}$ is the maximum (nonoverloaded) amplitude.

Substituting Equation 3.6 into Equation 3.4 produces the PCM performance equation for uniform coding:

$$\mathrm{SQR} = 1.76 + 6.02n + 20 \log_{10}\left(\frac{A}{A_{\max}}\right) \tag{3.7}$$

*This result is derived with the assumption of minimum performance requirements. Higher performance objectives (less quantization noise and greater dynamic range) require as many as 13 bits per sample for uniform PCM systems. This coding performance was established when it was likely that multiple conversions would occur in an end-to-end connection. Now that the possibility of multiple A/D and D/A conversions has been eliminated, end-to-end voice quality is much better than it was in the analog network.

The first two terms of Equation 3.7 provide the SQR when encoding a full-range sine wave. The last term indicates a loss in SQR when encoding a lower level signal. These relationships are presented in Figure 3.14, which shows the SQR of a uniform PCM system as a function of the number of bits per sample and the magnitude of an input sine wave.

Example 3.2. What is the minimum bit rate that a uniform PCM encoder must provide to encode a high-fidelity audio signal with a dynamic range of 40 dB? Assume the fidelity requirements dictate passage of a 20-kHz bandwidth with a minimum signal-to-noise ratio of 50 dB. For simplicity, assume sinusoidal input signals.

Solution. To prevent foldover distortion, the sampling rate must be at least 40 kHz. Assuming an excess sampling factor comparable to that used in D-type channel banks (4000/3400), we choose a sampling rate of 48 kHz as a compromise for a practical bandlimiting filter. By observing that a full-amplitude signal is encoded with an SQR of 40 + 50 = 90 dB, we can use Equation 3.7 to determine the number of bits n required to encode each sample:

$$n = \frac{90 - 1.76}{6.02}$$

$$= 15 \text{ bits}$$

Thus the required bit rate is

$$(15 \text{ bits/sample})(48{,}000 \text{ samples/sec}) = 720 \text{ kbps}$$

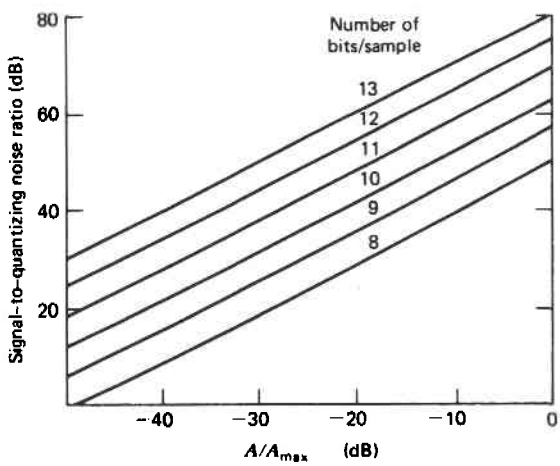


Figure 3.14 SQR of uniform PCM coding.

3.2.4 Companding

In a uniform PCM system the size of every quantization interval is determined by the SQR requirement of the lowest signal level to be encoded. Larger signals are also encoded with the same quantization interval. As indicated in Equation 3.7 and Figure 3.14, the SQR increases with the signal amplitude A. For example, a 26-dB SQR for small signals and a 30-dB dynamic range produces a 56-dB SQR for a maximum-amplitude signal. In this manner a uniform PCM system provides unneeded quality for large signals. Moreover, the large signals are the least likely to occur. For these reasons the code space in a uniform PCM system is very inefficiently utilized.

A more efficient coding procedure is achieved if the quantization intervals are not uniform but allowed to increase with the sample value. When quantization intervals are directly proportional to the sample value, the SQR is constant for all signal levels. With this technique fewer bits per sample provide a specified SQR for small signals and an adequate dynamic range for large signals. When the quantization intervals are not uniform, a nonlinear relationship exists between the codewords and the samples they represent.

Historically, the nonlinear function was first implemented on analog signals using nonlinear devices such as specially designed diodes [15]. The basic process is shown in Figure 3.15, where the analog input sample is first compressed and then quantized with uniform quantization intervals. The effect of the compression operation is shown in Figure 3.16. Notice that successively larger input signal intervals are compressed into constant-length quantization intervals. Thus the larger the sample value, the more it is compressed before encoding. As shown in Figure 3.15, a nonuniform PCM decoder expands the compressed value using an inverse compression characteristic to recover the original sample value. The process of first compressing and then expanding a signal is referred to as *companding*. When digitizing, companding amounts to assigning small quantization intervals to small samples and large quantization intervals to large samples.

Various compression–expansion characteristics can be chosen to implement a compandor. By increasing the amount of compression, we increase the dynamic range at the expense of the signal-to-noise ratio for large-amplitude signals. One family of compression characteristics used in North America and Japan is the μ-law characteristic, defined as

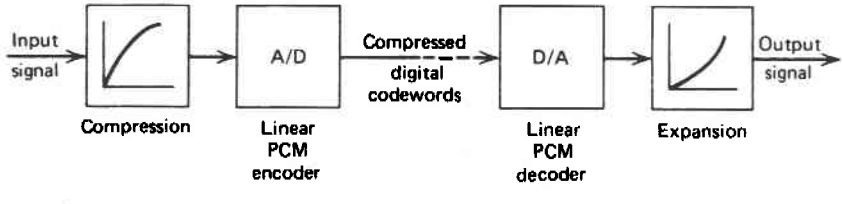


Figure 3.15 Companded PCM with analog compression and expansion.

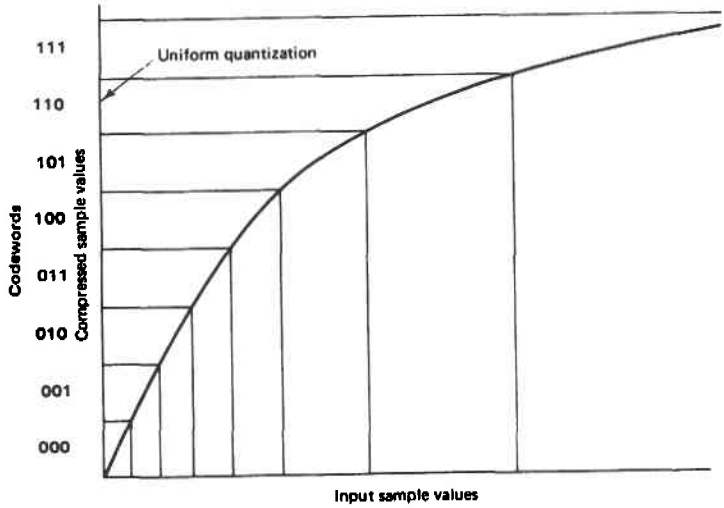


Figure 3.16 Typical compression characteristic.

$$F_{\mu}(x) = \text{sgn}(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)} \tag{3.8}$$

where $x =$ input signal amplitude ($-1 \le x \le 1$)
$\text{sgn}(x) =$ polarity of x
$\mu =$ parameter used to define amount of compression

Because of the mathematical nature of the compression curve, companded PCM is sometimes referred to as log-PCM. A logarithm compression curve is ideal in the sense that quantization intervals and, hence, quantization noise are proportional to the sample amplitude. The inverse or expansion characteristic for a μ-law compandor is defined as

$$F_{\mu}^{-1}(y) = \text{sgn}(y)\left(\frac{1}{\mu}\right)[(1+\mu)^{|y|} - 1] \tag{3.9}$$

where $y =$ the compressed value, $=F_{\mu}(x)$ $(-1 \le y \le 1)$
$\text{sgn}(y) =$ polarity of y
$\mu =$ companding parameter

The first T-carrier systems in the United States used D1 channel banks [16], which approximated Equation 3.8 for $\mu = 100$. The compression and expansion functions were implemented with the specially biased diodes mentioned previously. Figure 3.17 depicts a block diagram of a D1 channel bank. Notice that the time division multiplexing and demultiplexing functions are implemented on analog PAM samples. Thus the companding and encoding/decoding functions were shared by all 24 voice channels.

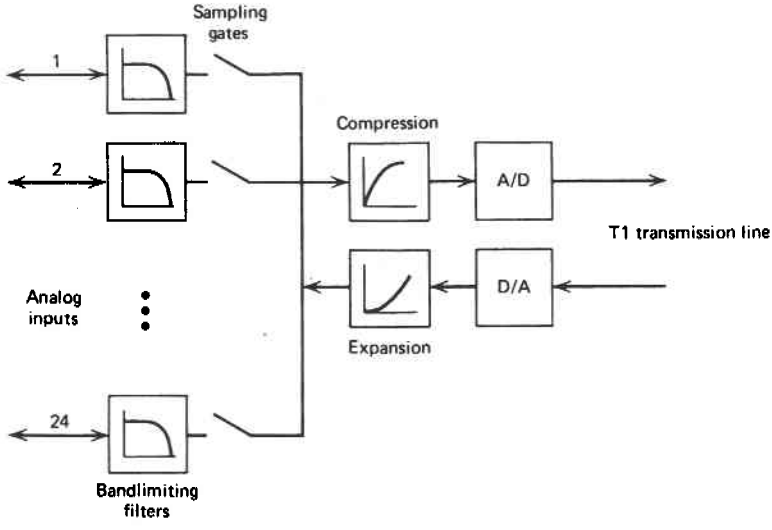


Figure 3.17 Functional block diagram of D1 channel bank.

The ability to share this relatively expensive equipment was one of the reasons that PCM was originally chosen as the means of digitally encoding speech. Subsequent development of integrated circuit PCM codecs diminished the need to share this equipment. Thus later generation systems could use per-channel codecs and provide more flexibility in implementing various system sizes. When most of the cost of a channel bank is in the common equipment, as in the original channel banks, less than fully provisioned systems are overly expensive.

Each sample produced by a D1 channel bank was encoded into 7 bits: 1 polarity bit and 6 compressed magnitude bits. In addition, 1 signaling bit was added to each channel to produce an 8-bit codeword for each time slot. Since the sampling rate was 8 kHz, a 64-kbps channel resulted. Even though the D1 channel banks have been superceded by newer channel banks utilizing a different coding format, the 64-kbps channel rate has persisted as a standard.

3.2.5 Easily Digitally Linearizable Coding

The success of the first T1 systems for interoffice exchange area transmission paved the way for further use of digital time division multiplex systems. As it became apparent that digital transmission was useful within the toll network, it also became clear that the form of PCM encoding used in the D1 channel banks was inadequate. In contrast to the exchange area, an end-to-end connection through the toll network could have conceivably involved as many as nine tandem connections. Since digital switching was not in existence at the time that T-carrier systems for the toll network were being developed, each of these tandem connections implied an extra D/A and A/D

conversion. Thus the quality of each conversion had to be improved to maintain the desired end-to-end quality. The D2 channel bank [17] was therefore developed with improved digitized voice quality.

When the D2 channel bank was being developed, digital switching was recognized as a coming technology, implying that channel banks would be paired on a dynamic basis, as opposed to the one-to-one basis in T-carrier systems. Thus a greater degree of uniformity in companding characteristics would be required to allow pairing of channel banks on a nationwide basis. The main features incorporated into the D2 channel bank (and ensuing channel banks such as D3, D4, and D5) to achieve the improved quality and standardization are:

1. Eight bits per PCM codeword
2. Incorporation of the companding functions directly into the encoder and decoder
3. A new companding characteristic (μ255)

The D1 channel banks use 1 bit per time slot for signaling and 7 bits for voice. Thus a signaling rate of 8 kbps was established, which was more than necessary for basic voice service. To provide a higher data rate for voice, signaling between D2 and all subsequent channel banks is inserted into the least significant bit position of 8-bit code-words in every sixth frame. Thus every sixth μ255 PCM codeword contains only 7 bits of voice information, implying that the effective number of bits per sample is actually $7\frac{5}{6}$ bits instead of 8. The use of 1 bit in every sixth frame for signaling is often referred to as "robbed bit signaling." When common-channel signaling is utilized, the associated T-carrier systems no longer need to carry signaling information on a per-channel basis and a full 8 bits of voice can be transmitted in every time slot of every frame.

The compression and expansion characteristics of the D1 channel banks were implemented separately from the encoders and decoders. The D2 channel bank incorporates the companding operations into the encoders and decoders themselves. In these channel banks a resistor array is used to establish a sequence of nonuniformly spaced decision thresholds. A sample value is encoded by successively comparing the input value to the sequence of decision thresholds until the appropriate quantization interval is located. The digital output becomes whatever code is used to represent the particular quantization interval. (See Appendix B for a detailed description of the direct encoding procedure used in the D2 channel banks.) By incorporating the companding functions directly into the encoders and decoders, D2 channel banks avoid certain problems associated with parameter variability and temperature sensitivity of companding diodes in D1 channel banks.

The D2 channel banks also introduced improved performance in terms of the effect of channel errors on the decoded outputs. Of paramount concern in a PCM system is the effect of a channel error in the most significant bit position of a codeword. Bit errors in other positions of a codeword are much less noticeable to a listener. A channel error in the most significant bit of a codeword produced by a D1 channel bank causes an output error that is always equal to one-half of the entire coding range of the coder.

The D2 channel bank, on the other hand, uses a sign-magnitude coding format. With this format, a channel error in the polarity bit causes an output error that is equal to twice the sample magnitude (i.e., the polarity is inverted). In the worst case this error may correspond to the entire encoding range. Maximum-amplitude samples are relatively rare, however, so most channel errors in a D2 coding format produce outputs with error magnitudes less than one-half the coding range. Thus, on average, the sign-magnitude coding format of the D2 and ensuing channel banks provide superior performance in the presence of channel errors [17].

In addition to a need for improved voice quality, it also became apparent that as more of the network began using digital techniques, it would be necessary, or at least desirable, to implement many signal processing functions directly on digital signals and not convert them to an analog format for processing. Most signal processing functions (such as attenuating a signal or adding signals together) involve linear operations. Thus before processing a log-PCM voice signal, it is necessary to convert the compressed transmission format into a linear (uniform) format.

To simplify the conversion process, the particular companding characteristic with $\mu = 255$ was chosen. This companding characteristic has the property of being closely approximated by a set of eight straight-line segments also referred to as chords. Furthermore, the slope of each successive segment is exactly one-half the slope of the previous segment. The first four segments of a μ255 approximation are shown in Figure 3.18. The overall result is that the larger quantization intervals have lengths that are binary multiples of all smaller quantization intervals. Because of this property, a compressed codeword is easily expanded into a uniform representation. Similarly, the uniform representation is easily converted into a compressed representation.* In fact, commercially available PCM codecs digitally compress uniform codewords instead of using direct compressed encoding, as in the D2 channel banks. These techniques use a uniform encoder with a relatively large number of bits to cover the entire dynamic range of the signal. As described in Appendix B, the least significant bits of large sample values are discarded when compressing the code. The number of insignificant bits deleted is encoded into a special field included in the compressed-code format. In this manner digital companding is analogous to expressing a number in scientific notation.

As shown in Figure 3.18, each major segment of the piecewise linear approximation is divided into equally sized quantization intervals. For 8-bit codewords the number of quantization intervals per segment is 16. Thus an 8-bit μ255 codeword is composed of 1 polarity bit, 3 bits to identify a major segment, and 4 bits for identifying a quantizing interval within a segment. Table 3.1 lists the major segment endpoints, the quantization intervals, and the corresponding segment and quantization interval codes.

The quantization intervals and decoded sample values in Table 3.1 have been expressed in terms of a maximum-amplitude signal of 8159 so that all segment endpoints and decoder outputs are integers. Notice that the quantizing step is doubled in each of

*The inexorable advance of semiconductor technology has obviated much of the ingenuity that went into selecting EDL coding formats. Brute-force table look-up conversion between codes using read-only memories (ROMs) is now the most cost-effective approach.

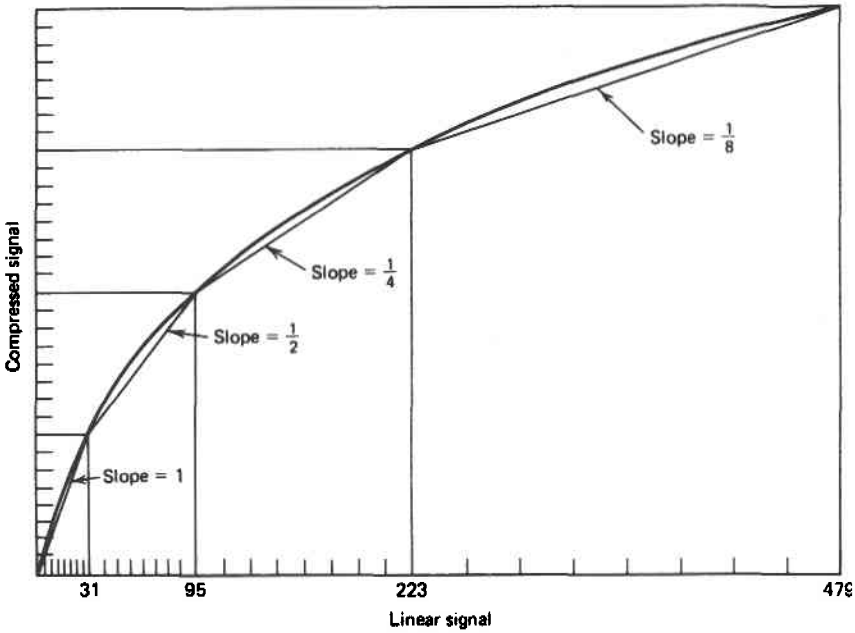


Figure 3.18 First four segments of straight-line approximation to μ255 compression curve.

eight successive linear segments. It is this property that facilitates the conversion to and from a uniform format. A complete encoding table is provided in Appendix B along with a detailed description of the conversion process to and from uniform codes.

The straight-line approximation of the μ255 companding curve is sometimes referred to as a 15-segment approximation. The 15 segments arise because, although there are 8 positive segments and 8 negative segments, the 2 segments nearest the origin are colinear and therefore can be considered as 1. When viewed in this manner, the middle segment contains 31 quantization intervals with 1 segment straddling the origin (from −1 to +1 in Table 3.1). Codewords for this middle quantization interval arise as a positive value less than +1 or a negative value greater than −1. There are, in effect, a positive zero and a negative zero. As represented in Table 3.1, these values are encoded as 00000000 and 10000000, respectively. However, μ255 PCM codecs invert all codewords for transmission. The smaller amplitude signals, with mostly 0 bits in the segment code, are most probable and would therefore cause less than 50% pulses on the transmission line. The density of pulses is increased by inversion of the transmitted data, which improves the timing and clock recovery performance of the receiving circuitry in the regenerative repeaters. Thus the actual transmitted codewords corresponding to a positive zero and a negative zero are respectively 11111111 and 01111111, indicating strong timing content for the line signal of an idle channel.

In the interest of ensuring clock synchronization in the T1 repeaters, μ255 PCM codecs alter the transmitted data in one other way. As indicated in Table 3.1, a maximum-amplitude negative signal is all 1's, which would normally be converted to all 0's for

TABLE 3.1 Encoding/Decoding Table for μ255 PCM[a]

Input Amplitude Range	Step Size	Segment Code S	Quantization Code Q	Code Value	Decoder Amplitude
0–1	1		0000	0	0
1–3			0001	1	2
3–5	2	000	0010	2	4
⋮			⋮	⋮	⋮
29–31			1111	15	30
31–35			0000	16	33
⋮	4	001	⋮	⋮	⋮
91–95			1111	31	93
95–103			0000	32	99
⋮	8	010	⋮	⋮	⋮
215–223			1111	47	219
223–239			0000	48	231
⋮	16	011	⋮	⋮	⋮
463–479			1111	63	471
479–511			0000	64	495
⋮	32	100	⋮	⋮	⋮
959–991			1111	79	975
991–1055			000	80	1023
⋮	64	101	⋮	⋮	⋮
1951–2015			1111	95	1983
2015–2143			0000	96	2079
⋮	128	110	⋮	⋮	⋮
3935–4063			1111	111	3999
4063–4319			0000	112	4191
⋮	256	111	⋮	⋮	⋮
7903–8159			1111	127	8031

[a]This table displays magnitude encoding only. Polarity bits are assigned a "0" for positive and a "1" for negative. In transmission all bits are inverted.

transmission. Instead, for the all-0 codeword only, the second least significant bit is set to 1 so that 00000010 is transmitted. In effect, an encoding error is produced to preclude an all-0 codeword. Fortunately, maximum-amplitude samples are extremely unlikely so that no significant degradation occurs. (If the least significant bit were forced to a 1, a smaller decoding error would result. However, in every sixth frame this bit is "stolen" for signaling purposes and therefore is occasionally set to 0 independently of the codeword. To ensure that "all-zero" codewords are never transmitted, the second least significant bit is forced to a 1 when necessary.)

Example 3.3. Determine the sequence of codewords for a μ255 PCM encoded channel bank format representing a half-maximum-power 1-kHz digital signal.

Solution. Since the sampling rate of the standard μ255 PCM channel bank is 8 kHz, a sequence of eight samples can be repeated in cyclic fashion to generate the 1-kHz waveform. For convenience, the phases of the samples are chosen to begin at 22.5°. Thus the eight samples correspond to 22.5°, 67.5°, 112.5°, 157.5°, 202.5°, 247.5°, 292.5°, and 337.5°. With these phases, only two different magnitudes corresponding to 22.5° and 67.5° are required. The maximum amplitude of a half-maximum-power sine wave is $0.707 \times 8159 = 5768$. Thus the two amplitudes contained in the sample sequences are

$$(5768)\sin(22.5°) = 2207 \qquad (5768)\sin(67.5°) = 5329$$

Using the encoding table in Appendix B, we determine the codes for these sample magnitudes to be 1100001 and 1110100, respectively. The sequence of eight samples can now be established as follows:

Sample Phase (deg)	Polarity	Segment			Quantization			
22.5	0	1	1	0	0	0	0	1
67.5	0	1	1	1	0	1	0	0
112.5	0	1	1	1	0	1	0	0
157.5	0	1	1	0	0	0	0	1
202.5	1	1	1	0	0	0	0	1
247.5	1	1	1	1	0	1	0	0
292.5	1	1	1	1	0	1	0	0
337.5	1	1	1	0	0	0	0	1

Note: This sequence defines a 1-kHz test signal at a power level of 1 mW at the transmission level point (0 dBm0). However, the actual transmitted data pattern is the complement of that provided above. Because only two amplitude samples are required to produce the test tone, this tone does not test all encoding/decoding circuitry. In general, a 1004-Hz tone is a better test tone since it is not harmonically related to an 8000-Hz sampling rate and will therefore exercise all encoder and decoder levels.

Performance of a μ255 PCM Encoder

As mentioned, the main motivation for changing the encoding algorithm of the D1 channel bank was to provide better speech quality for digital toll network transmission links. The SQR for a maximum-amplitude sine wave in the first segment of a μ255 codec is determined easily from Equation 3.4 as

$$\mathrm{SQR}(A = 31) = 7.78 + 20 \log_{10}\left(\frac{31}{2}\right)$$

$$= 31.6 \text{ dB}$$

The SQRs for larger amplitude sinusoids are not as easy to calculate since the lengths of the quantization intervals vary with the sample size. Thus a general calculation of the quantizing noise power involves finding the expected value of the power of the quantization errors:

$$\text{noise power} = \frac{1}{12}\sum_{i=0}^{7} p_i q_i^2 \tag{3.10}$$

where p_i = probability of a sample in ith segment
q_i = quantization size for segment i, = 2^{i+1} for segmented μ255 coding

Using Equation 3.10, we determine the SQR power for a full-range sinusoid as

$$\text{SQR}(A = 8159) = 39.3 \text{ dB}$$

For comparison, if all quantization intervals had the maximum length of 256 as in the upper segment, Equation 3.4 provides an SQR of 37.8 dB. The difference of only 1.5 dB reflects the fact that a full-scale sine wave spends 67% of the time in the upper segment where the quantization intervals are maximum (i.e., $p_7 = 0.67$). A voice signal, on the other hand, has a higher peak-to-average ratio than a sine wave. The average quantization error is smaller, but so is the average signal power. Hence the SQR is approximately the same.

The dynamic range of a segmented PCM encoder is determined as the signal power difference between a low-level signal occupying the entire range of the first segment and a high-level signal extending to the limits of the code. Thus the dynamic range of a segmented 255 coder is determined as

$$\text{DR}(A = 31 \text{ to } A = 8159) \doteq 20 \log_{10}(8159/31) = 48.4 \text{ dB}$$

In summary, an 8-bit μ255 PCM codec provides a theoretical SQR greater than 30 dB across a dynamic range of 48 dB. For comparison, Equation 3.4 or Figure 3.14 reveals that a uniform PCM encoder/decoder requires 13 bits for equivalent performance. (The extra quality of uniform encoding at high signal levels is unneeded.)

The theoretical performance of an 8-bit segmented μ255 coder is shown in Figure 3.19 as a function of the amplitude of a sine wave input. Also shown is the theoretical performance of an unsegmented μ255 coder and a 7-bit μ100 coder used in the D1 channel bank. Notice that the 8-bit coders provide about 5 dB improvement over the 7-bit coder for high-level signals and even more improvement for low-level signals. The performance shown for the 8-bit coders does not include the effect of using only 7 bits for voice coding in every sixth frame. When this effect is included, the 8-bit coders lose 1.76 dB in performance.

The scalloped effect of the segmented coder occurs because the quantization intervals change abruptly at the segment endpoints instead of continuously as in analog companding. Also as shown in Figure 3.19, note the required performance of a D3 channel bank and codecs designed to meet this specification [13]. This specification assumes all noise measurements are made using C-message weighting. C-message weighting reduces the effective noise level by 2 dB and therefore improves the SQR. Thus an ideal 8-bit μ255 coder actually exceeds the specification by more than that

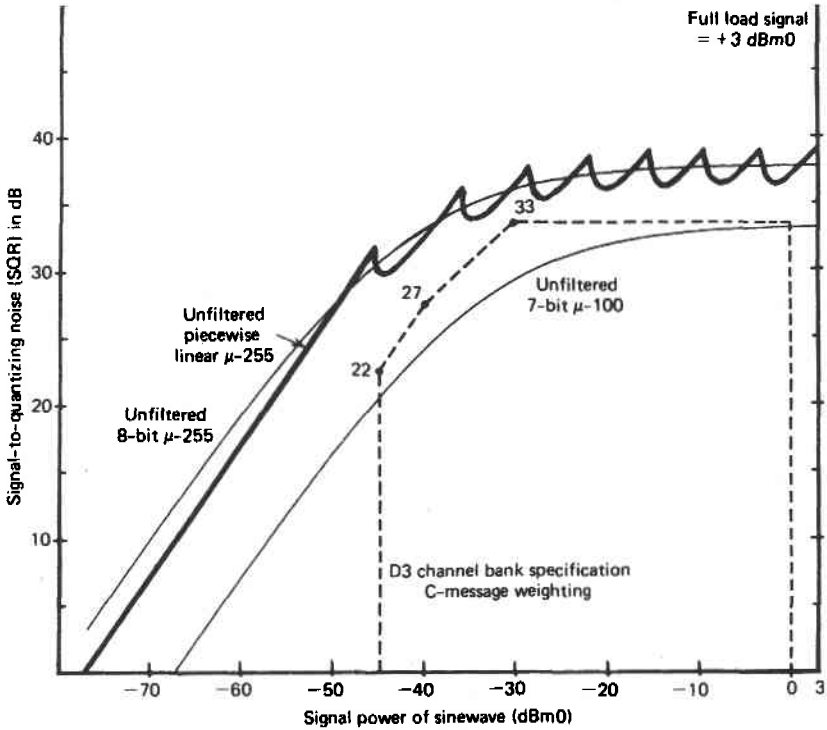


Figure 3.19 SQR of μ-law coding with sine wave inputs.

shown in Figure 3.19. When the least significant bit of every sixth frame is used for signaling, however, the SQR is reduced by a comparable 1.76 dB.

A-Law Companding

The companding characteristic recommended by ITU-T for Europe and most of the rest of the world is referred to as an A-law characteristic. This characteristic has the same basic features and implementation advantages as does the μ-law characteristic. In particular, the A-law characteristic can also be well approximated by straight-line segments to facilitate direct or digital companding and can be easily converted to and from a uniform format. The normalized A-law compression characteristic is defined as

$$F_A(x) = \begin{cases} \operatorname{sgn}(x)\left[\dfrac{A|x|}{1+\ln(A)}\right] & 0 \le |x| \le \dfrac{1}{A} \\ \operatorname{sgn}(x)\left[\dfrac{1+\ln|Ax|}{1+\ln(A)}\right] & \dfrac{1}{A} \le |x| \le 1 \end{cases} \tag{3.11}$$

The inverse or expansion characteristic is defined as

$$F_A^{-1}(y) = \begin{cases} \operatorname{sgn}(y)\dfrac{|y|[1+\ln(A)]}{A} & 0 \le |y| \le \dfrac{1}{1+\ln(A)} \\ \operatorname{sgn}(y)\dfrac{e^{|y|(1+\ln(A))-1}}{A} & \dfrac{1}{1+\ln(A)} \le |y| \le 1 \end{cases} \tag{3.12}$$

where $y = F_A(X)$ and $A = 87.6$.

Notice that the first portion of the A-law characteristic is linear by definition. The remaining portion of the characteristic ($1/A \le |x| \le 1$) can be closely approximated by linear segments in a fashion similar to the μ-law approximation. In all, there are eight positive and eight negative segments. The first two segments of each polarity (four in all) are colinear and therefore are sometimes considered as one straight-line segment. Thus the segmented approximation of the A-law characteristic is sometimes referred to as a "13-segment approximation." For ease in describing the coding algorithms of the segmented companding characteristic, however, a 16-segment representation is used, just as in the case of the segmented μ-law characteristic.

The segment endpoints, quantization intervals, and corresponding codes for an 8-bit segmented A-law characteristic are shown in Table 3.2. The values are scaled to a maximum value of 4096 for integral representations. Figure 3.20 displays the theoretical performance of the A-law approximation and compares it to the performance of a μ-law approximation presented in Figure 3.19. Notice that the A-law characteristic provides a slightly larger dynamic range. However, the A-law characteristic is inferior to the μ-law characteristic in terms of small-signal quality (idle channel noise). The difference in small-signal performance occurs because the minimum step size of the A-law standard is 2/4096 whereas the minimum step size of the μ-1aw is 2/8159. Furthermore, notice that the A-law approximation does not define a zero-level output for the first quantization interval (i.e., uses a midriser quantizer). However, the difference between midriser and midtread performance at 64 kbps is imperceptible [18].

3.2.6 Syllabic Companding

A significant attribute of the companding techniques described for PCM systems is that they instantaneously encompass the entire dynamic range of the coder on a sample-to-sample basis. Thus μ-law and A-law companding is sometimes referred to as instantaneous companding. However, because the power level of a speech signal remains fairly constant for 100 or more 8-kHz samples, it is wasteful of code space to be able to encode very low level signals adjacent to very high level signals. In other words, it is unnecessary to allow for instantaneous changes in the dynamic range of a signal. One technique of reducing the amount of transmission bandwidth allocated to dynamic range is to reduce the dynamic range of the signal at the source before encoding and then restore the original dynamic range of the signal at the receiver while decoding. When the adjustments to the dynamic range occur on a periodic basis that more or less corresponds to the rate of syllable generation, the technique is referred to as syllabic companding. Because the dynamic range adjustments occur only every 30 msec or so, there is very little bandwidth needed to communicate the adjustments.

TABLE 3.2 Segmented *A*-Law Encoding/Decoding Table[a]

Input Amplitude Range	Step Size	Segment Code S	Quantization Code Q	Code Value	Decoder Amplitude
0–2			0000	0	1
2–4		000	0001	1	3
⋮			⋮	⋮	⋮
30–32	2		1111	15	31
32–34			0000	16	33
⋮		001	⋮	⋮	⋮
62–64			1111	31	63
64–68			0000	32	66
⋮	4	010	⋮	⋮	⋮
124–128			1111	47	126
128–136			0000	48	132
⋮	8	011	⋮	⋮	⋮
248–256			1111	63	252
256–272			0000	64	264
⋮	16	100	⋮	⋮	⋮
496–512			1111	79	504
512–544			0000	80	528
⋮	32	101	⋮	⋮	⋮
992–1024			1111	95	1008
1024–1088			0000	96	1056
⋮	64	110	⋮	⋮	⋮
1984–2048			1111	111	2016
2048–2176			0000	112	2112
⋮	128	111	⋮	⋮	⋮
3968–4096			1111	127	4032

[a]In transmission, every other bit is inverted.

Syllabic companding was first developed for use on noisy analog circuits to improve the idle channel noise. As shown in Figure 3.21, the power level of low-level syllables is increased (compressing the dynamic range) for transmission but attenuated upon reception (expanding the dynamic range). The process of attenuating the received signal restores the low-power syllable to its original level but attenuates any noise arising on the transmission link. Thus the receiver signal-to-noise ratio (SNR) is improved for low-level signals. The amount of amplification applied at the source is dependent on the short-term (syllabic) power level of the transmitted signal. Similarly, the compensating attenuation applied at the receiving terminal is determined from the short-term power level of the received signal.

Syllabic companding on digital systems provides the same basic improvement in SQRs as it does on noisy analog transmission links. When the digital encoders and de-

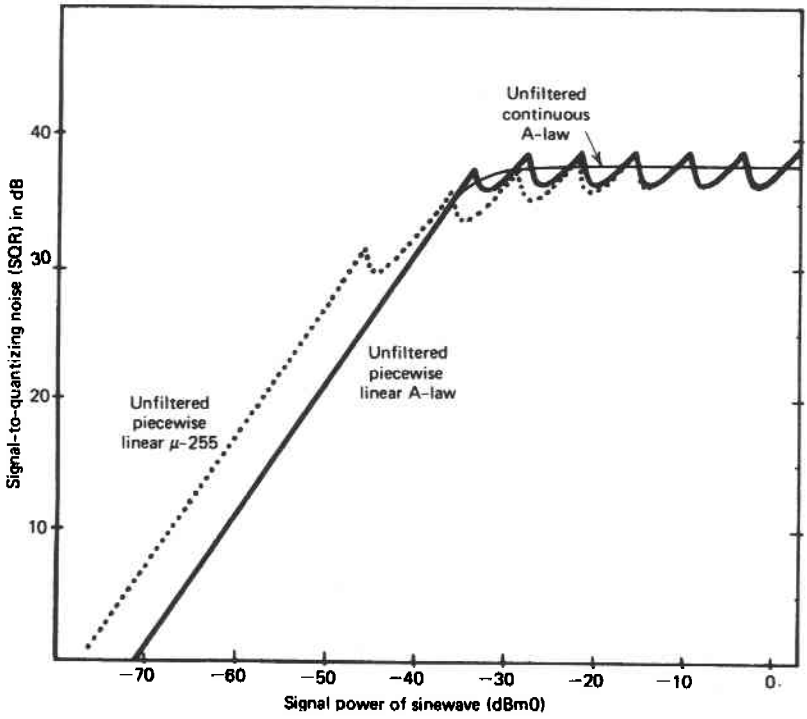


Figure 3.20 SQR of A-law PCM coding with sine wave inputs.

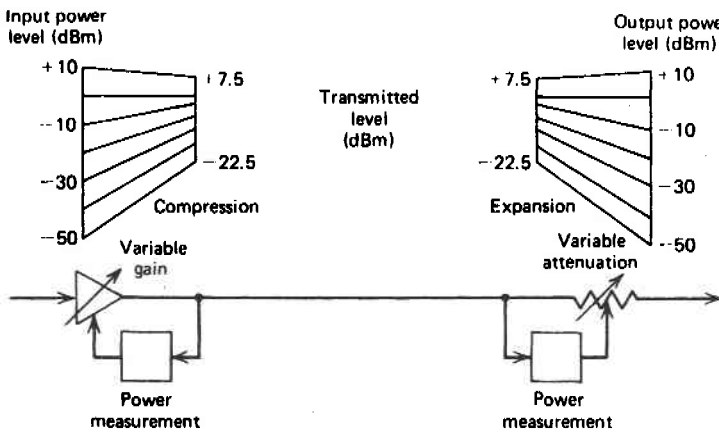


Figure 3.21 Syllabic companding of an analog signal.

coders are considered as part of the transmission link, the process of amplifying low-level signals before encoding and attenuating them after decoding effectively reduces the quantization noise with no net change in signal level. In practice, syllabic companding as implemented in digitized voice terminals does not amplify the signal at the source and attenuate it at the destination. Instead, an equivalent process of controlling the step sizes in the encoder and decoder is used. As far as the transmitted bit stream is concerned, it makes no difference if the signal is amplified and encoded with fixed quantization or if the signal is unmodified but encoded with smaller quantization intervals. Thus syllabic compandors in digital voice terminals typically reduce the quantization intervals when encoding and decoding low-power syllables but increase the quantization intervals for high-power syllables.

Although syllabic companding can be used in conjunction with any type of voice coding, the technique has been applied most often to differential systems described in the following sections. In many of the applications, the adaptation time has been reduced to 5 or 10 msec, which is somewhat shorter than the duration of a typical syllable (approximately 30 msec). The technique is still generally referred to as syllabic companding, however, to distinguish it from the instantaneous variety.

To adjust the step sizes in the decoder in synchronism with adjustments made in the encoder, some means must be established to communicate the step size information from the source to the destination. One method explicitly transmits the step size information as auxiliary information. A generally more useful approach is to derive the step size information from the transmitted bit stream. Thus, in the absence of channel errors, the decoder and the encoder operate on the same information. This procedure is analogous to syllabic companded analog systems in which the receiver determines its attenuation requirements from the short-term power level of the received signal. In a digital system the bit stream is monitored for certain data patterns that indicate the power level of the signal being encoded. Indications of high power level initiate an increase in the step size, whereas indications of low levels cause a decrease.

Determining the step size information from the transmitted bit stream is generally better than explicitly transmitting the step sizes for the following reasons. Because there is no explicit transmission of step size, the transmission of sampled speech information is never interrupted, and the speech sample rate is equal to the transmission rate. Also, the bit stream does not have to be framed to identify step size information separately from the waveform coding. Furthermore, if the step size adjustments are made on a more gradual basis, the individual increments are small enough that occasional incorrect adjustments in the receiver caused by channel errors are not critical. However, on transmission links with very high error rates (one error in a hundred bits or so), better decoded voice quality can be obtained if the step size is transmitted explicitly and redundantly encoded for error correction [19].

3.2.7 Adaptive Gain Encoding

In the syllabic companding example of Figure 3.21 the dynamic range of a signal is reduced by a factor of 2 (in dBm). Thus, if an uncompanded signal has 36 dB of dy-

namic range, the encoder sees only 18 dB. The 18-dB reduction implies three fewer bits are needed for dynamic range encoding. In the limit, if the power level of all input signals is adjusted to a single value, no bits of the encoder need to be allocated to dynamic range encoding. A process that adjusts all signals to a standard value is referred to as automatic gain control (AGC). AGC is traditionally used on carrier transmission systems to adjust all received signals to a standard value, thereby removing variations in propagation attentuation. AGC cannot be applied to a source voice signal without allowances for speech pauses when there is no signal present. Otherwise, idle channel noise would be amplified to the average level of active voice. Notice that with AGC there is no residual information in the power level of the encoded signal as there is in syllabic companding. To ascertain the original power level, AGC must be augmented with adaptive gain encoding (AGE), as indicated in Figure 3.22.

There are two basic modes of operation for gain encoding depending on how gain factors are measured and to which speech segments the factors are applied. One mode of operation, as implied in Figure 3.22, involves measuring the power level of one segment of speech and using that information to establish a gain factor for ensuing speech segments. Obviously, this mode of operation relies on gradually changing power levels. This mode of operation is sometimes referred to as "backward estimation."

Another mode of operation involves measuring the power level of a speech segment and using the gain factor thus derived to adapt the encoder to the same segment. This approach, referred to as "forward estimation," has the obvious advantage that the encoder and decoder use gain factors specifically related to the speech segments from which they are derived. The disadvantage is that each speech segment must be delayed while the gain factor is being determined. Although the availability of digital memory has made the cost of implementing the delay insignificant, the impact of the delay on echoes and singing in a partially analog network must be considered. (As long as the subscriber loops are analog, the network is partially analog.)

Adaptive gain control with explicit transmission of gain factors is not without shortcomings. First, when the periodic gain information is inserted into the transmitted bit stream, some means of framing the bit stream into blocks is needed so gain information can be distinguished from waveform coding. Second, periodic insertion of gain information disrupts information flow, causing higher transmitter clock rates that might be inconveniently related to the waveform sample clock. Third, correct recep-

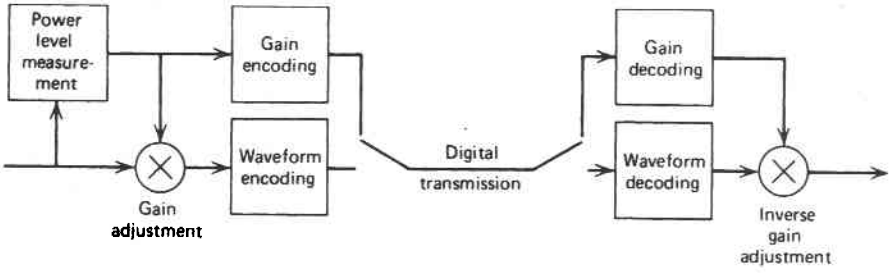


Figure 3.22 Adaptive gain encoding.

tion of gain factors is usually critical to voice quality, indicating a need to redundantly encode gain information.

Reference [20] describes a modified form of PCM using forward estimation of gain factors that is referred to as *nearly instantaneously companded PCM*. The need for transmitting speech segments in blocks is not a disadvantage in the application mentioned (mobile telephone) because repetitive bursts with error checking are used as a means of overcoming short-lived multipath fading. This system provides a bit rate reduction of 30% with respect to conventional PCM. Another example of the use of AGE is the subscriber loop multiplexer (SLM) system developed by Bell Labs [21]. The SLM system became obsolete when low-cost PCM codecs became available and the subscriber carrier systems could be integrated into digital end offices (with SLC 96 and later DLC systems). All of the encoding algorithms described in the following sections use syllabic companding or AGE in some form to reduce the bit rate.

3.3 SPEECH REDUNDANCIES

Conventional PCM systems encode each sample of an input waveform independently from all other samples. Thus a PCM system is inherently capable of encoding an arbitrarily random waveform whose maximum-frequency component does not exceed one-half the sampling rate. Analyses of speech waveforms, however, indicate there is considerable redundancy from one sample to the next. In fact, as reported in reference [10], the correlation coefficient (a measure of predictability) between adjacent 8-kHz samples is generally 0.85 or higher. Hence the redundancy in conventional PCM codes suggests significant savings in transmission bandwidths are possible through more efficient coding techniques. All of the digitization techniques described in the rest of this chapter are tailored, in one degree or another, to the characteristics of speech signals with the intent of reducing the bit rate.

In addition to the correlation existing between adjacent samples of a speech waveform, several other levels of redundancy can be exploited to reduce encoded bit rates. Table 3.3 lists these redundancies. Not included are higher level redundancies related to context-dependent interpretations of speech sounds (phonemes), words, and sen-

TABLE 3.3 Speech Redundancies

Time-domain redundancies
1. Nonuniform amplitude distributions
2. Sample-to-sample correlations
3. Cycle-to-cycle correlations (periodicity)
4. Pitch-interval to pitch-interval correlations
5. Inactivity factors (speech pauses)
Frequency-domain redundancies
6. Nonuniform long-term spectral densities
7. Sound-specific short-term spectral densities

tences. These topics are not covered because techniques that analyze speech waveforms to extract only information content eliminate subjective qualities essential to general telephony.

3.3.1 Nonuniform Amplitude Distributions

As mentioned in the introduction to companding, lower amplitude sample values are more common than higher amplitude sample values. Most low-level samples occur as a result of speech pauses in a conversation. Beyond this, however, the power levels of active speech signals also tend to occur at the lower end of the encoding range. The companding procedures described in the previous section provide slightly inferior quality (i.e., lower signal-to-noise ratios) for small signals compared to large signals. Thus the average quality of PCM speech could be improved by further shortening lower level quantization intervals and increasing upper level quantization intervals. The amount of improvement realized by such a technique is minimal and probably would not justify the additional complexities. The most beneficial approach to processing signal amplitudes in order to reduce encoder bit rates involves some form of adaptive gain control, as discussed earlier.

3.3.2 Sample-to-Sample Correlation

The high correlation factor of 0.85 mentioned in Section 3.3 indicates that any significant attempt to reduce transmission rates must exploit the correlation between adjacent samples. In fact, at 8-kHz sampling rates, significant correlations also exist for samples two to three samples apart. Naturally, samples become even more correlated if the sampling rate is increased.

The simplest way to exploit sample-to-sample redundancies in speech is to encode only the differences between adjacent samples. The difference measurements are then accumulated in a decoder to recover the signal. In essence these systems encode the slope or derivative of a signal at the source and recover the signal by integrating at the destination. Digitization algorithms of this type are discussed at length in later sections.

3.3.3 Cycle-to-Cycle Correlations

Although a speech signal requires the entire 300–3400-Hz bandwidth provided by a telephone channel, at any particular instant in time certain sounds may be composed of only a few frequencies within the band. When only a few underlying frequencies exist in a sound, the waveform exhibits strong correlations over numerous samples corresponding to several cycles of an oscillation. The cyclic nature of a voiced sound is evident in the time waveform shown in Figure 3.23. Encoders exploiting the cycle-to-cycle redundancies in speech are markedly more complicated than those concerned only with removing the redundancy in adjacent samples. In fact, these encoders more

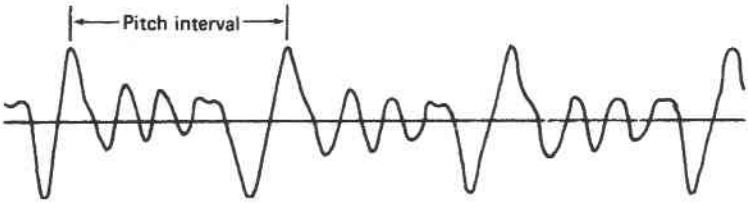


Figure 3.23 Time waveform of voiced sound.

or less represent a transition from the relatively high rate, natural-sounding waveform encoders to the relatively low rate, synthetic-sounding vocoders.

3.3.4 Pitch-Interval-to-Pitch-Interval Correlations

Human speech sounds are often categorized as being generated in one of two basic ways. The first category of sounds encompasses "voiced" sounds that arise as a result of vibrations in the vocal cords. Each vibration allows a puff of air to flow from the lungs into the vocal tract. The interval between puffs of air exciting the vocal tract is referred to as the pitch interval or, more simply, the rate of excitation is the pitch. Generally speaking, voiced sounds arise in the generation of vowels and the latter portions of some consonants. An example of a time waveform for a voiced sound is shown in Figure 3.23.

The second category of sounds includes the fricatives, or "unvoiced" sounds. Fricatives occur as a result of continuous air flowing from the lungs and passing through a vocal tract constricted at some point to generate air turbulence (friction). Unvoiced sounds correspond to certain consonants such as *f*, *j*, *s*, and *x*. An example of a time waveform of an unvoiced sound is shown in Figure 3.24. Notice that an unvoiced sound has a much more random waveform than a voiced sound.

As indicated in Figure 3.23, not only does a voiced sound exhibit the cycle-to-cycle redundancies mentioned in Section 3.3.3, but also the waveform displays a longer term repetitive pattern corresponding to the duration of a pitch interval. Thus one of the most efficient ways of encoding the voiced portions of speech is to encode one pitch interval waveform and use that encoding as a template for each successive pitch interval in the same sound. Pitch intervals typically last from 5 to 20 msec for men and from 2.5 to 10 msec for women. Since a typical voiced sound lasts for approximately 100 msec, there may be as many as 20–40 pitch intervals in a single sound. Although pitch interval encoding can provide significant reductions in bit rates, the pitch is sometimes very difficult to detect. (Not all voiced sounds produce a readily identifi-

Figure 3.24 Time waveform of unvoiced sound.

able pitch interval as in Figure 3.23.) If the pitch gets encoded erroneously, strange sounds result.

An interesting aspect of pitch interval encoding is that it provides a means of speeding up speech while maintaining intelligibility. By deleting some percentage of pitch intervals from each sound (phoneme), the rate of sound generation is effectively increased in a manner analogous to more rapid word formation. The pitch of the sounds remains unchanged. In contrast, if the rate of reconstruction is merely increased, all frequencies including the pitch increase proportionately. Moderate speedups produce obvious distortion while greater speedups become unintelligible. Devices designed to simulate faster word formation have demonstrated that we are capable of assimilating spoken information much faster than we can generate it.

3.3.5 Inactivity Factors

Analyses of telephone conversations have indicated that a party is typically active about 40% of a call duration. Most inactivity occurs as a result of one person listening while the other is talking. Thus a conventional (circuit-switched) full-duplex connection is significantly underutilized. Time assignment speech interpolation (TASI) described in Chapter 1 is a technique to improve channel utilization on expensive analog links. Digital speech interpolation (DSI) is a term used to refer to a digital circuit counterpart of TASI. DSI involves sensing speech activity, seizing a channel, digitally encoding and transmitting the utterances, and releasing the channel at the completion of each speech segment.

Digital speech interpolation is obviously applicable to digital speech storage systems where the duration of a pause can be encoded more efficiently than the pause itself. In recorded messages, however, the pauses are normally short since a "half-duplex" conversation is not taking place. DSI techniques have been used to expand the voice channel capacity of digital TDM links. The inputs are standard PCM signals that are digitally processed to detect speech activity. The DSI operation is often combined with speech compression algorithms to implement digital circuit multiplication (DCM) equipment. When a 2 : 1 voice compression algorithm is combined with a 2.5 : 1 DSI concentration factor, an overall 5 : 1 circuit expansion is achieved. Depending on the quality of speech desired, even greater concentration factors are possible. The use of such equipment in a network must be carefully managed to ensure that voiceband data and digital data channels bypass the DCM operations.

3.3.6 Nonuniform Long-Term Spectral Densities

The time-domain redundancies described in the preceding sections exhibit characteristics in the frequency domain that can be judiciously processed to reduce the encoded bit rate. Frequency-domain redundancies are not independent of the redundancies in the time domain. Frequency-domain techniques merely offer an alternate approach to analyzing and processing the redundancies.

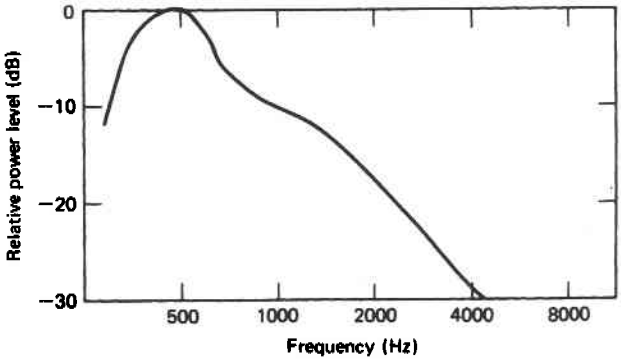


Figure 3.25 Long-term power spectral density of speech.

A totally random or unpredictable signal in the time domain produces a frequency spectrum that is flat across the bandwidth of interest. Thus a signal that produces uncorrelated time-domain samples makes maximum use of its bandwidth. On the other hand, a nonuniform spectral density represents inefficient use of the bandwidth and is indicative of redundancy in the waveform.

Figure 3.25 shows the long-term spectral density of speech signals averaged across two populations: men and women [22]. Notice that the upper portions of the 3-kHz bandwidth passed by the telephone network have significantly reduced power levels. The lower power levels at higher frequencies are a direct consequence of the time-domain sample-to-sample correlations discussed previously. Large-amplitude signals cannot change rapidly because, on average, they are predominantly made up of lower frequency components.

A frequency-domain approach to more efficient coding involves flattening the spectrum before encoding the signal. The flattening process can be accomplished by passing the signal through a high-pass filter to emphasize the higher frequencies before sampling. The original waveform is recovered by passing the decoded signal through a filter with a complementary, low-pass characteristic. An important aspect of this process is that a high-pass filter exhibits time-domain characteristics of a differentiator and a low-pass filter has time-domain characteristics analogous to an integrator. Thus the spectrum-flattening process essentially means the slope of the signal is encoded at the source, and the signal is recovered by integrating at the destination—the basic procedure described previously for sample-to-sample redundancy removal in the time domain.

In studying Figure 3.25 it is natural to think that the remarkably low levels of signal energy at the higher frequencies (2–3.4 kHz) means that more bandwidth is being allocated to a voice signal than is really necessary. The error in such a conclusion, however, lies in the distinction between energy content and information content of the voice frequency spectrum. As any beginning computer programmer soon learns,* the

*I am assuming that beginning programmers still encounter older languages or file systems with seven-character name limitations.

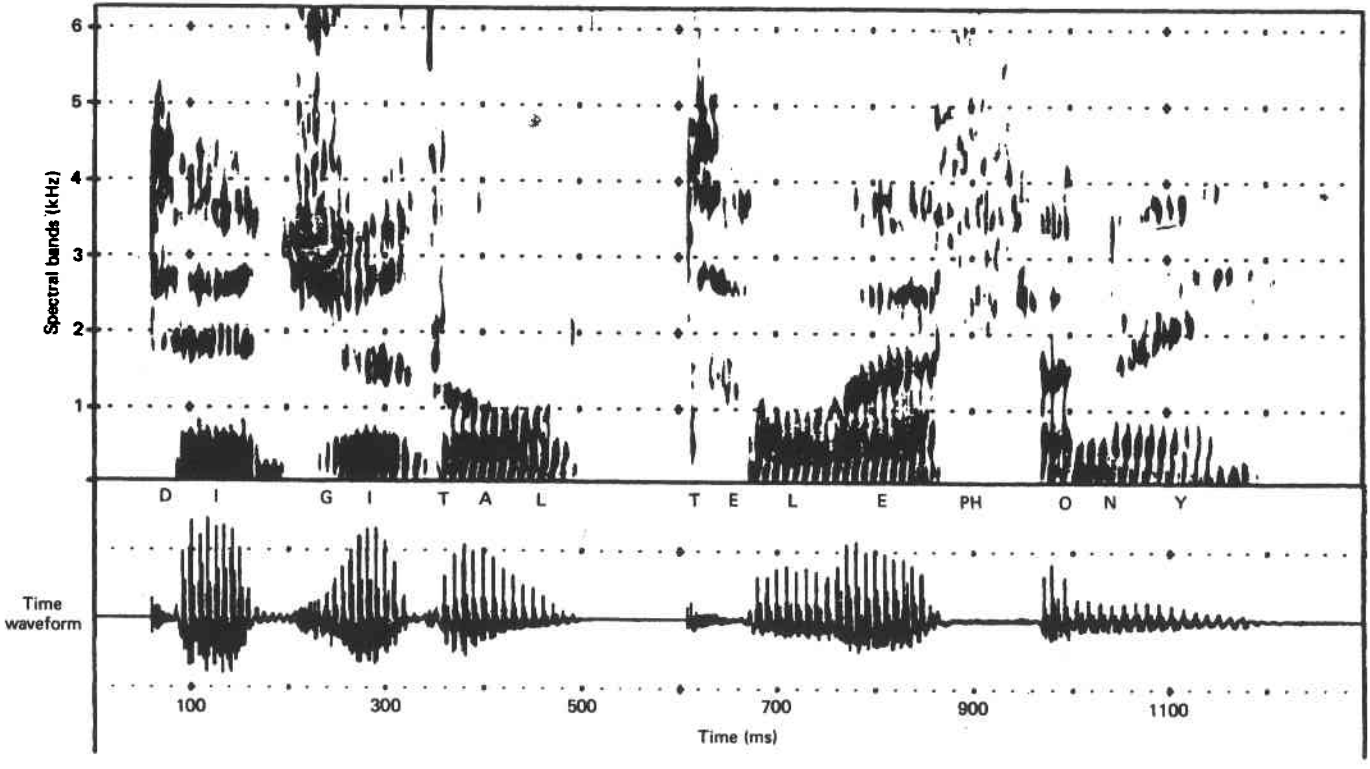


Figure 3.26 Specrogram of phrase: "digital telephony." (Courtesy of Robert L. Davis of Texas Instruments, Inc.)

meaning of a program variable can often be retained even though it is shortened by deleting all of the vowels. In speech the vowels require most of the energy and primarily occupy the lower portion of the frequency band. The consonants, on the other hand, contain most of the information but use much less power and generally higher frequencies. Hence merely reproducing a high percentage of the original speech energy is an inadequate goal for a digital speech transmission or storage system.

3.3.7 Short-Term Spectral Densities

The speech spectrums shown in Figure 3.25 represent long-term averages of the spectral densities. Over shorter periods of time the spectral densities vary considerably and exhibit sound-specific structures with energy peaks (resonances) at some frequencies and energy valleys at others. The frequencies at which the resonances occur are called formant frequencies, or simply formants. Voiced speech sounds typically contain three to four identifiable formants. These features of the short-term spectral density are illustrated in the spectogram of Figure 3.26. A spectogram is a display of speech spectral energy as a function of time and frequency. The horizontal axis represents time, the vertical axis represents frequency, and the shadings represent energy levels. Thus the darker portions in Figure 3.26 indicate relatively high energy levels (formants) at particular instants in time.

Frequency-domain voice coders provide improved coding efficiencies by encoding the most important components of the spectrum on a dynamic basis. As the sounds change, different portions (formants) of the frequency band are encoded. The period between formant updates is typically 10–20 msec. Instead of using periodic spectrum measurements, some higher quality vocoders continuously track gradual changes in the spectral density at a higher rate. Frequency-domain vocoders often provide lower bit rates than the time-domain coders but typically produce less natural sounding speech.

3.4 DIFFERENTIAL PULSE CODE MODULATION

Differential pulse code modulation (DPCM) is designed specifically to take advantage of the sample-to-sample redundancies in a typical speech waveform. Since the range of sample *differences* is less than the range of individual samples, fewer bits are needed to encode difference samples. The sampling rate is often the same as for a comparable PCM system. Thus the bandlimiting filter in the encoder and the smoothing filter in the decoder are basically identical to those used in conventional PCM systems.

A conceptual means of generating the difference samples for a DPCM coder is to store the previous input sample directly in a sample-and-hold circuit and use an analog subtracter to measure the change. The change in the signal is then quantized and encoded for transmission. The DPCM structure shown in Figure 3.27 is more complicated, however, because the previous input value is reconstructed by a feedback loop that integrates the encoded sample differences. In essence, the feedback signal is an

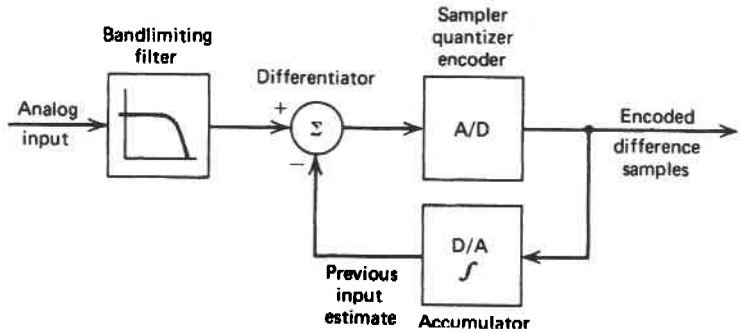


Figure 3.27 Functional block diagram of differential PCM.

estimate of the input signal as obtained by integrating the encoded sample differences. Thus the feedback signal is obtained in the same manner used to reconstruct the waveform in the decoder.

The advantage of the feedback implementation is that quantization errors do not accumulate indefinitely. If the feedback signal drifts from the input signal, as a result of an accumulation of quantization errors, the next encoding of the difference signal automatically compensates for the drift. In a system without feedback the output produced by a decoder at the other end of the connection might accumulate quantization errors without bound.

As in PCM systems, the analog-to-digital conversion process can be uniform or companded. Some DPCM systems also use adaptive techniques (syllabic companding) to adjust the quantization step size in accordance with the average power level of the signal. (See reference [9] for an overview of various techniques.)

Example 3.4. Speech digitization techniques are sometimes measured for quality by use of an 800-Hz sine wave as a representative test signal. Assuming a uniform PCM system is available to encode the sine wave across a given dynamic range, determine how many bits per sample can be saved by using a uniform DPCM system.

Solution. A basic solution can be obtained by determining how much smaller the dynamic range of the difference signal is in comparison to the dynamic range of the signal amplitude. Assume the maximum amplitude of the sine wave is A, so that

$$x(t) = A \sin(2\pi \cdot 800t)$$

The maximum amplitude of the difference signal can be obtained by differentiating and multiplying by the time interval between samples:

$$\frac{dx}{dt} = A(2\pi)(800) \cos(2\pi \cdot 800t)$$

$$|\Delta x(t)|_{max} = A(2\pi)(800)\left(\frac{1}{8000}\right) = 0.628A$$

The savings in bits per sample can be determined as

$$\log_2\left(\frac{1}{0.628}\right) = 0.67 \text{ bits}$$

Example 3.4 demonstrates that a DPCM system can use $\frac{2}{3}$ bit per sample less than a PCM system with the same quality. Typically DPCM systems provide a full 1-bit reduction in codeword size. The larger savings is achieved because, on average, speech waveforms have a lower slope than an 800-Hz tone (see Figure 3.25).

3.4.1 DPCM Implementations

Differential PCM encoders and decoders can be implemented in a variety of ways depending on how the signal processing functions are partitioned between analog and digital circuitry. At one extreme the differencing and integration functions can be implemented with analog circuitry, while at the other extreme all signal processing can be implemented digitally using conventional PCM samples as input. Figure 3.28 shows block diagrams of three different implementations with differing amounts of digital signal processing.

Figure 3.28*a* depicts a system using analog differencing and integration. Analog-to-digital conversion is performed on the difference signal, and D/A conversion for the feedback loop is immediately performed on the limited-range difference code. Analog summation and storage in a sample-and-hold (S/H) circuit is used to provide integration.

Figure 3.28*b* shows a system that performs the integration function digitally. Instead of immediately converting the difference code back to analog for feedback, the difference code is summed and stored in a data register to generate a digital representation of the previous input sample. A full-scale D/A converter is then used to produce the analog feedback signal for differencing. Notice that the D/A converters in Figure 3.28*b* must convert the full amplitude range whereas the D/A converters in Figure 3.28*a* convert the more limited difference signal.

Figure 3.28*c* shows a system where all signal processing is performed by digital logic circuits. The A/D converter produces full-amplitude-range sample codes, which are compared to digitally generated approximations of the previous amplitude code. Notice that the A/D converter in this case must encode the entire dynamic range of the input whereas the A/D converters in the other two versions operate on only the difference signals.

Due to the availability of digital signal processing components, some of which contain internal A/D converters, digital processing (as in Figure 3.28*c*) is generally the most effective means of implementing a DPCM algorithm. In fact, most DPCM applications involve processing speech signals that have already been digitized into standard 64-kbps PCM formats. Thus the DPCM implementation usually requires no

analog processing. As an aid in processing log-PCM signals, some DSP components provide internal μ-law and *A*-law conversion functions.

The decoders in all three implementations shown in Figure 3.28 are exactly like the feedback implementations in the corresponding encoder. This reinforces the fact that the feedback loop generates an approximation of the input signal (delayed by one sample). If no channel errors occur, the decoder output (before filtering) is identical to the feedback signal. Thus the closer the feedback signal matches the input, the closer the decoder output matches the encoder input.

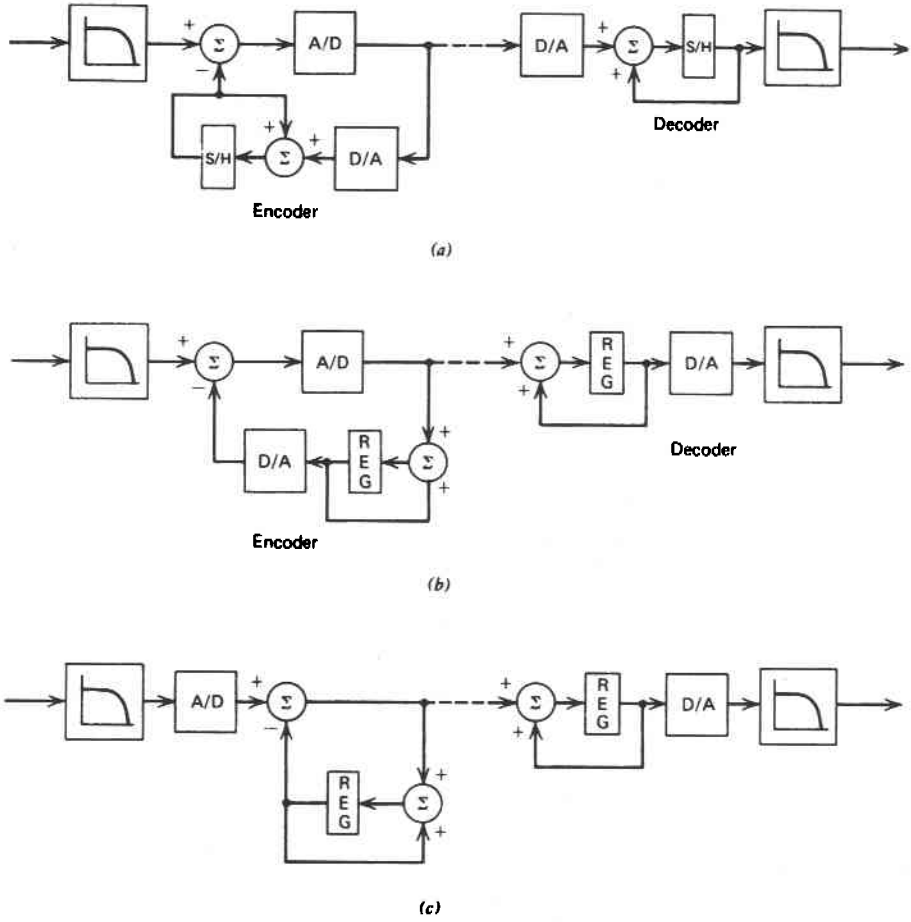


Figure 3.28 DPCM implementations: (*a*) analog integration; (*b*) digital integration; (*c*) digital differencing.

3.4.2 Higher Order Prediction

A more general viewpoint of a DPCM encoder considers it as a special case of a linear predictor with encoding and transmission of the prediction error. The feedback signal of a DPCM system represents first-order prediction of the next sample value, and the sample difference is a prediction error. Under this viewpoint the DPCM concept can be extended to incorporate more than one past sample value into the prediction circuitry. Thus the additional redundancy available from all previous samples can be weighted and summed to produce a better estimate of the next input sample. With a better estimate, the range of the prediction error decreases to allow encoding with fewer bits. For systems with constant predictor coefficients, results have shown that most of the realizable improvement occurs when using only the last three sample values. The basic implementation of linear prediction using the last three sample values is shown in Figure 3.29. For conceptual purposes this implementation shows analog differencing and integration as in Figure 3.28*a*. The most effective implementations use digital memory, multiplication, and addition in a DSP component in lieu of the analog processing shown, particularly because most applications involve already digitized (PCM) signals.

As mentioned in Section 3.4, analysis of differential PCM systems with first-order predication typically provides a 1-bit-per-sample reduction in code length relative to PCM systems with equivalent performance. Extended DPCM systems utilizing third-order prediction can provide reductions of $\frac{11}{2}$ –2 bits per sample [23]. Thus a standard DPCM system can provide 64-kbps PCM quality at 56 kbps, and third-order linear prediction can provide comparable quality at 48 kbps. However, some subjective evaluations have indicated that somewhat higher bit rates are needed to match 64-kbps PCM quality.

3.4.3 Adaptive Differential PCM

Relatively straightforward implementations of DPCM can provide savings of 1–2 bits per sample with respect to standard PCM encoding. Even greater savings can be

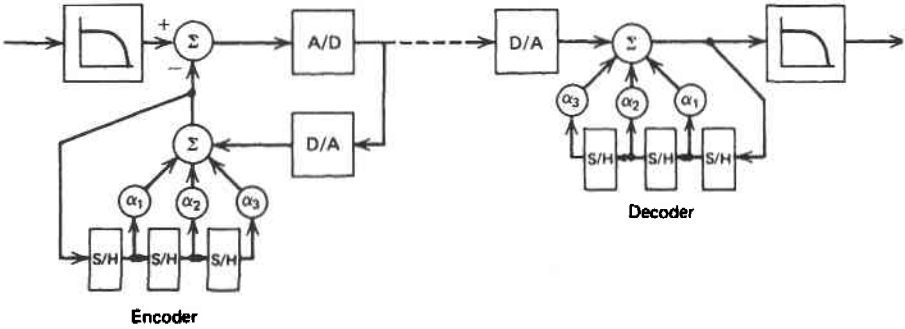


Figure 3.29 Extension of DPCM to third-order prediction.

achieved by adding adaptation logic to the basic DPCM algorithm to create what is referred to as adaptive differential PCM (ADPCM). Many forms of ADPCM have been investigated and used in various applications. Two of the most prevalent applications are voice messaging and DCM equipment for increasing the number of voice channels on a T1 line. With respect to the latter application, ITU-T has established a 32-kbps ADPCM standard (Recommendation G.721) [24]. This algorithm has been extensively tested and characterized to not significantly degrade toll quality voice circuits when inserted into the internal portions of the network. Design considerations of the standard are:

1. Multiple tandem encodings and decodings between both PCM and analog interfaces
2. End-to-end signal quality for voice, voiceband data, and facsimile
3. Effects of random and bursty channel errors
4. Performance on analog signals degraded by loss, noise, amplitude distortion, phase distortion, and harmonic distortion
5. Easy transcoding with μ-law and A-law PCM

The 32-kbps rate implies a 2 : 1 savings in channel bandwidth with respect to standard PCM. A significant impairment introduced by implementations of the ADPCM standard is the corruption of modem signals carrying data rates greater than 4800 bps [24]. Voiceband data at rates of 4800 bps and below are adequately supported.

The G.721 ADPCM algorithm is conceptually similar to that shown in Figure 3.29 but more sophisticated in that it uses an eighth-order predictor, adaptive quantization, and adaptive prediction. Furthermore, the algorithm is designed to recognize the difference between voice or data signals and use a fast quantizer adaptation mode for voice and a slow adaptation mode for data.

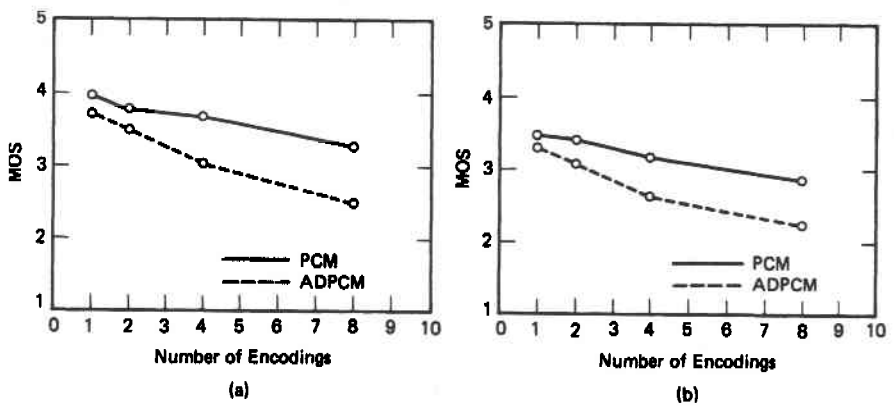


Figure 3.30 Average MOS versus number of encodings for PCM and ADPCM: (*a*) linear microphone; (*b*) carbon microphone.

Subjective evaluation [25] of the G.721 algorithm using the mean opinion score (MOS) method of evaluating speech quality is shown in Figure 3.30. The MOS method uses trained listeners to evaluate the speech quality on a scale of 1 : 5. Figure 3.30 shows the average scores of 32-kbps ADPCM and 64-kbps PCM as a function of the number of analog tandem encodings. The speech quality associated with multiple digital transcodings (to and from PCM) does not degrade beyond the first transcoding as long as there are no digital transmission impairments such as channel errors or slips (Chapter 7). The performance of ADPCM (or any other coder) used to be an important consideration when the public network was composed of mixtures of analog and digital equipment. Because the internal portions of the network are all digital, performance with multiple analog conversions is no longer a consideration. (Multiple PCM-to-ADPCM conversions can still occur when multiple DCM systems are encountered within a network.)

Because ADPCM at 32 kbps provides good quality at a moderate cost and power consumption, it is used in several cordless telephone or low-tier* digital cellular systems:

Personal Access Communications System (PACS) (North America)
Second Generation Cordless Telephones (CT2) (Europe)
Digital European Cordless Telephones (DECT) (Europe)
Personal Handyphone System (PHS) (Japan)

3.5 DELTA MODULATION

Delta modulation (DM) is another digitization technique that specifically exploits the sample-to-sample redundancy in a speech waveform. In fact, DM can be considered as a special case of DPCM using only 1 bit per sample of the difference signal. The single bit specifies merely the polarity of the difference sample and thereby indicates whether the signal has increased or decreased since the last sample. An approximation to the input waveform is constructed in the feedback path by stepping up one quantization level when the difference is positive ("one") and stepping down when the difference is negative ("zero"). In this way the input is encoded as a sequence of "ups" and "downs" in a manner resembling a staircase. Figure 3.31 shows a DM approximation of a typical waveform. Notice that the feedback signal continues to step in one direction until it crosses the input, at which time the feedback step reverses direction until the input is crossed again. Thus, when tracking the input signal, the DM output "bounces" back and forth across the input waveform, allowing the input to be accurately reconstructed by a smoothing filter.

Since each encoded sample contains a relatively small amount of information (1 bit), DM systems require a higher sampling rate than PCM or multibit DPCM systems. In fact, the sampling rate is necessarily much higher than the minimum (Nyquist) sam-

*A low-tier cellular system utilizes simple, low-power mobile units, and small cells and only supports pedestrian speeds.

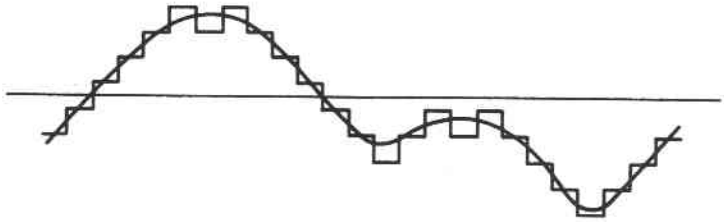


Figure 3.31 Waveform encoding by delta modulation.

pling rate of twice the bandwidth. From another viewpoint, "oversampling" is needed to achieve better prediction from one sample to the next.

The main attraction of DM is its simplicity. Figure 3.32 shows a basic implementation of a DM encoder and decoder. Notice that the A/D conversion function is provided by a simple comparator. A positive-difference voltage produces a 1, and a negative-difference voltage produces a 0. Correspondingly, the D/A function in the feedback path, and in the decoder, is provided by a two-polarity pulse generator. In the simplest form the integrator can consist of nothing more than a capacitor to accumulate the charge from the pulse generator.

In addition to these obvious implementation simplicities, a delta modulator also allows the use of relatively simple filters for bandlimiting the input and smoothing the output [26]. As discussed in Section 3.1, the spectrum produced by a sampling process consists of replicas of the sampled spectrum centered at multiples of the sampling frequency. The relatively high sampling rate of a delta modulator produces a wider separation of these spectrums, and, hence, foldover distortion is prevented with less stringent roll-off requirements for the input filter.

3.5.1 Slope Overload

The conceptual operation of a delta modulator shown in Figure 3.31 indicates that the encoded waveform is never much more than a step size away from the input signal. Sometimes a delta modulator, or any differential system such as DPCM, may not be able to keep up with rapid changes in the input signal and thus fall more than a step

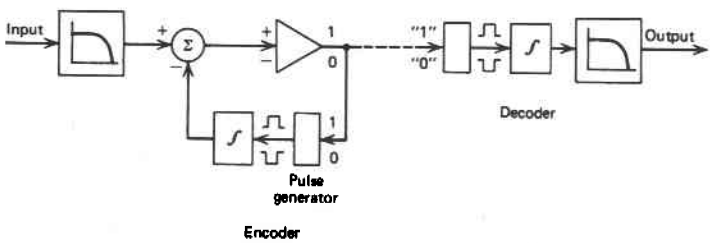


Figure 3.32 Delta modulation encoder and decoder.

size behind. When this happens, the delta modulator is said to be experiencing "slope overload." A slope overload condition is shown in Figure 3.33.

Basically, slope overload occurs when the rate of change of the input exceeds the maximum rate of change that can be generated by the feedback loop. Since the maximum rate of change in the feedback loop is merely the step size times the sampling rate, a slope overload condition occurs if

$$\left|\frac{dx(t)}{dt}\right| > qf_s \tag{3.13}$$

where $x(t)$ = input signal
q = step size
f_s = sampling frequency

The design of a DM (or DPCM) necessarily involves a trade-off between two types of distortion: the more or less random quantization noise, sometimes referred to as granular noise, and the slope overload noise. As indicated in Figure 3.33, granular noise is a predominant consideration for slowly changing signals, whereas slope overload is dominant during rapidly changing signals. Obviously, granular noise is small if step sizes are small, but small step sizes increase the likelihood of slope overload. The optimum DM step size in terms of minimizing the total of granular and slope overload noise has been considered by Abate [27].

The perceptual effects of slope overload on the quality of a speech signal are significantly different from the perceptual effects produced by granular noise. As indicated in Figure 3.33, the slope overload noise reaches its peaks just before the encoded signal reaches its peaks. Hence, slope overload noise has strong components identical in frequency and approximately in phase with a major component of the input. Distortion that is correlated in this manner to the speech signal is effectively "masked" by the speech energy and therefore is less noticeable than "uncorrelated" distortion. In fact, overload noise is much less objectionable to a listener than random or granular noise at an equivalent power level [28]. Hence, from the point of view of perceived speech quality, the optimum mix of granular and slope overload noise is difficult to determine.

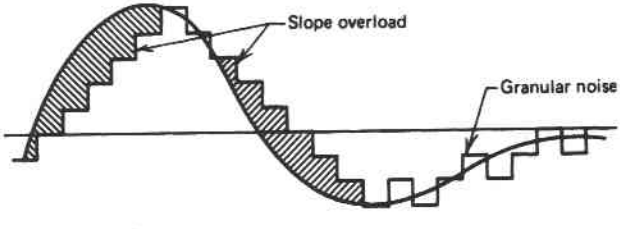


Figure 3.33 Slope overload and granular noise of delta modulation system.

Slope overload is not a limitation of just a DM system, but an inherent problem with any system, such as DPCM in general, that encodes the difference in a signal from one sample to the next. A difference system encodes the slope of the input with a finite number of bits and hence a finite range. If the slope exceeds that range, slope overload occurs. In contrast, a conventional PCM system is not limited by the rate of change of the input, only by the maximum encodable amplitude. Notice that a differential system can encode signals with arbitrarily large amplitudes, as long as the large amplitudes are attained gradually.

Many versions of DM for voice encoding were investigated in the 1960s and 1970s [29, 30]. All of these efforts focused on ways of implementing adaptive delta modulation (ADM) to improve the performance at a given bit rate. The intense interest at that time was related to the simplicity, good tolerance of channel errors, and relatively low cost implementation. The cost factor is no longer relevant because even relatively complicated coding algorithms now have insignificant costs compared to most system costs. ADM is still used in some old PBXs, in some military secure voice radio systems, and as a means of encoding the residual error signal of some predictive coders described in the following sections.

3.6 ADAPTIVE PREDICTIVE CODING

The differential systems described in the previous sections (DPCM, ADPCM, ADM) operate with lower data rates than PCM systems because they encode a difference signal that has lower average power than the raw input signal. The ratio of the input signal power to the power of the difference signal is referred to as the prediction gain. Simple DPCM systems (first-order predictors) provide about 5 dB of prediction gain. ADPCM provides greater levels of prediction gain depending on the sophistication of the adaptation logic and the number of past samples used to predict the next sample. The prediction gain of ADPCM is ultimately limited by the fact that only a few past samples are used to predict the input and the adaptation logic only adapts the quantizer, not the prediction weighting coefficients (the α's in Figure 3.29).

Adaptive predictive coding (APC) [31, 32] provides greater levels of prediction gain by adapting the prediction coefficients to individual speech segments and, in most cases, using higher orders of prediction (e.g., up to 12). If the coefficients are determined from past history and used to predict subsequent speech segments (backward estimation), 13 dB of prediction gain is possible [10]. If speech segments are delayed so predictor coefficients can be used on the same speech segments from which they were derived (forward estimation), 20 dB of prediction gain is possible [33].

A block diagram of a basic APC encoder/decoder is shown in Figure 3.34. The input to the encoder and the output from the decoder are assumed to be uniform PCM, most likely representing conversions from and to log PCM. The transmitted data stream is necessarily composed of blocks containing three types of information: (1) the encoded difference signal (residual), (2) a gain factor, and (3) the predictor coefficients. The most significant difference between this coder and a DPCM or ADPCM coder involves the periodic determination and transmission of the predictor coeffi-

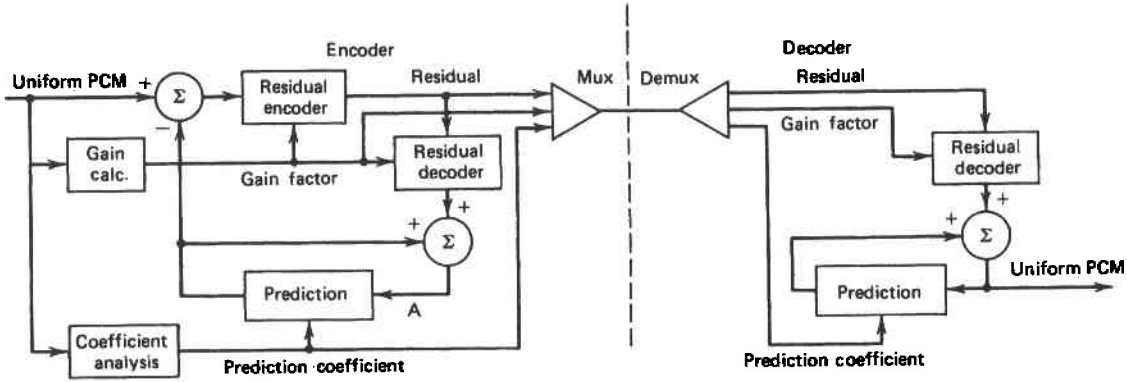


Figure 3.34 Adaptive predictive encoder/decoder.

cients. Notice that the integrated residual signal at point A is identical to the input signal delayed by one sample (except for residual quantization error). Thus the corresponding point in the decoder is the reconstructed PCM output.

Explicit transmission of a gain factor, as opposed to deriving a gain factor from the transmitted residual, is useful in this application because a block-structured transmission format is required for transmitting the predictor coefficients anyway. Residual encoding may use any of the waveform algorithms described previously. Arbitrarily accurate reconstruction of the input can be achieved if enough bits are used in encoding the residual. Of course, the purpose of the adaptive prediction is to achieve a low data rate. Single-bit PCM that conveys only the polarity of the residual provides adequate performance for low-data-rate applications [33]. Notice that because of the feedback path, single-bit encoding of the residual essentially produces a delta modulator with very sophisticated feedback logic that is periodically changed to match the shape and energy level of corresponding (forward-estimated) speech segments.

A large variety of algorithms have been investigated for calculating the predictor coefficients, most of which involve extensive computation. If the application is for voice response systems, the computations do not have to occur in real time and are therefore sometimes done on a large mainframe computer. Real-time encoding restricts the algorithm to one that can be realized with a DSP or special-purpose IC designed to implement a specific coefficient determining algorithm. A linear predictive coding (LPC) algorithm as described in Section 3.8.3 is a common algorithm because it provides good performance and is available in custom parts. The U.S. defense department adopted an APC algorithm using a fourth-order LPC coefficient analysis as a government standard for secure voice transmission at 9.6 kbps [34]. This system transmits blocks of 240 bits containing 180 one-bit samples of the residual, 54 bits for parameters and gain factors, 5 bits for error corrections of critical most significant bits, and 1 framing bit.

3.7 SUBBAND CODING

A subband coder is one form of coder using a frequency-domain analysis of the input signal instead of a time-domain analysis as in previously described coders. As shown in Figure 3.35, the coder first divides the input spectrum into separate bands using a bank of bandpass filters. The signal passing through each of the relatively narrow subbands is individually encoded with separate adaptive APCM, PCM, or ADPCM encoders. After each subband is encoded, the individual bit streams are multiplexed for transmission to the decoder, where they are demultiplexed, decoded, and combined to reconstruct the input.

Separately encoding each subband is advantageous for several reasons. First, by using separate adaptation for each band, the quantization step sizes can be adjusted according to the energy level in each band. Those bands with relatively high energy levels can be encoded with relatively coarse quantization. In this manner the spectrum of the quantization noise is matched to the short-term spectrum of the signal. This property is very desirable in a perceptual sense because it allows the speech signal to

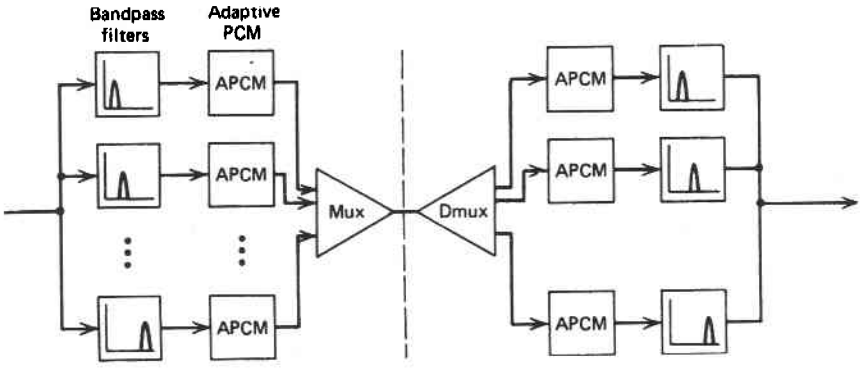


Figure 3.35 Subband coder.

mask the quantization noise. (The human ear perceives speech by measuring the short-term energy level of individual frequency bands. Hence, relatively low noise in a band with no speech energy is perceptually more significant than greater noise in a band with significant speech energy.)

A second advantage of subband coding is that the bit rate (quality) assigned to each band can be optimized according to the perceptual importance of each individual band. In particular, a relatively large number of bits per sample can be used for low frequencies where it is important to preserve the pitch and formant structure of voiced sounds. At higher frequencies, however, fewer bits per sample can be used because noiselike fricatives do not require comparable quality in reproduction.

As reported in reference [10], subband coders provide significant bit rate reductions compared to the more common and simpler coding algorithms: adaptive delta modulation and adaptive differential PCM. Specifically, subband coders at 16 kbps are reported to be perceptually equivalent to ADPCM coders at 22 kbps. A subband coder at 9.6 kbps is reported to be equivalent to an ADM coder at 19.5 kpbs. Extensive description and performance analysis of subband coding are available in reference [35].

A particularly significant example of subband coding is the ITU-T recommendation G.722 for wideband speech coding [36, 37]. This standard provides for encoding 7-kHz speech bandwidth with 64 kbps. Thus it provides a means of significantly improving voice fidelity when end-to-end digital channels are available. Applications that can benefit most from the higher fidelity are teleconferencing and speakerphones. Because it is not intended for internal equipment of the network, it does not have to process voiceband data signals or support tandem analog conversions. Transcodings to and from uniform PCM are required in support of conference bridge arrangements.

As shown in Figure 3.36 [36], the G.722 algorithm divides the input speech bandwidth from 50 to 7000 Hz into two approximately equal-sized subbands. Both subbands are encoded with ADPCM: the lower subband at 48 kbps and the upper subband

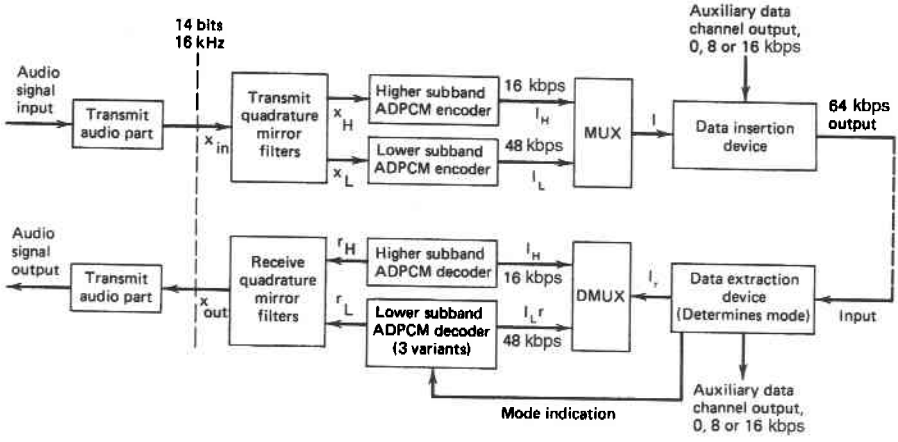


Figure 3.36 G.722 7-kHz audio codec. (From reference [36], p. 10.)

at 16 kbps. The upper band does not require as high a data rate because it is not as important perceptually and has significantly less energy. The algorithm permits the lower band to be encoded at 40 or 32 kbps, which allows 8 or 16 kbps for an auxiliary data channel in teleconferencing applications or for the use of 56-kbps channels when 64 kbps is not available. Figure 3.37 shows the MOS performance of the G.722 algorithm for speech and music at three basic rates [38].

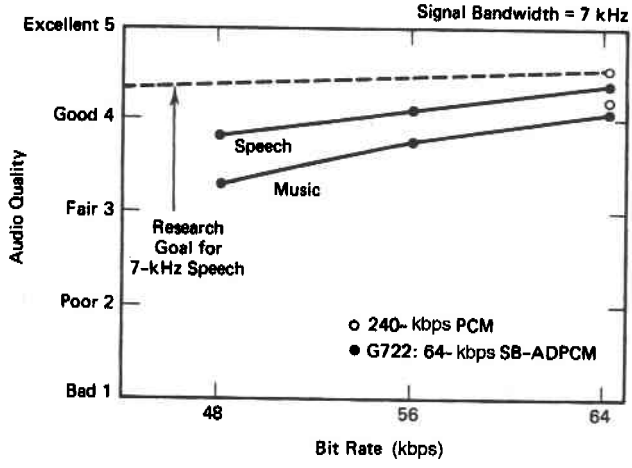


Figure 3.37 Average MOS versus bit rate of G.722 *k*-kHZ audio codec. (From reference [38], p. 15.)

3.8 VOCODERS

For the most part, the encoding/decoding algorithms described previously have been concerned primarily with reproducing the input waveform as accurately as possible. Thus they assume little or no knowledge of the nature of the signal they process and are basically applicable to any signal occurring in a voice channel. Exceptions occur when subband coding and adaptive predictive coding are designed for particularly low bit rates (20 kbps or less). At these bit rates the encoders have been closely tailored to the statistics of a speech signal and cannot provide comparable quality for other signals. Differential systems, such as DPCM and DM, also exhibit a certain amount of speech-specific properties by virtue of their high-frequency encoding deficiencies (slope overload).

The digitization procedures described in this section very specifically encode speech signals and speech signals only. For this reason these techniques are referred to collectively as "vocoders," an acronym for voice coders. Since these techniques are designed specifically for voice signals, they are not applicable to portions of the public telephone network in which other analog signals (such as modem signals) must be accommodated.*

The basic goal of a vocoder is to encode only the perceptually important aspects of speech with fewer bits than the more general waveform encoders. Thus they can be used in limited-bandwidth applications where the other techniques cannot.

Some of the main applications for vocoders are recorded (e.g., "wrong number") messages, encrypted voice transmission over narrowband HF radio, digital cellular telephony, digital circuit multiplication, computer output, games, and telephony over the Internet. A particularly interesting, early use of an LPC vocoder provided multiple voice channels over a single voice frequency leased line. Using a well-conditioned leased line to obtain a 9600-bps circuit, four 2400-bps voice signals were time division multiplexed into a single line [39]. This is one of the first cases (early 1980) where digitization was used to actually decrease the bandwidth of a voice signal. This system provided intelligible voice, but the overall quality was below telephone standards. This particular system became obsolete when long-distance leased-line costs dropped to the point that the sacrifice in voice quality was unjustified.

This section describes three of the most basic vocoding techniques: the channel vocoder, the formant vocoder, and the aforementioned LPC. Many other forms and variations of vocoders have been proposed and studied. For a discussion of some of the other techniques and an extensive bibliography on the subject, see reference [10]. Most commercial applications for vocoders have concentrated on adaptations of the LPC algorithm, particularly for digital cellular and voice over data networks.

Phase Insensitivity

A fundamental requirement for maintaining good speech quality involves preserving the short-term power spectrum of the signal. The phase relationship between individ-

*Vocoders can be inserted into internal portions of a network if the interfaces are equipped to detect the presence of voiceband modem or fax signals and process them accordingly.

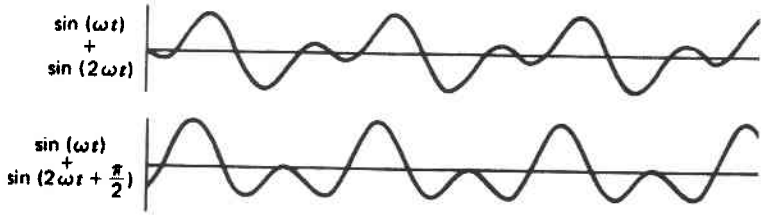


Figure 3.38 Effect of phase shift in the superposition of two tones.

ual frequency components is perceptually much less important. One of the best examples of the ear's insensitivity to phase is demonstrated when two notes are played simultaneously, or nearly simultaneously, on a piano. The composite sound, as perceived by a listener, is seemingly no different if one note is struck slightly later than the other. In essence, the ear senses how much energy is present at various frequencies in the speech spectrum but does not sense the phase relationship between individual frequency components.

The effect of a phase shift in one component of a composite signal is shown in Figure 3.38. The first composite waveform is produced when two individual frequency components have identical starting phases. The second composite waveform occurs when the two frequency terms have starting phases shifted by 90° with respect to each other. Notice that the composite waveforms are markedly different even though the difference is imperceptible to the ear. For these reasons the time waveform produced by a vocoder generally bears little resemblance to the original input waveform. Instead, the emphasis of a vocoder is in reproducing the short-term power spectrum of the input.

3.8.1 Channel Vocoder

Channel vocoders were first developed in 1928 by Homer Dudley [40]. Dudley's original implementation compressed speech waveforms into an analog signal with a total bandwidth of about 300 Hz. Based on the original concept, digital channel vocoders have been developed operating in the range of 1–2 kbps.

A major part of the encoding process of a channel vocoder involves determining the short-term signal spectrum as a function of time. As indicated in Figure 3.39, a bank of bandpass filters is used to separate the speech energy into subbands that are full wave rectified and filtered to determine relative power levels. The individual power levels are encoded and transmitted to the destination. Notice that this much of a channel vocoder is very similar to the subband coder discussed previously. A subband coder, however, typically uses wider bandpass filters, which necessitate sampling the subband waveforms more often (determining a waveform instead of just a power level). Since a subband coder encodes waveforms, it also includes phase information that is ignored by a channel vocoder.

In addition to measuring the signal spectrum, modern channel vocoders also determine the nature of speech excitation (voice or unvoiced) and the pitch frequency of

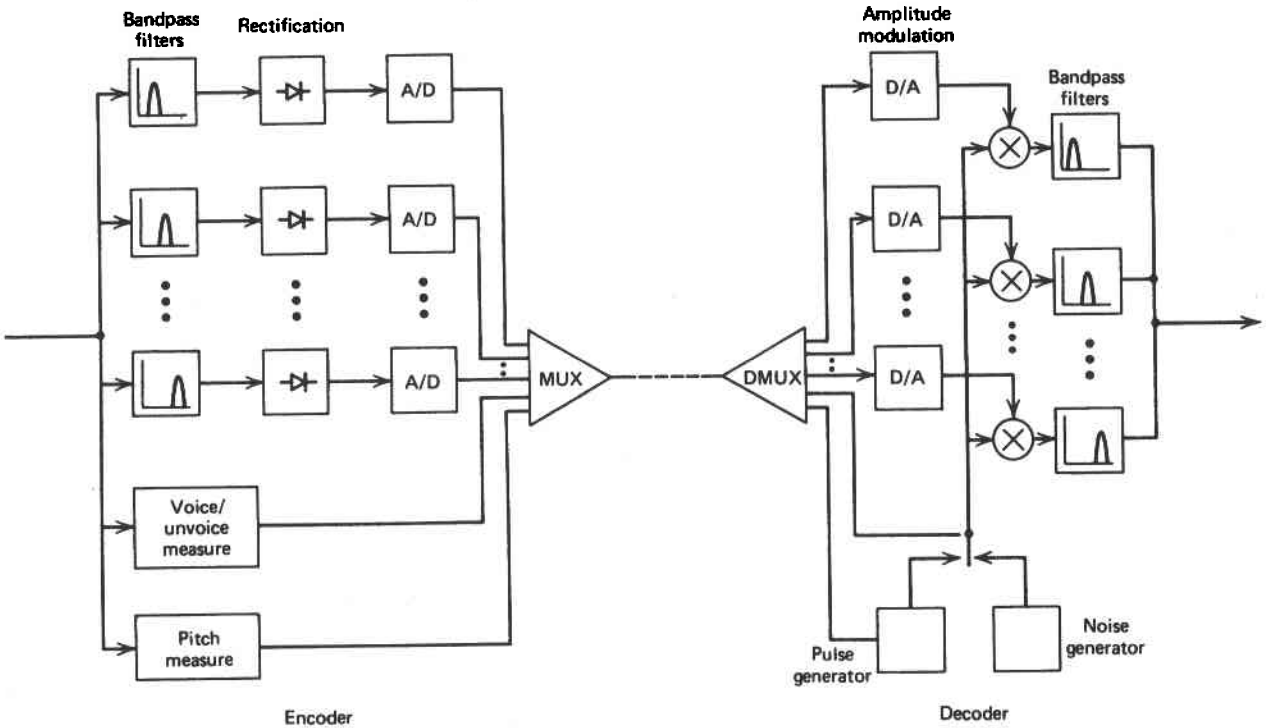


Figure 3.39 Channel vocoder.

voiced sounds. The excitation measurements are used to synthesize the speech signal in the decoder by passing an appropriately selected source signal through a frequency-domain model of the vocal tract transfer function. Voiced excitation is simulated by a pulse generator using a repetition rate equal to the measured pitch period. Unvoiced excitation is simulated by a noise generator. Owing to the synthesized nature of the excitation, this form of a vocoder is sometimes referred to as a pitch-excited vocoder.

As indicated in Figure 3.39, a decoder implements a vocal tract transfer function as a bank of bandpass filters whose input power levels are determined by respective subband power levels in the encoder. Thus outputs of each bandpass filter in the decoder correspond to outputs of respective bandpass filters in the encoder. Superposing the individual bands re-creates, in a spectral sense, the original signal.

Many variations in the basic channel vocoder have been developed involving the nature of the excitation and the means of encoding the power levels. Recent advances in digital technology have introduced the use of digital signal processing to determine the input spectrum by way of Fourier transform algorithms in lieu of the bank of analog filters. All forms of vocoders that measure the power spectral density are sometimes referred to collectively as *spectrum channel vocoders* to distinguish them from time-domain vocoders such as the LPC described later.

The most difficult aspect of most vocoder realizations involves determining the pitch of voiced sounds. Furthermore, certain sounds are not clearly classifiable as purely voiced or purely unvoiced. Thus a desirable extension of the basic vocoder involves more accurate characterization of the excitation. Without accurate excitation information, vocoder output quality is quite poor and often dependent on both the speaker and the particular sounds being spoken. Some of the more advanced channel vocoders have produced highly intelligible, although somewhat synthetic sounding, speech at 2400 bps [23].

3.8.2 Formant Vocoder

As indicated in the spectogram of Figure 3.26, the short-term spectral density of speech is rarely distributed across the entire voice band (200–3400 Hz). Instead, speech energy tends to be concentrated at three or four peaks called formants. A formant vocoder determines the location and amplitude of these spectral peaks and transmits this information instead of the entire spectrum envelope. Thus a formant vocoder produces lower bit rates by encoding only the most significant short-term components in the speech spectrum.

The most important requirement for achieving useful speech from a formant vocoder involves accurately tracking changes in the formants. Once this is accomplished, a formant vocoder can provide intelligible speech at less than 1000 bps [10].

3.8.3 Linear Predictive Coding

A linear predictive coder is a popular vocoder that extracts perceptually significant features of speech directly from a time waveform rather than from frequency spectra, as does a channel vocoder and formant vocoder. Fundamentally, LPC analyzes a

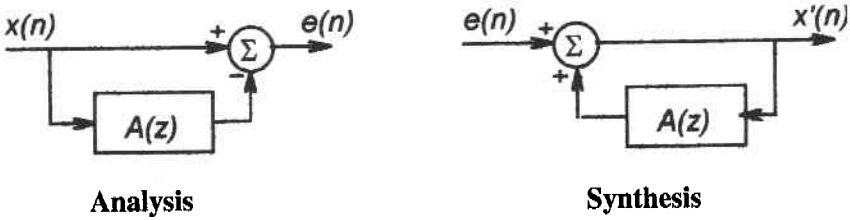


Figure 3.40 Basic model of LPC analysis and synthesis relationship.

speech waveform to produce a time-varying model of the vocal tract excitation and transfer function. A synthesizer in the receiving terminal re-creates the speech by passing the specified excitation through a mathematical model of the vocal tract. By periodically updating the parameters of the model and the specification of the excitation, the synthesizer adapts to changes in either. During any one specification interval, however, the vocal tract is assumed to represent a linear time-invariant process.

The analysis–synthesis relationship between the encoder/decoder is shown in Figure 3.40. The analysis process determines the coefficients of the $A(z)$ matrix to minimize the error $e(n)$ with a given set of speech samples $x(n)$. If the decoder knows both $A(z)$ and $e(n)$, it can re-create (synthesize) the input samples $x(n)$. In the most basic form of LPC $e(n)$ is not transmitted to the decoder. Instead, $e(n)$ is indirectly specified as the excitation of a vocal tract model. Notice that $e(n)$ is the excitation of the synthesizer in Figure 3.40.

A block diagram of the basic model for speech generation is shown in Figure 3.41, which is also a model of the most basic form of LPC decoder/synthesizer. The equation of the vocal tract model shown in Figure 3.44 is defined as:

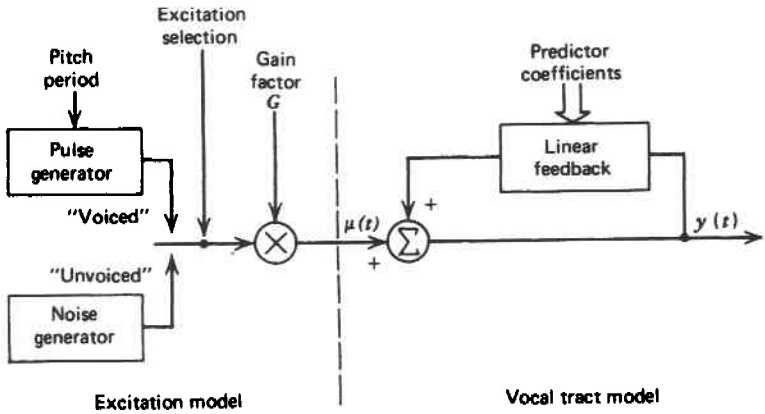


Figure 3.41 Speech generation model of linear predictive coding.

$$y(n) = \sum_{k=1}^{p} a_k y(n-k) + Gx(n) \tag{3.14}$$

where $y(n)$ = nth output sample
a_k = kth predictor coefficient
G = gain factor
$x(n)$ = input at sample time n
p = order of the model

Notice that the speech output in Equation 3.14 is represented as the present input value plus a linear combination of the previous p outputs of the vocal tract. The model is adaptive in that the encoder periodically determines a new set of predictor coefficients corresponding to successive speech segments. Basic LPC does not measure and encode difference waveforms or error signals. Instead, the error signals are minimized in a mean-squared sense when the predictor coefficients are determined. The ability to avoid encoding the predictor errors comes from the fact that LPC uses prediction parameters based on the actual input segments to which the parameters are applied (forward estimation). In contrast, most of the predictive coders mentioned previously base their prediction on past measurements only (backward estimation). Rapid changes in the vocal tract or excitation cause models based on past measurements to be less accurate. The negative aspect of forward estimation is the delay it inserts into the encoding process, which, depending on the application, may or may not be a consideration.

The information that a LPC encoder/analyzer determines and transmits to the decoder/synthesizer consists of

1. Nature of excitation (voiced or unvoiced)
2. Pitch period (for voiced excitation)
3. Gain factor
4. Predictor coefficients (parameters of vocal tract model)

The nature of the excitation is determined, as in other vocoders, by determining whether strong periodic components exist in the waveform. Pitch is determined by measuring periodicity when it exists. In addition to measuring pitch with techniques similar to those used by other vocoders, an LPC encoder/analyzer has particular properties that aid in pitch determination [41].

The predictor coefficients can be determined using one of several different computational procedures. All procedures use actual waveform samples as the desired outputs of the synthesizer. Using these sample values, a set of p linear equations in p unknown coefficients is produced. Thus the coefficients are determined by inverting a p-by-p matrix. Since the order of p may vary from 6 to 12, depending on the speech quality desired, straightforward matrix inversion represents a significant amount of computation. Depending on specific assumptions made in the model, however, the

matrices have special properties that greatly simplify the solution of the equations [41].

Although linear predictive coders process time-domain representations of speech, their operation is such that they provide good estimates of the peaks of the speech spectrum. Furthermore, an LPC is capable of effectively tracking gradual changes in the spectrum envelope. The overall result is that LPCs provide more natural sounding speech than the purely frequency-domain-based vocoders [23]. Most LPC research has concentrated on encoding speech in the range of 1.2–2.4 kbps.

The U.S. Department of Defense established a 2400-bps, 10th-order LPC (LPC-10) as a standard for secure communications over dial-up telephone lines [42, 43]. This system transmits blocks of 54 bits every 22.5 msec. The 54 bits are allocated as 7 bits for pitch, 5 bits for energy level (gain factor), 41 bits for 10 coefficients (voiced excitation), and 1 bit for framing.

The performance of the coder with respect to natural speech is provided in Table 3.4. A diagnostic rhyme test (DRT) [44] is a means of determining the intelligibility of a speech system as a percentage of correct word recognitions from a standardized list of pairs of rhyming words (e.g., goat vs. coat or thick vs. tick). The DRT-with-noise test measures the word recognition percentage when background noise appropriate to the application is added to the test words before encoding. The diagnostic acceptability measure (DAM) [45] rates intelligibility and subjective acceptability using procedures that eliminate much of the dependence on personal preferences of the listeners. The scores are normalized from 0 to 100.

3.8.4 Enhanced-Excitation Linear Predictive Coding

The basic LPC algorithm described in the previous section synthesizes speech in the decoder using a very simple model for the excitation that requires only about 10% of the aggregate data rate. Because voiced signals are modeled by a simple pulse generator, an LPC coder is sometimes referred to as a pulse-excited linear prediction (PELP) coder. The simplicity of the model inevitably produces synthetic sounding speech. To overcome this shortcoming, numerous techniques have been developed to enhance the excitation. Three enhanced-excitation algorithms are multipulse LPC (MPLPC), residual excited linear prediction (RELP), mixed-excitation linear prediction (MELP), and various versions of code-excited linear prediction (CELP). Because the enhanced-excitation LPC algorithms involve encoding a residual prediction error

TABLE 3.4 Performance Comparison of Natural Speech and LPC-10

Test	Natural Speech (3.6 kHz)	LPC-10
DRT	95%	90%
DRT with noise	92%	82%
DAM	65	48

in some form or another, these algorithms are often referred to as hybrid codecs: a combination of vocoding and waveform coding.

Multipulse LPC

As shown in Figure 3.42, where it can be compared with a conventional LPC system, MPLPC [46] is a conceptually simple extension of conventional LPC. Where LPC uses the residual to determine the pitch period of voiced signals, MPLPC uses the residual to determine a sequence of pulses to more closely match the residual. In the simplest case MPLPC uses a fixed number of pulses in a pulse train and determines only the starting phase of the train and the amplitudes and polarities of each pulse within the train. A significant side benefit of MPLPC is that it does not have to identify whether a speech segment is voiced or not and consequently determine the pitch period of voiced signals. Instead, the multipulse determination in each analysis segment automatically adapts to the nature of the actual excitation. MPLPC at 9600 bps is used in AT&T's 14A Announcement System for recorded messages to telephone users [46].

MPLPC is also the algorithm chosen for the Skyphone Aeronautical Telephone service using 9.6-kbps channels on the INMARSAT maritime satellite [47]. Another application of an MPLPC algorithm is the PanEuropean digital mobile radio system (GSM) [48] that allocates a 13-kbps data rate to voice encoding. The speech codec that has been standardized by CEPT for this application is referred to as regular-pulse excitation LPC with long-term prediction (RPE-LTP) [49]. This system transmits 260 bit blocks consisting of 72 bits of prediction parameters and 188 bits of excitation encoding. Because the codec does not support voiceband data rates at 1200 bps and above, the system has provisions for data bypass of the codec.

Mixed-Excitation LPC

As the name implies, mixed-excitation LPC (MELP) [50] uses a more sophisticated model for the excitation than either LPC or MPLPC. An MELP algorithm has been

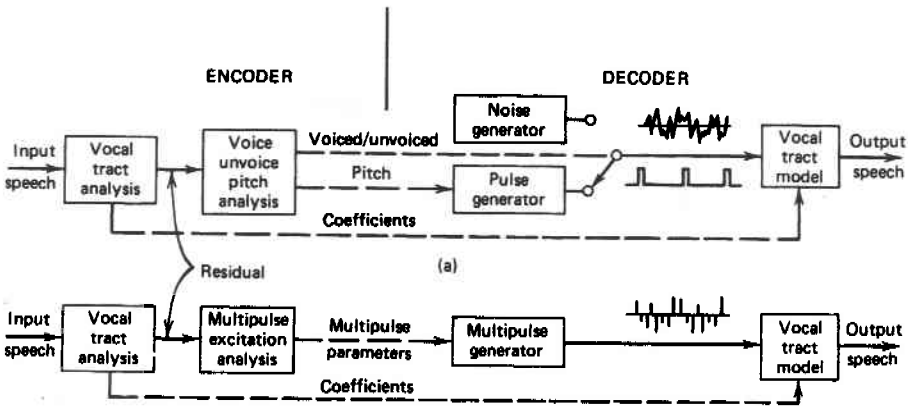


Figure 3.42 Multipulse linear predictive coding. Comparison of LPC (*a*) and MPLPC (*b*) encoder/decoder.

selected by the U.S. government for secure voice applications at 2.4 kbps [51]. Development of this algorithm has paid particular attention to background noise (as might arise in a tank). Enhanced excitation features of this version of LPC are:

1. Mixed pulse and noise excitation
2. Periodic or nonperiodic pulse generation
3. Adaptive spectral enhancement (to improve formant reproduction)
4. Pulse dispersion (spreading the impulses across more of a pitch interval)
5. Fourier analysis of residual (to improve pitch determination)

Residual Excited LPC

The APC algorithm described in Section 3.6 transmits the encoded difference signal (prediction error or residual) determined in the encoder so the decoder can integrate it to recover the original input. Notice that the decoded residual waveform in the APC decoder of Figure 3.34 acts as the excitation of the prediction loop. If the prediction loop uses an LPC formulation, APC essentially becomes residual-waveform excited LPC.

The formal term *residual excited LPC* (RELP) refers to a structure that is identical to the APC block diagram shown in Figure 3.34 but differs in the manner in which the residual is encoded. An RELP encoder does not encode the residual directly but preprocesses it to allow transmission at a lower data rate. The fundamental premise of the preprocessing is that the residual consists of a fundamental frequency and multiples (harmonics) of the fundamental. Thus, an RELP encoder encodes only the fundamental. The decoder reconstructs the residual (in a frequency-domain sense) by decoding the fundamental and adding in the harmonics. In the sense that the process of extracting and encoding the fundamental is "residual encoding" and that of decoding and adding harmonics is "residual decoding," the diagram in Figure 3.34 serves as a basic diagram of RELP.

As commonly implemented, an RELP encoder [52] extracts the fundamental by low-pass filtering the residual and down sampling to reduce the sampling frequency to the Nyquist rate. As an example, if a (DSP-based) low-pass filter has a cutoff of 800 Hz, only every fifth sample of the filtered residual is needed to meet the 1600-Hz Nyquist rate. In this manner, an RELP decoder is excited by samples occurring at a 1600-Hz rate—approximately the same rate as in MPLPC. Thus the residual decoder in an RELP receiver is similar to the pulse generator of an MPLPC synthesizer. In fact, the regular pulse excitation LPC of the European digital mobile system determines its pulse excitation by testing each of four sequences obtained by taking every fourth sample of a filtered residual and selecting the sequence with maximum correlation to the residual sequence. Thus even the encoding algorithms of MPLPC and RELP are sometimes similar. In a comparison of three specific coders (subband, MPLPC, and RELP), MPLPC coders provided the best performance [53].

RELP encoding has been used in various mobile radio and satellite applications. An example of the latter is a 9.6-kbps RELP codec designed to support conversion of analog FM voice channels to digital voice channels [54].

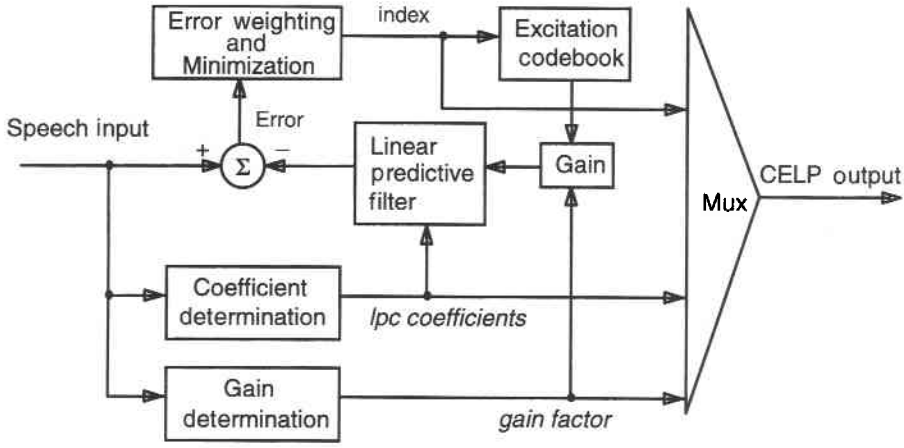


Figure 3.43 CELP encoder block diagram.

Code-Excited LPC

Code-excited LPC (CELP) [55] is yet another example of enhanced-excitation LPC whose basic block diagram can be loosely represented by Figure 3.34. As indicated in the CELP specific block diagram of Figure 3.43, CELP uses a codebook for encoding residuals. Instead of encoding a residue waveform on a sample-by-sample basis and using that as excitation in the decoder, CELP assumes residuals are not random waveforms with independent samples but rather that a block of residue samples can be represented by one of a manageable number of waveform templates. "Excitation encoding" in this case involves selecting a codeword in a codebook that minimizes the overall error in the reconstructed (synthesized) signal. "Residual decoding" in this context implies using the received codeword as an index into the table of codewords to obtain the residual sequence chosen in the encoder. Because a block of residue samples can be considered as a residue vector, this form of coding is also referred to as vector quantization (VQ) excited LPC [56].

Maximum coding efficiency is achieved by encoding long sequences (i.e., vectors with many dimensions), which, of course, implies large codebooks. Thus, much of the research effort of this class of codecs involves establishing a large enough set of vectors in the codebook that all residue vectors can be adequately matched. Next, an efficient procedure for searching the codebook is determined to find the vector with the best match. Because the number of vectors is necessarily large, tree-structured searches are required, which implies the entries in the codebook are grouped into hierarchical families of vectors.

Because CELP coders provide good quality at low bit rates, an extensive amount of research has been undertaken to produce a variety of CELP algorithms. A particularly efficient implementation determines the excitation codeword as a sum of sequentially determined codewords called vectors. Thus, this technique is referred to as vector sum excited LPC (VSELP) [56]. The following list identifies prominent ver-

sions of VSELP and CELP in North American applications. Other versions of CELP coders are included in the list of ITU standard coding algorithms in Section 3.10.

VSELP (vector sum excited LPC); used in North American digital cellular systems (IS-54/IS-136) at 7.95 kbps [57]

QSELP (Qualcomm vector sum excited LPC): used in CDMA digital cellular systems (IS-95) at 13 kbps

DoD-CELP Federal standard (FS-1016) at 4.8 kbps [58]

3.9 ENCODER/DECODER SELECTION CONSIDERATIONS

This chapter has described several techniques for digitizing voice signals and has indicated that many other variations are possible. This section compares the various algorithms in terms of (1) voice quality, (2) transparency for nonvoice signals, (3) tolerance of transmission errors, and (4) delay.

3.9.1 Voice Quality

Voice coding applications and corresponding quality levels have been traditionally classified as broadcast quality, toll quality, communications quality, and synthetic quality. These categories and their respective ranges of data rates are qualitatively shown in Figure 3.44 as obtained from reference [11]. The relatively new category of hybrid coding (e.g., MPLPC or CELP) has been added. Table 3.5, obtained from references [38] and [59], compares various coders in terms of the quality measures: DRT [44], DAM [45], and MOS [9, 11, 18]. Notice that the low-rate coders do reasonably well on intelligibility (DRT) but fall off significantly in terms of subjective measures. Another presentation of speech quality of various coders is shown in Figure 3.45, obtained from reference [60] (which contains a good overview of ITU standard coding algorithms). This figure essentially contains specific MOS ratings of specific coders.* Additional comparisons of low-rate coders (APC, RELP, LPC) are available in references [61] and [62].

The category of toll quality, which is traditionally used for public telephone network voice quality objectives, is no longer a well-defined category. Toll quality in the analog network could be quantified in terms of parameters like frequency response and noise level. Because new, low-rate coders (e.g., CELPs) do not reproduce waveforms, a quantitative analytical measure of quality is not possible. Suppliers of highly compressed voice for applications like digital cellular and voice over the Internet commonly use DRT or MOS scores to support the claim of "toll quality" under ideal transmission environments (e.g., no bit errors or packet loss). This practice has contributed to a relaxation of a "toll quality" standard.

*MOS scores may vary by 5% or so from one study to another.

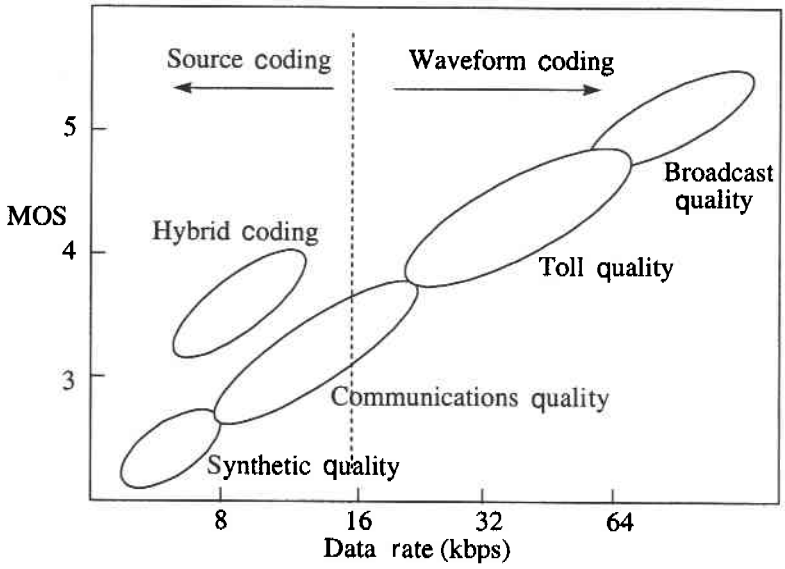


Figure 3.44 General speech quality versus transmission rate.

3.9.2 Transparency for Nonvoice Signals

The separation between communications quality and toll quality coders in Figure 3.44 also separates those algorithms that do not perform well on nonvoice signals from those that do. The lower data rates of communications quality coders are achieved by tailoring the algorithms to the specifics of voice signals, thereby sacrificing modem and facsimile performance. For a comprehensive theoretical treatment of voiceband

TABLE 3.5 Speech Quality Comparisons of Common Coders

Coder	DRT	DAM	MOS
64-kbps PCM	95	73	4.3
14.4-kbps QCELP13			4.2
32-kbps ADPCM (G.721)	94	68	4.1
16-kbps LD-CELP (G.728)	94	70	4.0
6.4 kbps MP-MLQ (G.723.1)			3.9/3.4[a]
13-kbps RPE-LTP (GSM)			3.5
9.6-kbps MPLPC (Skyphone)			3.4
9.6-kbps QCELP			3.4
8-kbps CELP	93	68	3.7
4.8-kbps DoD-CELP	93	67	3.0
2.4-kbps LPC	90	54	2.5

[a]The lower performance corresponds to 3% packet loss.

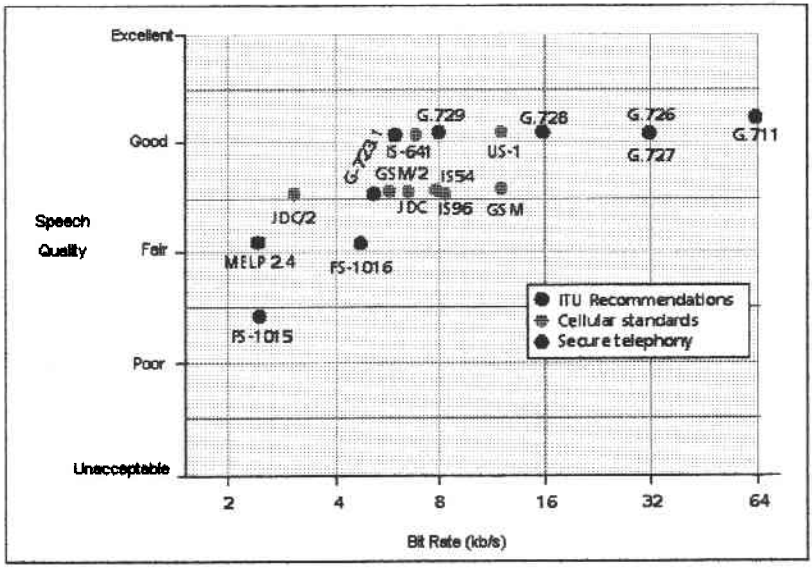


Figure 3.45 Speech quality of standard encoding algorithms. (From R. V. Cox, "Three New Speech Coders from the ITU Cover a Range of Applications," *IEEE Communications Magazine*, September 1997.)

data signal digitization, see reference [63]. As end-to-end digital channels become more available, there will be less need to support nonvoice applications as analog voiceband signals. (See the discussion of V.90 PCM modems in Chapter 11.)

In addition to end user applications, coders installed in the internal portions of a network must provide acceptable quality for network-related signaling tones such as DTMF, MF, and SF. DTMF tones, in particular, must be accurately reconstructed because they are used for end-to-end control by users. G.729 at 8 kbps has special provisions for carrying DTMF tones. Support for MF and SF signaling is less of a problem because their use is confined to specific transmission links that have been mostly replaced by newer transmission systems using common-channel signaling.

An additional consideration for voice quality is the performance of some very low bit rate vocoders in the presence of audio background noise. If a coder is tuned too tightly to process voice signals and voice signals only, it can go off into left field when speech is superimposed on background sounds such as loud music. The low-bit-rate coders used in digital cellular applications have to be particularly sensitive to this situation.

3.9.3 Tolerance of Transmission Errors

Of the waveform coders, the differential systems (DPCM, ADPCM, DM) are the most tolerant of transmission errors and PCM the least. The threshold of perceptibility of random errors on delta modulation is 10^{-3}. For PCM the threshold is 10^{-4}. Delta modulation is intelligible at random error rates as high as 10%, but PCM is unintelli-

gible at 1% error rates. In burst error environments the tolerances of PCM and differential systems are more nearly the same. (If the most significant bit of a PCM codeword is in error, it does not matter if the least significant bits are also in error.)

Those parameters of synthetic quality coders that are critical for voice reconstruction are usually redundantly encoded. Error correction of critical parameters in digital mobile systems allows intelligible voice at error rates as high as 1%.

3.9.4 Delay

The effect of encoding and decoding delay of a voice digitization/compression algorithm must be considered in the context of the particular application. If the application involves inserting artificial delay of more than 10 msec into local (analog) connections, echo/singing control will have to be added. Furthermore, if a significant amount of delay is added into a long-distance circuit, existing echo cancelers may not have enough delay capacity to accommodate the delay insertion.

Experience with satellite-based voice connections indicates that roundtrip delays on the order of 200 msec can be tolerated without significant user dissatisfaction. As indicated in the following, coding/compression algorithms, in themselves, do not approach this limit.

ADPCM (G.726)	0.125 msec
LD-CELP (G.728)	2.5 msec
CS-CELP (G.729) [64]	10 msec
ACELP (G.723.1)	30 msec (plus 7.5 msec of look-ahead)

Although the roundtrip encoding/decoding delays (which are double the above numbers) do not approach the 200-msec threshold, they can add to other system delays to exceed the maximum desirable delay. System factors such as interleaving for error correction and packet delay on TDMA mobile systems can add another 30 msec or so in each direction. If individual compressed voice packets are carried through an ATM or packet-switched network, delays in excess or 200 msec are easily produced. (See Chapter 10 for a discussion of these applications.)

3.10 ITU-T CODING STANDARDS

The following list identifies various digital voice coding standards of the ITU:

G.711 Standard for speech codecs that provides toll quality audio at 64 kbps using either *A*-law or μ-law PCM.

G.721 Standard for speech codecs that provides toll quality audio at 32 kbps using ADPCM.

G.722 Standard for speech codecs that provides high-quality (program) audio at 64 kbps using subband ADPCM (SB-ADPCM). The algorithm uses a 16-kHz sample rate to capture audio frequencies between 50 and 7000 Hz. Two ADPCM subbands are used by this standard to give audio performance superior to a single-band ADPCM algorithm operating at the same bit rate.

G.723 Standard for speech codecs that provides toll quality audio at 20 or 40 kbps using ADPCM.

G.723.1 Standard for speech codecs optimized for modems. It provides toll quality audio at 6.4 kbps (MP-MLQ) or 5.3 kbps (ACELP).

G.726 Adaptive differential PCM at 16, 24, 32, or 40 kbps (using 2-, 3-, 4-, or 5-bit samples).

G.727 An extension of Recommendation G.726 for use in packetized speech systems. The ADPCM samples are divided into 2, 3, or 4 core bits and 0, 1, 2, or 3 enhancement bits. The core bits provide for the basic functioning of the algorithm while the enhancement bits add quality to that provided by the core bits. In overload situations, the enhancement bits can be discarded while the core bits provide basic quality.

G.728 Standard for speech codecs that provides near toll quality audio at 16 kbps using low-delay CELP (LD-CELP). G.728 encodes five μ-law or *A*-law PCM samples into 10-bit, linear predictive codewords at 1600 codewords per second.

G.729 Standard for speech codecs that provides toll quality audio at 8 kbps using CELP.

REFERENCES

1 J. M. Kasson, "The Rolm Computerized Branch Exchange: An Advanced Digital PBX," *Computer Magazine*, June 1979, p. 24.

2 G. Frantz and K.-S. Lin, "Speech Technology in Consumer Products," *Speech Technology*, Apr. 1982, pp. 25–27.

3 S. Finkelstein, "Customized LPC Vocabulary Lets Cars Talk," *Speech Technology*, Sept./Oct. 1983, pp. 65–69.

4 M. G. Carasso, J. B. H. Peek, and J. P. Sinjou, "The Compact Disc Digital Audio System," *Phillips Technical Review*, Vol. 40, No. 6, 1982, pp. 151–155.

5 L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

6 Y. Sagiska, "Speech Synthesis from Text," *IEEE Communications Magazine*, Jan. 1990, pp. 35–41.

7 J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed. Springer-Verlag, New York, 1972.

8 K. F. Lee, H. W. Hon, and R. Reddy, "An Overview of the SPHINX Speech Recognition System," *IEEE Transactions on Acoustics Speech and Signal Processing*, Jan. 1990, pp. 35–45.

9 N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

10 J. Flanagan, M. Schroeder, B. Atal, R. Crochiere, N. Jayant, and J. Tribolet, "Speech Coding," *IEEE Transactions on Communications*, Apr. 1979, pp. 710–737.

11 N. Kitawaki and H. Nagabuchi, "Quality Assessment of Speech Coding and Speech Synthesis Systems," *IEEE Communications Magazine*, Oct. 1988, pp. 36–44.

12 B. J. McDermott, C. Scagliola, and D. J. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM," *Bell System Technical Journal*, May/June 1978, pp. 1597–1618.

13 "The D3 Channel Bank Compatibility Specification—Issue 3," Technical Advisory No. 32, American Telephone and Telegraph Company, Basking Ridge, NJ, Oct. 1977.

14 M. B. Akgun, "Transmission Objectives for a Subscriber Loop System Digital Voice Codec," paper presented at the International Symposium—Subscriber Loops and Services, Ottawa, 1974, pp. 7.1.1–7.1.7.

15 H. Nann, H. M. Straube, and C. P. Villars, "A Companded Coder for an Experimental PCM Terminal," *Bell System Technical Journal*, Jan. 1962, pp. 173–226.

16 K. E. Fultz and D. B. Penick, "The T1 Carrier System," *Bell System Technical Journal*, Sept. 1965, pp. 1405–1451.

17 H. H. Henning and J. W. Pan, "D2 Channel Bank: System Aspects," *Bell System Technical Journal*, Oct. 1972, pp. 1641–1657.

18 W. R. Daumer, "Subjective Evaluation of Several Efficient Speech Coders," *IEEE Transactions on Communications*, Apr. 1982, pp. 655–662.

19 N. S. Jayant, "Step-Size Transmitting Differential Coders for Mobile Telephony," *Bell System Technical Journal*, Nov. 1975, pp. 1557–1581.

20 D. L. Duttweiler and D. G. Messerschmitt, "Nearly Instantaneous Companding and Time Diversity as Applied to Mobile Radio Transmission," IEEE International Communications Conference, 1975, pp. 4–12–40–15.

21 I. M. McNair, Jr., "Digital Multiplexer for Expanded Rural Service," *Bell Labs Record*, Mar. 1972, pp. 80–86.

22 H. K. Dunn and S. D. White, "Statistical Measurements on Conversational Speech," *Journal of Acoustic Society of America*, Jan. 1940, pp. 278–288.

23 J. W. Bayless, S. J. Campanella, and A. J. Goldberg, "Voice Signals: Bit by Bit," *IEEE Spectrum*, Oct. 1973, pp. 28–34.

24 M. Decina and G. Modena, "CCITT Standards on Digital Speech Processing," *IEEE Journal on Selected Areas in Communications*, Feb. 1988, pp. 227–234.

25 G. Williams and H. Suyderhoud, "Subjective Performance Evaluation of the 32 kbit/s ADPCM Algorithm," *IEEE Globecom Proceedings*, 1984, pp. 23.2.1–23.2.1.8.

26 J. A. Greefkes and K. Riemens, "Code Modulation with Digitally Controlled Companding for Speech Transmission," *Phillips Technical Review*, Vol. 31, 1970, pp. 335–353.

27 J. E. Abate, "Linear and Adaptive Delta Modulation," *Proceedings of IEEE*, Mar. 1967, pp. 298–308.

28 N. S. Jayant and A. E. Rosenberg, "The Preference of Slope Overload to Granularity in the Delta Modulation of Speech," *Bell System Technical Journal*, Dec. 1971, pp. 3117–3125.

29 J. A. GreeEkes and F. de Jager, "Continuous Delta Modulation," *Phillips Research Reports*, Vol. 23, 1968, pp. 233–246.

30 J. A. Greefkes, "A Digitally Companded Delta Modulation Modem for Speech Transmission," *Proceedings of IEEE International Communications*, 1970, p. 7-37 to 7-48.

31 B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," *Bell System Technical Journal*, Oct. 1970, pp. 1973–1986.

32 P. Noll, "On Predictive Quantizing Schemes," *Bell System Technical Journal*, May/June 1978, pp. 1499–1532.

33 M. R. Schroeder, "Linear Predictive Coding of Speech: Review and Current Directions," *IEEE Communications Magazine*, Aug. 1985, pp. 54–61.

34 T. Tremain, "The Government Standard Adaptive Predictive Coding Algorithm: APC-04," *Speech Technology*, Feb./Mar. 1985, pp. 52–62.

35 R. E. Crochiere, "An Analysis of 16 kb/s Sub-Band Coder Performance, Dynamic Range, Tandem Connections, and Channel Errors," *Bell System Technical Journal*, Oct. 1978, pp. 2927–2951.

36 P. Mermelstein, "G.722 A New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals," *IEEE Communications Magazine*, Jan. 1988, p. 815.

37 X. Maitre, "7 kHz Audio Coding within 64 kbit/s," *IEEE Journal on Selected Areas in Communications*, Feb. 1988, pp. 283–298.

38 N. S. Jayant, "High-Quality Coding of Telephone Speech and Wideband Audio," *IEEE Communications Magazine*, Jan. 1990, pp. 10–20.

39 L. Radin, "Low-Bit-Rate Voice Digitizer Hits the Market at 2,400 b/s," *Speech Technology*, Sept./Oct. 1983, pp. 92–95.

40 H. Dudley, "Remaking Speech," *Journal of the Acoustic Society of America*, Oct. 1939, pp. 169–177.

41 L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

42 T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology*, Apr. 1982, pp. 40–49.

43 J. P. Campbell and T. E. Tremain, "Voiced/Unvoiced Classification of Speech with Applications to the U.S. Government LPC-10E Algorithm," IEEE International Conference on Acoustics, Speech, and Signal Processing, 1986, pp. 473–476.

44 W. D. Voiers, "Evaluating Processed Speech Using the Diagnostic Rhyme Test," *Speech Technology*, Jan./Feb. 1983, pp. 30–39.

45 W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," *Proceedings International Conference on Acoustics, Speech and Signal Processing*, May 1977, pp. 204–207.

46 B. S. Atal and R. Q. Hofacker, "The Telephone Voice of the Future," *Bell Labs Record*, July 1985, pp. 4–10.

47 I. Boyd, C. B. Southcott, D. P. Crowe, and P. J. Bolingbroke, "Speech Codec for the Skyphone Aeronautical Telephone Service," *British Telecommunications Engineering*, July 1986, pp. 83–91.

48 J. E. Natig, S. Hansen, and J. deBrito, "Speech Processing in the Pan European Digital Mobile Radio System (GSM)-System Overview," *IEEE Globecom Proceedings*, 1989, pp. 29B.1.1–29B.1.1.5.

49 K. Hellwig, P. Vary, D. Massaloux, J. P. Petit, C. Galand, and M. Russo, "Speech Codec for the European Mobile Radio System," *IEEE Globecom Proceedings*, 1989, pp. 29B.2.1–29B.2.1.5.

50 A. V. McCree and T. P. Barnwell, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Transactions on Speech and Audio Processing*, July 1995, pp. 242–250.

51 "Specifications for the Analog to Digital Conversion of Voice by 2,400 Bit/Second Mixed Excitation Linear Prediction," Department of Defense Draft, May 28, 1998 (www.plh.af.mil/ddvpc/melp.htm).

52 C. K. Un and D. T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbot/s," *IEEE Transactions on Communications*, Dec. 1975, pp. 1466–1474.

53 K. K. Paliwal and T. Svendsen, "A Study of Three Coders (Sub-band, RELP, and MPE) for Speech with Additive White Noise," *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985, pp. 43.5.1–43.5.1.4.

54 A. J. Morris and S. Kay, "Telephony Earth Station," *IEEE Globecom Proceedings*, 1989, pp. 48.1.1–48.1.1.4.

55 M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985, pp. 25.1.1–25.1.1.4.

56 J. H. Chen and A. Gersho, "Gain-Adaptive Vector Quantization with Application to Speech Coding," *IEEE Transactions on Communications*, Sept. 1987, pp. 918–930.

57 "Vector Sum Excited Linear Prediction (VSELP) 7950 Bit Per Second Voice Coding Algorithm," Technical Description, Motorola, Shaumburg, IL, Nov. 14, 1989.

58 J. P. Campbell, V. C. Welch, and T. E. Tremain, "An Expandable Error-Protected 4800 BPS CELP Coder (U.S. Federal Standard 4800 BPS Voice Coder)," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1989, pp. 735–737.

59 T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, Upper Saddle River, NJ, 1996.

60 R. V. Cox, "Three New Speech Coders from the ITU Cover a Range of Applications," *IEEE Communications Magazine*, Sept. 1997, pp. 40–47.

61 R. D. Hoyle and D. D. Falconer, "A Comparison of Digital Speech Coding Methods for Mobile Radio Systems," *IEEE Journal on Selected Areas in Communications*, June 1987, pp. 915–920.

62 N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems," *IEEE Journal on Selected Areas in Communications*, Feb. 1988, pp. 242–247.

63 J. B. O'Neal, Jr., "Waveform Encoding of Voiceband Data Signals," *Proceedings of the IEEE*, Feb. 1980, pp. 232–247.

64 G. Schröder and M. H. Sherif, "The Road to G.729: ITU 8-kb/s Speech Coding Algorithm with Wireline Quality," *IEEE Communications Magazine*, Sept. 1997, pp. 48–54.

PROBLEMS

3.1 Assume a signal consists of three tones: one at 1 kHz, one at 10 kHz, and one at 21 kHz. What tones will be present at the output of a PAM decoder if the sam-

pling rate is 12 kHz and the input is unfiltered? (Assume the output filter cutoff frequency is 6 kHz.)

3.2 Derive an expression for the average quantization noise power that occurs when the decoder output samples are offset from the center of a quantization interval by a distance equal to 25% of the interval. (The output values are at the 75% point instead of the 50% point.) How much degradation in decibels does this offset represent (assuming uncorrelated offsets)?

3.3 How much does the signal-to-noise ratio of a uniform PCM encoder improve when 1 bit is added to the codeword?

3.4 A black-and-white television signal has a bandwidth of about 4.2 MHz. What bit rate is required if this signal is to be digitized with uniform PCM at an SQR of 30 dB? Use a sampling-rate-to-Nyquist-rate ratio comparable to that used for PCM voice encoding.

3.5 How much dynamic range is provided by a uniform PCM encoder with 12 bits per sample and a minimum SQR of 33 dB?

3.6 What is the signal-to-quantizing-noise ratio produced by a segmented μ255 PCM coder when the input signal is a full-range triangular wave? (Assume the repetition frequency is low enough that the bandlimiting filter does not change the waveform significantly.)

3.7 Given a sample value of 420 mV for a μ255 PCM encoder capable of encoding a maximum level of 2 V, determine each of the following:

(a) The compressed μ255 codeword
(b) The linear representation of the compressed code
(c) The A-law code obtained by converting from the μ255 code
(d) The μ-law code obtained by converting back from the A-law code

3.8 Given the following μ255 codewords, determine the (noninverted) codeword that represents the linear sum (0 110 1001), (1 011 0111).

3.9 Generate an encoding table (i.e., list the quantization intervals and corresponding codes) for the magnitude of a piecewise linear code with segment slopes 1, $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$. Assume four equally spaced intervals are in each segment. Assume all intervals in the first segment are of equal length (as in A-law coding).

3.10 What is the signal-to-noise ratio of a full-amplitude sample for the coder of Problem 3.9?

3.11 What is the signal-to-noise ratio of a maximum-amplitude sample in the first linear segment of the preceding coder?

3.12 What is the dynamic range implied by Problems 3.10 and 3.11?

3.13 For the encoder in Problem 3.9, how many bits are required for uniform encoding of the same dynamic range and same minimum quantization interval?

3.14 A uniform PCM system is defined to encode signals in the range of −8159 to +8159 in quantization intervals of length 2. (The quantization interval at the origin extends from −1 to +1.) Signals are encoded in sign-magnitude format with a polarity bit = 1 denoting a negative signal.

(a) How many bits are required to encode the full range of signals?

(b) How many unused codes are there?

(c) Determine the quantization noise, noise power, and signal-to-noise ratio (in decibels) of each of the following sample values: 30.2, 123.2, –2336.4, and 8080.9.

3.15 Repeat part (c) in Problem 3.14 for piecewise linear μ255 PCM.

3.16 Given two *A*-law piecewise linear (noninverted) codewords (00110110 and 00101100), determine their linear representations, add them, and convert back to compressed representation.

3.17 A D3 channel bank uses "robbed digit" signaling wherein the least significant bit of every sixth frame is stolen for signaling. Determine the relative increase in overall quantization noise produced by this process under the following conditions:

(a) The decoder merely treats the signaling bit as a voice bit and decodes the PCM sample accordingly.

(b) The decoder knows which bits are signaling bits and generates an output sample corresponding to the middle of the double-length quantization interval defined by the 7 available bits. (This is the actual technique specified for D3 channel bank decoders.)

3.18 If 2 bits per sample are added to a PCM bit stream, how much can the dynamic range be increased if the quantization intervals are adjusted to improve the SQR by 3 dB?

3.19 How much can the SQR of a digitized voice signal be increased if the bandwidth is reduced by 30% and the dynamic range is unchanged?

3.20 An *A*-law PCM encoder with a perfect zero-level threshold detects an arbitrarily small input sine wave because the first quantization interval does not straddle the origin. What is the power level of the decoded output signal? Assume the power level of a full amplitude sinewave is +3 dBm.

3.21 Determine the sequence of four μ255 PCM samples for a 2000-Hz tone at –6 dBm when the first sample occurs at a phase of 45°.

3.22 Repeat Problem 3.21 for *A*-law signals.

3.23 Determine the *A*-law codewords that correspond to the μ-law codewords of the digital milliwatt signal given in Example 3.3. Assume both systems are scaled to the same maximum signal power.

4

DIGITAL TRANSMISSION AND MULTIPLEXING

A fundamental consideration in the design of digital transmission systems is the selection of a finite set of discrete electrical waveforms for encoding the information. In the realm of digital communications theory these discrete waveforms are referred to as signals. The same terminology is used in this chapter with the understanding that signals in the present context refer to internal waveforms (pulses) of a transmission system and not the control information (signaling) used to set up and monitor connections within a network. In communications theory terminology, *signal processing* refers to filtering, shaping, and transformations of electrical waveforms, not the interpretation of control signals by the processor of a switching machine.

A second aspect of digital transmission involves defining certain time relationships between the individual transmission signals. The source terminal transmits the individual signals using predefined time relationships so the receiving terminal can recognize each discrete signal as it arrives. Invariably the establishment of a proper time base at the receiver requires transmission capacity above that needed for the digital information itself. Over relatively short distances (as within a switching system or a computer complex), the timing information (clock) is usually distributed separately from the information-bearing signals. Over long distances, however, it is more economical to incorporate the timing information into the signal format itself. In either case the timing information requires channel capacity in terms of bandwidth, data rate, or code space.

This chapter discusses the most common digital signaling techniques for wireline transmission systems. Thus the subject of this chapter is often referred to as line coding. These techniques are generally applicable to any transmission system transmitting digital signals directly in the form of pulses (such as coaxial cable or optical fiber). The fact that these techniques include low-frequency components in their transmitted spectrum leads to their also being called low-pass or baseband transmission systems. In Chapter 6 we discuss bandpass transmission systems, that is, radio systems or voiceband modems that require modulation and carrier frequency transmission. Chap-

ter 8 discusses some additional baseband transmission formats commonly used in optical fiber systems.

The following discussions concentrate on system and application level considerations of digital transmission. Analytic details of pulse transmission systems are not emphasized since material of this nature is available in all works on digital communications theory. Some fundamentals of pulse transmission systems are presented in Appendix C, where many of the equations presented in this chapter are derived.

4.1 PULSE TRANSMISSION

All digital transmission systems are designed around some particular form of pulse response. Even carrier systems must ultimately produce specific pulse shapes at the detection circuitry of the receiver. As a first step, consider the perfectly square pulse shown in Figure 4.1. The frequency spectrum corresponding to the rectangular pulse is derived in Appendix A and shown in Figure 4.2. It is commonly referred to as a $\sin(x)/x$ response:

$$F(\omega) = (T)\frac{\sin(\omega T/2)}{\omega T/2} \tag{4.1}$$

where ω = radian frequency $2\pi f$,
T = duration of a signal interval

Notice that Figure 4.2 also provides the percentage of total spectrum power at various bandwidths. As indicated, 90% of the signal energy is contained within the first spectral null at $f = 1/T$. The high percentage of energy within this band indicates that the signal can be confined to a bandwidth of $1/T$ and still pass a good approximation to the ideal waveform. In theory, if only the sample values at the middle of each signal interval are to be preserved, the bandwidth can be confined to $1/2T$. From this fact the maximum baseband signaling rate in a specified bandwidth is determined as

$$R_{\max} = 2\text{BW} \tag{4.2}$$

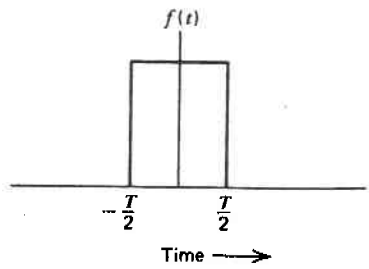


Figure 4.1 Definition of a square pulse.

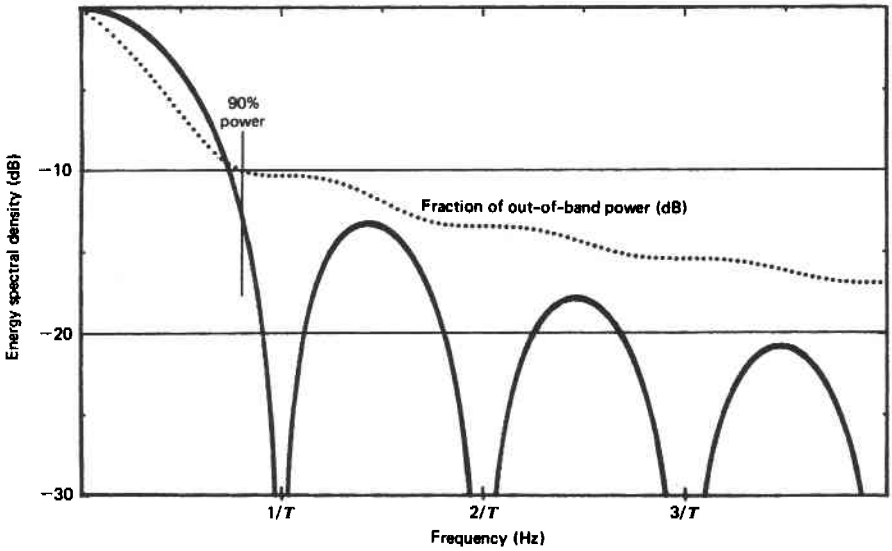


Figure 4.2 Spectrum of square pulse with duration T.

where R = signaling rate, = $1/T$
BW = available bandwidth

Equation 4.2 states a fundamental result from communications theory credited to Harry Nyquist: The maximum signaling rate achievable through a low-pass bandwidth with no intersymbol interference is equal to twice the bandwidth. This rate $R_{\max}$ is sometimes referred to as the *Nyquist rate*.

Although discrete, square-shaped pulses are easiest to visualize, preservation of the square shape requires wide bandwidths and is therefore undesirable. A more typical shape for a single pulse is shown in Figure 4.3. The ringing on both sides of the main part of the pulse is a necessary accompaniment to a channel with a limited bandwidth.

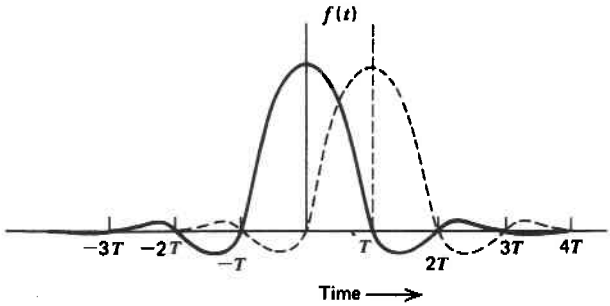


Figure 4.3 Typical pulse response of a bandlimited channel.

Normally, a digital transmission link is excited with square pulses (or modulated equivalents thereof), but bandlimiting filters and the transmission medium itself combine to produce a response like the one shown. Figure 4.3 shows pulse output in negative time so the center of the pulse occurs at $t = 0$. Actually, the duration of the preringing is limited to the delay of the channel, the filters, and the equalizers.

An important feature of the pulse response shown in Figure 4.3 is that, despite the ringing, a pulse can be transmitted once every T seconds and be detected at the receiver without interference from adjacent pulses. Obviously, the sample time must coincide with the zero crossings of the adjacent pulses. Pulse responses like the one shown in Figure 4.3 can be achieved in channel bandwidths approaching the minimum (Nyquist) bandwidth equal to one-half of the signaling rate. Appendix C describes pulse transmission design in more detail.

4.1.1 Intersymbol Interference

As the signaling rate of a digital transmission link approaches the maximum rate for a given bandwidth, both the channel design and the sample times become more critical. Small perturbations in the channel response or the sample times produce nonzero overlap at the sample times called intersymbol interference. The main causes of intersymbol interference are:

1. Timing inaccuracies
2. Insufficient bandwidth
3. Amplitude distortion
4. Phase distortion

4.1.2 Timing Inaccuracies

Timing inaccuracies occurring in either the transmitter or the receiver produce intersymbol interference. In the transmitter, timing inaccuracies cause intersymbol interference if the rate of transmission does not conform to the ringing frequency designed into the channel. Timing inaccuracies of this type are insignificant unless extremely sharp filter cutoffs are used while signaling at the Nyquist rate.

Since timing in the receiver is derived from noisy and possibly distorted receive signals, inaccurate sample timing is more likely than inaccurate transmitter timing. Sensitivity to timing errors is small if the transmission rate is well below the Nyquist rate (e.g., if the transmission rate is equal to or less than the channel bandwidth, as opposed to being equal to the theoretical maximum rate of twice the bandwidth; see Appendix C).

4.1.3 Insufficient Bandwidth

The ringing frequency shown in Figure 4.3 is exactly equal to the theoretical minimum bandwidth of the channel. If the bandwidth is reduced further, the ringing frequency is reduced and intersymbol interference necessarily results.

Some systems purposely signal at a rate exceeding the Nyquist rate, but do so with prescribed amounts of intersymbol interference accounted for in the receiver. These systems are commonly referred to as partial-response systems—so called because the channel does not fully respond to an input during the time of a single pulse. The most common forms of partial-response systems are discussed in a later section.

4.1.4 Amplitude Distortion

Digital transmission systems invariably require filters to bandlimit transmit spectrums and to reject noise and interference in receivers. Overall, the filters are designed to produce a specific pulse response. When a transmission medium with predetermined characteristics is used, these characteristics can be included in the overall filter design. However, the frequency response of the channel cannot always be predicted adequately. A departure from the desired frequency response is referred to as amplitude distortion and causes pulse distortions (reduced peak amplitudes and improper ringing frequencies) in the time domain. Compensation for irregularities in the frequency response of the channel is referred to as *amplitude equalization*.

4.1.5 Phase Distortion

When viewed in the frequency domain, a pulse is represented as the superposition of frequency components with specific amplitude and phase relationships. If the relative amplitudes of the frequency components are altered, amplitude distortion results as above. If the phase relationships of the components are altered, phase distortion occurs. Basically, phase distortion results when the frequency components of a signal experience differing amounts of delay in the transmission link. Compensation of phase distortion is referred to as *phase equalization*. For a good tutorial on equalization, including a description of an automatic equalizer for data transmission, see reference [1]. A good description of DSP based equalizers is available in reference [2].

4.2 ASYNCHRONOUS VERSUS SYNCHRONOUS TRANSMISSION

There are two basic modes of digital transmission involving two fundamentally different techniques for establishing a time base (sample clock) in the receiving terminal of a digital transmission link. The first of these techniques is asynchronous transmission, which involves separate transmissions of groups of bits or characters. Within an individual group a specific predefined time interval is used for each discrete signal. However, the transmission times of the groups are unrelated to each other. Thus the sample clock in the receiving terminal is reestablished for reception of each group. With the second technique, called synchronous transmission, digital signals are sent continuously at a constant rate. Hence the receiving terminal must establish and maintain a sample clock that is synchronized to the incoming data for an indefinite period of time.

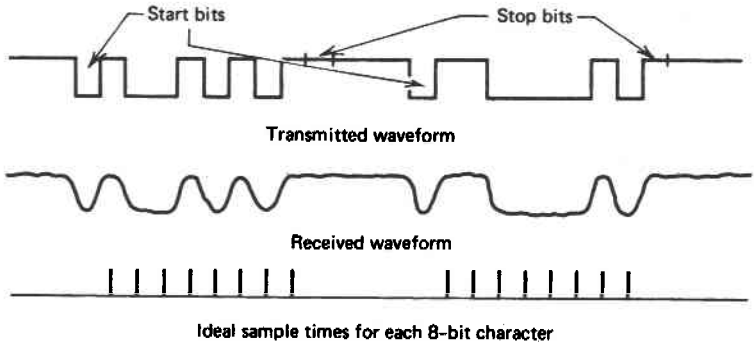


Figure 4.4 Asynchronous transmission.

4.2.1 Asynchronous Transmission

Between transmissions an asynchronous line is in an inactive or idle state. The beginning of each transmission group is signified by a start bit. The middle of the start bit is determined, and succeeding information bits are sampled at a nominal rate beginning at the middle of the second bit interval. Following the information symbols, one or more stop bits are transmitted to allow the line to return to the inactive state.* Figure 4.4 shows an asynchronous mode of operation commonly used for low-speed data communications.

As shown in Figure 4.4, the detection of each information bit is accomplished by ideally sampling the input waveform at the middle of each signal interval. In practice, sample times depart from the ideal depending on how much the start bit is corrupted by noise and distortion. Since the sample time for each information bit is derived from a single start bit, asynchronous systems do not perform well in high-noise environments. Of course, more than one start bit could be used to improve the accuracy of the starting phase of the sample clock, but this would complicate the receiver and add more overhead for transmission of timing information.

Sample timing errors also arise if the nominal rate of the sample clock in the receiver is different from the nominal rate of transmission at the source. Even though the start bit might define the proper starting phase for the sample clock, an offset in the clock frequency of the receiver causes each successive sample time to drift farther from the center of the respective signal intervals. Since the very use of the term "*asynchronous*" implies a free-running clock in the receiver, a certain amount of drift is inevitable in all asynchronous systems. The maximum length of each symbol group or character is determined by the limits of the initial phase inaccuracies and the maximum expected frequency difference between the transmitter and receiver clocks.

*Originally, stop bits were inserted to allow electromechanical equipment enough time to reset before the next character arrived. With electronic equipment the only purpose of stop bits is to allow a start bit to always be a transition to a space (logic 0).

The main attraction of asynchronous transmission is the ease with which it determines the sample times in the receiver. In addition, asynchronous transmission automatically provides character framing and is inherently flexible in the range of average data rates that can be accommodated. For high rates, one character after another is transmitted. Lower data rates are automatically accommodated by increasing the idle time between characters. In contrast, a synchronous receiver must track changes in a transmitter rate before it can sample the incoming signals properly. Normally the receive clock of a synchronous system can be adjusted only quite slowly and only over a narrow range. Hence an asynchronous system is more naturally suited to applications where the data rate varies.

Synchronous transmission systems can support variable information rates, but the task of adjusting the information rate falls upon higher level processes (data link protocols) that insert null codes into the bit stream. The null codes are used as filler when a source has nothing to send. This form of transmission is sometimes referred to as "isochronous." An isochronous mode of operation is required whenever a synchronous line carries data from an asynchronous source.

The major drawback of asynchronous transmission is its poor performance in terms of error rates on noisy lines. Thus asynchronous transmission evolved for applications where implementation costs dominated performance considerations. Asynchronous transmission has been used in voiceband data sets (modems) for transmission rates up to 1200 bps. For digital telephony, something similar to asynchronous transmission was proposed for two-wire digital subscriber loops. These systems provided a full-duplex (four-wire) circuit by transmitting bursts of data alternately in each direction on a single pair of wires. Thus these systems are sometimes referred to as "ping-pong" transmission systems; they are not truly asynchronous since each transmission in each direction occurs at prescribed times allowing timing information in one burst to carry over into the next burst. The use of asynchronous transmission on long-distance transmission links is obsolete because the cost of electronics for better performing synchronous transmission is no longer a consideration.

4.2.2 Synchronous Transmission

T1 lines and all other interoffice digital transmission links used in the telephone networks use synchronous transmission exclusively. Thus the line-coding format for these systems must incorporate special considerations to ensure that each regenerative repeater or receiver can synchronize a local sample clock to the incoming signaling rate. Generally speaking, the synchronization requirements imply that a certain minimum density of signal transitions is required to provide continuous indication of signaling boundaries. Often, purely random data cannot be relied upon to provide a sufficient number of transitions. In these cases certain provisions must be made to insert artificial transitions into the transmitted waveforms. Although these extra transitions imply a certain amount of transmission overhead, the loss in capacity can be relatively small. Following are descriptions of five techniques for ensuring the existence of signal transitions for timing recovery:

1. Source code restriction
2. Dedicated timing bits
3. Bit insertion
4. Data scrambling
5. Forced bit errors

A sixth technique—inserting the transitions into the signal waveforms themselves—is discussed in Section 4.3.

Source Code Restriction

One means of ensuring a sufficient number of signal transitions is to restrict the code set or data patterns of the source so that long, transition-free data sequences do not occur. Historically, PCM channel banks in the North American telephone network precluded all-0's codewords since the original line code of T1 lines produced no transitions for 0's. In the case of 8-bit PCM codewords, the exclusion of a single codeword represents a loss in transmission capacity of only one part in 256.*

Relying on source coding to ensure sufficient transitions in the line code has one very significant drawback: The transmission link cannot be used for new applications where the source does not exclude the unwanted data patterns. For this reason, the total capacity of a T1 line could not be used for random data until a new line code [binary eight-zero substitution (B8ZS)] was introduced.

Dedicated Timing Bits

As an alternative to excluding transition-free data patterns, the line itself can periodically insert transition-bearing bits into the data stream. These bits are inserted at regular intervals, independently of the source data, to ensure the existence of a minimum number of signal transitions. Thus some fraction of the channel capacity is dedicated to timing bits.

As an example, the original Dataphone Digital Service (DDS) offering for data communications over T1 lines provides a maximum capacity of 56 kbps for each channel. When carried within a 64-kbps T1 channel, only 7 of the 8 bits in each time slot of the channel are available for the user. Among other functions, the unused bit in each time slot provides an assurance that all 8 bits of a time slot are not 0. Several fiber systems described in Chapter 8 also use explicit timing bits as inherent aspects of their line codes. The density of timing pulses in these systems ranges from 1 in 5 bits to 1 in 20 bits. Notice that insertion of a dedicated timing bit is essentially the same procedure used to establish timing for asynchronous transmission. In a synchronous receiver, however, a sample clock is obtained by averaging the timing information over a large number of timing transitions, not just one.

Bit Insertion

In the preceding DDS example 1 bit of every 8 bits in a time slot is dedicated to ensuring sufficient timing information in the bit stream. Another possibility for preclud-

*From an information theory point of view the loss in voice information is even lower since the probability of occurrence of the all-0's codeword is much less than 1 in 256.

ing unwanted line patterns is to use "bit insertion" only when necessary. As an example, the source data over a T1 line could be monitored for all 0's in the first 7 bits of a time slot. Whenever the 0's occur, a 1 could be inserted into the data stream as the eighth bit of the time slot. At the other end of the line the 1 following seven 0's is removed automatically. Each insertion delays the source data by the time of 1 bit, but otherwise the full capacity of the channel is available.

This procedure is directly analogous to the "zero-bit insertion" algorithm used in the High Level Data Link Control (HDLC) protocol. In this protocol a specific data pattern called a "flag" is used to indicate the end of a data block. The transmitter must be precluded from inadvertently sending the flag as part of the source data. The means of preventing inadvertent flags is to insert a 0 following a string of five 1's in the user data. Since a flag contains a string of six 1's, zero-bit insertion precludes unintended flag transmissions. The receiving node of an HDLC data link removes any 0 following five 1's. The receipt of six 1's, however, can only be part of a legitimate flag (01111110). If HDLC data are inverted for transmission, it never produces all 0's (except during the idle state).

Although a bit insertion algorithm allows for more efficient use of a channel than dedicated timing bits, the insertion procedure has a number of drawbacks. First, this process causes the source data to be delayed every time an insertion is made. Hence a continuous, real-time transmission application (such as voice) requires smoothing the arrival rate with data buffers at the destination. Second, the bit insertion process causes any character structure in the user's data to be become unrelated to the time slot structure of a time division multiplexed transmission link. Thus if user data consist of 8-bit characters (like PCM voice samples), character boundaries cannot be maintained with respect to 8-bit time slots in a T1 transmission link.

Date Scrambling

Many digital transmission systems use data scramblers to randomize the data patterns on their transmission links. Although these data scramblers are similar to those used for encryption, the fundamental purpose of these scramblers is to prevent the transmission of repetitive data patterns, not to encrypt the traffic. Repetitive data patterns generate line spectra that can be significantly more degrading from an interference point of view than continuously distributed spectra produced by random data patterns. Voiceband data modems, for example, are allowed to operate at higher power levels if they include scramblers to randomize the data traffic. Also, digital radio systems are required by the FCC to not transmit line spectra, which essentially means that repetitive data patterns must be excluded.

Even when not required, data scramblers are useful in transforming data sequences with low transition densities into sequences with strong timing components. Scrambling is not used on lower rate T-carrier systems (T1 and obsolete T2) but was used on the 274-Mbps T4M coaxial transmission system [3] and current optical fiber systems.

Data scramblers (with equal input and output bit rates) do not prevent long strings of 0's in an absolute sense. They merely ensure that relatively short repetition patterns are transformed to randomized traffic with a minimum density of transitions. If purely

random input data are scrambled, the output data are also purely random and therefore have a certain statistical probability of producing arbitrarily long strings of 0's. The probability of a random occurrence, however, is acceptable when compared to the probability of nonrandom sequences representing speech pauses or idle data terminals. To determine which seemingly random data sequence produces all 0's at the output of a scrambler, apply all 0's to the corresponding descrambler.

The T4M coaxial transmission system used a data scrambler as the basic means of producing adequate timing information. This system could tolerate much longer strings of 0's because the timing recovery circuits in the regenerative repeaters used phase-locked loops that maintain timing over relatively long periods of time. In contrast, the original T1 systems recovered timing with tuned circuits that resonated at the desired clock frequency (1.544 MHz) when excited by a received pulse. Because the tuned circuits have lower effective Q's than a phase-locked loop, the oscillations drift from the proper frequency and die out more rapidly. Hence the original T1 receivers could not tolerate as long a string of 0's as could T4M receivers. Phase-locked loop clock recovery circuits are used on all later generation wireline and fiber transmission systems so much longer strings of 0's can be tolerated.

Forced Bit Errors

A fifth method of maintaining sufficient timing information in the line signals involves having the transmission terminal at the source force an occasional bit error in order to interrupt a long, transition-free data pattern. If the transition-free sequences are long enough and uncommon enough, the intentional bit errors might be less frequent than random channel errors on the digital transmission link. Thus the intentional errors may not represent a significant degradation over and above that that already exists. Nevertheless, forced bit errors are not generally recommended as part of a line-coding procedure but are mentioned in the interest of completeness.

As mentioned previously, North American channel banks force a bit error in the second least significant bit of an all-0's transmission code to ensure sufficient signal transitions. An important aspect of this procedure is that it is performed by the source where the significance of the bit error is known. If the transmission link itself inserted the bit errors, the effects would not be as controllable, particularly when a variety of traffic types are being serviced.

A more subtle problem with forced transmission errors arises if the digital transmission link is used for automatic repeat request (ARQ) data transmission. An ARQ data communication link is designed to provide error-free transmission, despite random channel errors, by inserting redundancy into the data stream and checking the received data for error-free reception. If errors are detected, a retransmission is requested. When the errors are not random, but forced by the transmission link, the ARQ system will become frustrated if it ever encounters the restricted sequence, no matter how unlikely it is.* Once again, if forced errors are used, they should be incorporated into the source as part of the source code restriction process and not as a function of the transmission link.

*If the transmission link uses a scrambler, the unlikely sequence will not be repeated.

4.3 LINE CODING

In the preceding section various techniques for establishing timing information are described in general terms. The choice of any particular technique is dependent on the specific line code in use. This section describes the most common line codes used for digital transmission and indicates what additional steps, if any, are needed to maintain synchronization between the transmitter and receiver. Some line-coding techniques provide adequate timing information themselves and do not require any of the procedures discussed previously. In addition to synchronization requirements, other considerations for choosing a line code are the spectrum of the line code and the available bandwidth (particularly at low frequencies), noise and interference levels, synchronization acquisition times, performance monitoring, and implementation costs.

4.3.1 Level Encoding

Conceptually, the simplest form of line coding uses a different signal level to encode each discrete symbol transmitted. Within a computer system the most common form of coding is an on–off code using a 3-V level for a "1" and near 0 V for a "0." Over a transmission link, however, it is more efficient in terms of power to encode binary data with an equivalent difference in levels but symmetrically balanced about 0 V. For example, the average power required for equally likely +3- and 0-V encodings is 4.5 W (assuming 1 Ω resistance). With +1.5- and −1.5-V encodings, the same error distance is achieved with half the power requirements (2.25 W). Communications engineers commonly refer to the unbalanced code as a unipolar code and the balanced code as a polar code. A representative sequence of binary data and its corresponding balanced and unbalanced level encoding are shown in Figure 4.5. Notice that the level of each signal is maintained for the duration of a signal interval. For this reason

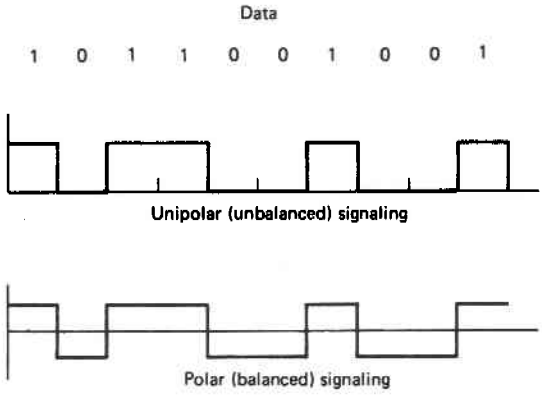


Figure 4.5 Unipolar and polar (NRZ) line codes.

Figure 4.6 Direct-current wander of NRZ signal.

the balanced (polar) encoding is also referred to as a nonreturn-to-zero (NRZ) code.*

As indicated in Figure 4.5, an NRZ signal contains no transitions for long strings of 1's or 0's. Hence one of the procedures described previously to ensure timing transitions must be employed if NRZ encoding is used on a synchronous transmission link.

An NRZ line code is a pulse transmission system wherein the pulse (before filtering) lasts for the duration of a signaling interval T. Hence the frequency spectrum (assuming random data) of an NRZ code is the $\sin(x)/x$ spectrum of Equation 4.1 and shown in Figure 4.2. As indicated, the frequency spectrum is significantly nonzero at zero frequency (dc). Most wireline transmission links, however, do not pass dc signals by virtue of their being alternating current coupled with transformers or capacitors to eliminate dc ground loops. Furthermore, some systems purposely remove dc components from the signal to allow line powering of repeaters or to facilitate single-sideband transmission. The elimination of the low-frequency components in the waveform causes long strings of 1's or 0's to decay gradually in amplitude. Hence a receiver not only would lose timing information during these strings but also would lose its amplitude reference for optimally discriminating between a "1" level and a "0" level. The effect of low-frequency cutoff, called dc wander, is shown in Figure 4.6 for a typical transmission sequence. Notice that following the long string of 1's, the output of the link is such that 1-to-0 errors are more likely than 0-to-1 errors. Similarly, long strings of 0's increase the likelihood of a 0-to-1 error. This problem arises not only for long strings of 1's or 0's, but whenever there is an imbalance in the number of 1's and 0's. Hence periodic timing pulses are not sufficient to remove dc wander.

The existence of low frequencies in a random data signal is the basic reason why modems are needed for data communications over the analog telephone network. (Analog telephone circuits do not pass direct current.) It is also the reason that NRZ coding is not often used for long-distance transmission. Direct-current wander is not unique to data transmission systems. It is a phenomenon that must be reconciled in television receivers, radar receivers, or radiation detectors.

One technique of offsetting dc wander is referred to as dc or baseline restoration [4]. As illustrated in Figure 4.7, dc restoration involves passing received pulses through a capacitor, detecting them, and then removing the charge on the capacitor before the next pulse arrives. Charge on the capacitor is removed by driving the voltage

*Some communications theorists refer to a balanced two-level code as a "bipolar code." The North American telephone industry, however, uses the term *bipolar* to refer to a three-level code described in the next section.

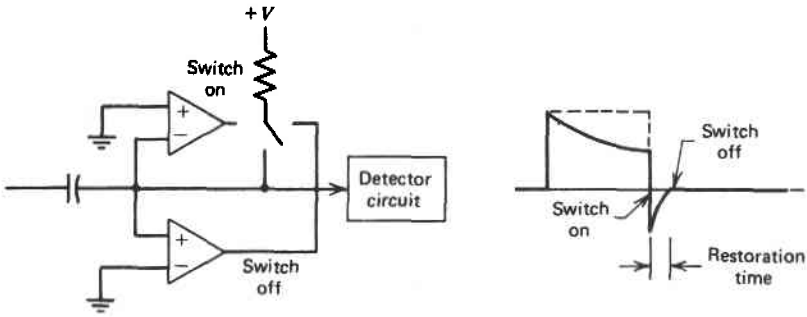


Figure 4.7 Direct-current restoration for unipolar pulses.

to a specific threshold (0 V in Figure 4.7) and then removing the driving voltage when the threshold is reached. Since all charge on the capacitor is removed after each pulse, the baseline or decision reference level is constant at the beginning of a signal interval. An obvious disadvantage of this technique is that the signal input must have zero amplitude or be disabled during the reset time.

A generally more useful technique for overcoming baseline wander is to use decision feedback, also called quantized feedback equalization [5–7]. In contrast to dc restoration, which drives the capacitor voltage to a constant, predetermined level, quantized feedback compensates for dc wander by locally generating the unreceived low-pass response and adding it to the received signal. To accomplish this, the original data stream is reconstructed. As shown in Figure 4.8, the reconstructed data stream is passed through a low-pass filter that generates a pulse equal to the tail or droop characteristic of the channel. The feedback signal adds to the received signal to eliminate the droop (intersymbol interference). Using a frequency-domain analysis, the feedback response is complimentary to the channel response. Quantized feedback is used in ISDN basic rate line interfaces [8].

4.3.2 Bipolar Coding

The dc restoration techniques mentioned in the preceding section simplify pulse detection by creating a low-pass pulse response in the receiver. There are numerous line

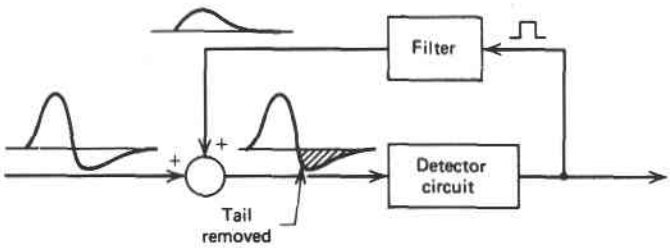


Figure 4.8 Decision feedback equalization.

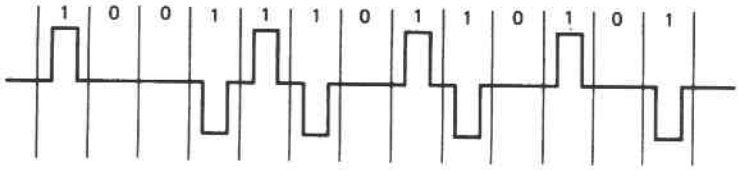


Figure 4.9 Bipolar (AMI) coding.

codes that are specifically designed to not contain dc energy and thereby be unaffected by dc removal. Bipolar coding solves the dc wander problem by using three levels to encode binary data. Specifically, a logic 0 is encoded with zero voltage while a logic 1 is alternately encoded with positive and negative voltages. Hence the average voltage level is maintained at zero—to eliminate dc components in the signal spectrum. Since bipolar coding uses alternate polarity pulses for encoding logic 1's, it is also referred to as *alternate mark inversion* (AMI).*

Bipolar coding is the basic line-coding procedure used by T1 lines in the telephone network. Rather than using full-period pulses, however, T1 lines use a 50% duty cycle pulse to encode each logic 1. Return-to-zero (RZ) pulses (Figure 4.9) were selected to minimize intersymbol interference and simplify timing recovery in the regenerative repeaters of a T1 line [9]. The power spectrum of a bipolar code is obtained from [10] as

$$S(\omega) = \frac{2p(1-p)}{T}\,|G(\omega)|^2\,\frac{1-\cos\omega T}{1-2(2p-1)\cos\omega T+(2p-1)^2} \tag{4.3}$$

where p = probability of a 1

$G(\omega)$ = spectrum of an individual pulse

$$G(\omega) = \left(\frac{T}{2}\right)\frac{\sin(\omega T/4)}{\omega T/4} \quad \text{for 50\% duty cycle pulses}$$

Equation 4.3 is plotted in Figure 4.10 for various values of p. For purely random data, $p = \frac{1}{2}$. Recall, however, that source coding for μ255 PCM codecs produces more 1's than 0's in the interest of establishing a strong clock signal. Hence the appropriate value of p for a T1 voice line is normally somewhat larger than 0.5 and depends on the amplitude of the voice signal. Low-level signals that remain in the first encoding segment produce a value of p approximately equal to 0.65. On the other hand, full-scale sine waves produce a value for p that is somewhat below 0.5 since most of the samples occur near maximum amplitude.

Because a bipolar code uses alternating polarities for encoding 1's, strings of 1's have strong timing components. However, a string of 0's contains no timing information and therefore must be precluded by the source. The specifications for T1 line re-

*A mark is a term arising from telegraphy to refer to the active, or 1, state of a level encoded transmission line.

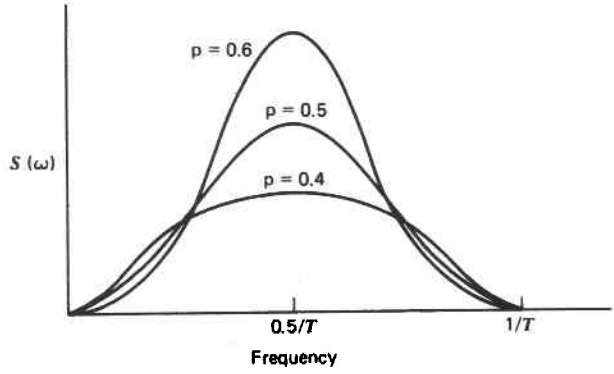


Figure 4.10 Spectral density of bipolar coding.

peaters state that the repeaters will maintain timing as long as no string of greater than fifteen 0's is allowed to occur [11]. A string of fifteen 0's can only occur if a 0 framing bit falls between a 10000000 code in time slot 24 and a 00000001 code in time slot 1.

Code Space Redundancy

In essence, bipolar coding uses a ternary code space but only two of the levels during any particular signal interval. Hence bipolar coding eliminates dc wander with an inefficient and redundant use of the code space. The redundancy in the waveform also provides other benefits. The most important additional benefit is the opportunity to monitor the quality of the line with no knowledge of the nature of the traffic being transmitted. Since pulses on the line are supposed to alternate in polarity, the detection of two successive pulses of one polarity implies an error. This error condition is known as a *bipolar violation*. No single error can occur without a bipolar violation also occurring. Hence the bipolar code inherently provides a form of line code parity. The terminals of T1 lines are designed to monitor the frequency of occurrence of bipolar violations, and if the frequency of occurrence exceeds some threshold, an alarm is set.

In T-carrier systems, bipolar violations are used merely to detect channel errors. By adding some rather sophisticated detection circuitry, the same redundancy can be used for correcting errors in addition to detecting them. Whenever a bipolar violation is detected, an error has occurred in one of the bits between and including the pulses indicating the violation. Either a pulse should be a 0 or an intervening 0 should have been a pulse of the opposite polarity. By examining the actual sample values more closely, a decision can be made as to where the error was most likely to have occurred. The bit with a sample value closest to its decision threshold is the most likely bit in error. This technique belongs to a general class of decision algorithms for redundant signals called maximum likelihood or Viterbi decoders [12]. Notice that this method of error correction requires storage of pulse amplitudes. If decision values only are stored, error correction cannot be achieved (only error detection).

An additional application of the unused code space in bipolar coding is to purposely insert bipolar violations to signify special situations such as time division mul-

tiplex framing marks, alarm conditions, or special codes to increase the timing content of the line signals. Since bipolar violations are not normally part of the source data, these special situations are easily recognized. Of course, the ability to monitor the quality of the line is compromised when bipolar violations occur for reasons other than channel errors.

4.3.3 Binary *N*-Zero Substitution

A major limitation of bipolar (AMI) coding is its dependence on a minimum density of 1's in the source code to maintain timing at the regenerative repeaters. Even when strings of 0's greater than 14 are precluded by the source, a low density of pulses on the line increases timing jitter and therefore produces higher error rates. Binary *N*-zero substitution (BNZS) [13] augments a basic bipolar code by replacing all strings of *N* 0's with a special *N*-length code containing several pulses that purposely produce bipolar violations. Thus the density of pulses is increased while the original data are obtained by recognizing the bipolar violation codes and replacing them at the receiving terminal with *N* 0's.

As an example, a three-zero substitution algorithm (B3ZS) is described. This particular substitution algorithm is specified for the standard DS-3 signal interface in North America [14]. It was also used in the LD-4 coaxial transmission system in Canada [15].

In the B3ZS format, each string of three 0's in the source data is encoded with either 00V or B0V. A 00V line code consists of 2-bit intervals with no pulse (00) followed by a pulse representing a bipolar violation (V). A B0V line code consists of a single pulse in keeping with the bipolar alternation (B), followed by no pulse (0), and ending with a pulse with a violation (V). With either substitution, the bipolar violation occurs in the last bit position of the three 0's replaced by the special code. Thus the position of the substitution is easily identified.

The decision to substitute with 00V or B0V is made so that the number of B pulses (unviolated pulses) between violations (V) is odd. Hence if an odd number of 1's has been transmitted since the last substitution, 00V is chosen to replace three 0's. If the intervening number of 1's is even, B0V is chosen. In this manner all purposeful violations contain an odd number of intervening bipolar pulses. Also, bipolar violations alternate in polarity so that dc wander is prevented. An even number of bipolar pulses between violations occurs only as result of a channel error. Furthermore, every purposeful violation is immediately preceded by a 0. Hence considerable systematic redundancy remains in the line code to facilitate performance monitoring. Table 4.1 summarizes the substitution algorithm.

Example 4.1. Determine the B3ZS line code for the following data sequence: 10100011000000010001. Use + to indicate a positive pulse, – to indicate a negative pulse, and 0 to indicate no pulse.

TABLE 4.1 B3ZS Substitution Rules

Polarity of Preceding Pulse	Number of Bipolar Pulses (1's) Since Last Substitution: Odd	Even
−	00−	+0+
+	00+	−0−

Solution. There are two possible sequences depending on whether an odd or even number of pulses has been transmitted following the previous violation:

Substitutions													
	101		$\overbrace{000}$		11		$\overbrace{000}$		$\overbrace{000}$		001		$\overbrace{000}$		1
Case 1 (odd):	+0−		00−		+−		+0+		−0−		00+		00+		−
Case 2 (even):	+0−		+0+		−+		−0−		+0+		00−		00−		+
Violations															

Example 4.1 indicates that the process of breaking up strings of 0's by substituting with bipolar violations greatly increases the minimum density of pulses in the line code. In fact, the minimum density is 33% while the average density is just over 60%. Hence the B3ZS format provides a continuously strong timing component. Notice that all BNZS coding algorithms guarantee continuous timing information with no restrictions on source data. Hence BNZS coding supports any application in a completely transparent manner.

Another BNZS coding algorithm is the B6ZS algorithm used on obsolete T2 transmission lines [16]. The B6ZS algorithm is defined in Table 4.2. This algorithm produces bipolar violations in the second and fifth bit positions of the substituted sequence.

ITU recommends another BNZS coding format referred to as high-density bipolar (HDB) coding [17]. As implemented in the E1 primary digital signal, HDB coding replaces strings of four 0's with sequences containing a bipolar violation in the last bit position. Since this coding format precludes strings of 0's greater than three, it is referred to as HDB3 coding. The encoding algorithm is basically the same as the B3ZS algorithm described earlier. Table 4.3 presents the basic algorithm. Notice that substitutions produce violations only in the fourth bit position, and successive substitutions produce violations with alternating polarities.

TABLE 4.2 B6ZS Substitution Rules

Polarity of Pulse Immediately Preceding Six 0's to be Substituted	Substitution
−	0 − + 0 + −
+	0 + − 0 − +

Example:

1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1

+ − (0 − + 0 + −) + 0 − + (0 + − 0 − +) (0 + − 0 − +) 0 0 0 −

− + (0 + − 0 − +) − 0 + − (0 − + 0 + −) (0 − + 0 + −) 0 0 0 +

A fundamental feature of end-to-end digital connectivity as provided by ISDN is 64-kbps transparent channels referred to as clear-channel capability (CCC) [18]. Two aspects of a bipolar/AMI line code as used on T1 lines preclude CCC: robbed signaling in the least significant bit of every sixth frame and the need to avoid all-0's codewords on the channel. Bit robbing for signaling is avoided with common-channel signaling (also an inherent requirement for ISDN deployment). Two means of augmenting T1 lines to allow transparent channels have been developed. The first method is zero-byte time slot interchange (ZBTSI) described in reference [19]. ZBTSI was used for upgrading some T1 facilities for CCC but is not a good long-term solution.

The second method, which is most desirable in the long run, involves the use of a B8ZS line code for T1 lines. As such, the B8ZS line code became the new line code standard for T1 lines with the introduction of D5 channel banks. (D5 channel banks can also operate with bipolar/AMI line codes for backward compatibility.) As indicated in Table 4.4, the B8ZS algorithm is similar to the B6ZS algorithm in that each individual substitution is dc balanced. Notice that there are two bipolar violations, in positions 4 and 7, in every substitution. The purposeful introduction of bipolar violations requires replacement of any equipment that monitors all violations for performance monitoring.

TABLE 4.3 HDB3 Substitution Rules

	Number of Bipolar Pulses (1's) Since Last Substitution	
Polarity of Preceding Pulse	Odd	Even
−	000−	+00+
+	000+	−00−

TABLE 4.4 BBZS Substitution Rules

Polarity of Preceding Pulse	Substitution
–	000–+0+–
+	000+–0–+

4.3.4 Pair Selected Ternary

The BNZS substitution algorithms described in the preceding section are examples of selecting codes in the ternary code space to increase the timing content of a binary signal. Pair selected ternary (PST) [20] is another example.

The PST coding process begins by pairing the binary input data to produce sequences of 2-bit codewords. These codewords are then translated into two ternary digits for transmission. Since there are nine two-digit ternary codes but only four two-digit binary codes, there is considerable flexibility available in selecting the codes. The most useful of the possible coding formats is shown in Table 4.5. This particular format not only ensures a strong timing component but also prevents dc wander by switching modes to maintain a balance between positive and negative pulses. The codes are selected from one column until a single pulse is transmitted. At that time the encoder switches modes and selects codes from the other column until another single pulse (of the opposite polarity) is transmitted.

Example 4.2. Encode the following binary data stream into a PST line code: 01001110101100.

Solution. There are two possible solutions depending on whether the encoder is in the positive or negative mode at the beginning of the sequence:

	01	00	11	10	10	11	00
Case 1 (+ mode):	0+	–+	+–	–0	+0	+–	–+
Case 2 (– mode):	–+	–+	+–	+0	–0	+–	–+

One potential drawback of the PST coding algorithm is that the binary data stream must be framed into pairs. Hence a PST decoder must recognize and maintain pair

TABLE 4.5 Pair Selected Ternary Encoding

Binary Input	+ Mode	– Mode
00	–+	–+
01	0+	0–
10	+0	–0
11	+–	+–

boundaries. Recognition of the boundaries is not difficult if random data are being transmitted since a pairwise misframe eventually produces unallowed codes (00, + +, −−). Furthermore, time division multiplex formats typically provide character and pairwise framing automatically.

The power spectrum of a PST line code with equal probabilities for 1's and 0's is obtained from reference [20] and plotted in Figure 4.11. Also shown is the B6ZS power spectrum [16] and the conventional bipolar power spectrum.

An important point to notice in Figure 4.11 is that bipolar coding and its extensions require equal bandwidths. Their only significant difference is that B6ZS and PST have higher energy levels as a result of greater pulse densities. The higher energy levels have the undesirable effect of increasing crosstalk interference in multipair cables. However, the degradation from the increased crosstalk is somewhat offset by improved accuracy of the recovered sample clock (assuming all signals in the cable are using the same line code).

4.3.5 Ternary Coding

Since bipolar and PST coding use a ternary code space to transmit binary data, they do not achieve as high an information rate as is possible with more efficient use of the code space. For example, an eight-element ternary code is capable of representing $3^8 = 6561$ different codes. In contrast, 8 bits of binary data produce only $2^8 = 256$ different codes. The previously described line codes do not take advantage of the higher information content of ternary codes; they select codes for their timing content and spectral properties.

One ternary encoding procedure involves mapping successive groups of 4 bits into three ternary (4B3T) digits. Since binary words of 4 bits require only 16 of the 27 possible three-digit ternary codewords, considerable flexibility exists in selecting the ternary codes. Table 4.6 presents one possible encoding procedure. Ternary words in the middle column are balanced in their dc content. Codewords from the first and third

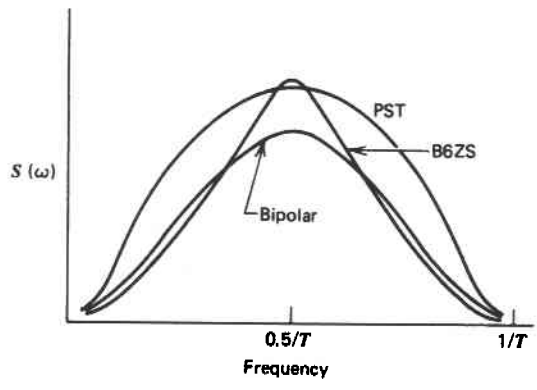


Figure 4.11 Spectrum of bipolar, B3ZS, and PST line codes for equally likely 1's and 0's.

TABLE 4.6 Encoding Table for 4B3T Line Code

	Ternary Word (Accumulated Disparity)		
Binary Word	−	0	+
0000	− − −		+ + +
0001	− − 0		+ + 0
0010		− 0 −	+ 0 +
0011		0 − −	0 + +
0100		− − +	+ + −
0101	− + −		+ − +
0110	+ − −		− + +
0111	− 0 0		+ 0 0
1000	0 − 0		0 + 0
1001	0 0 −		0 0 +
1010		0 + −	
1011		0 − +	
1100		+ 0 −	
1101		− 0 +	
1110		+ − 0	
1111		− + 0	

columns are selected alternately to maintain dc balance. If more positive pulses than negative pulses have been transmitted, column 1 is selected. When the disparity between positive and negative pulses changes, column 3 is chosen. Notice that the all-0's codeword is not used. Hence a strong timing content is maintained. Because of the higher information efficiency, however, the ability to monitor performance is sacrificed, and framing is required on three-digit boundaries. 4B3T coding is used on the T148 span line developed by ITT Telecommunications [21]. This system provides T-carrier transmission for two DS1 signals (48 channels) using a bandwidth that is only 50% greater than a T1 bandwidth (carrying 24 channels). A generalized discussion of 4B3T coding and other ternary coding techniques is contained in reference [22].

4.3.6 Digital Biphase

Bipolar coding and its extensions BNZS and PST use extra encoding levels for flexibility in achieving desirable features such as timing transitions, no dc wander, and performance monitorability. These features are obtained by increasing the code space and not by increasing the bandwidth. (The first spectral null of all codes discussed so far, including an NRZ code, is located at the signaling rate $1/T$.)

Many varieties of line codes achieve strong timing and no dc wander by increasing the bandwidth of the signal while using only two levels for binary data. One of the most common of these codes providing both a strong timing component and no dc wander is the digital biphase code, also referred to as "diphase" or a "Manchester" code.

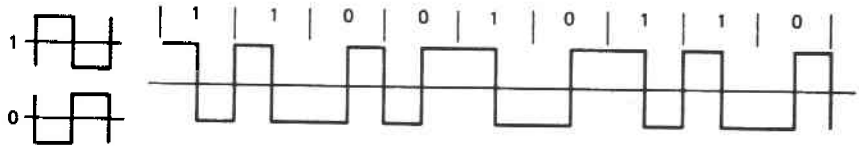


Figure 4.12 Digital biphase (Manchester) line code.

A digital biphase code uses one cycle of a square wave at a particular phase to encode a 1 and one cycle of an opposite phase to encode a 0. An example of a digital biphase coding sequence is shown in Figure 4.12. Notice that a transition exists at the center of every signaling interval. Hence strong timing components are present in the spectrum. Furthermore, logic 0 signals and logic 1 signals both contain equal amounts of positive and negative polarities. Thus dc wander is nonexistent. A digital biphase code, however, does not contain redundancy for performance monitoring. If in-service performance monitoring is desired, either parity bits must be inserted into the data stream or pulse quality must be monitored. (A later section of this chapter discusses performance monitoring in more detail.)

The frequency spectrum of a digital biphase signal is derived in Appendix C and plotted in Figure 4.13, where it can be compared to the spectrum of an NRZ signal. Notice that a digital biphase signal has its first spectral null at $2/T$. Hence the extra timing transitions and elimination of dc wander come at the expense of a higher frequency signal. In comparison to three-level bipolar codes, however, the digital biphase code has a lower error rate for equal signal-to-noise ratios (see Appendix C).

Examination of the frequency spectra in Figure 4.13 shows that the diphase spectrum is similar to an NRZ spectrum but translated so it is centered about $1/T$ instead of direct current. Hence digital biphase actually represents digital modulation of a square wave carrier with one cycle per signal interval. Logic 1's cause the square wave to be multiplied by +1 while logic 0's produce multiplication by −1. Diphase is pri-

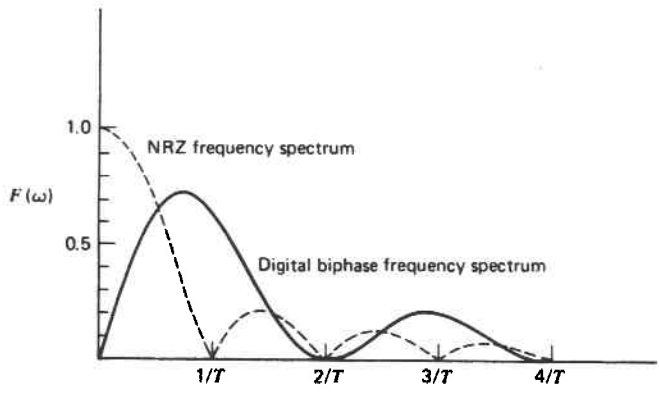


Figure 4.13 Power spectral density of digital biphase.

marily used on shorter links where terminal costs are more significant than bandwidth utilization. The "Ethernet" IEEE 802.3 local area data network uses digital biphase (Manchester) coding.

4.3.7 Differential Encoding

One limitation of NRZ and digital biphase signals, as presented up to this point, is that the signal for a 1 is exactly the negative of a signal for a 0. On many transmission media, it may be impossible to determine an absolute polarity or an absolute phase reference. Hence the decoder may decode all 1's as 0's and vice versa. A common remedy for this ambiguity is to use differential encoding that encodes a 1 as a change of state and encodes a 0 as no change in state. In this manner no absolute reference is necessary to decode the signal. The decoder merely detects the state of each signal interval and compares it to the state of the previous interval. If a change occurred, a 1 is decoded. Otherwise, a 0 is determined.

Differential encoding and decoding do not change the encoded spectrum of purely random data (equally likely and uncorrelated 1's and 0's) but do double the error rate. If the detector makes an error in estimating the state of one interval, it also makes an error in the next interval. An example of a differentially encoded NRZ code and a differentially encoded diphase signal is shown in Figure 4.14. All signals of differentially encoded diphase retain a transition at the middle of an interval, but only the 0's have a transition at the beginning of an interval.

4.3.8 Coded Mark Inversion

A variety of line codes have evolved similar to the digital biphase code described previously. One of these is referred to as coded mark inversion (CMI) in ITU recommendations [17]. CMI encodes 1's (marks) as an NRZ level opposite to the level of the

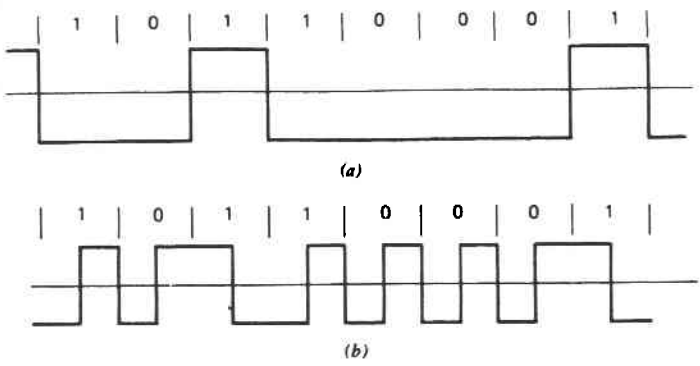


Figure 4.14 Differentially encoded NRZ and diphase signals: (*a*) differentially encoded NRZ; (*b*) differentially encoded diphase.

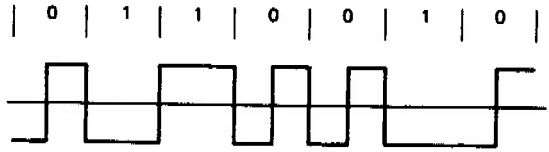


Figure 4.15 Coded mark inversion.

previous one and 0's as a half-cycle square wave of one particular phase. Figure 4.15 shows a sample encoding for CMI. There is no dc energy in the signal and an abundance of signal transitions exist as in diphase. Furthermore, there is no ambiguity between 1's and 0's. Elimination of the ambiguity actually leads to a major drawback of CMI coding: Its error performance is 3 dB worse than diphase.* CMI is the specified interface code for the fourth-level (E4) ITU multiplex signal at 139.264 Mbps. CMI is also used for electrical interfaces of SONET STS-3c signals described in Chapter 8.

4.3.9 Multilevel Signaling

In the line codes discussed so far, two-level (binary) signaling has been assumed. In applications where the bandwidth is limited but higher data rates are desired, the number of levels can be increased while maintaining the same signaling rate. The data rate R achieved by a multilevel system is given as

$$R = \log_2(L)\left(\frac{1}{T}\right) \tag{4.4}$$

where L = number of levels that can be freely chosen during each interval
T = signaling interval

The signaling rate $1/T$ is often referred to as the symbol rate and is measured in bauds. Within the data communications industry it is common practice to use "baud" as being synonymous with bit rate. Strictly speaking, however, the bit rate is only equal to the baud rate if binary signaling (1 bit per signal interval) is used. Figure 4.16 shows an example of an eight-level transmission format that achieves 3 bits per signal interval (i.e., 3 bits per baud).

Multilevel transmission systems achieve greater data rates within a given bandwidth but require much greater signal-to-noise ratios for a given error rate. One aspect of wireline transmission that favors multilevel line coding is the lower baud rate for a given data rate, which in turn reduces the crosstalk. Hence, in crosstalk-limited sys-

*The error performance of CMI is 3 dB worse than diphase when bit-by-bit detection is used. The inefficiency arises because for one-half of an interval a 1 looks like a 0. Because CMI has redundancy, some of the inefficiency can be recovered with maximum-likelihood (Viterbi) detection.

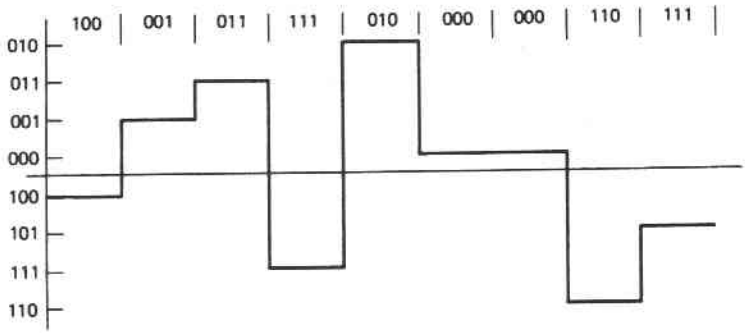


Figure 4.16 Multilevel transmission with 3 bits per signal interval.

tems the signal-to-noise ratio penalty of a multilevel line code is not as significant. The T1G T-carrier system developed by AT&T [23] uses a four-level line code at the T1C baud rate (3.152 MHz) to double the capacity of a T1C system (from 48 channels to 96).

Another example of multilevel transmission of particular significance is the ISDN basic rate digital subscriber line (DSL), which uses four-level transmission at a signaling rate of 80 kilobauds to achieve 160 kbps. The primary factors that led to selecting a multilevel line code in this application are (1) near-end crosstalk that cannot be eliminated by pair isolation* as in T1 systems and (2) high levels of intersymbol interference caused by bridged tap reflections. Both of these factors are easier to control when lower frequency signals are used [24]. Additional aspects of this application are described in the ISDN section of Chapter 11.

4.3.10 Partial-Response Signaling

Conventional bandlimiting filters of a digital transmission system are designed to restrict the signal bandwidth as much as possible without spreading individual symbols so they interfere with sample values of adjacent intervals. One class of signaling techniques, variously referred to as duobinary [25], correlative level encoding [26], or partial-response signaling [27], purposely introduces a prescribed amount of intersymbol interference that is accounted for in the detection circuitry of the receivers. By overfiltering the encoded signal, the bandwidth is reduced for a given signaling rate, but the overlapping pulses produce multiple levels that complicate the detection process and increase the signal power requirements for a given error rate.

Figure 4.17 shows the pulse response of a typical partial-response system. If the channel is excited by a pulse of duration T, channel filters (defined in Appendix C) limit the spectrum such that the main part of the pulse extends across three signal intervals and contributes equally to two sample times. The reason for the term *partial response* is now apparent: The output only responds to one-half the amplitude of the input.

*Pair isolation involves separating go and return pairs into separate cables.

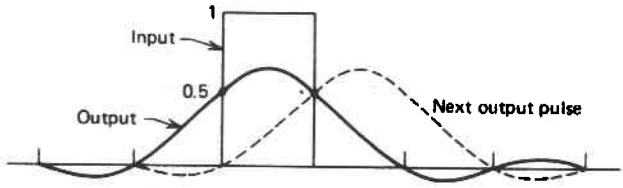


Figure 4.17 Output pulse of partial-response channel.

If the input pulse of Figure 4.17 is followed by another pulse of the same amplitude, the output will reach full amplitude by virtue of the overlap between pulses. However, if the next input pulse has a negative polarity, the overlap produces zero amplitude at the sample time. Thus, as shown in Figure 4.18, a partial-response system with two-level inputs (+1, −1) produces an output with three levels (+1 , 0, −1). In similar fashion, Figure 4.19 shows a system with four input levels (+3, +1, −1, −3) and seven output levels (+3, +2, +1, 0, −1, −2, −3).

Detection of a partial response signal (PRS) is complicated by the additional levels produced by the channel filters and the fact that sample values are dependent on two adjacent pulse amplitudes. One method of detecting a PRS involves subtracting the overlap of a previously detected pulse from the incoming signal to generate a difference signal representing the channel response to the new pulse. Only the overlap (intersymbol interference) at the sample times needs to be subtracted. The difference samples are nominally equal to one-half the amplitude of the unknown input pulse. This technique doubles the error rate in the same manner that differentially encoded error rates are doubled.

Another method of detecting partial-response systems involves a technique called *precoding* at the source. Precoding transforms the input data in such a manner that the output level at the detector directly indicates the original data without comparison to the previous sample value. In a binary system, for example, a 1 is encoded with a pulse of the same polarity as the previous pulse. Hence logic 1's show up at the detector as either a +1 or a −1 sample. Transmission of a 0 is encoded with a pulse of opposite polarity to the previous pulse. Hence logic 0's always show up at the detector as a zero-level signal. Similar precoding techniques exist for multilevel systems [28].

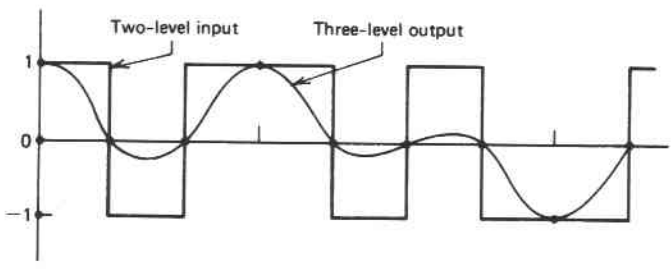


Figure 4.18 Three-level partial-response inputs and outputs.

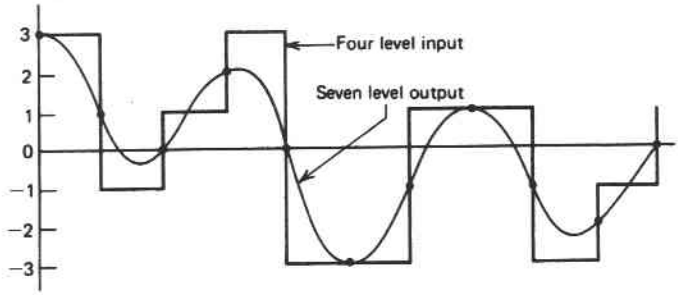


Figure 4.19 Seven-level partial-response inputs and outputs.

The partial-response technique just described is actually a special case of a more general class of signaling techniques referred to as *correlative level encoding* [26]. To describe the properties of generalized partial-response or correlative level systems, it is convenient to introduce a delay operator D to denote delay equal to one signal interval T. Physically, D can be thought of as a delay line of length T. Two units of delay is implemented with two delay lines in series and is denoted as D^2.

Using this notation, the partial-response system described above is referred to as $1+D$ PRS: The output represents the superposition of the input with a delayed version of the same input. Other forms of overlap are possible. These systems do not necessarily produce the overlapping pulses by overfiltering an input. An alternative approach is to overlap and add the pulses directly in the encoding process (hence the term correlative level encoding). An interesting special case of correlative level encoding is the $1-D$ system shown in Figure 4.20 and extended in Figure 4.21 to show the effect of overlapping pulses.

The $1-D$ encoder uses a single cycle of a square wave across two signal intervals to encode each bit. Since neither of the two individual signals (+− or −+) produce dc energy, the encoded signal has no dc wander. Furthermore, the positive and negative levels of the composite signal alternate in a manner reminiscent of bipolar coding. In fact, if differential encoding is used (i.e., if a 0 is encoded with the same phase

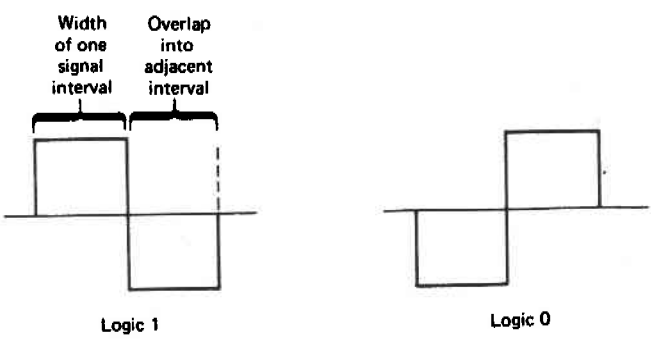


Figure 4.20 Individual signal elements for $1-D$ correlative level coding.

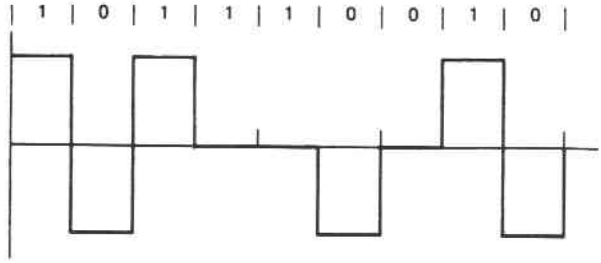


Figure 4.21 Representative waveform of $1 - D$ correlative level encoding.

as the previous interval and a 1 as the opposite phase), this form of encoding is identical to bipolar coding (assuming the NRZ levels are replaced by 50% RZ pulses). Thus $1 - D$ correlative level encoding is used to shape the bandwidth rather than to limit it. The spectra of unfiltered $1 + D$, $1 - D$, and $1 - D^2$ signals are obtained from reference [27] and plotted in Figure 4.22. The T1D T-carrier system of AT&T uses precoded 1 + D level encoding (also called duobinary).

The spectrum of the $1 - D^2$ signal is particularly interesting. It has no dc component and an upper limit equal to $1/2T$: the same upper limit as a maximally filtered NRZ code with no intersymbol interference. The $1 - D^2$ partial-response systems have been

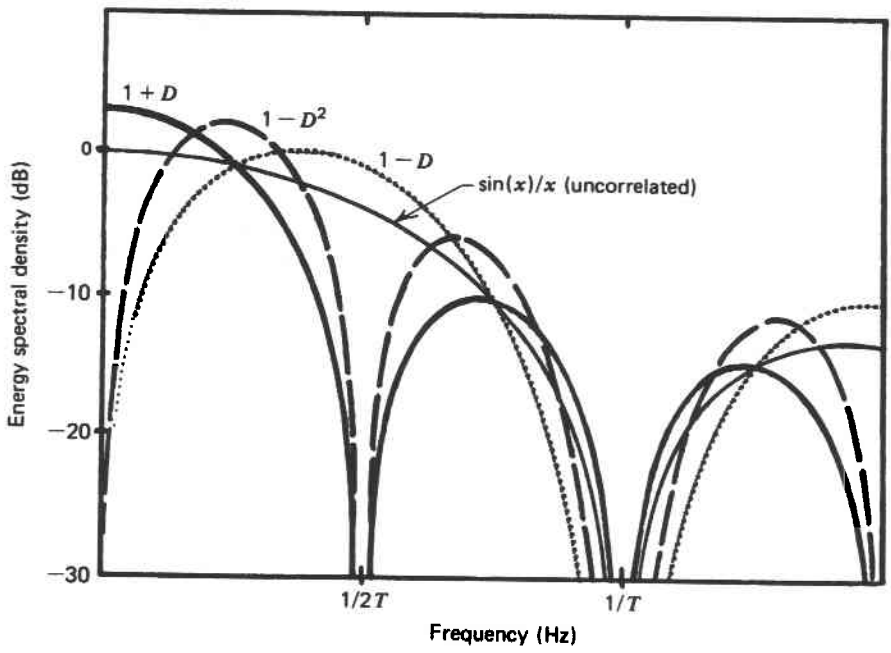


Figure 4.22 Spectra of unfiltered $1 + D$, $1 - D$, and $1 - D^2$ correlative encoded signals.

used by GTE in a modified T-carrier system providing 48 voice channels [29] and for digital transmission over analog microwave radios using very low baseband frequencies for pilots and service channels (where no message spectrum exists) [30]. The obsolete 1A-RDS (data-under-voice) system of AT&T also used $1 - D^2$ coding with four-level inputs to get 1.544 Mbps into 500 kHz of bandwidth [31].

Notice that the $1 - D^2$ spectrum is essentially the product of the $1 + D$ and the $1 - D$ spectra. Indeed, a $1 - D^2$ system can be implemented by concatenating a $1 - D$ encoder with a $1 + D$ channel response: $(1 - D)(1 + D) = 1 - D^2$. Thus correlative level system polynomials are very useful for simultaneously representing correlations and spectrum shaping. Since all correlative level systems use more levels than necessary for encoding the data, they are inferior to uncorrelated or full-response systems in terms of error performance.

A summary of digital transmission systems identifying various line codes is provided in Table 4.7. Other line codes used specifically in fiber systems are described in Chapter 8.

4.4 ERROR PERFORMANCE

The preceding sections of this chapter emphasized the timing and power spectrum requirements of various transmission codes. Another fundamental consideration in choosing a line code is the error rate performance in the presence of noise. Except on relatively short lines, where noise may be insignificant, error performance requirements can impact system cost significantly. If a certain minimum error rate is specified, those coding schemes providing the desired error rate at lower signal-to-noise ratios allow regenerative repeaters to be spaced farther apart, thereby reducing instal-

TABLE 4.7 Digital Transmission Systems

Designation	Country or Administration	Bit Rate (Mbps)	Line Code	Media	Repeater Spacing
T1	AT&T	1.544	AMI/B8ZS	Twisted pair	6000 ft
E1	ITU-T	2.048	HDB3(B4ZS)	Twisted pair	2000 m
T1C	AT&T	3.152	Bipolar	Twisted pair	6000 ft
T148	ITT	2.37, ternary	4B3T	Twisted pair	6000 ft
9148A	GTE	3.152	$1 - D^2$, duobinary	Twisted pair	6000 ft
T1D	AT&T	3.152	$1 + D$, duobinary	Twisted pair	6000 ft
T1G	AT&T	6.443	Four-level	Twisted pair	6000 ft
T2	AT&T	6.312	B6ZS	Low-capacitance twisted pair	4800 ft
LD-4	Canada	274.176	B3ZS	Coax	1900 m
T4M	AT&T	274.176	Polar	Coax binary (NRZ)	5700 ft

lation and maintenance.* Repeater spacing is an important economic factor in wireline transmission, in optical fiber transmission, and on point-to-point radio links.

The error rate results and comparisons presented in this section are based on white Gaussian noise. This is the most common form of noise and the best understood. In some applications, as within the older electromechanical telephone network, impulse noise may be more prevalent. Thus these analyses do not provide a complete error rate analysis for some applications. The intent of this section is to present relative error performance comparisons of various line-coding techniques. For this purpose a white noise analysis is most appropriate. If impulses are large enough, they cause errors independent of the coding scheme in use.

For the most part, the following sections present only the results of the error analyses in the form of graphs of error rate as a function of signal-to-noise ratios. Appendix C derives the basic equations used to produce the results.

4.4.1 Signal Detection

Invariably, the detection circuitry of a digital receiver processes incoming signal waveforms to measure each possible discrete signal. In most cases the measures are nothing more than samples of a filtered receive signal. Depending on the signal shape and the level of performance desired, the receiver uses more sophisticated processing. In any case, the end measurement of a binary signal nominally produces one voltage level for a 0 and another voltage level for a 1. A decision of which signal was transmitted is made by comparing the measurement (at the appropriate time) to a threshold located halfway between these nominal voltages. Naturally, the error probability depends on the nominal distance between the voltages and the amount of fluctuation in the measurements caused by noise.

Since signal measurements are normally linear in nature, the error distance between 1's and 0's is proportional to the received signal voltage. Equivalently, the amount of noise power required to produce an error is a direct function of the signal-to-noise ratio. The optimum detector for a particular signal set maximizes the signal-to-noise ratio at the input to the decision circuit.

4.4.2 Noise Power

White Gaussian noise is characterized as having a uniform frequency spectrum across an arbitrarily large bandwidth and an amplitude distribution that varies according to a normal (Gaussian) probability distribution. A parameter N_0 conventionally represents the power spectral density of white noise and is the amount of power measured in a bandwidth of 1 Hz. Hence the rms power of white noise coming out of a filter with a bandwidth BW is $(N_0)(\text{BW})$.†

*On T1 lines the repeater spacing was predetermined by the locations where loading coils needed to be removed. Thus the error rate performance essentially determined the maximum data rate.

†The power spectral density of white noise is also specified as $\frac{1}{2}N_0$ for a two-sided spectral density. As a practical matter, there is no difference in the specifications, since a real filter has the mathematical equivalent of identical positive- and negative-frequency bands. Thus the measured power coming through a filter with a one-sided (positive-frequency) bandwidth BW is N_0 BW in either case.

To determine the amount of noise power present at a decision circuit, it is necessary to determine N_0 and the effective bandwidth of the detection circuitry. When the detection circuitry consists of nothing more than a filter, the effective bandwidth is usually very close to the 3 dB bandwidth of the filter. With other more sophisticated detectors, the effective bandwidth can be more difficult to determine. As derived in Appendix C, the effective bandwidth is usually referred to as the effective noise bandwidth (NBW) of the receiver. Hence the noise power at the decision circuit is (N_0) (NBW).

4.4.3 Error Probabilities

An error in detection occurs whenever noise causes the signal measurement to cross the threshold between the two nominal output levels of the detection circuitry. Appendix C derives this probability for white Gaussian noise as

$$\text{prob(error)} = \frac{1}{\sqrt{2\pi}\sigma} \int_{v}^{\infty} e^{-t^2/2\sigma^2}\, dt \tag{4.5}$$

where $v =$ nominal distance (voltage) to a decision threshold
$\sigma^2 =$ noise power at detector, $= (N_0)(\text{NBW})$

Equation 4.5 is nothing more than the area under the probability density function of a normal distribution. As shown in Figure 4.23, the equation represents the error probability as the probability of exceeding v/σ standard deviations in a normal distribution with zero mean and unit variance $N(1, 0)$.

The error rate is completely determined by the ratio of v to σ. Since v is the noise-free sample voltage and σ^2 is the rms noise power, v^2/σ^2 is a signal-power-to-noise-power ratio at the detector. This ratio is sometimes referred to as a postdetection SNR, since it is measured after the detection circuitry. It is usually more important to express error rates in terms of an SNR at the input to the receiver. Figure 4.24 depicts a basic channel, a basic detection model, and the relation between a predetection SNR and a postdetection SNR. For reasons discussed in Appendix C, the most appropriate predetection SNR for comparing line codes and digital modulation formats is an energy-

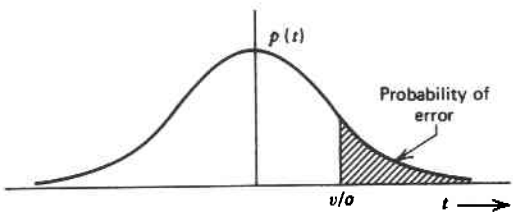


Figure 4.23 Probability of error for binary signaling.

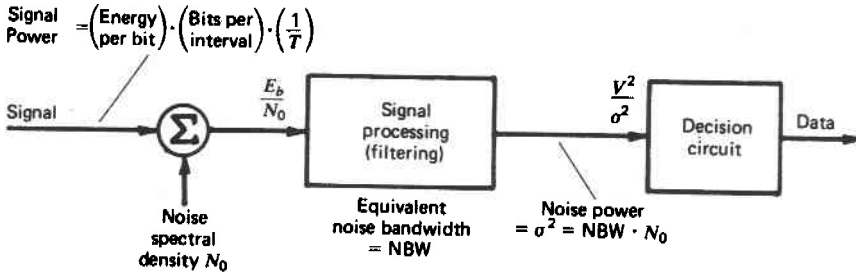


Figure 4.24 Signal detection model.

per-bit-to-noise-density ratio E_b/N_0. The relationship between E_b/N_0 and the signal-power-to-noise-power ratio is

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}}$$

$$= \frac{dE_s(1/T)}{N_0\,\text{NBW}}$$

$$= \frac{dE_b \log_2 L\,(1/T)}{N_0\,\text{NBW}} \tag{4.6}$$

where d = pulse density
E_s = energy per symbol
E_b = energy per bit
$\log_2 L$ = number of bits per symbol (i.e., L = number of levels)
$1/T$ = signaling rate
NBW = effective noise bandwidth of receiver

In determining the signal power in Equation 4.6, notice the dependence on the pulse density d. In a balanced NRZ line code the pulse density is 1, but in many of the other line codes the pulse density is dependent on the data and any substitution algorithms that may be in use. In these codes, increasing the pulse density increases the SNR but does not reduce the error rate. The error rate is determined by the energy-per-bit-to-noise-density ratio. In fact, increasing the pulse density might worsen the error rate if interference between cable pairs is a significant consideration. (Interference is directly proportional to the signal power.)

Antipodal Signaling

The optimum signaling format for binary signaling maximizes the error distance for a given received signal power and simultaneously minimizes the noise bandwidth. This condition arises only when two signal levels are allowed and only when one sig-

nal is the exact negative of the other. Since the signal for a 1 is the exact opposite (the antipode) of the signal for a 0, optimum signaling is often referred to as *antipodal signaling*. Since no other bit-by-bit signaling scheme is any better, antipodal performance is often used as a basis for comparisons. Of the line codes described previously, only balanced two-level (NRZ) encoding and digital biphase can provide antipodal performance. Figure 4.25 shows the optimum error performance provided by antipodal signaling as a function of E_b/N_0 and the SNR.

Error Rate of Level Encoded Signals

As already mentioned, a balanced, two-level line code is capable of providing optimum error rate performance. If an unsymmetric level code is used, such as unipolar, the same basic detector is used as in symmetric level encoding. The only difference is that the decision threshold must be moved from zero to half the amplitude of the "on" signal. To maintain the same error distance, the average transmit power is increased by a factor of 2. Hence a unipolar code carries a performance penalty of 3 dB with respect to antipodal performance.

Figure 4.25 shows the ideal performance of a unipolar (on–off) code where it is compared to antipodal performance. Notice that for all SNRs the error rate of the on–off system is exactly equal to the error rate of an antipodal system with 3 dB less SNR.

Bipolar Signaling

With respect to error performance, bipolar signaling is basically identical to a unipolar code. During any particular signal interval, the receiver must decide between one of two possible levels: zero or a pulse with the appropriate polarity. Hence the decision threshold pertinent to a particular signal interval lies halfway between zero and the amplitude level of the allowed pulse. Narrowing the pulse to 50% of the signal interval does not change the theoretical error performance (with respect to average pulse energy). Thus the error rate curve for on–off keying in Figure 4.25 can be used to determine theoretical bipolar error rates.

One consideration in a bipolar line code contributes to a slightly higher error rate than in on–off keying. This increase occurs because both positive and negative noise can cause an erroneous threshold crossing when a zero-level signal is transmitted. In contrast, a unipolar code is affected only by positive noise when a lower level signal is transmitted and only by negative noise when the upper level signal is transmitted. If the bipolar detector treats the erroneous pulse as a 1 (despite a bipolar violation), the error probability when 0's are transmitted is doubled. Thus the overall error probability is increased by 50% if 0's and 1's are equally likely.

Because of the steepness of the curve, an increase in the error rate of 50% does not represent much of a performance penalty. For example, if the error rate is increased from 1×10^{-6} to 1.5×10^{-6}, the source power of a line code needs to increase by only 0.2 dB to get back to a 1×10^{-6} error rate. At higher error rates, a larger penalty occurs because the curve is not as steep. These effects are demonstrated in Figure 4.25, where the ideal performance of a bipolar code (with 50% 0's) can be compared to the performance of a unipolar code.

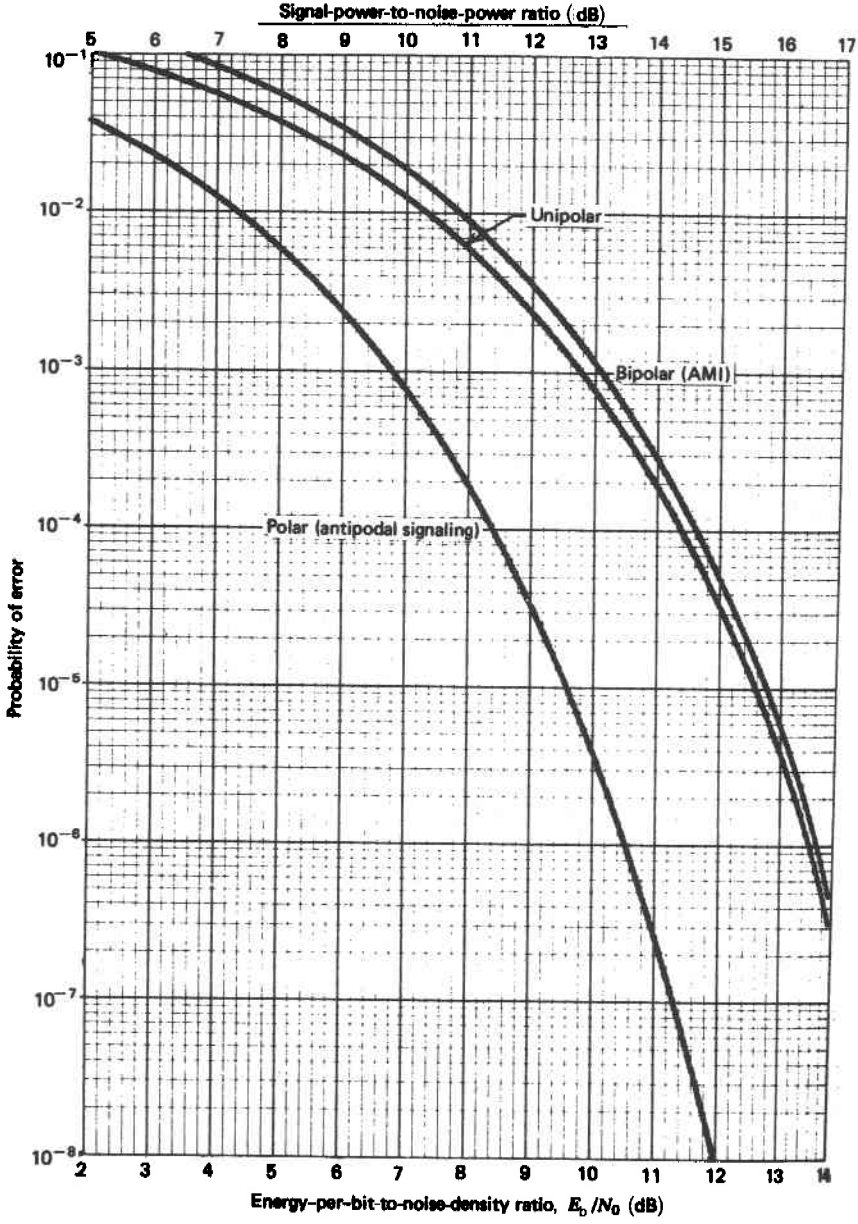


Figure 4.25 Error rates of polar (NRZ), unipolar, and bipolar line codes.

The fact that bipolar coding incurs a 3.2-dB penalty (at 10^{-6} error rate) with respect to digital biphase is indicative that timing and dc wander problems are solved by increasing the number of signal levels. In contrast, digital biphase incurs a bandwidth penalty. Not all of the 3.2-dB penalty for bipolar coding, and its extensions BNZS and PST, can be attributed to removing direct current and adding a consistent timing pattern. Bipolar coding contains considerable redundancy for performance monitoring or possible error correction.*

Example 4.3. Assume each section of a T1 transmission system on 22-gauge cable is near-end crosstalk limited and operating with a 10^{-6} error rate. What design changes are needed to reduce the error rate to 10^{-8}?

Solution. Since the system is crosstalk limited, the error rate cannot be improved by increasing the power out of the regenerative repeaters. The solution is to space the repeaters closer together (ignore the impractical aspects of this solution). From Figure 4.25 it can be determined that the signal power must be increased by 1.6 dB to improve the error rate from 10^{-6} to 10^{-8}. In Figure 4.10 it is seen that the main lobe of the bipolar spectrum extends up to 1.544 MHz. However, most of the energy in the spectrum lies below 1 MHz. Using Figure 1.14, we determine that the attenuation of 22-gauge cable is 5 dB/kft at 1 MHz. Hence 1.6/5 = .32 kft or 320 ft is the reduction in repeater spacing required.

Example 4.3 demonstrates a number of important aspects of digital transmission systems. First, arbitrarily good transmission quality can be obtained with only a small penalty in transmit power or repeater spacing.† Thus, as mentioned in Chapter 2, a digital transmission and switching network can be readily designed to impart no degradation to voice quality, except where analog-to-digital and digital-to-analog conversions take place.

Second, the dramatic improvement in error rate for a relatively small increase in the SNR implies extreme sensitivity in the opposite direction also. A slight increase in noise power or signal attenuation would cause a large increase in the error rate. Hence the nominal design of a digital link often provides considerably better performance than normally necessary.

Third, the solution to Example 4.3 applies to any digital transmission format using the same band of frequencies and, hence, the same attenuation per mile. Since all line codes have approximately the same steepness at a 10^{-6} error rate, the relative change in SNR is the same for all systems. Thus a T1 system with suboptimum detection would exhibit the same performance improvement if the repeaters were moved 320 ft closer together. (T1 lines are designed to provide less than a 10^{-6} error rate to begin with. See reference [32] for a thorough report of error rate surveys on T1 lines.)

*In terms of logic-level decisions, a bipolar code cannot provide error correction. If a pulse is removed to eliminate a bipolar violation, half of the time the error is corrected but half of the time another error is made. Error correction is possible only with a Viterbi-like detector described previously. In this case a bipolar code provides better performance than a unipolar code.

†Some systems, such as the radio systems discussed in Chapter 6, require a greater decrease in repeater spacing to achieve the same improvement in performance.

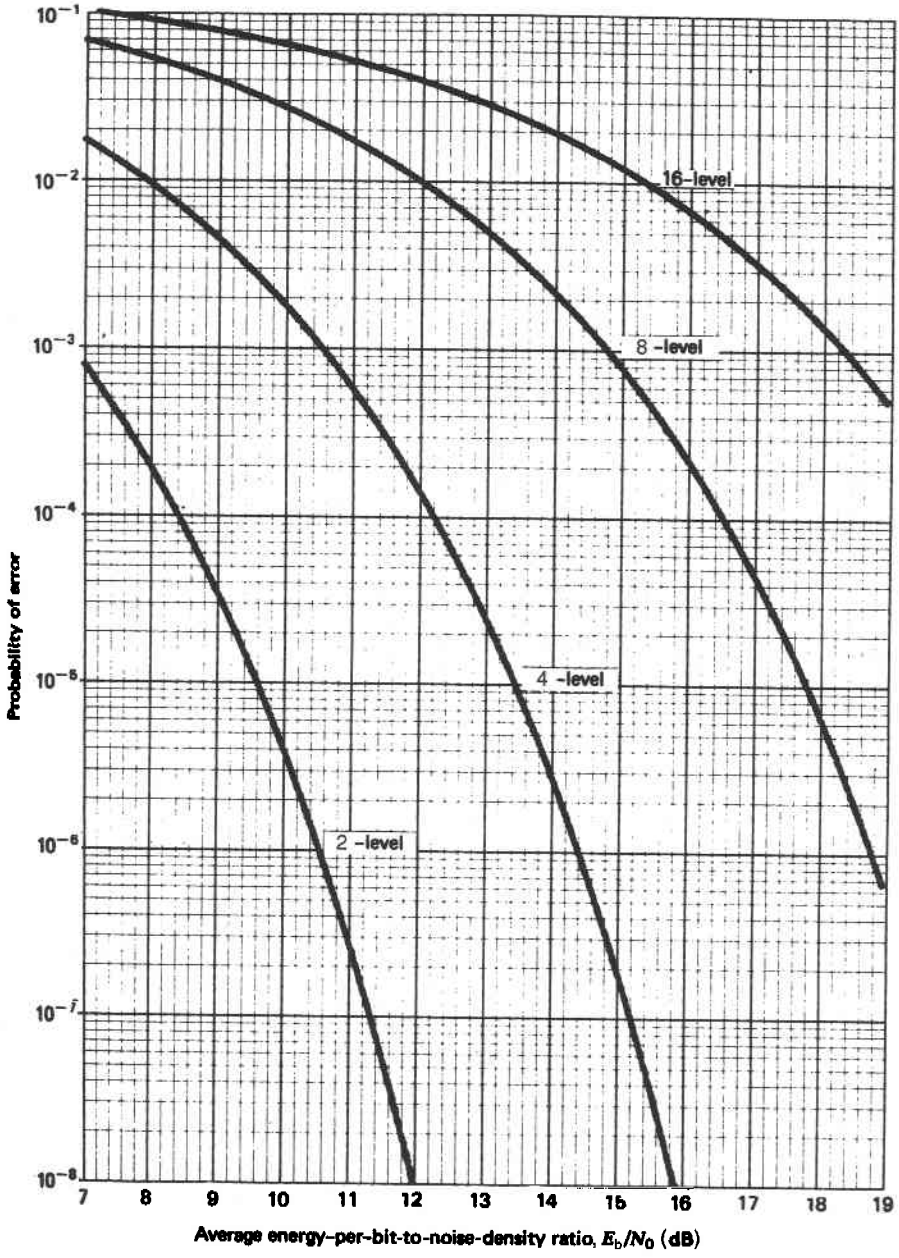


Figure 4.26 Error rate of balanced multilevel signals (all systems providing an identical data rate).

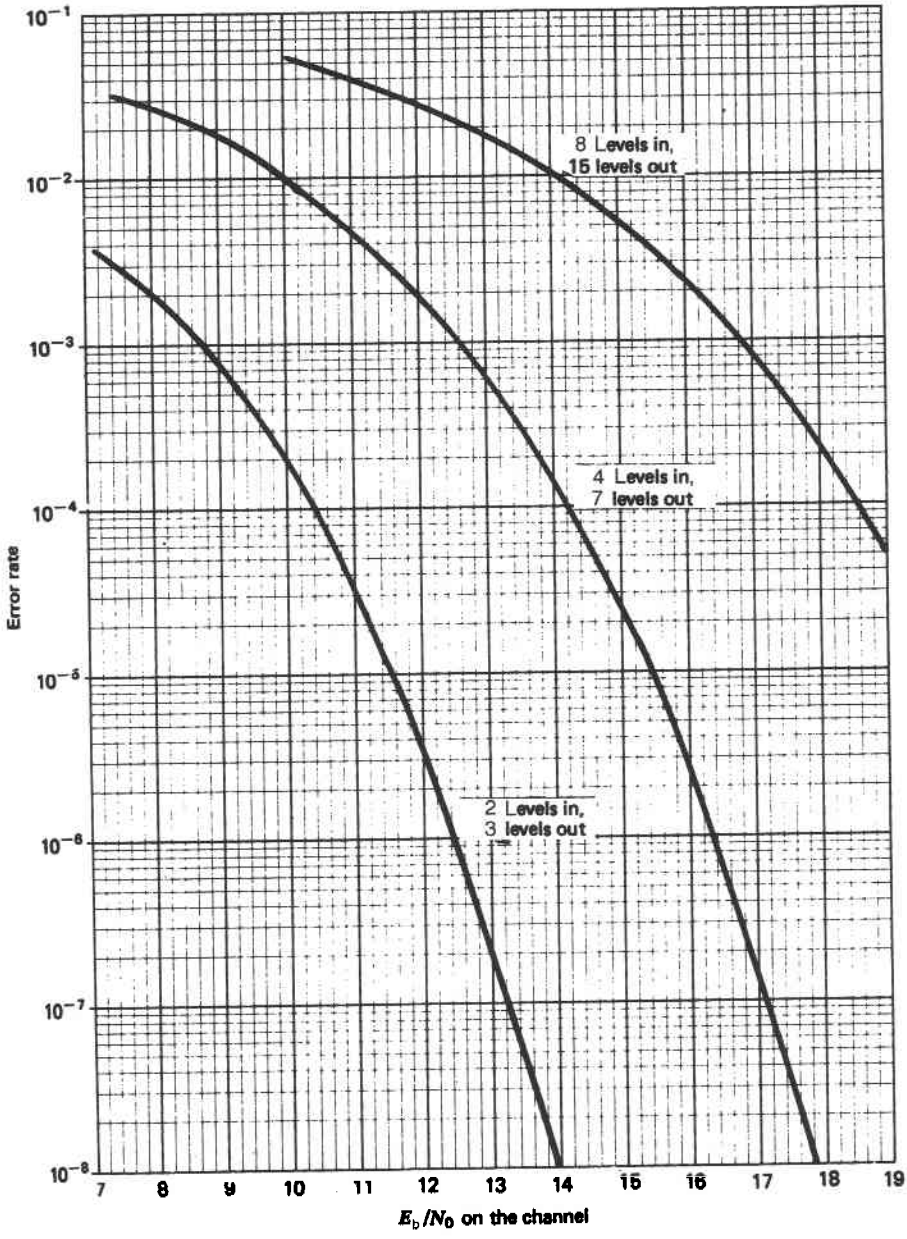


Figure 4.27 Error rates of $1 + D$ partial-response systems.

Multilevel Error Rates

The multilevel transmission system shown in Figure 4.16 does not require bandwidth in excess of a binary system using the same signaling rate, yet it achieves three times the data rate. The penalty for multilevel transmission comes in the form of greatly increased signal power for a given error rate. For example, the average power of the eight-level system in Figure 4.16 is 8.7 dB above the average power of a symmetric two-level system with the same error distance. To make matters worse, some systems are peak power limited, in which case the eight-level system has a 12.4-dB disadvantage with respect to a two-level system. The error rates of multilevel systems are derived in Appendix C and plotted in Figure 4.26 as a function of E_b/N_0. For those cases where the peak power is of interest, the relationship between the peak and average power of a multilevel system is derived in Appendix C as

$$\text{Peak-to-average (dB)} = 10 \log_{10} \frac{(L-1)^2}{(2/L)\sum_{i=1}^{L/2} (2i-1)^2} \tag{4.7}$$

where L is the number of equally spaced levels centered about zero [e.g., $\pm 1, \pm 3, \pm 5, \ldots, \pm (L-1)$].

The error rates of $1 + D$ partial-response systems are also derived in Appendix C and plotted in Figure 4.27. These error rates are derived under the assumption that bit-by-bit detection is used. Since partial-response systems contain redundancy (correlation) in adjacent samples, better performance can be achieved with Viterbi decoders [12].

4.5 PERFORMANCE MONITORING

Two basic techniques exist for directly monitoring the quality of a digital transmission link: redundancy checks and pulse quality measurements. Both techniques are designed to provide an indication of the bit error rate (BER) of the channel.

4.5.1 Redundancy Checks

Redundancy can be incorporated into a digital signal using one of two common methods. First, the line code itself may contain redundancy as in bipolar coding. In a random, independent error environment the frequency of bipolar violations is very nearly equal to the channel BER (except for extremely high BERs). Second, logic-level redundancy can be inserted into the data stream in the form of overhead bits. For example, parity bits are inserted into DS3 and DS4 signals for the express purpose of monitoring the channel error rate. (The frame structures of these and other higher level multiplex signals are provided in Chapter 7.) Cyclic redundancy check (CRC) codes are also incorporated into a number of transmission systems as a means of monitoring BERs and validating framing acquisition. Two prevalent examples of CRC use are (1) extended superframe (ESF) on T1 lines [33] introduced with the D4 channel banks and (2) optical fiber transmission using the SONET standard described in Chapter 8.

The use of logic-level redundancy (either parity bits or CRC codes) is generally better than the use of line code redundancy (such as bipolar violations) because the latter is dependent on the technology of the transmission link itself. Logic-level redundancy, on the other hand, is unaffected by a change from cable pairs to fiber or radio. As described later in this chapter, a major impetus for the introduction of ESF into T1 transmission systems was to get performance monitorability independent of bipolar line codes. DS1 signals are often multiplexed into higher level digital signals and carried on a variety of transmission systems such as fiber and radio. The CRC codes in the ESF framing format provide the means for end-to-end performance monitoring independent of whatever transmission systems carry the data stream.

In contrast to line code redundancies, parity bits and CRC codes do not provide a one-to-one indication of the BER. The following equation relates the parity error rate (PER) to the channel BER:

$$\mathrm{PER} = \sum_{i=1}^{N} \binom{N}{i} \rho^i (1-\rho)^{N-i} \qquad (i \text{ odd}) \tag{4.8}$$

where N = length of a parity field (number of bits over which parity is generated)
p = BER assuming random, independent errors

The relationship between the PER and the BER is plotted in Figure 4.28 for DS3 and DS4 signals. Notice that at low error rates the PER is essentially N times the BER (Np). At high error rates, however, this relationship changes because any odd number of errors in a frame is indistinguishable from a single error. When the BER is high enough that more than one error in a parity field is likely, the PER is useless as an absolute estimate of the error rate. In these cases the PER indicates only that the BER is above a threshold approximately equal to $1/N$. Because a DS4 frame format contains a higher density of parity bits, DS4 signals can be measured for higher error rates than can DS3 signals.

Determination of a CRC error rate is simplified because it is unlikely for multiple errors to not produce a CRC error. Thus the probability of a CRC error (CRCER) is 1 minus the probability that no errors occur:

$$\mathrm{CRCER} = 1 - (1-p)^N \tag{4.9}$$

where N = length of the CRC field (including CRC bits)
p = BER assuming random, independent errors

Again, at low error rates Equation 4.9 reverts to N times the BER (Np). Equation 4.9 is plotted in Figure 4.28 for ESF frames where the CRC field is 4614 bits. (The CRC field does not include F bits in Table 4.8 except the C bits themselves.)

To have confidence in error rate measurements, the sample sequence must be long enough to allow an average of about 10 errors in the sample size (e.g., the sample size

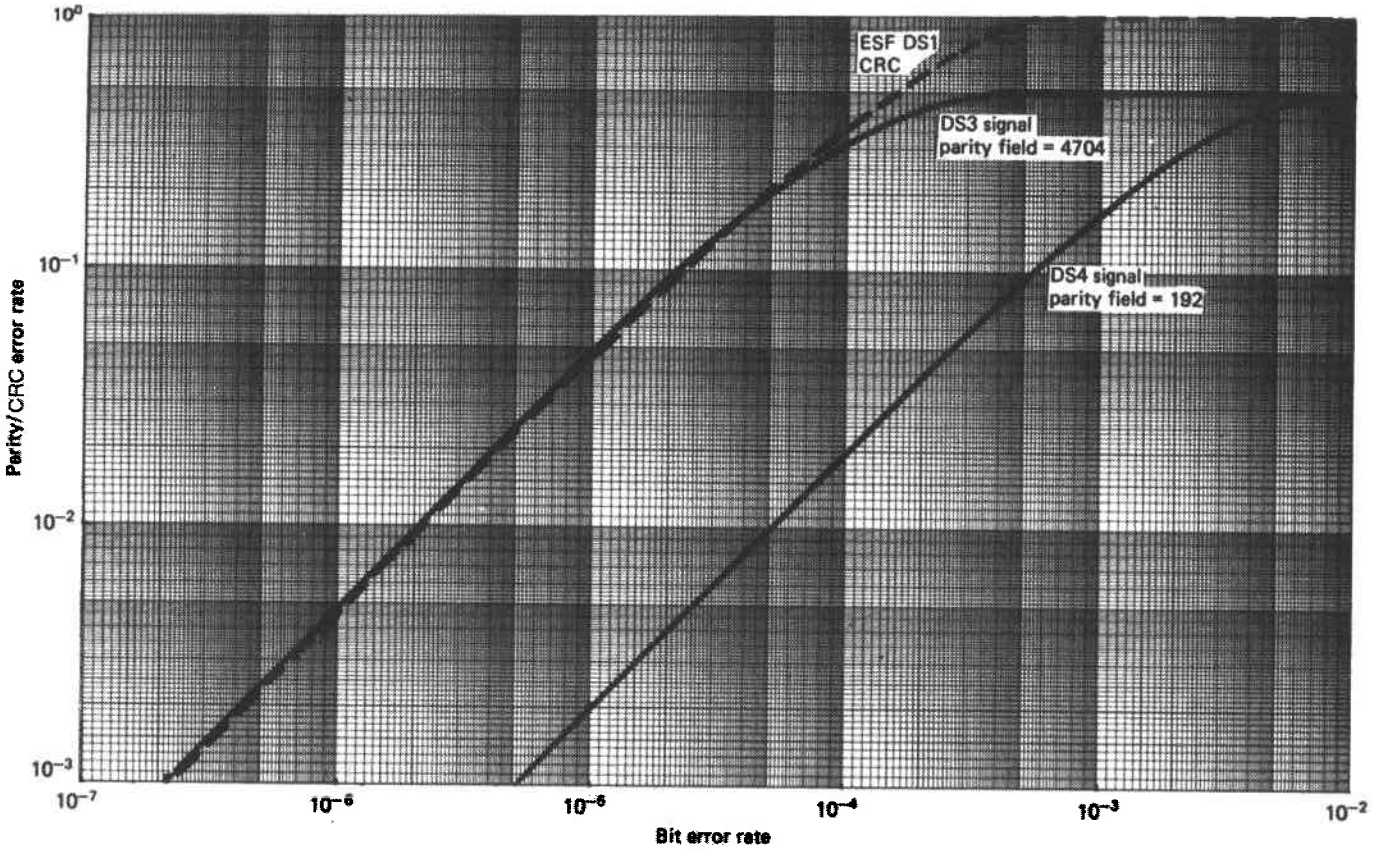


Figure 4.28 Parity error rates of DS3 and DS4 signals and CRC error rate of ESF DS1 signal versus bit error rate.

TABLE 4.8 Extended Superframe Framing Channel Format[a]

		F-Bit Assignment		
ESF Frame Number	ESF Bit Number	FPS	FDL	CRC
1	0	—	m	—
2	193	—	—	CB1
3	386	—	m	—
4	579	0	—	—
5	772	—	m	—
6	965	—	—	CB2
7	1158	—	m	—
8	1351	0	—	—
9	1544	—	m	—
10	1737	—	—	CB3
11	1930	—	m	—
12	2123	1	—	—
13	2316	—	m	—
14	2509	—	—	CB4
15	2702	—	m	—
16	2895	0	—	—
17	3088	—	m	—
18	3281	—	—	CB5
19	3474	—	m	—
20	3667	1	—	—
21	3860	—	m	—
22	4053	—	—	CB6
23	4246	—	m	—
24	4439	1	—	—

[a]FPS, framing pattern sequence (. . .001011. . .); FDL, 4 kbps facility data link (message bits m); CRC, CRC-6 cyclic redundancy check (check bits CB1–CB6).

must be 10/BER). Hence, when trying to measure low BERs (e.g., 10^{-6} or 10^{-7}), the measurement time may be too long to respond to changing channel conditions such as radio channel fading.

4.5.2 Signal Quality Measurements

The second basic technique for monitoring digital transmission quality is to process the digital signal directly and measure certain properties related to the error rate. A simple approach involves merely measuring the received signal power, a common technique in analog systems. In a fixed-noise environment this approach is adequate. However, on transmission links where the noise level can vary or where signal distortions can arise, the quality of the pulses themselves must be measured.

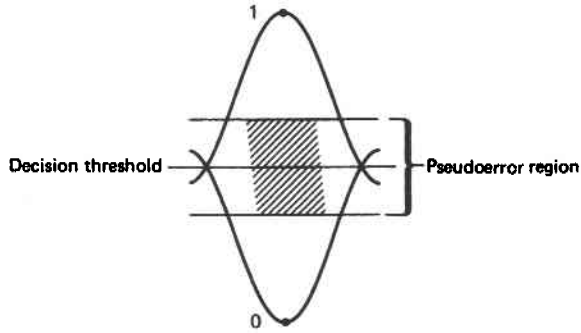


Figure 4.29 Pseudoerror detection.

Figure 4.29 demonstrates the operation of a "pseudo" error detector designed to detect received pulses with abnormal amplitudes at the sample times. In the example shown, binary data is detected by use of a single threshold located midway between the normal pulse amplitudes. Two additional thresholds are included to detect the presence of pulses with abnormal amplitudes. Sample values falling into the central decision region are not necessarily data errors, but a high pseudo-error rate is a good indication that the channel is not performing properly.

In a random (Gaussian) noise environment the occurrence rate of pseudo-errors is directly related to the actual error rate. Figure 4.30 shows a Gaussian noise distribution and decision thresholds chosen to produce pseudo-errors at 100 times a 10^{-6} error rate. Hence an attractive feature of this error rate measurement is that it can measure very low error rates using comparatively short test intervals. Note, however, that the error multiplication factor is dependent on the error rate. The technique of estimating a very low error rate by extrapolating from an artificially generated high error rate is sometimes referred to as a Q-factor technique [34]. In essence, the parameters (mean and variance) of a Gaussian distribution are estimated with easily obtained measurements so the much lower error rates can be estimated from the parameters. This technique is particularly useful in optical fiber systems where extremely low error rates are to be determined that otherwise require extremely long measurement times.

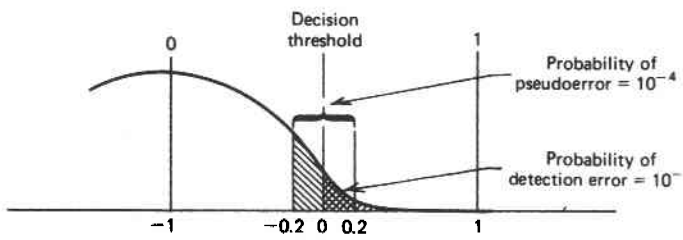


Figure 4.30 Pseudoerror decision thresholds for error multiplication of 100.

A major disadvantage of pseudo-error detection is that pseudo-error rates are not accurately related to actual error rates when the noise is non-Gaussian. Channel distortions, in particular, cause high pseudo-error rates when the actual error rates are small. Even though abnormal pulse amplitudes are present, errors do not occur unless noise adds to the distortion. On the other hand, impulse noise might produce high actual error rates but relatively low pseudoerror rates.

The correspondence between measured and deduced error rates is improved when the decision threshold is narrowed. Unfortunately, narrower pseudo-error decision regions are more difficult to implement and provide lower "error" rate multiplication factors.

4.5.3 Framing Channel Errors

As described in Section 4.6.2, identification of individual channels of a TDM data stream requires overhead bits with a unique data pattern for frame identification. After a receiving terminal has "locked" onto the framing pattern, it ignores occasional bit errors in the framing pattern but declares an out-of-frame (OOF) condition when the errors are consistent. Thus framing bit errors can be used to determine low error rates but not high error rates that cause false OOF indications and subsequent loss of framing. Error rates that are high enough to cause false OOF indications are usually so high that the link is effectively out of service.

4.5.4 Performance Objectives

The most basic measure of quality of a digital transmission link is its BER. If the errors occur in a truly random and independent manner (i.e., they conform to a simple Poisson probability model), the average BER completely specifies the performance. Unfortunately, error rate observations [32] show that errors often occur in bursts, and because bursts affect different services in different ways, specifying the quality of a transmission link requires more than just the average BER. For example, data traffic is generally transmitted in blocks that are retransmitted no matter how many errors occur in a block. Thus, a burst has little more effect than a single bit error. The rate of retransmissions of data blocks is more a function of the frequency of the bursts than of the long-term average BER. Voice traffic, on the other hand, is increasingly degraded by all bit errors whether or not they occur in bursts.

Channel errors that occur in bursts can also be much more deleterious to certain operations of a network than are distributed errors at a similar average rate. A loss of framing, for example, occurs more often in a burst error environment.* Certain timing operations (pulse stuffing) described in Chapter 7 are also similarly affected by bursts. Losses of framing and time synchronization both produce outages (e.g., continuous er-

*Errors in a burst are not necessarily contiguous. They merely occur in a short time interval at a rate that is too high to be reasonably explained as chance independent events.

rors) in the associated traffic channels or even inadvertent disconnects if the conditions last too long.

Common performance parameters used to characterize error rates are:

1. *Error-Free Second* (EFS): Because data blocks transmitted at 64 kbps require much less than a second, the percentage of EFS essentially specifies the percentage of time that the channel is available for data transmission applications. (When the block transmissions are much shorter than a second, a shorter measurement interval is more appropriate.)
2. *Burst Errored Second* (BES): An errored second with at least 100 errors.
3. *Errored Second* (ES): A second with at least one error.
4. *Severely Errored Second* (SES): A second with a BER greater than $(10)^{-3}$.
5. *Degraded Minute* (DM): A minute with a BER greater than 10^{-6}.

The last three parameters are defined in ITU recommendation G.821 for performance objectives of a digital network. The objectives for a 27,500-km reference connection (implying error contributions from many sources) are ES < 8%, SES < 0.2%, and DM < 10%.

The availability of fiber optic transmission technology with extremely low error rates (e.g., 1 error in 10^{12} bits) and higher level communications protocols that rely on very low error rates have created a situation where the very low error rates are relied on, even though they might not be obtained by metallic or radio systems. As discussed in Chapter 10, contemporary data networks [frame relay, asynchronous transfer mode (ATM), Internet protocol (IP)] do not provide error control at the data link layer but defer the problem to a higher level process.

4.5.5 Forward Error Correction

On transmission channels such as digital subscriber lines and digital radio links, relatively high error rates are sometimes unavoidable. When real-time applications like voice or video are involved, the effects of channel errors cannot be accommodated by periodically adding redundant check-sum bits and requesting retransmission of corrupted blocks of data as is commonly done in many data communications protocols. As mentioned in the previous chapter, particularly critical bits of compressed digital mobile speech are redundantly encoded with enough bits to permit error correction of some number of bit errors. Because this basic process does not require retransmission of data, it is known as forward error correction (FEC) [35].

FEC involves adding redundant data bits, like parity bits or CRC bits, but in sufficient quantity to enable error correction instead of just error detection. Two basic forms of FEC are block encoding and convolutional encoding. A block encoder attaches error correction redundancy to fixed-length blocks of data, which implies the need for establishing block framing to identify the fields of redundant bits. A convolutional encoder, on the other hand, continuously inserts error correction bits that are continuously processed by the decoding circuitry while recovering the original data.

Reed–Solomon Codes

Although a wide variety of block-coding algorithms have been developed, Reed–Solomon (RS) coding is the most popular form. When a block of source data symbols of length M is input to an RS coder, an output block of length N symbols is produced, where $N - M$ is the number of check symbols R. An RS code with these parameters is commonly referred to as a RS(N, M) = RS(N, $N - R$) code. An RS(N, $N - R$) code can correct $\frac{1}{2}R$ symbol errors. In most cases a symbol consists of an 8-bit byte of data,* so multiple bit errors in a single byte are no worse than a single bit error. Digital video broadcasting (DVB) systems use RS(204,188) codes, which means that as many as eight bytes can be corrupted and be corrected.

Example 4.4. Determine the probability of failure of an RS(204, 188) code operating in a random error environment with a bit error probability of 10^{-3}. Assume each symbol is an 8-bit byte.

Solution. Because the probability of multiple bit errors in a single byte is small, the probability of a symbol error p_s is very close to 8×10^{-3}. Using p_s, the probability of failure (the probability of more than eight symbol errors), p_F can be determined as

$$p_F = \sum_{i=9}^{204} \binom{204}{i} p_s^i (1 - p_s)^{204-i}$$

$$= 1 - \sum_{i=0}^{8} \binom{204}{i} p_s^i (1 - p_s)^{204-i}$$

$$= 0.00005$$

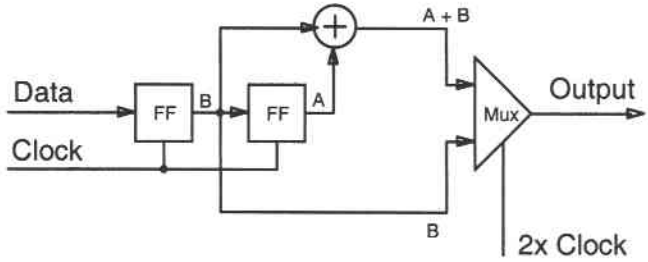
Convolutional Coding

Block diagrams of two basic convolutional encoders are shown in Figure 4.31.

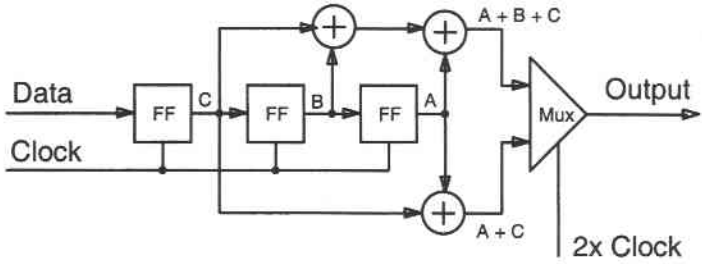
Both of these encoders are referred to as rate 1/2 encoders because the source data rate is half the channel data rate. During each bit time of a source two output bits are generated. As illustrated in Figure 4.32, the constraint-2 length coder in Figure 4.31*a* outputs odd parity over bits A and B along with the value of B directly. In Figure 4.31*b*, odd parity across overlapping fields (A, B, C and A, C) are generated. In the case of odd parity over A, B, C an output value is a 1 if and only if an odd number of A, B, C are 1.

An important consideration of a convolutional coder is the *constraint length*, which specifies the number of instances that a particular source bit gets mapped into a chan-

*Reed–Solomon codes are sometimes denoted as RS(N, $N - R$, B), where B designates the length of a symbol in bits.



(*a*)



(*b*)

Figure 4.31 Rate 1/2 convolutional encoders: (*a*) source constraint length 2; (*b*) source constraint length 3.

nel bit. In Figure 4.31*a* the length of the shift registers is 2 bits, which means that this coder has a source constraint length of 2. Because the output clock is twice the input rate, the channel constraint length is 4. Similarly, the encoder shown in Figure 4.31*b* has a source constraint length of 3 and a channel constraint length of 6. Longer constraint lengths produce better performance. Because the encoder shown in Figure 4.31*a* has the shortest possibly constraint length, it is not commonly used but is convenient for illustrating the basic operation of a convolutional decoder. Notice that each

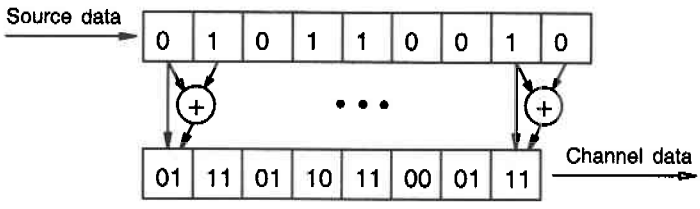


Figure 4.32 Example input and output data sequences for convolutional encoder in Figure 4.31*b*.

source bit in Figure 4.31*a* is mapped into 3 channel bits. It is this redundancy that gets processed to recover the original data.

Example 4.5. Determine the decoding logic to decode received data for the convolutional encoder of Figure 4.31*a*. Assume bit-by-bit decisions are made on each received channel bit and only consider isolated channel errors.

Solution. Two cases are possible for isolated channel errors: an error in a source bit B or an error in a parity bit. If a source bit is received in error, two parity errors result. Thus, when two adjacent parity errors occur, the most likely scenario is that the implied data bit needs to be inverted. If a single parity bit is received in error, the most likely scenario is that the parity bit is itself in error and no source errors have occurred. Thus, isolated parity errors are ignored.

Example 4.5 illustrates that a convolutional encoder/decoder can easily correct isolated channel errors—at the expense of doubling the data rate. In contrast to the previous example and the normal implementation of an RS decoder, a convolutional decoder usually processes the received signal with maximum-likelihood sequence (Viterbi) detectors [36]. In essence, a sequence of received signal samples is compared to all possible channel input sequences with the closest allowable sequence chosen to determine the source data. The added complexity provides significant benefits. Constraint length 7 convolutional coding, for example, provides better than 5 dB of improvement in error performance for a given SNR. Since a doubling of the necessary bandwidth (for rate 1/2 coding) implies a relative noise increase of 3 dB, a net gain of 2 dB is achieved. Notice further that a convolutional encoder can correct a much greater number of corrupted bits than can an RS code—as long as the corrupted bits are separated by something greater than the channel constraint length. Hence, convolutional coding is well suited to correct independent errors while RS codes work well in burst error environments. For these reasons combinations of the two codes are often used.

Maximum-likelihood detection of convolutional code sequences is very similar to trellis-coded modulation (TCM) detection discussed in Chapter 6. The TCM discussion provides some example measurements that involve the same basic process required for Viterbi detection of convolution codes.

4.6 TIME DIVISION MULTIPLEXING

Although FDM of digital transmission signals is possible and is used in some special situations, TDM is by far the most common and economical means of subdividing the capacity of a digital transmission facility. One application where FDM techniques have been used for digital signals is on multidrop data communications lines, where the sources and destinations of the data are distributed along the line. Most telephone network applications, however, involve clusters of channels in the form of trunk groups between switching offices. In local digital access applications, where sub-

scriber locations are distributed throughout a service area, channels are sometimes added and dropped on a distributed basis, but only with a limited number of access points because of maintenance and reliability considerations.*

There are two basic modes of operation for TDM: those that repeatedly assign a portion of the transmission capacity to each source and those that assign capacity only as it is needed. The first form of operation is referred to as *synchronous* time division multiplexing (STDM) when necessary, to distinguish it from the "as-needed" mode of operation. Otherwise, TDM is generally understood to imply the synchronous variety. The as-needed form of TDM is variously referred to as asynchronous time division multiplexing (ATDM), Asynchronous Transfer Mode (ATM), statistical time division multiplexing (Stat-Mux), or packet switching. Circuit-switched telephone networks use STDM, whereas data networks typically use ATDM. Discussion of these latter techniques is deferred to Chapter 10.

The term *asynchronous* is also used in another TDM context to refer to the multiplexing of multiple, independent tributary signals into a single higher rate signal. In this context, "asynchronous" is totally unrelated to either asynchronous transmission as described at the beginning of this chapter or packetlike multiplexing mentioned in the previous paragraph. In this third context asynchronous refers to multiplexing of multiple unsynchronized tributaries into a single high-rate serial stream. This chapter is concerned only with multiplexing of synchronized tributaries. TDM of unsynchronized tributaries is described in Chapter 7.

4.6.1 Bit Interleaving Versus Word Interleaving

Two different structures of (synchronous) time division multiplex frames are shown in Figure 4.33. In the first instance each channel is assigned a time slot corresponding to a single bit—hence the term *bit interleaving*. In the second instance each channel is assigned a longer time slot corresponding to some larger number of bits, referred to as a word—hence the term *word interleaving*.

The decision as to which structure to use in a particular application is primarily dependent on the nature of the sources. In T-carrier systems each channel produces a complete 8-bit word at a time. Hence word interleaving is desirable so all bits can be transmitted as generated.

In contrast, higher level digital TDM multiplexers typically use bit interleaving of the lower level multiplex signals since the lower level signals represent continuous, 1-bit-at-a-time, data streams. The specific formats of the higher level multiplex signals are described in Chapter 7 when synchronization of bit streams is considered. One notable exception to higher level bit interleaving is the word-interleaved structure of fiber-based SONET or SDH multiplexing described in Chapter 8.

*The digital network has evolved with ever-increasing levels of multiplexing creating serial bit streams with ever-increasing data rates. The use of WDM on optical fiber systems represents a departure from the single-stream paradigm stimulated by two factors: (1) it is impractical to transmit the ultrahigh speed serial channels and (2) separate wavelengths provide transparency for diverse applications.

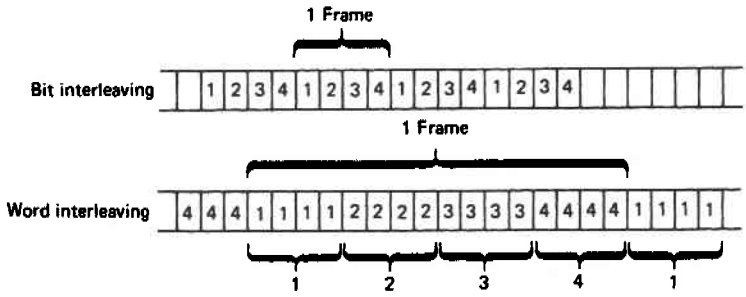


Figure 4.33 Bit interleaving and word interleaving of four-channel TDM multiplexers.

4.6.2 Framing

To identify individual time slots within a TDM frame, a receiving terminal uses a counter synchronized to the frame format of the transmitter. Just as for synchronization of sample clocks, a certain amount of transmission overhead is required to establish and maintain frame synchronization. In fact, most of the techniques used for frame synchronization are directly analogous to clock synchronization techniques discussed previously. Specifically, the basic means of establishing frame synchronization are:

1. Added-digit framing
2. Added-channel framing
3. Unique line signal framing
4. Statistical framing

The main considerations in choosing a framing procedure are time required to establish framing, effects of channel errors in maintaining frame synchronization, relationships between the line clock and sample clocks derived from the line clock, transmission overhead, and complexity of the framing circuitry.

The severity of a loss of framing and the time required to reestablish synchronization depend on the nature of the traffic. Since a loss of framing implies a loss of data on all channels, the mean time between misframes must be as long as possible. For voice traffic infrequent misframes can be tolerated if frame synchronization is reestablished rapidly enough to minimize the duration of the "glitch" in the output speech. For data traffic the duration of reframe times is not as critical as the frequency of occurrence since most data communications protocols initiate recovery procedures and retransmit a message no matter how many data bits are lost.

A critical requirement for reframe time in the telephone network comes from the possibility that various in-channel control signals may be lost and interpreted as disconnects. Thus the maximum reframe time on a particular digital transmission link is often determined from analog network signaling conventions. A loss of framing is also used as a performance monitor for the transmission link and usually sets alarm condi-

tions, which in turn cause automatic switching to spare terminals or transmission facilities. Generally speaking, terminals serving 600 or more channels require automatic protection switching [37].

Out-of-frame conditions occur in two ways. First, the locally derived sample clock may lose synchronization with the line clock and produce a slip in the counter sequence. Normally, the timing information in the line clock is sufficiently strong to prevent misframes of this type. An exception occurs on radio links when deep fades reduce the signal power to the point that clock synchronization is impossible. Receiving terminals can also lose clock synchronization through no fault of their own or of the transmission link. If the source clock at the transmitting terminal has too much jitter or generates abrupt phase shifts, receiving terminals may be unable to track the clock signal. Both phenomena are related to network synchronization practices, as discussed in Chapter 7.

A second source of misframes is from channel errors creating false out-of-frame framing patterns. Thus considerable redundancy in the framing pattern is required to minimize the probability of false misframes. A loss of framing is determined when the occurrence of framing pattern violations exceeds some short-term density threshold.

In all of the framing techniques discussed in the following paragraphs, special framing bits or codes are inserted into the information stream. These insertions do not have to occur with each frame. Instead, they can be sent only once for a predetermined number of information frames. In this manner the transmission overhead for framing can be reduced, which is particularly important in bit-interleaved systems. An individual information frame contained between framing indicators is sometimes referred to as a "subframe."

Added-Digit Framing

One common technique of framing a digital TDM information stream is to periodically insert a framing bit with an identifiable data sequence. Usually the framing bit is added once for every frame and alternates in value. This particular format is the procedure used to establish framing in the original D1 channel banks. When the T1 line carries only voice traffic, this framing format is particularly useful since no information bits can sustain an alternating 1, 0 pattern. (An alternating pattern represents a 4-kHz signal component, which is rejected by the bandlimiting filter in the PCM codecs.)

Framing is established in a receiving D1 channel bank by monitoring first one bit position within a 193-bit frame and then another, until the alternating pattern is located. With this framing strategy, the expected framing time from a random starting point with random data is derived in Appendix A as

$$\text{Frame time} = \begin{pmatrix} \text{average number} \\ \text{of bit positions} \\ \text{before} \\ \text{framing bit} \\ \text{is found} \end{pmatrix} \times \begin{pmatrix} \text{average number of} \\ \text{bitsto} \\ \text{determine that an} \\ \text{information position is} \\ \text{not a framing position} \end{pmatrix}$$

$$= (\tfrac{1}{2}N)(2N+1)$$

$$= N^2 + \tfrac{1}{2}N \text{ bit times} \tag{4.10}$$

where N is the number of bits in a frame including the framing bit.

For D1 channel banks, $N = 193$ so that the framing time is 37,346 bits, or 24.188 msec. Also of interest is the maximum framing time. Unfortunately, there is no absolute maximum framing time for a T1 system with random data. It is very unlikely, however, that the framing time would ever exceed the average search time for all bit positions, or 48.25 msec. This latter measure of framing time is referred to as the *maximum average frame time*. It is the average time required to establish framing, but with the assumption that all bit positions must be tested for the framing sequence before the actual framing bit is found. Obviously, the maximum average frame time is twice the average value from a random starting point defined in Equation 4.10.

The framing time can be reduced by using more sophisticated frame search strategies. One approach examines one bit at a time, as before, but during a reframe the search begins a few bit positions in front of the present position under the assumption that short lapses of clock synchronization cause small counter offsets. A second approach [38, 39] uses a parallel search by monitoring all bit positions simultaneously for the framing pattern. With this framing procedure, framing is established when the last of the $N-1$ information bit positions finally produces a framing pattern violation. The probability that all information bit positions produce a framing violation in n or less frames is derived in Appendix A as follows:

$$\text{prob(frame time} < n) = [1 - (\tfrac{1}{2})^n]^{N-1} \tag{4.11}$$

where N is the number of bits in a frame and the probability of 1's is $\frac{1}{2}$. Using Equation 4.11, we determine the median framing time by setting prob(frame time $< n$) $= \frac{1}{2}$. Hence

$$n = -\log_2[1 - (\tfrac{1}{2})^{1/(N-1)}] \tag{4.12}$$

Setting $N = 193$ for the frame length of the D1 channel bank produces the result that $n = 8.1$ frames, or approximately 1 msec.

An even more sophisticated framing strategy involves continually monitoring all bit positions, even while the system is synchronized. Then, when a misframe is detected, a new frame position is immediately established or the frame search time is significantly reduced. Furthermore, the continuous search for framing patterns provides additional information for declaring an out-of-frame condition. (There is little point is discarding the present frame position unless another bit position exhibits a more consistent framing pattern.)

The framing pattern of second-generation channel banks (D3, D4) from AT&T was changed from the alternating 1, 0 pattern to establish a longer sequence for identification of signaling frames. As mentioned previously, these channel banks provide 8 bits of voice for all time slots except in every sixth frame, which uses the least significant PCM bit for signaling. The signaling channels thus derived are divided into an A and a B subchannel, implying each subchannel sends a bit in every twelfth frame. Hence a 12-bit framing sequence is needed to identify the signaling bits and the two signaling subchannels. The 12-frame superframe (SF) structure and associated framing sequence are shown in Figure 4.34.

When describing or processing the framing sequence shown in Figure 4.34, it is convenient to divide the framing bits into two separate sequences. During the odd frames the framing bit alternates, while during the even frames the framing bit sequence is 000111000111000. Figure 4.34 indicates that the A signaling frame can be identified by a 0-to-1 transition in the even-numbered frame sequence. Correspondingly, a 1-to-0 transition in the even-numbered frame sequence signifies a B signaling frame. Frame acquisition begins by finding the alternating bit sequence (with 385 intervening bits). Then, the 000111 framing pattern is located. Another framing sequence, extended superframe (ESF), for DS1 signals is described in Section 4.6.3.

Added-Channel Framing

Added-channel framing is basically identical to added-digit framing except that framing digits are added in a group such that an extra channel is established. Hence the

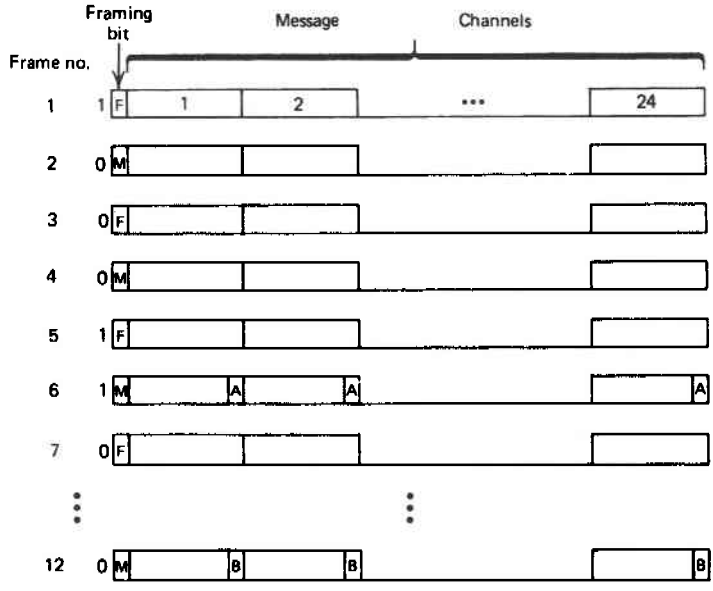


Figure 4.34 Twelve-frame superframe structure of DS1 signal with robbed digit signaling. Frame alignment signal (F) = 101010; multiframe alignment signal (M) = 001110.

transmission rate of individual channels is integrally related to the line rate of the entire multiplex.

The fact that the frame boundaries are identified by whole codewords adds considerable performance and flexibility to the framing process. First, framing can be established more rapidly since random 8-bit codewords are very unlikely to appear as framing codes. (See the problems at the end of the chapter.) Second, the larger code space simplifies identification of auxiliary functions such as superframe boundaries, parity bits, or equipment status. In most systems the added channel contains more than framing bits.

The first-level digital multiplex signal of ITU (E1) is an example of a system using added-channel framing. The E1 standard establishes 32 channels per frame with one channel providing framing and one more channel dedicated to signaling. Thus, 30 of 32 channels are available for message channels. Figure 4.35 shows the frame structure of the E1 signal. The frame alignment signal (FAS) is inserted into the framing channel of every even-numbered frame and a 1 bit inserted into the second bit position in every odd-numbered frame (to preclude FAS generation in those frames). The first bit of every FAS frame may carry a cyclic redundancy check (CRC-4) for additional frame alignment integrity and error rate monitoring. The remaining bits of the framing channel are used for CRC-4 alignment or alarm indications or reserved for other uses. (See ITU recommendation G.704 for more details.)

The signaling channel of Figure 4.35 depicts the use of channel-associated signaling (CAS), which implies that 4 bits per 16-frame multiframe are allocated to each message channel. The positions of the associated signaling channels are determined with respect to the multiframe alignment signal (MAS). The 4 bits of each CAS signaling channel should never be 0000 to preclude inadvertent generation of the

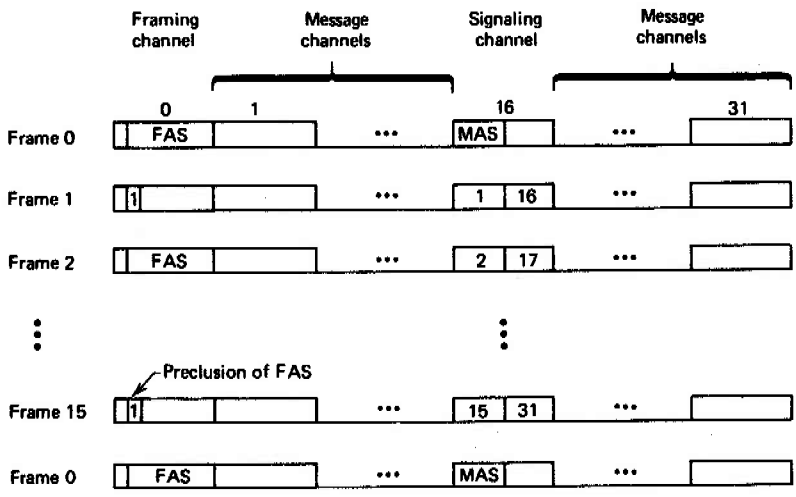


Figure 4.35 Channel format of ITU primary digital signal (E1).

MAS. The use of these bits is similar to the "bit-robbed" ABCD bits of the North American DS1 signal. When common channel signaling is carried on an E1 link, the CAS channels are replaced by a 64-kbps HDLC signaling link in time slot 16 of the E1 frame.

The average frame acquisition time of a multibit frame code is derived in Appendix A as

$$\text{Frame time} = \frac{N^2}{2(2^L - 1)} + \frac{N}{2} \quad \text{(in bits)} \tag{4.13}$$

where N is the length of a frame including the frame code, L is the length of the frame code, and it is assumed that 1's and 0's are equally likely.

From Figure 4.35 it can be seen that for the E1 signal $N = 512$ and $L = 7$. Thus the average frame time from a random starting point is determined from Equation 4.13 as 0.5 msec. Again, the "maximum average" frame time is twice the average from a random starting point, or 1 msec. Notice that these frame times are much faster than DS1 added-digit frame times because a higher percentage of bits are allocated to framing.

Unique Line Code Framing

Bipolar coding manages to shape the spectrum (remove the dc component) of the line code by adding extra signal levels to provide more flexibility in selecting signals. The same technique can be used to establish frame synchronization. In fact, with bipolar coding and added-digit framing, bipolar violations can be used to identify the framing boundaries uniquely and rapidly. A particularly significant example of using line code violations for framing is the ISDN S/T basic rate interface described in Chapter 11.

Even without added-digit framing, unique line codes can be used to carry information while simultaneously conveying frame positioning. If the number of signal levels is doubled for the framing bit only, the extra levels provide sufficient code space for the information but uniquely identify that bit position as a frame boundary. This procedure does not necessarily increase the error rate, since for any particular bit (information or framing), half of the levels can be disallowed.

The main advantage of using unique line codes for framing is that the information bit positions cannot generate framing patterns. Hence framing is established as soon as a frame bit occurs, and misframes are detected almost immediately.

The main disadvantages of unique line codes are the added signal processing requirements (if new levels are established just for framing) and the dependence on the transmission terminals to locate framing. With the other framing techniques described in this section, the framing patterns are represented in the data of the bit stream. Hence the transmission equipment can be changed independently of the multiplex equipment. For additional information on framing techniques and performance analyses see references [40] and [41].

Statistical Framing

Statistical framing relies on the statistics of data within individual bits of a transmission scheme. Assuming the source of the data is known, it may be possible to ascertain information such as which bit is a most significant bit (MSB) of a PCM codeword and thereby recover byte framing without an explicit framing bit of any type. Obviously, this method of framing is only applicable to special applications. One such application is determining word alignment in ADPCM 7-kHz audio [42], as in ITU-T Recommendation G.722.

4.6.3 DS1 Extended Superframe

When a T1 line is used as an interoffice transmission link, the operating telephone companies have access to both ends of the line for maintenance and performance monitoring. Furthermore, most installations include protection switching facilities that can also be used to routinely test a line while the traffic normally assigned to that line is diverted to a spare line. A significantly different situation arises when a T1 line is used by a customer as access to the public switched network or as a leased line in a private network. The explosive use of T1 lines in such applications created the need for enhanced maintenance features for T1 customers. The ESF as provided in the D5 channel bank provides the desired features.

Foremost among the customer needs is nonintrusive performance monitoring of an end-to-end T1 circuit. Monitoring bipolar violations at the customer service unit (CSU) provides nonintrusive monitoring of the receive signal but provides no information regarding the quality of the signal at the other end of the line (which terminates at the provider's facilities). Furthermore, leased T1 circuits generally involve intermediate transmission, multiplexing, and cross-connect equipment. Because the interfaces of each of these equipments remove bipolar violations, the customer has no access to performance statistics internal to the provider's facilities. Determination of end-to-end error rate performance in this environment requires taking at least a portion of the T1 circuit out of service (for which there is usually no spare).

ESF allows customers to achieve end-to-end, in-service performance monitoring by redefining the 8-kbps F bit of each DS1 frame to consist of a 2-kbps framing channel, a 4-kbps data link channel, and a 2-kbps CRC channel representing a check sum over all information bits. Because the CRC channel is carried intact from one customer location to another, channel errors occurring in any intermediate facility produce CRC errors at the far end.

The data link supports numerous services, one of which is the ability to interrogate the far end, or any intermediate equipment, for performance statistics. Thus the data link provides both the customer and the service provider with an extremely useful tool for isolating faulty span lines or equipment.

As indicated in Table 4.8 obtained from reference [33], the three subchannels are established by "extending" the D4 superframe format to encompass 24 DS1 frames. Because the actual framing bits occur only once every four DS1 frames, there are 771 bits intervening between bits carrying the frame pattern sequence (FPS) 001011.

The 6 CRC bits (CB1 to CB6) of each extended superframe represent a CRC check of all 4608 information bits in the previous superframe.* Besides providing end-to-end performance monitoring, the CRC virtually precludes the chances of false framing on a data bit position. Even though static user data can easily simulate the FPS, it is extremely unlikely that user data can spuriously generate valid CRC codes in successive superframes. (Six bits of random data match a 6-bit code with a probability of 1 in 64.)

The performance parameters measured and reported by the 4-kbps data link (DL) are framing bit errors, CRC errors, out-of-frame (OOF) events, line code (bipolar) violations, and controlled slip events (described in Chapter 7). Individual events are reported as well as event summaries. The four performance summaries reported are:

1. Errored seconds (ESs) (ES = at least one CRC event)
2. Bursty seconds (BSs) (BS = 2–319 ESs)
3. Severely errored seconds (SESs) (SES = >319 ESs or OOFs)
4. Failed seconds (FSs) (10 consecutive SESs)

ESF CSUs typically determine the above parameters on 15-min intervals and store them for up to 24 hr for polling by a controller [43]. The SES report conforms to ITU recommendation G.821. In addition to supporting remote interrogation of performance statistics, the data link carries alarm information, loopback commands, and protection switching commands.

In addition to the previously mentioned features, ESF introduces a new option for per-channel signaling via the robbed signaling bits in every sixth frame. Because an ESF is 24 frames long, there are four signaling bits in every channel in every superframe as opposed to 2 bits in SF format (Figure 4.34). Whereas the two signaling bits in the SF format are designated as A and B bits, the four bits in the ESF case are designated A, B, C, and D. Three signaling modes are defined: 2-state where all bits are A bits, 4-state where the signaling bits are ABAB, and 16-state where the signaling bits are ABCD. The SF format provides the first two signaling modes but not the last.

4.7 TIME DIVISION MULTIPLEX LOOPS AND RINGS

In Chapter 2 it is mentioned that TDM is not as amenable to applications with distributed sources and sinks of traffic as is FDM. In this section a particular form of a TDM network is described that is quite useful in interconnecting distributed nodes. The basic structure of interest is referred to as a TDM loop or TDM ring and is shown in Figure 4.36.

Basically, a TDM ring is configured as a series of unidirectional (two-wire) links arranged to form a closed circuit or loop. Each node of the network is implemented with two fundamental operational features. First, each node acts as a regenerative repeater merely to recover the incoming bit stream and retransmit it. Second, the net-

*Calculation of the CRC actually includes F bits that are set to 1 for purposes of CRC calculation only. Thus, channel errors in the F bits do not create CRC errors (unless they occur in the CRC bits themselves).

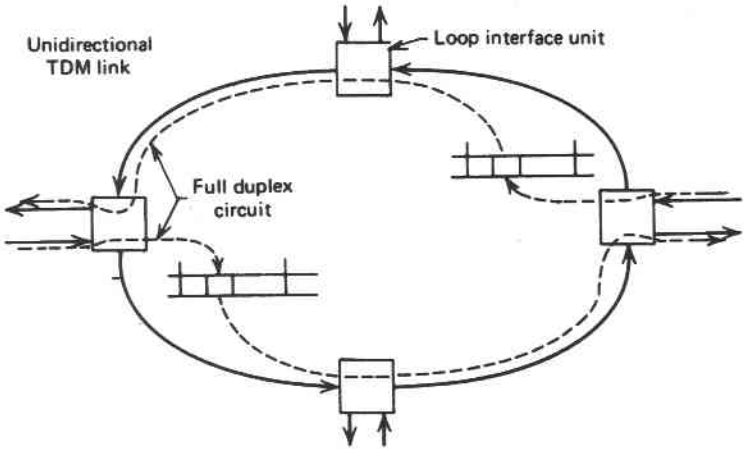


Figure 4.36 Time division multiplex loop.

work nodes recognize the TDM frame structure and communicate on the loop by removing and inserting data into specific time slots assigned to each node. As indicated in Figure 4.36, a full-duplex connection can be established between any two nodes by assigning a single time slot or channel for a connection. One node inserts information into the assigned time slot that propagates around the loop to the second node (all intervening nodes merely repeat the data in the particular time slot). The destination node removes data as the assigned time slot passes by and inserts return data in the process. The return data propagates around the loop to the original node where it is removed and replaced by new data, and so forth.

Since other time slots are not involved with the particular connection shown, they are free to be used for other connections involving arbitrary pairs of nodes. Hence a TDM loop with C time slots per frame can support C simultaneous full-duplex connections.

If, as channels become available, they are reassigned to different pairs of nodes, the transmission facilities can be highly utilized with high concentration factors and provide low blocking probabilities between all nodes. Thus a fundamental attraction of a loop network is that the transmission capacity can be assigned dynamically to meet changing traffic patterns. In contrast, if a star network with a centralized switching node is used to interconnect the nodes with four-wire links, many of the links to particular nodes would be underutilized since they cannot be shared as in a loop configuration.

Another feature of the loop-connected network is the ease with which it can be reconfigured to accommodate new nodes in the network. A new access node is merely inserted into the nearest link of the network and the new node has complete connectivity to all other nodes by way of the TDM channels. In contrast, a star structured network requires transmission to the central node and expansion of the centralized switching facilities.

The ability to reassign channels to arbitrary pairs of nodes in a TDM loop implies that the loop is much more than a multiplexer. It is, in fact, a distributed transmission and switching system. The switching capabilities come about almost as a by-product of TDM transmission. TDM loops represent the epitome of integrated transmission and switching.

TDM loops have been used within computer complexes to provide high capacity and high interconnectivity between processors, memories, and peripherals [44]. The loop structure in this application is sometimes more attractive than more conventional bus structures since all transmission is unidirectional and therefore avoids timing problems on bidirectional buses that limit their physical length. Furthermore, as more nodes are added to a bus, the electrical loading increases, causing a limitation on the number of nodes that can be connected to a bus. Loops, on the other hand, have no inherent limits of transmission length or numbers of nodes.

The loop structure of Figure 4.36 is topologically identical to the token-passing ring developed by IBM and standardized by the IEEE as a 802.5 local area network. However, a token-passing ring operates differently than a TDM loop in that there is only one channel. When a node on a ring becomes active, it uses the entire capacity of the outgoing link until it is through sending its message. In contrast, a node on a loop uses only specific time slots in the TDM structure, allowing other nodes to be simultaneously "connected" using other time slots. In essence, a TDM loop is a distributed-circuit switch and an 802.5 ring is a distributed-packet switch.

A particularly attractive use of a loop with high-bandwidth links is shown in Figure 4.37. This figure illustrates the use of add–drop multiplexers (ADMs) that access whatever bandwidth is needed at a local node but pass the rest on to other nodes. In typical applications the amount of bandwidth allocated to each node is quasi-static: It is changed only in response to macroscopic changes in traffic patterns, possibly as a function of the time of day. This basic operation is generally referred to as a cross-con-

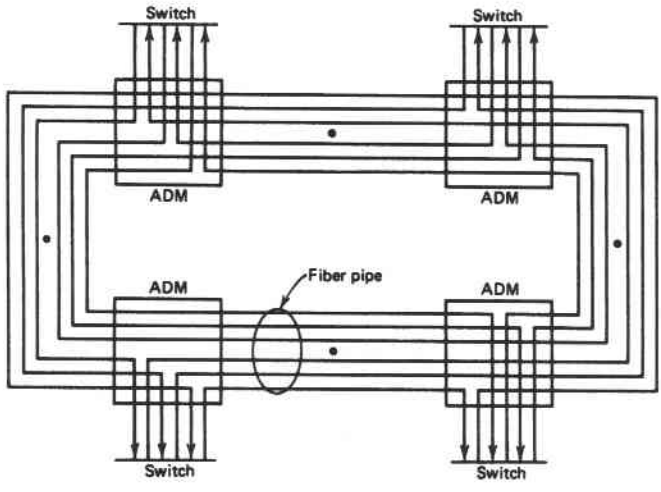


Figure 4.37 Functional mesh, fiber loop, and ADMs.

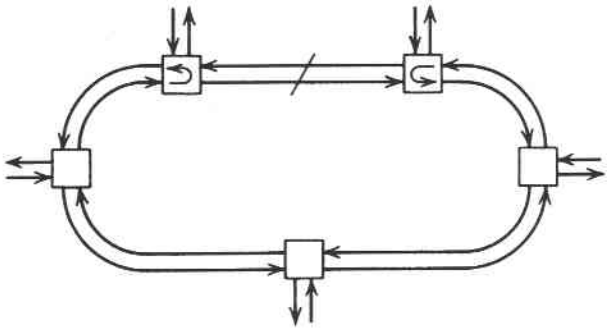


Figure 4.38 Use of reverse loop to circumvent link failures in TDM loops.

nect function as opposed to a switching function, which involves call-by-call reconfigurations. Both types of switching operations are discussed in the next chapter. An important point to note about Figure 4.37 is the ability to utilize a general-purpose physical topology but define an arbitrary functional topology on top of it.

One obvious limitation of a loop is its vulnerability to failures of any link or node. The effect of a node failure can be minimized by having bypass capabilities included in each node. When bypassed, a node becomes merely a regenerative repeater, as on T-carrier transmission links. Link failures can be circumvented by providing alternate facilities. Figure 4.38 shows one particular structure using a second, reverse-direction loop to provide backup capabilities in the case of failures. When fully operational, the network can use the reverse loop as a separate, independent network for traffic as needed. Whenever a failure occurs, the nodes adjacent to the break establish a new loop by connecting the forward path to the reverse path at both places. Hence all nodes continue to have full connectivity to any node on the new loop.

A particular example of the use of the dual reverse loop for both protection and distributed queued access to the channels is the distributed queued dual-bus (DQDB) [45] system developed by QPSX in Australia and standardized by the IEEE as an 802.6 metropolitan area network. Further examples of loop structures and applications are provided in Chapter 8 where SONET rings are discussed.

REFERENCES

1 F. deJager and M. Christiaens, "A Fast Automatic Equalizer for Data Links," *Philips Technical Review*, Vol. 36, 1977, pp. 10–24.

2 K. Azadet and C. J. Nicole, "Low-Power Equalizer Architectures for High-Speed Modems," *IEEE Communications Magazine*, October, 1998, pp. 118–126.

3 F. D. Waldhauer, "A 2-Level, 274 Mb/s Regenerative Repeater for T4M," *IEEE International Communications Conference*, 1975, pp. 48-13–48-17.

4 N. Karlovac and T. V. Blalock, "An Investigation of the Count Rate Performance of Baseline Restorers," paper presented at the Nuclear Sciences Symposium, Washington, DC, 1974.

5 F. D. Waldhauer, "Quantized Feedback in an Experimental 280-Mb/s Digital Repeater for Coaxial Transmission," *IEEE Transactions on Communications*, Jan. 1974, pp. 1–5.

6 J. Salz, "Optimum Mean Square Decision Feedback Equalization," *Bell System Technical Journal*, Oct. 1973, pp. 1341–1373.

7 J. M. Cioffi, W. L. Abbott, H. K. Thapar, C. M. Melas, and K. D. Fisher, "Adaptive Equalization in Magnetic-Disk Storage Channels," *IEEE Communications Magazine*, Feb. 1990, pp. 15–29.

8 S. Sugimoto, K. Hayashi, and F. Mano, "Design of 2B1Q Transceiver for ISDN Subscriber Loops," *IEEE International Conference on Communications*, June 1989, pp. 8.1.1–8.1.5.

9 M. R. Aaron, "PCM Transmission in the Exchange Plant," *Bell System Technical Journal*, Jan. 1962, pp. 99–141.

10 Technical Staff, Bell Telephone Laboratories, *Transmission Systems for Communications*, Western Electric Co., Winston Salem, NC, 1971, p. 668.

11 "1.544 Mbps Digital Service," Bell System Technical Reference Publication No. 41451, May 1977.

12 G. D. Forney, "The Viterbi Algorithm," *Proceedings of IEEE*, Mar. 1973, pp. 268–278.

13 V. I. Johannes, A. G. Kaim, and T. Walzman, "Bipolar Pulse Transmission with Zero Extraction," *IEEE Transactions on Communications*, Apr. 1969, pp. 303–310.

14 "The D3 Channel Bank Compatibility Specification—Issue 3," Technical Advisory No. 32, American Telephone and Telegraph Company, Basking Ridge, NJ, Oct. 1977.

15 B. Johnston and W. Johnston, "LD-4 A Digital Pioneer in Action," *Telesis*, Vol. 5, No. 3, June 1977, pp. 66–72.

16 J. H. Davis, "T2: A 6.3 Mb/s Digital Repeatered Line," *IEEE International Communications Conference*, 1969, pp. 34-9–34-16.

17 Recommendation G.703, CCITT *Orange Book*, Vol. 3, No. 2.

18 R. M. Wienski, "Evolution to ISDN within the Bell Operating Companies," *IEEE Communications Magazine*, Jan. 1984, pp. 33–41.

19 American National Standard: Digital Hierarchy—Formats Specifications, ANSI T1.107-1988, Editorial Revised Apr. 1989.

20 J. M. Sipress, "A New Class of Selected Ternary Pulse Transmission Plans for Digital Transmission Lines," *IEEE Transactions on Communication Technology*, Sept. 1965, pp. 366–372.

21 E. E. Schnegelberger and P. T. Griffiths, "48 PCM Channels on T1 Facilities," *National Electronics Conference*, 1975, pp. 201–205.

22 P. A. Franaszek, "Sequence-State Coding for Digital Transmission," *Bell System Technical Journal*, Dec. 1967, pp. 143–157.

23 J. O. Azaret, J. F. Graczyk, W. M. Hauser, and J. V. Mahala, "New Improved T-Carrier Doubles Capacity, Cuts Costs," *Bell Labs Record*, July 1985, pp. 26–31.

24 J. W. Lechleider, "Line Codes for Digital Subscriber Lines," *IEEE Communications Magazine*, Sept. 1989, pp. 25–31.

25 A. Lender, "The Duobinary Technique for High Speed Data Transmission," *IEEE Transactions on Communication Electronics*, May 1963, pp. 214–218.

26 A. Lender, "Correlative Level Coding for Binary Data Transmission," *IEEE Spectrum*, Feb. 1960, pp. 104–110.

27 P. Kabal and S. Pasupathy, "Partial Response Signaling," *IEEE Transactions on Communications*, Sept. 1975, pp. 921–934.

28 A. M. Gerrish and R. D. Howson, "Multilevel Partial Response Signaling," *International Communications Conference*, 1967, p. 186.

29 D. W. Jurling and A. L. Pachynski, "Duobinary PCM System Doubles Capacity of T1 Facilities," *International Communications Conference*, 1977, pp. 32.2-297–32.2-301.

30 T. Seaver, "An Efficient 96 PCM Channel Modem for 2 GHz FM Radio," *National Telecommunications Conference*, 1978, pp. 38.4.1–38.4.5.

31 "Special Issue: 1ARDS," *Bell System Technical Journal*, Mar. 1984.

32 M. B. Brilliant, "Observations of Errors and Error Rates on T1 Digital Repeatered Lines," *Bell System Technical Journal*, Mar. 1978, pp. 711–746.

33 "Carrier-to-Customer Installation—DS1 Metallic Interface," ANSI T1.403-1989, American National Standards Institute.

34 N. S. Bergano, F. W. Kerfoot, and C. R. Davidson, *IEEE Photonics Technology Letters*, Vol. 5, 1993, pp. 304–306.

35 S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

36 L. H. Charles Lee, *Convolutional Coding: Fundamentals and Applications*, Artech House, Norwood, MA, 1997.

37 Technical Staff, Bell Telephone Laboratories, *Engineering and Operations in the Bell System*, Western Electric Co., Indianapolis, 1978.

38 D. E. Dodds, L. R. Button, and S.-M. Pan, "Robust Frame Synchronization for Noisy PCM Systems," *IEEE Transactions on Communications*, May 1985, pp. 465–469.

39 D. E. Dodds, S.-M. Pan, and A. G. Wacker, "Statistical Distribution of PCM Framing Times," *IEEE Transactions on Communications*, Nov. 1988, pp. 1236–1241.

40 R. A. Scholtz, "Frame Synchronization Techniques," *IEEE Transactions on Communications*, Aug. 1980, pp. 1204–1212.

41 D. T. R. Munhoz, J. R. B. de Marca, and D. S. Arantes, "On Frame Synchronization of PCM Systems," *IEEE Transactions on Communications*, Aug. 1980, pp. 1213–1218.

42 M. Anderson and O. Petruchka, "Apparatus and Methods for Recovering Alignment from a Non-Ideally Encoded Multi-bit Digital Signal," U.S. Patent No. 4,955,037, Sept. 4, 1990.

43 K. Stauffer and A. Brajkovic, "DS-1 Extended Superframe Format and Related Performance Issues," *IEEE Communications Magazine*, Apr. 1989, pp. 19–23.

44 *C-System General Description*, Radio Company Report No. 523-0561697-20173R, Collins Radio Company, Dallas, TX, May 1, 1970.

45 "New Proposal Extends the Reach of Metro Area Nets," *Data Communications*, Feb. 1988, pp. 139–145.

PROBLEMS

4.1 If the transmitter and receiver of an asynchronous transmission system utilize clock sources that are accurate to one part in 10^3, determine the maximum number of bits in a codeword if the maximum sample timing error is 20% of a pulse

interval. Assume the sample clock is eight times the bit rate and that the start bit is determined by counting four sample times from the beginning of the start bit.

4.2 Using the symbols +, 0, and – to represent a positive pulse, no pulse, and a negative pulse, respectively, determine the following line code sequences of the binary data sequence

0110 10000100011000000010

(a) Bipolar with most recent pulse being positive
(b) Bipolar with most recent pulse being negative
(c) Pair selected ternary beginning in the negative mode
(d) B3ZS with a +0+ substitution having just been made
(e) B6ZS with the most recent pulse being positive

4.3 Assume that two identical cable systems are used for digital transmission with equal pulse amplitudes. One system uses conventional bipolar signaling and the other uses PST. Compare the crosstalk levels of the two systems. (Assume that 1's and 0's are equally probable.)

4.4 A digital transmission system operating at an error rate of 10^{-6} is to have its data rate increased by 50%. Assuming that the same transmitted power is to be used in the second system, what is the new error rate?

4.5 What is the average reframe time of a D3 channel bank (using SF framing) from a random starting point? What is the maximum average reframe time of a D3 channel bank? (Assume 1's and 0's in the message traffic are equally likely.)

4.6 Repeat Problem 4.5 for the primary TDM multiplex signal specified by CCITT.

4.7 A TDM system operating at 2 Mbps is to have a maximum average reframe time of 20 msec. What is the maximum possible frame length if framing is established with a bit-by-bit frame search? (Assume that 1's and 0's in message channels are equally likely.)

4.8 A T1 transmission system using a D1 frame format is to have an average reframe time (from a random starting point) of 10 msec. How large a block of bit positions must be examined in parallel to achieve the desired result?

4.9 What is the expected framing time for a T1 line (D3 frame format) if the framing strategy is bit by bit and the data stream has 60% 1's and 40% 0's?

4.10 What is the average pulse density of 4B3T coding (Assume 1's and 0's are equally likely.)

4.11 A TDM transmission link using 4B3T coding can transmit 32 64-kbps voice channels using the same symbol rate as a T1 line (1544 kbps). Assuming a fixed Gaussian noise environment, how much must the average transmit power of the 4B3T system be increased to provide the equivalent error rate of a bipolar code?

4.12 Assume that crosstalk interference in a multipair cable system produces an effect equivalent to Gaussian noise at an equal power level. Using an error rate of 10^{-6} as a design objective, determine the effective degradation of the crosstalk on binary (polar) NRZ coding under each of the following conditions. (The effective

degradation is determined as the increased transmit power, in decibels, required to achieve the desired error rate.)

(a) The crosstalk level is 16 dB below the average signal level, but the crosstalk is to be overcome on only one pair (i.e., all other pairs stay at a power level for 10^{-6} BER with no crosstalk).

(b) The crosstalk level is 16 dB below the average signal level, but the effects of the crosstalk are to be overcome on all pairs. (*Hint*: Use signal-power-to-noise-power ratios, not E_b/N_0.)

4.13 Repeat Problem 4.12 for bipolar coding.

4.14 How much does −18 dB of crosstalk degrade the error rate performance of a polar binary NRZ signal for a BER of 10^{-7}? (Assume that all transmitters are at equal power levels.) If crosstalk increases 15 dB per decade in frequency, what is the relative crosstalk level of a four-level NRZ code carrying the same data rate? What is the overall performance penalty of the four-level system compared to the two-level system?

4.15 How many distinct codewords of length 4 can be constructed from ternary symbols? How many of these codewords contain an equal number of positive and negative pulses? How many codewords can be used to ensure a minimum of one timing pulse occurs in each codeword?

4.16 Can a set of ternary codewords of length 8 be constructed to encode binary data using exactly four pulses per 8-bit word and containing equal numbers of positive and negative pulses?

4.17 Given the input sequence +1, −3, +1, −1, +3, +3, −3 of signal levels, determine the sequence of output signal levels for each of the following correlative encodings.

(a) $1 + D$ encoder
(b) $1 - D$ encoder
(c) $1 - D^2$ encoder

4.18 What is the probability of a CRC error in a DS1 signal with ESF framing if the random BER is 10^{-7}?

4.19 What is the minimum and maximum length of a correctable burst of errors in an RS(204, 188) code where each symbol is an 8-bit byte?

4.20 Determine the probability of failure of an RS(7, 2) code operating in a random-error environment with a bit error probability of 10^{-4}. Assume each symbol is an 8-bit byte.

4.21 Determine the probability of failure of an RS(31, 15) code operating in a random-error environment with a bit error probability of 10^{-3}. Assume each symbol is an 8-bit byte.

4.22 Determine the parity error pattern that results in a bit-by-bit constraint length convolutional decoder if two data (B) bits in a row are corrupted and detected in error but no other received bits are in error. Use the following bit sequence for reference, where P_0 is the first received data bit and B_1 and B_2 are in error. P_0B_0, P_1B_1, P_2B_2, P_3B_3.

5

DIGITAL SWITCHING

Of the three basic elements in a communications network (terminals, transmission media, and switches), switches are the most invisible to the users yet represent the most important elements in terms of available service offerings. As mentioned in Chapter 1, a major milestone was established in 1965 when stored-program control switching was first introduced into the U.S. public telephone network. Stored-program control provides the means for implementing many innovative user services and for greatly simplifying switch administration and maintenance.

The use of computers to control the switching functions of a central office led to the designation "electronic" switching [e.g., electronic switching system (ESS) or electronic automatic exchange (EAX)]. However, the switching matrices of these first-generation electronic switches are actually electromechanical in nature. The first use of electronic switching matrices occurred in France in 1971, when digital switching was applied to an end office environment. Ironically, these first digital switches did not use stored-program control. Digital electronic switching matrices were first introduced into the U.S. public network in 1976 with AT&T's No. 4 ESS digital toll switch.

In the late 1970s digital class 5 switches began to be installed in the United States, mostly in smaller end offices as replacements for step-by-step switches. At that time most major metropolitan switching offices already had stored-program control from No. 1 ESS or No. 1 EAX machines. Because the digital toll and end office switching machines were initially installed in predominantly analog transmission environments, their digital matrix provided no direct benefit to network customers. The motivation for the digital machines was reduced costs for the operating companies: lower maintenance, reduced floor space, simplified expansion, and lower manufacturing costs [1].

By the mid-1980s the interoffice transmission environment had changed to be almost exclusively digital. Thus analog-to-digital conversion costs moved from being associated with digital transmission links to being associated with analog switches, thereby further sealing the fate of analog toll or tandem switching technology. At this time frame, end office transmission environments also began the swing to a digital environment. Interoffice trunks were already digital, digital loop carrier systems became

cost effective in metropolitan applications, remotely controlled switching modules with digital fiber interconnect became common for service to outlying communities, digital cross-connect systems (DCSs) were being deployed, and the feeder portion of the subscriber loop plant began to use fiber.

The cost penalties of interfacing to a digital transmission environment and higher maintenance costs led to the older electromechanical class 5 switches being replaced by digital machines when expansion or consolidation of an office occurred. The ability to offer ISDN services was a lesser motivation for changing to digital switches because low-cost fiber transmission in conjunction with DCS systems allows provisioning of these services from other offices.

This chapter describes the basic operation and implementation of digital time division switching as applied to PBXs, end offices, toll switches, and cross connects. Before digital switching is discussed, certain basic switching concepts and terminology are introduced.

5.1 SWITCHING FUNCTIONS

Obviously, the basic function of any switch is to set up and release connections between transmission channels on an "as-needed basis." The structure and operation of a switch depend on particular applications. Three switching categories for voice circuits are local (line-to-line) switching, transit (tandem) switching, and call distribution.

The most common switching function involves direct connections between subscriber loops at an end office or between station loops at a PBX. These connections inherently require setting up a path through the switch from the originating loop to a specific terminating loop. Each loop must be accessible to every other loop. This level of switching is sometimes referred to as line switching.

Transit connections require setting up a path from a specific incoming (originating) line to an outgoing line or trunk group. Normally, more than one outgoing circuit is acceptable. For example, a connection to an interoffice trunk group can use any one of the channels in the group. Hence transit switching structures can be simplified because alternatives exist as to which outgoing line is selected. Furthermore, it is not even necessary that every outgoing line be accessible from every incoming line. Transit switching functions are required by all switching machines in the telephone network. Some machines such as remote concentrators and toll or tandem switches service only transit traffic (e.g., do not provide local connections). These concepts are illustrated in Figure 5.1.

Call distributors are often implemented with the same basic equipment as PBXs. The mode of operation (software) is significantly different, however, in that incoming calls can be routed to any available attendant. Normally, the software of an automatic call distributor (ACD) is designed to evenly distribute the arriving calls among the attendants. Although it is not an inherent requirement that every incoming line (trunk) be connectable to every attendant, call distributors are normally designed to provide accessibility to all attendants. Furthermore, it is often desirable that nonblocking op-

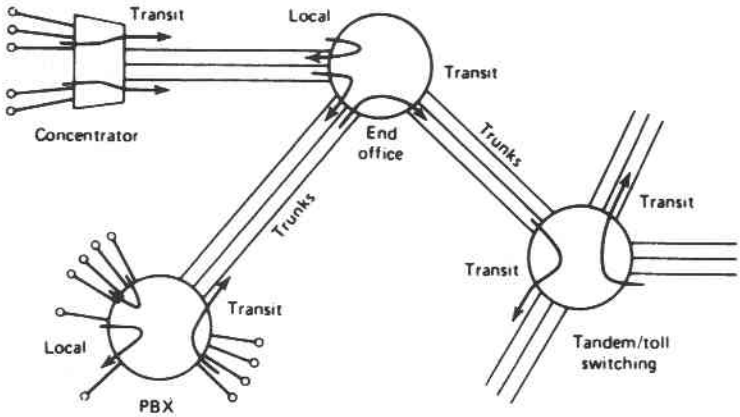


Figure 5.1 Local and transit traffic switching examples.

erations be provided. (No matter what switch paths are in use, a new request can be serviced by any available attendant.)

5.2 SPACE DIVISION SWITCHING

Conceptually, the simplest switching structure is a rectangular array of crosspoints, as shown in Figure 5.2. This switching matrix can be used to connect any one of N inlets to any one of M outlets. If the inlets and outlets are connected to two-wire circuits, only one crosspoint per connection is required.*

Rectangular crosspoint arrays are designed to provide intergroup (transit) connections only, that is, from an inlet group to an outlet group. Applications for this type of an operation occur in the following:

1. Remote concentrators
2. Call distributors
3. Portion of a PBX or end office switch that provides transit switching
4. Single stages in multiple-stage switches

In most of the foregoing applications, it is not necessary that the inlets be connectable to every outlet. In situations involving large groups of outlets, considerable savings in total crosspoints can be achieved if each inlet can access only a limited number of outlets. When such a situation occurs, "limited availability" is said to exist. By overlapping the available outlet groups for various inlet groups, a technique called "grading" is established. An example of a graded switching matrix is shown in Figure 5.3. Notice

*In fact, two (and sometimes three) switching contacts are associated with each crosspoint of a two-wire switch. Since these contacts are part of a single unit and operate in unison, they are considered a single crosspoint.

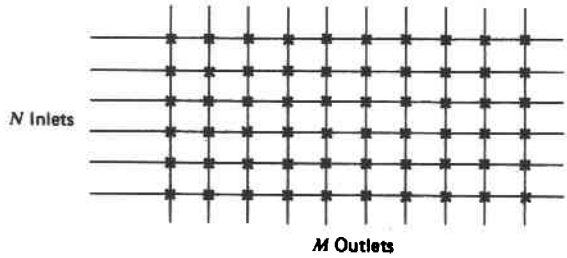


Figure 5.2 Rectangular crosspoint array.

that if outlet connections are judiciously chosen, the adverse effect of limited availability is minimized. For example, if inlets 1 and 8 in Figure 5.3 request a connection to the outlet group, outlets 1 and 3 should be chosen instead of outlets 1 and 4 to avoid future blocking for inlet 2.

Graded switching structures were often used for access to large trunk groups in electromechanical switches where crosspoints were expensive and individual switching modules were limited in size. Gradings were also used in individual switching stages of large multiple-stage switches where more than one path to any particular outlet exists. Because very large digital matrices can be implemented with full accessibility, graded switch structures are no longer necessary.

Intragroup switching, as in line-to-line switching, requires each line to be connectable to every other line. Thus full availability from all inlets to all outlets of the switching matrix is required. Figure 5.4 shows two matrix structures that can be used to fully interconnect two-wire lines. The dashed lines indicate that corresponding inlets and outlets of two-wire switching matrices are actually connected together to provide bidirectional transmission on two-wire circuits. For purposes of describing switching matrices, however, it is convenient to consider the inlets and outlets of two-wire switching matrices as being distinct.

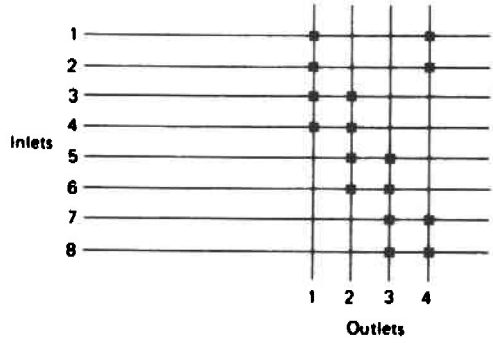


Figure 5.3 Graded rectangular switching matrix.

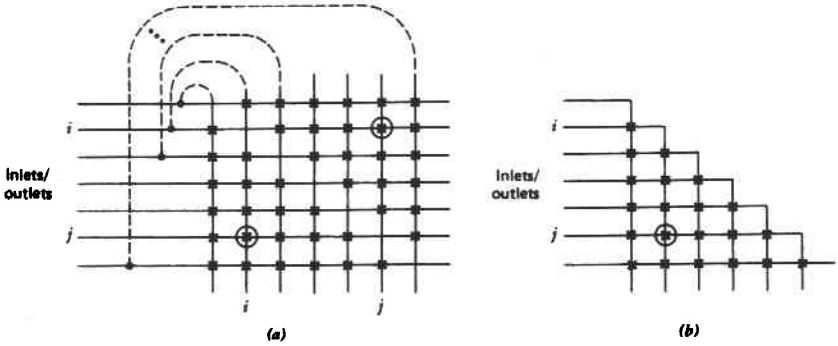


Figure 5.4 Two-wire switching matrices: (*a*) square; (*b*) triangular (folded).

Both structures in Figure 5.4 allow any connection to be established by selecting a single crosspoint. However, the square matrix, which is also called a two-sided matrix, allows any particular connection to be established in two ways. For example, if input link i is to be connected to input link j, the selected crosspoint can be at the intersection of inlet i and outlet j—or at the intersection of inlet j and outlet i. For simplicity these crosspoints are referred to as (i, j) and (j, i), respectively. In a typical implementation, crosspoint (i, j) is used when input i requests service, and crosspoint (j, i) is used when input j requests service.

In the triangular matrix of Figure 5.4 the redundant crosspoints are eliminated. The crosspoint reduction does not come without complications, however. Before setting up a connection between switch input i and switch input j, the switch control element must determine which is larger: i or j. If i is larger, crosspoint (i, j) is selected. If i is smaller, crosspoint (j, i) must be selected. With computer-controlled switching, the line number comparison is trivial. In the older, electromechanically controlled switches, however, the added complexity of the switch control is more significant.

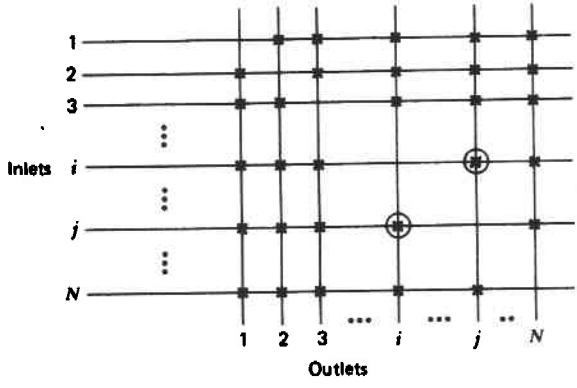


Figure 5.5 Four-wire switching matrix.

Switching machines for four-wire circuits require separate connections for the go and return branches of a circuit. Thus two separate connections must be established for each service request. Figure 5.5 depicts a square-matrix structure used to provide both connections. The structure is identical to the square matrix shown in Figure 5.4 for two-wire switching. The difference, however, is that corresponding inlets and outlets are not connected to a common two-wire input. All of the inlets of the four-wire switch are connected to the wire pair carrying the incoming direction of transmission, and all of the outlets are connected to the outgoing pairs. When setting up a connection between four-wire circuits i and j, the matrix in Figure 5.5 must select both crosspoints (i, j) and (j, i). In actual operation these two crosspoints may be selected in unison and implemented as a common module.

5.2.1 Multiple-Stage Switching

In the switching structures described to this point, an inlet is connected directly to an outlet through a single crosspoint. (Four-wire switches use two crosspoints per connection, but only one for an inlet-to-outlet connection.) For this reason, these switching structures are referred to as "single-stage" switches. Single-stage switches have the property that each individual crosspoint can only be used to interconnect one particular inlet–outlet pair. Since the number of inlet–outlet pairs is equal to $N(N-1)/2$ for a triangular array, and $N(N-1)$ for a square array, the number of crosspoints required for a large switch is prohibitive. Furthermore, the large number of crosspoints on each inlet and outlet line imply a large amount of capacitive loading on the message paths. Another fundamental deficiency of single-stage switches is that one specific crosspoint is needed for each specific connection. If that crosspoint fails, the associated connection cannot be established. (An exception is the square, two-wire switch that has a redundant crosspoint for each potential connection. Before the redundant crosspoint could be used as an alternate path, however, the inlet-oriented selection algorithm would have to be modified to admit outlet-oriented selection.)

Analysis of a large single-stage switch reveals that the crosspoints are very inefficiently utilized. Only one crosspoint in each row or column of a square switch is ever in use, even if all lines are active. To increase the utilization efficiency of the crosspoints, and thereby reduce the total number, it is necessary that any particular crosspoint be usable for more than one potential connection. If crosspoints are to be shared, however, it is also necessary that more than one path be available for any potential connection so that blocking does not occur. The alternate paths serve to eliminate or reduce blocking and also to provide protection against failures. The sharing of crosspoints for potential paths through the switch is accomplished by multiple-stage switching. A block diagram of one particular form of a multiple-stage switch is shown in Figure 5.6.

The switch of Figure 5.6 is a three-stage switch in which the inlets and outlets are partitioned into subgroups of N inlets and N outlets each. The inlets of each subgroup are serviced by a rectangular array of crosspoints. The inlet arrays (first stage) are $n \times k$ arrays, where each of the k outputs is connected to one of the k center-stage ar-

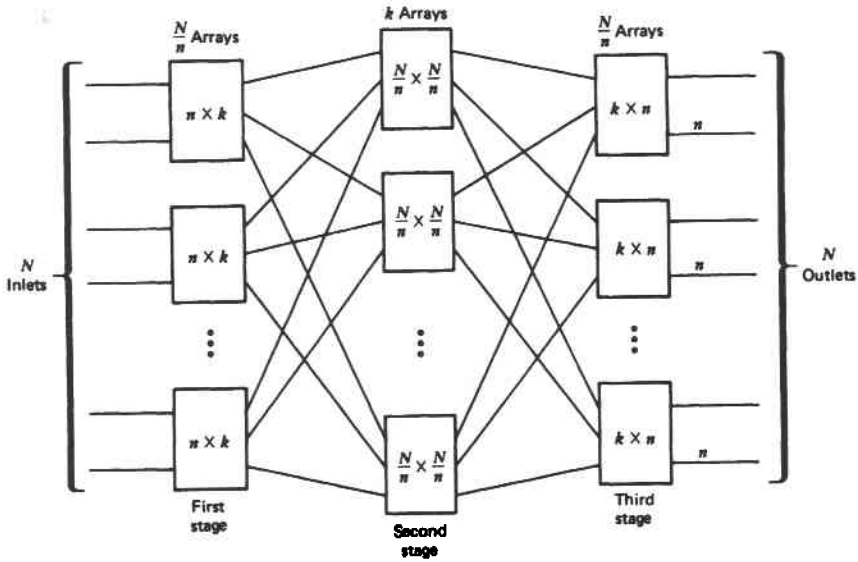


Figure 5.6 Three-stage switching matrix.

rays. The interstage connections are often called junctors. The third stage consists of $k \times n$ rectangular arrays that provide connections from each center-stage array to the groups of n outlets. All center-stage arrays are $N/n \times N/n$ arrays that provide connections from any first-stage array to any third-stage array. Notice that if all arrays provide full availability, there are k possible paths through the switch for any particular connection between inlets and outlets. Each of the k paths utilizes a separate center-stage array. Thus the multiple-stage structure provides alternate paths through the switch to circumvent failures. Furthermore, since each switching link is connected to a limited number of crosspoints, capacitive loading is minimized.

The total number of crosspoints N_X required by a three-stage switch, as shown in Figure 5.6, is

$$N_X = 2Nk + k\left(\frac{N}{n}\right)^2 \tag{5.1}$$

where $N =$ number of inlets–outlets
$n =$ size of each inlet–outlet group
$k =$ number of center-stage arrays

As is demonstrated shortly, the number of crosspoints defined in Equation 5.1 can be significantly lower than the number of crosspoints required for single-stage switches. First, however, we must determine how many center-stage arrays are needed to provide satisfactory service.

Nonblocking Switches

One attractive feature of a single-stage switch is that it is strictly nonblocking. If the called party is idle, the desired connection can always be established by selecting the particular crosspoint dedicated to the particular input–output pair. When crosspoints are shared, however, the possibility of blocking arises. In 1953 Charles Clos [2] of Bell Laboratories published an analysis of three-stage switching networks showing how many center-stage arrays are required to provide a strictly nonblocking operation. His result demonstrated that if each individual array is nonblocking, and if the number of center stages k is equal to $2n - 1$, the switch is strictly nonblocking.

The condition for a nonblocking operation can be derived by first observing that a connection through the three-stage switch requires locating a center-stage array with an idle link from the appropriate first stage and an idle link to the appropriate third stage. Since the individual arrays themselves are nonblocking, the desired path can be set up any time a center stage with the appropriate idle links can be located. A key point in the derivation is to observe that since each first-stage array has n inlets, only $n - 1$ of these inlets can be busy when the inlet corresponding to the desired connection is idle. If k is greater than $n - 1$, it follows that, at most, $n - 1$ links to center-stage arrays can be busy. Similarly, at most $n - 1$ links to the appropriate third-stage array can be busy if the outlet of the desired connection is idle.

The worst-case situation for blocking occurs (as shown in Figure 5.7) if all $n - 1$ busy links from the first-stage array lead to one set of center-stage arrays and if all $n - 1$ busy links to the desired third-stage array come from a separate set of center-stage arrays. Thus these two sets of center-stage arrays are unavailable for the desired connection. However, if one more center-stage array exists, the appropriate input and output links must be idle, and that center stage can be used to set up the connection. Hence if $k = (n - 1) + (n - 1) + 1 = 2n - 1$, the switch is strictly nonblocking. Substituting

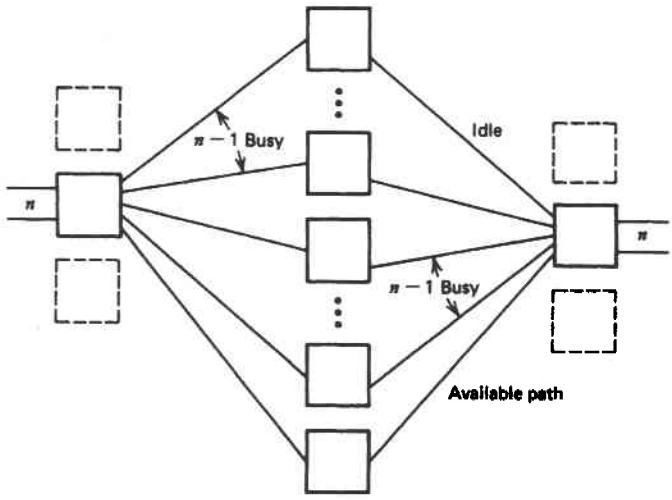


Figure 5.7 Nonblocking three-stage switching matrix.

this value of k into Equation 5.1 reveals that for a strictly nonblocking operation of a three-stage switch

$$N_X = 2N(2n - 1) + (2n - 1)\left(\frac{N}{n}\right)^2 \tag{5.2}$$

As expressed in Equation 5.2, the number of crosspoints in a nonblocking three-stage switch is dependent on how the inlets and outlets are partitioned into subgroups of size n. Differentiating Equation 5.2 with respect to n and setting the resulting expression equal to 0 to determine the minimum reveal that (for large N) the optimum value of n is $(N/2)^{1/2}$. Substituting this value of n into Equation 5.2 then provides an expression for the minimum number of crosspoints of a nonblocking three-stage switch:

$$N_X(\text{min}) = 4N(\sqrt{2N} - 1) \tag{5.3}$$

where N = total number of inlets–outlets.

Table 5.1 provides a tabulation of N_X(min) for various-sized nonblocking three-stage switches and compares the values to the number of crosspoints in a single-stage square matrix. Both switching structures inherently provide four-wire capabilities, a requirement for digital switches because voice digitization implies four-wire circuits.

As indicated in Table 5.1, a three-stage switching matrix provides significant reductions in crosspoints, particularly for large switches. However, the number of crosspoints for large three-stage switches is still quite prohibitive. Large switches typically use more than three stages to provide greater reductions in crosspoints. For example, the No. 1 ESS uses an eight-stage switching matrix that can service up to 65,000 lines. The most significant reductions in crosspoint numbers are achieved not so much from additional stages but by allowing the switch to introduce acceptably low probabilities of blocking.

TABLE 5.1 Crosspoint Requirements of Nonblocking Switches

Number of Lines	Number of Crosspoints for Three-Stage Switch	Number of Crosspoints for Single-Stage Switch
128	7,680	16,256
512	63,488	261,632
2,048	516,096	4.2 million
8,192	4.2 million	67 million
32,768	33 million	1 billion
131,072	268 million	17 billion

5.2.2 Blocking Probabilities: Lee Graphs

Strictly nonblocking switches are rarely needed in most voice telephone networks. Both the switching systems and the number of circuits in interoffice trunk groups are sized to service most requests as they occur, but economics dictates that network implementations have limited capacities that are occasionally exceeded during peak-traffic situations. Equipment for the public telephone network is designed to provide a certain maximum probability of blocking for the busiest hour of the day. The value of this blocking probability is one aspect of the telephone company's grade of service. (Other aspects of grades of service are availability, transmission quality, and delay in setting up a call.)

A typical residential telephone is busy 5–10% of the time during the busy hour. Business telephones are often busy for a larger percentage of their busy hour (which may not coincide with a residential busy hour). In either case, network-blocking occurrences on the order of 1%* during the busy hour do not represent a significant reduction in the ability to communicate since the called party is much more likely to have been busy anyway. Under these circumstances, end office switches and, to a lesser degree, PBXs can be designed with significant reductions in crosspoints by allowing acceptable blocking probabilities.

There are a variety of techniques that can be used to evaluate the blocking probability of a switching matrix. These techniques vary according to complexity, accuracy, and applicability to different network structures. One of the most versatile and conceptually straightforward approaches of calculating blocking probabilities involves the use of probability graphs as proposed by C. Y. Lee [3]. Although this technique requires several simplifying approximations, it can provide reasonably accurate results, particularly when comparisons of alternate structures are more important than absolute numbers. The greatest value of this approach lies in the ease of formulation and the fact that the formulas directly relate to the underlying network structures. Thus the formulations help provide insight into the network structures and how these structures might be modified to change the performance.

In the following analyses we are determining the blocking probabilities of various switching structures using utilization percentages, or "loadings," of individual links. The notation p will be used, in general, to represent the fraction of time that a particular link is in use (i.e., p is the probability that a link is busy). In addition to a utilization percentage or loading, p is also sometimes referred to as an occupancy. The probability that a link is idle is denoted by $q = 1 - p$.

When any one of n parallel links can be used to complete a connection, the composite blocking probability B is the probability that all links are busy†:

$$B = p^n \tag{5.4}$$

*Transmission and switching equipment in the public network is normally designed for even lower blocking probabilities to provide for growth in the traffic volume.

†Equations 5.4 and 5.5 assume each link is busy or idle independently of other links.

When a series of n links are all needed to complete a connection, the blocking probability is most easily determined as 1 minus the probability that they are all available:

$$B = 1 - q^n \tag{5.5}$$

A probability graph of a three-stage network is shown in Figure 5.8. This graph relates the fact that any particular connection can be established with k different paths: one through each center-stage array. The probability that any particular interstage link is busy is denoted by p'. The probability of blocking for a three-stage network can be determined as

$$\begin{aligned} B &= \text{probability that all paths are busy} \\ &= (\text{probability that an arbitrary path is busy})^k \\ &= (\text{probability that at least one link in a path is busy})^k \\ &= (1 - q'^2)^k \end{aligned} \tag{5.6}$$

where $k =$ number of center-stage arrays
$q' =$ probability that an interstage link is idle, $= 1 - p'$.

If the probability p that an inlet is busy is known, the probability p' that an interstage link is busy can be determined as

$$p' = \frac{p}{\beta} \qquad (p < \beta) \tag{5.7}$$

where $\beta = k/n$. Equation 5.7 presents the fact that when some number of inlets (or outlets) are busy, the same number of first-stage outputs (or third-stage inputs) are also busy. However, there are $\beta = k/n$ times as many interstage links as there are inlets or outlets. Hence the percentage of interstage links that are busy is reduced by the factor β.

The factor β is defined as though $k > n$, which implies that the first stage of the switch is providing space expansion (i.e., switching some number of input links to a larger number of output links). Actually, β may be less than 1, implying that the first

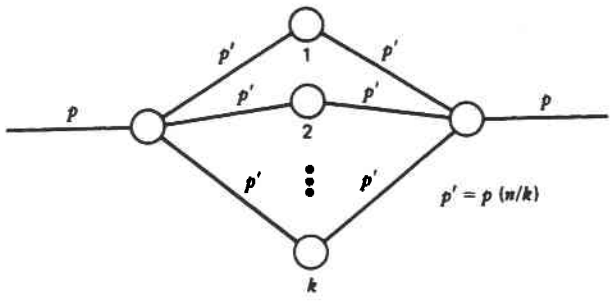


Figure 5.8 Probability graph of three-stage network.

stage is concentrating the incoming traffic. First-stage concentration has been used in end office or large PBX switches where the inlets are lightly loaded (5–10%). In tandem or toll offices, however, the incoming trunks are heavily utilized, and expansion is usually needed to provide adequately low-blocking probabilities.

Substituting Equation 5.7 into Equation 5.6 provides a complete expression for the blocking probability of a three-stage switch in terms of the inlet utilization p:

$$B = \left[1 - \left(1 - \frac{p}{\beta}\right)^2\right]^k \tag{5.8}$$

Table 5.2 tabulates numbers of crosspoints obtained from Equation 5.8 for the same switch sizes presented in Table 5.1. The number of center arrays was chosen in each case to provide a blocking probability on the order of 0.002. The inlet utilization in each example was assumed to be 10%. Notice that the designs with small but finite blocking probabilities are significantly more cost effective than nonblocking designs.

The switch designs in Table 5.2 assume that the inlets are only 10% busy, as might be the case for an end office switch or a PBX. The dramatic savings in crosspoints for large switches is achieved by introducing significant concentration factors ($1/\beta$) into the middle stage. When the inlet utilization is higher (as typically occurs in tandem switches), high concentration factors are not acceptable, and the crosspoint requirements therefore increase. Table 5.3 lists corresponding crosspoint requirements and implementation parameters for inlet loadings of 70%.

The results presented in Tables 5.2 and 5.3 indicate that very large switches still require prohibitively large numbers of crosspoints, even when blocking is allowed. As mentioned previously, very large switches use more than three stages to provide further reductions in crosspoints. Figure 5.9 shows a block diagram of a five-stage switch obtained by replacing every center-stage array in Figure 5.6 with a three-stage array. This particular structure is not optimum in terms of providing a given level of performance with the fewest crosspoints, but it is a useful design because of its modularity. (Furthermore, it is a lot easier to analyze than some other five-stage structures.)

If the middle three stages of a five-stage switch as shown in Figure 5.9 are strictly nonblocking ($k_2 = 2n_2 - 1$), the design provides a savings of 8704 crosspoints in each

TABLE 5.2 Three-Stage Switch Designs for Blocking Probabilities of 0.002 and Inlet Utilization of 0.1

Switch Size, N	n	k	β	Number of Crosspoints	Number of Crosspoints in Nonblocking Design	
128	8	5	0.625	2,560	7,680	($k = 15$)
512	16	7	0.438	14,336	63,488	($k = 31$)
2,048	32	10	0.313	81,920	516,096	($k = 63$)
8,192	64	15	0.234	491,520	4.2 million	($k = 127$)
32,768	128	24	0.188	3.1 million	33 million	($k = 255$)
131,072	256	41	0.160	21.5 million	268 million	($k = 511$)

TABLE 5.3 Three-Stage Switch Designs for Blocking Probabilities of 0.002 and Inlet Utilizations of 0.7

Switch Size N	n	k	β	Number of Crosspoints	Number of Crosspoints in Nonblocking Design	
128	8	14	1.75	7,168	7,680	(k = 15)
512	16	22	1.38	45,056	63,488	(k = 31)
2,048	32	37	1.16	303,104	516,096	(k = 63)
8,192	64	64	1.0	2.1 million	4.2 million	(k = 127)
32,768	128	116	0.91	15.2 million	33 million	(k = 255)
131,072	256	215	0.84	113 million	268 million	(k = 511)

center-stage array of the 32,768-line, three-stage switch designs presented earlier. Hence a little over 1 million crosspoints are saved in the 32,768-line tandem switch design of Table 5.3. Since the middle stages do not introduce blocking, the performance of this five-stage switch is identical to the performance of the three-stage design. Naturally, a more cost-effective design could be obtained by allowing small amounts of blocking in the middle stages. The probability graph of the five-stage switch is shown in Figure 5.10. From this graph, the blocking probability is determined as follows:

$$B = \{1 - (q_1)^2 [1 - (1 - q_2^2)^{k_2}]\}^{k_1} \tag{5.9}$$

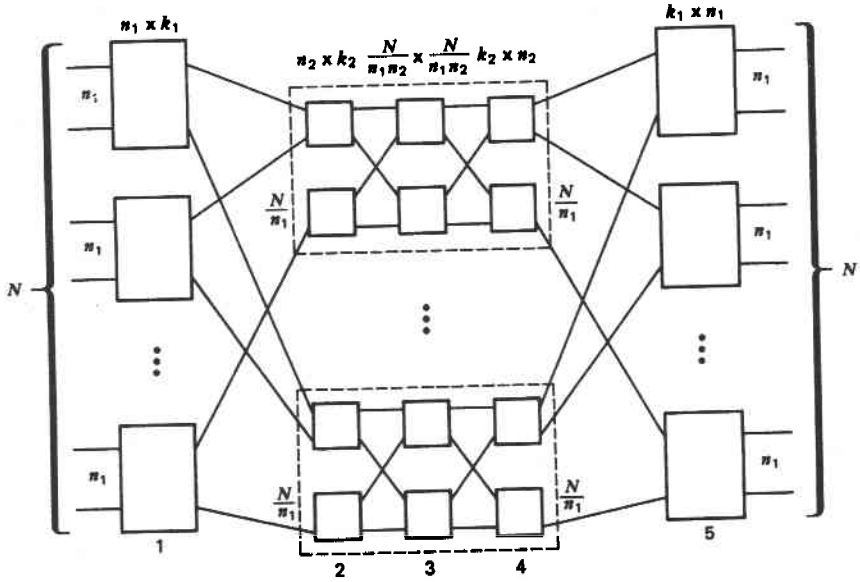


Figure 5.9 Five-stage switching network.

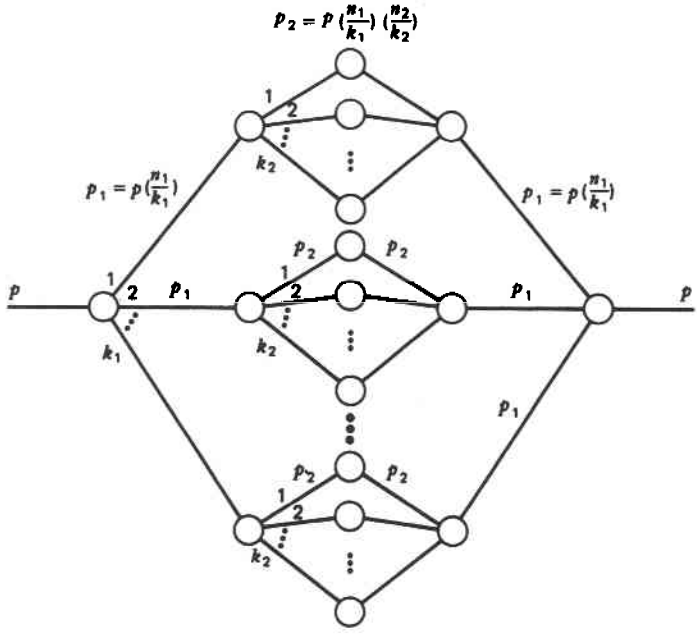


Figure 5.10 Probability graph of five-stage network.

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Even greater crosspoint reductions can, of course, be achieved by using more stages to replace the rather large first- and third-stage arrays. For example, the total number of crosspoints in the 32,000-line switch can be reduced to less than 3 million. The 130,000-line switch is not practical with electromechanical switching matrices but is well within the capabilities of a digital time division switch.

5.2.3 Blocking Probabilities: Jacobaeus

The formulations of blocking probability obtained from probability graphs entail several simplifying assumptions. One of these assumptions involves expressing the composite blocking probability of the alternate paths as the product of the blocking probabilities of each individual path. This step assumes that the individual probabilities are independent. In fact, the probabilities are not independent, particularly when significant amounts of expansion are present. Consider a switching matrix with $k = 2n - 1$. Equation 5.8 produces a finite blocking probability even though the switch is known to be strictly nonblocking. The inaccuracy results because when $2n - 2$ paths are busy, the remaining path is assumed to be busy with a probability of $1 - (q')^2$. In fact, the remaining path is necessarily idle.

In general, when space expansion exists, the assumption of independent individual probabilities leads to an erroneously high value of blocking. The inaccuracy results

because as more and more paths in a switch are found to be busy, the remaining paths are less and less likely to be in use (only a subset of n of the interstage links can ever be busy at any one time).

A more accurate but not exact analysis of multistage switching matrices was presented in 1950 by C. Jacobaeus [4]. Although the analysis is conceptually straightforward, it does involve a considerable amount of manipulation that is not presented here. The resulting equation for a three-stage switch is obtained from reference [5] as

$$B = \frac{(n!)^2}{k!(2n-k)!} p^k (2-p)^{2n-k} \tag{5.10}$$

where $n =$ number of inlets (outlets) per first- (third-) stage array
$k =$ number of second-stage arrays
$p =$ inlet utilization

In the interest of comparing the two methods, Equations 5.6 and 5.10 have been evaluated for three-stage switches with varying amounts of space expansion. The results were obtained for inlet utilization of 0.7 and are presented in Table 5.4.

Table 5.4 reveals that the two analyses are in close agreement for near-unity expansion factors. In fact, if $\beta \equiv 1$, the two formulations produce identical results. As expected, the Lee graph analysis (Equation 5.8) produces overly pessimistic values for the blocking probability when $\beta > 1$.

As another comparison between the two approaches, Table 5.5 is included to demonstrate the use of Equations 5.8 and 5.10 for switches with significant amounts of concentration made possible by a relatively low inlet utilization of 0.1.

Table 5.5 reveals that a Lee graph analysis (Equation 5.8) consistently underestimates the blocking probability when concentration exists. Actually, the Jacobaeus analysis presented in Equation 5.10 also underestimates the blocking probability if large concentration factors and high blocking probabilities are used. When necessary, more accurate techniques can be used for systems with high concentration and high

TABLE 5.4 Comparison of Blocking Probability Analyses ($p = 0.7$)[a]

Number of Center Stages, k	Space Expansion, β	Lee Equation 5.8	Jacobaeus Equation 5.10
14	0.875	0.548	0.598
16	1.0	0.221	0.221
20	1.25	0.014	0.007
24	1.50	3.2×10^{-4}	2.7×10^{-5}
28	1.75	3.7×10^{-6}	7.7×10^{-9}
31[b]	1.94	8.5×10^{-8}	0.1×10^{-12}

[a]Switch size $N = 512$; inlet group size $n = 16$; inlet utilization $p = 0.7$.
[b]Nonblocking.

TABLE 5.5 Comparison of Blocking Probability Analyses ($p = 0.1$)[a]

Number of Center Stages, k	Space Expansion, β	Lee Equation 5.8	Jacobaeus Equation 5.10
6	0.375	0.0097	0.027
8	0.5	2.8×10^{-4}	8.6×10^{-4}
10	0.625	4.9×10^{-6}	1.5×10^{-5}
12	0.75	5.7×10^{-8}	1.4×10^{-7}
14	0.875	4.0×10^{-10}	7.8×10^{-10}
16	1.0	2.9×10^{-12}	2.9×10^{-12}

[a]Switch size $N = 512$; inlet group size $n = 16$; inlet utilization $p = 0.1$.

blocking. However, switches with high blocking probabilities normally have no practical interest so they are not considered here.

Users of PBXs sometimes experienced high blocking probabilities, but blocking in these cases usually arises from too few tie lines to other corporate locations or too few trunk circuits to the public network. The subject of blocking in trunk groups is treated in Chapter 12.

Up to this point, the blocking probability analyses have assumed that a specific inlet is to be connected to a specific outlet. Also, it has been assumed that the requests for service on the individual lines are independent. These assumptions are generally valid for switching one subscriber line to another in an end office switch or for connecting one station to another in a PBX. Neither of these assumptions apply to connections to or from a trunk group.

When connecting to a trunk circuit, any circuit in a trunk group is acceptable. Thus the blocking probability to a specific circuit is only as important as its significance in the overall blocking to the trunk group. The blocking probability to any particular circuit in a trunk group can be relatively large and still achieve a low composite blocking probability to the trunk group as a whole. If the blocking probabilities to the individual trunks are independent, the composite blocking probability is the product of the individual probabilities. However, the paths to the individual trunk circuits normally involve some common links (e.g., the junctors from a first-stage array to all second-stage arrays). For this reason the individual blocking probabilities are normally dependent, which must be considered in an accurate blocking probability analysis.

As an extreme example, consider a case where all trunks in a trunk group are assigned to a single outlet array in a three-stage switch. Since the paths from any particular inlet to all trunks in the group are identical, the ability to select any idle trunk is useless. In practice, the individual circuits of a trunk group should be assigned to separate outlet arrays.

Another aspect of trunk groups that must be considered when designing a switch or analyzing the blocking probabilities involves the interdependence of activity on the

individual circuits within a trunk group. In contrast to individual subscriber lines or PBX stations, individual circuits in a trunk group are not independent in terms of their probabilities of being busy or idle. If some number of circuits in a trunk group are tested and found to be busy, the probability that the remaining circuits are busy is increased. The nature of these dependencies is discussed more fully in Chapter 12. At this point it is only necessary to point out that these dependencies cause increased blocking probabilities if the individual trunks are competing for common paths in the switch. Again, the effect of these dependencies is minimized by assigning the individual trunks to separate inlet–outlet arrays so that independent paths are involved in connections to and from the trunk group. This process is sometimes referred to as decorrelating the trunk circuits.

One last aspect of the blocking probability as a grade of service parameter that must be mentioned involves variations in the loading of the network by individual users. In the design examples for an end office presented earlier, it was tacitly assumed that all subscribers are busy 10% of the time during a busy hour. In fact, some subscribers are active much more than 10% of the time, and other subscribers are active less than 10% of the time. In terms of traffic theory, some subscribers present more than 0.1 erlangs* of traffic to the network, whereas others present less.

When a switch is partitioned into subgroups (as all large switches must be) and the traffic is concentrated by first-stage switching arrays, a few overactive subscribers in one subgroup can significantly degrade service for the other subscribers in the subgroup. It does not matter that the subscribers in some other subgroup may be experiencing lower than average blocking. Their essentially nonblocking service is no compensation for those subscribers experiencing a relatively poor grade of service.

Operating companies have traditionally solved the problem of overactive subscribers by specifically assigning the most active lines (businesses) to separate inlet groups of the switch. Sometimes this procedure requires making traffic measurements to determine which lines are most active and reassigning these lines to different parts of the switch. These procedures fall into the general category of line administration. If the subgroups are large enough, or if the designs provide adequate margin for overactive users, this aspect of line administration can be minimized. One feature of a digital switch that can be utilized in this regard is the ability to design economical switches with very low nominal blocking probabilities so that wide variations in traffic intensities can be accommodated.

Even a modern digital switch can experience loading problems when confronted with extreme traffic conditions. An Internet service provider (ISP) in a metropolitan area may attract an extremely large amount of traffic that all passes through a single class 5 switch. Although the connections to the ISP are lines, as far as the switch is concerned, they actually represent a trunk group with very heavy traffic, so much so that special line administration is required.

*An erlang is a measure of traffic intensity specifying the proportion of time that a device is busy. A circuit is said to carry 0.1 erlang if it is busy 10% of the time.

5.2.4 Folded Four-Wire Switches

Multiple-stage switches can be used for either two- or four-wire switching operations. Figure 5.11 depicts a four-wire connection through a four-stage switch. Notice that two paths must be established for the complete connection. A two-wire connection requires only one path since each outlet is externally connected to its corresponding inlet.

The two paths shown in Figure 5.11 demonstrate a particularly useful relationship: One path is a mirror image of the other path. If the switch diagram is folded about the vertical center line, the paths coincide. Hence this method of setting up connections is sometimes referred to as a folded operation. When all connections in the switch are set up with a folded relationship, several benefits result.

First of all, only one pathfinding operation is needed since the reverse path is automatically available as a mirror image of the forward path. In essence, every crosspoint on one side is paired with another crosspoint in a corresponding array on the opposite side of the switch. Whenever one crosspoint of a pair is used in a connection, the other crosspoint in the pair is also used. For example, the third inlet array in the first stage uses crosspoint (6, 4) to connect its sixth inlet to its fourth outlet (leading to the fourth array of the second stage). The corresponding crosspoint in the third outlet array of the last stage connects its fourth inlet (coming from the fourth array in the fourth stage) to its sixth outlet. In general, crosspoint i, j in one array is paired with crosspoint j, i in the corresponding array on the opposite side of a switch. Since the availability of one crosspoint in a pair ensures the availability of the other, the reverse path is automatically specified and available.

A second advantage of the folded four-wire operation results because the amount of information specifying the status of the switch can be cut in half. Only the status of

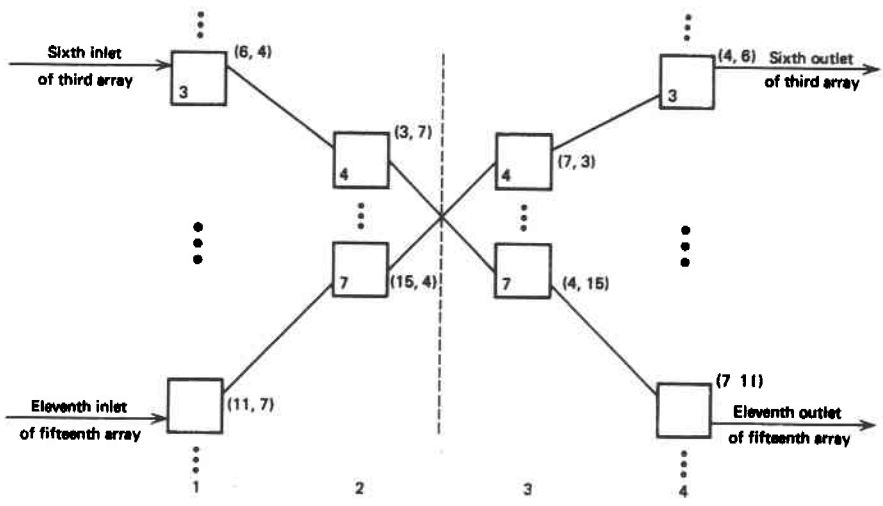


Figure 5.11 Four-wire connection through four-stage switch.

each pair of crosspoints or associated junctor is needed to find an available path through the switch.

A third benefit of the folded structure occurs because the blocking probability is one-half of the probability of finding two paths independently. It might seem that pairing the crosspoints in the described manner would restrict the paths available for a particular connection. On the contrary, the crosspoint pairings merely guarantee that a reverse path is automatically available for any selected path in the forward direction.

The folded operation in the preceding paragraphs referred to a switch with an even number of switching stages. An even number was chosen because the concept is easiest to demonstrate when no center stage is present. The basic approach can be extended to odd numbers of switching stages if the center stage contains an even number of arrays and is folded about a horizontal line at the center of the stage. In this manner, crosspoint i, j in the top center-stage array is paired with crosspoint j, i in the bottom center-stage array, and so on.

5.2.5 Pathfinding

Determining a path through a single-stage switch is virtually automatic since the necessary crosspoint is uniquely specified by the inlet–outlet pair to be connected. In contrast, availability of more than one path in a multiple-stage switch complicates the path selection process. The call processor of the switch must keep track of which potential paths for a particular connection are available in a state store. A pathfinding routine processes the state store information to select an available path. Whenever a new connection is established or an old one released, the state store is updated with the appropriate information.

Pathfinding Times

Pathfinding operations require the use of common equipment and must therefore be analyzed to determine the rate at which connect requests can be processed. The time required to find an available path is directly dependent on how many potential paths are tested before an idle one is found. Some systems can test a number of paths in parallel and thereby shorten the processing time. Since the expected number of potential paths that must be tested to find an idle path is a function of link utilization, pathfinding times unfortunately increase when the common control equipment is busiest.

Assume that the probability of a complete path through the switch being busy is denoted by p. If each of k possible paths through the switch has an equal and independent probability of being busy, the expected number of paths N_p that must be tested before an idle path is found is determined in Appendix A as follows:

$$N_p = \frac{1 - p^k}{1 - p} \tag{5.11}$$

Example 5.1. What is the expected number of potential paths that must be tested to find an idle path in the three-stage, 8192-line switch defined in Table 5.2?

Solution. As indicated in the table, a space expansion factor of 0.234 is used to provide a blocking probability of 0.002. Hence the utilization of each interstage link is 0.1/0.234 = 0.427. The blocking probability of each path through the switch is merely the probability that one of two links in series is busy. Hence the probability $p = 1 - (1 - 0.427)^2 = 0.672$, and the expected number of paths to be tested is

$$N_p = \frac{1 - (0.672)^{15}}{1 - 0.672}$$

$$= 3.04$$

Example 5.1 demonstrates that, on average, only 3 of the 15 potential paths need to be tested before an idle path is found. However, when the switch is experiencing greater than normal traffic loads, the average number of paths tested increases. For example, if the input line utilization increases from 10 to 15%, the blocking probability increases from 0.002 to 0.126, and the expected number of paths to be tested in the pathfinding operation increases from 3 to 4.9.

Although this simple pathfinding example is, by itself, not particularly important, it demonstrates a very important aspect of the loading of common equipment in a switching system (or a network as a whole): Linear increases in the offered traffic produce geometric increases in the utilization of the network resources. If a system is designed and analyzed under nominal traffic conditions and the results are extrapolated to determine total capacity, greatly optimistic estimates often occur. In the pathfinding example the processing requirements increase just when the call processor (stored-program control) is already loaded with greater than average traffic volumes. For further discussion of the effects of traffic loads on the common equipment of a switching system see Schwartz [6].

5.2.6 Switch Matrix Control

When an available path through a common control switching network is determined, the control element of the switch transfers the necessary information to the network to select the appropriate crosspoints. Crosspoint selection within a matrix is accomplished in one of two ways. The control may be associated with the output lines and therefore specify which inputs are to be connected to the associated outputs or the control information may be associated with each input and subsequently specify to which outputs the respective inputs are to be connected. The first approach is referred to as *output-associated control* while the second is called *input-associated control*. These two control implementations are presented in Figure 5.12.

Input-associated control was inherently required in step-by-step switches where the information (dial pulses) arrived on the input link and was used to directly select the output links to each successive stage. In common control systems, however, the address information of both the originating line and the terminating line is simultaneously available. Hence the connection can be established by beginning at the desired

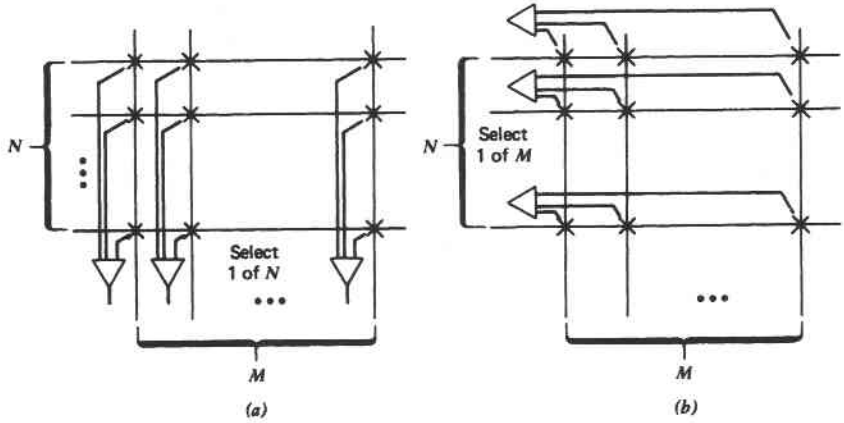


Figure 5.12 Switch matrix control: (*a*) output associated; (*b*) input associated.

outlet and proceeding backward through the switch while selecting inputs to each stage.

The implementation of both types of digital crosspoint arrays using standard components is shown in Figure 5.13. Output-associated control uses a conventional data selector/multiplexer for each matrix output. The number of bits required to control each data selector is $\log_2 N$, where N is the number of inlets. Thus the total number of bits required to completely specify a connection configuration is $M \log_2 N$.

Input-associated control can be implemented using conventional line decoders/demultiplexers. The outputs are commoned using a "wired-or" logic function. Thus the output gates of each decoder circuit must be open-collector or tristate devices if transistor–transistor–logic (TTL) is used. The total number of bits required to specify a connection configuration in this case is $N \log_2 M$.

A significant drawback of input-associated control arises from the need to disable unused inputs to prevent cross connects when another input selects the same output.

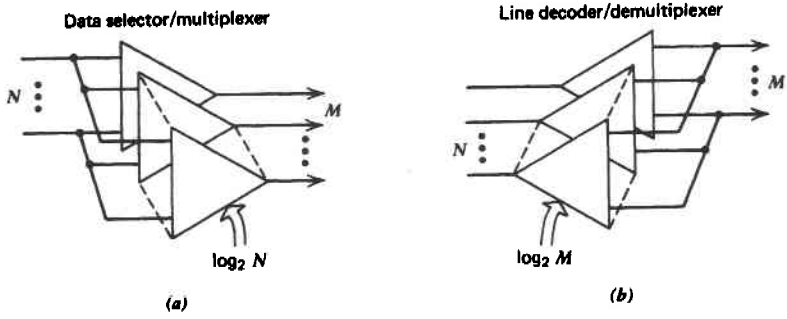


Figure 5.13 Standard component implementation of digital crosspoint array: (*a*) output-associated control; (*b*) input-associated control.

With output-associated control, unused outputs can remain connected to an input without preventing that input from being selected by another output. For this reason and for generally greater speeds of operation, digital switching networks typically use output-associated control. Notice, however, that the total amount of information needed to specify a connection configuration with input-associated control is less than that with output control if the number of inputs N is much smaller than the number of outputs M ($N \log_2 M < M \log_2 N$). Furthermore, input-associated control is more flexible in terms of wired-or (e.g., bus) expansion.

5.3 TIME DIVISION SWITCHING

As evidenced by multiple-stage switching, sharing of individual crosspoints for more than one potential connection provides significant savings in implementation costs of space division switches. In the cases demonstrated, the crosspoints of multistage space switches are shared from one connection to the next, but a crosspoint assigned to a particular connection is dedicated to that connection for its duration. Time division switching involves the sharing of crosspoints for shorter periods of time so that individual crosspoints and their associated interstage links are continually reassigned to existing connections. When the crosspoints are shared in this manner, much greater savings in crosspoints can be achieved. In essence, the savings are accomplished by time division multiplexing the crosspoints and interstage links in the same manner that transmission links are time division multiplexed to share interoffice wire pairs.

Time division switching is equally applicable to either analog or digital signals. At one time, analog time division switching was attractive when interfacing to analog transmission facilities, since the signals are only sampled and not digitally encoded. However, large analog time division switches had the same limitations as do analog time division transmission links: the PAM samples are particularly vulnerable to noise, distortion, and crosstalk. Thus, large electronic switching matrices have always incorporated the cost of digitizing PAM samples to maintain end-to-end signal quality. The low cost of codecs and the prevalence of digital trunk interconnections imply that analog switching is now used in only the smallest of switching systems (e.g., electronic key systems).

5.3.1 Analog Time Division Switching

Although analog time division switching has become obsolete, it is a good starting point to establish the basic principle of time division switching. Figure 5.14 depicts a particularly simple analog time division switching structure. A single switching bus supports a multiple number of connections by interleaving PAM samples from receive line interfaces to transmit line interfaces. The operation is depicted as though the receive interfaces are separate from the respective transmit interfaces. When connecting two-wire analog lines, the two interfaces are necessarily implemented in a common

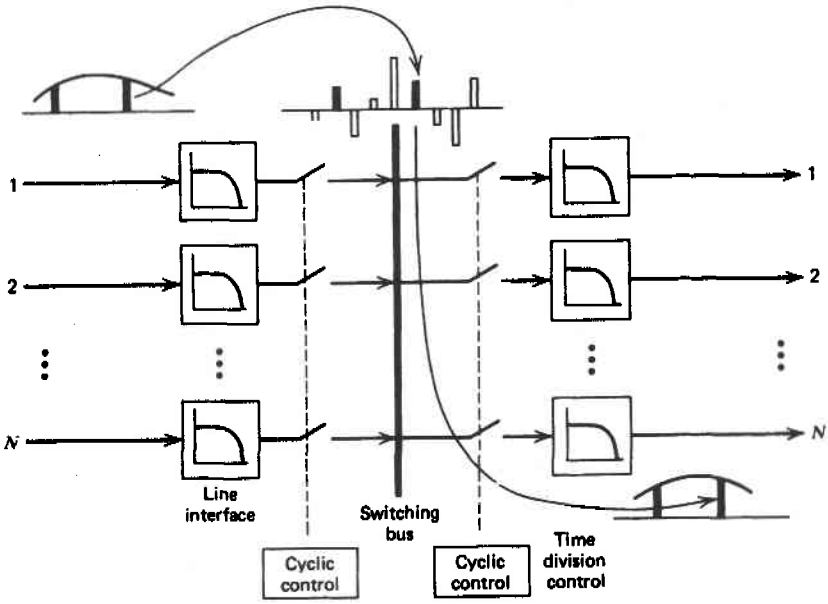


Figure 5.14 Analog time division switching.

module. Furthermore, in some PAM-PBX systems, analog samples were simultaneously transferred in both directions between the interfaces [7].

Included in Figure 5.14 are two cyclic control stores. The first control store controls gating of inputs onto the bus one sample at a time. The second control store operates in synchronism with the first and selects the appropriate output line for each input sample. A complete set of pulses, one from each active input line, is referred to as a frame. The frame rate is equal to the sample rate of each line. For voice systems the sampling rate ranges from 8 to 12 kHz. The higher sampling rates were sometimes used to simplify the bandlimiting filter and reconstructive filters in the line interfaces.

5.3.2 Digital Time Division Switching

The analog switching matrix described in the preceding section is essentially a space division switching matrix. By continually changing the connections for short periods of time in a cyclic manner, the configuration of the space division switch is replicated once for each time slot. This mode of operation is referred to as time multiplexed switching. While this mode of operation can be quite useful for both analog and digital signals, digital time division multiplexed signals usually require switching between time slots as well as between physical lines. Switching between time slots represents a second dimension of switching and is referred to as time switching.

In the following discussion of digital time division switching it is assumed, unless otherwise stated, that the switching network is interfaced directly to digital time divi-

sion multiplex links. This assumption is generally justified since, even when operating in an analog environment, the most cost-effective switch designs multiplex groups of digital signals into TDM formats before any switching operations take place. Thus most of the following discussion is concerned with the internal structures of time division switching networks and possibly not with the structure of an entire switching complex.

The basic requirement of a time division switching network is shown in Figure 5.15. As an example connection, channel 3 of the first TDM link is connected to channel 17 of the last TDM link. The indicated connection implies that information arriving in time slot 3 of the first input link is transferred to time slot 17 of the last output link. Since the voice digitization process inherently implies a four-wire operation, the return connection is required and realized by transferring information from time slot 17 of the last input link to time slot 3 of the first output link. Thus each connection requires two transfers of information, each involving translations in both time and space.

A variety of switching structures are possible to accomplish the transfers indicated in Figure 5.15. All of these structures inherently require at least two stages: a space division switching stage and a time division switching stage. As discussed later, larger switches use multiple stages of both types. Before discussing switching in both dimensions, however, we discuss the characteristics and capabilities of time switching alone.

A Digital Memory Switch

Primarily owing to the low cost of digital memory, time switching implementations provide digital switching functions more economically than space division implementations. Basically, a time switch operates by writing data into and reading data out of a single memory. In the process, the information in selected time slots is interchanged, as shown in Figure 5.16. When digital signals can be multiplexed into a single TDM format, very economical switches can be implemented with time switching alone. However, practical limitations of memory speed limit the size of a time switch so that some amount of space division switching is necessary in large switches. As demonstrated in later sections, the most economical multistage designs usually perform as much switching as possible in the time stages.

The basic functional operation of a memory switch is shown in Figure 5.17. Individual digital message circuits are multiplexed and demultiplexed in a fixed manner

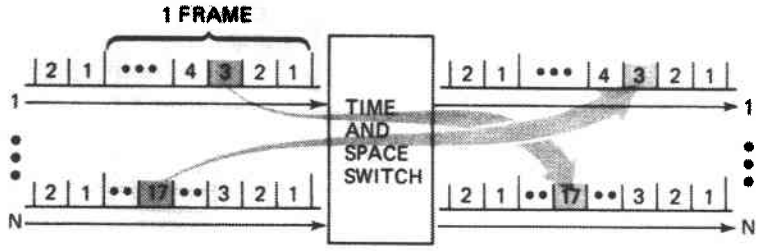


Figure 5.15 Time and space division switching.

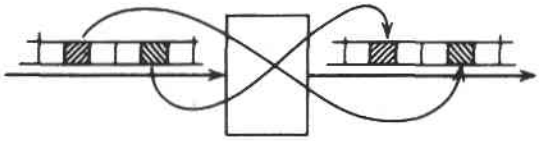


Figure 5.16 Time slot interchange operation.

to establish a single TDM link for each direction of travel. The multiplexing and demultiplexing functions can be considered as part of the switch itself, or they may be implemented in remote transmission terminals. In either case, a byte-interleaved form of multiplexing is required. The asynchronous transmission hierarchy (DS2, DS3, DS4) uses bit interleaving and therefore requires back-to-back demultiplexing and multiplexing operations before switching can be accomplished. In contrast, the SONET transmission format described in Chapter 8 can provide byte interleaving specifically so it can be more directly interfaced to a digital switching system.

The exchange of information between two different time slots is accomplished by a time slot interchange (TSI) circuit. In the TSI of Figure 5.17 data words in incoming time slots are written into sequential locations of the data store memory. Data words for outgoing time slots, however, are read from addresses obtained from a control store. As indicated in the associated control store, a full-duplex connection between TDM channel i and TDM channel j implies that data store address i is read during outgoing time slot j and vice versa. The data store memory is accessed twice during each link time slot. First, some control circuitry (not shown) selects the time slot number

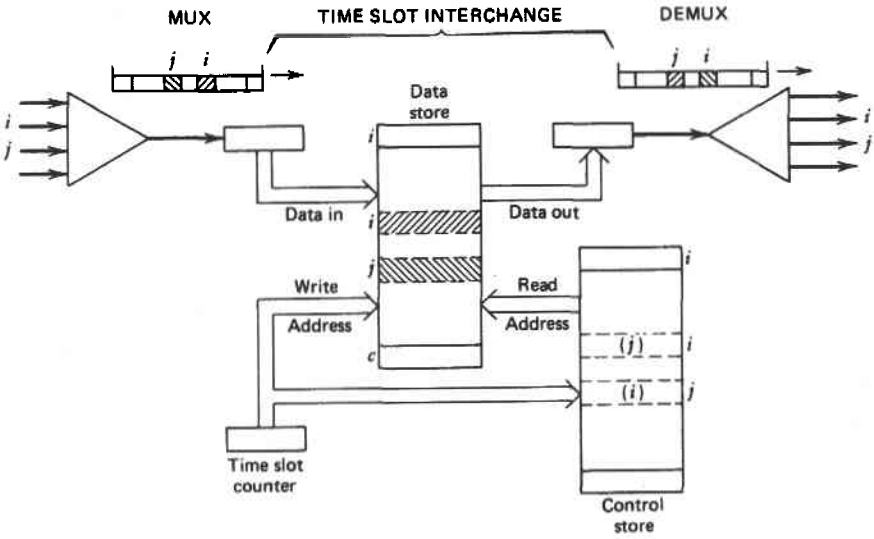


Figure 5.17 MUX/TSI/DEMUX memory switch.

as a write address. Second, the content of the control store for that particular time slot is selected as a read address.

Since a write and a read are required for each channel entering (and leaving) the TSI memory, the maximum number of channels c that can be supported by the simple memory switch is

$$c = \frac{125}{2t_c} \tag{5.12}$$

where 125 is the frame time in microseconds for 8 kHz sampled voice and t_c is the memory cycle time in microseconds.

As a specific example, consider the use of a 15.2 nsec of memory. Equation 5.12 indicates that the memory switch can support 4096 channels (2048 full duplex connections) in a strictly nonblocking mode of operation. The complexity of the switch (assuming digitization occurs elsewhere) is quite modest: The TSI memory stores one frame of data organized as 4096 words by 8 bits each. The control store also requires 4096 words, but each word has a length equal to $\log_2(c)$ (which is 12 in the example). Thus the memory functions can be supplied by 4096×8 and 4096×12 bit random-access memories (RAMs). The addition of a time slot counter and some gating logic to select addresses and enable new information to be written into the control store can be accomplished with a handful of conventional integrated circuits (ICs).

This switch should be contrasted to a space division design (Equation 5.3) that requires more than 1.5 million crosspoints for a nonblocking three-stage switch. Although modern IC technology might be capable of placing that many digital crosspoints in a few ICs, they could never be reached because of pin limitations. As mentioned in Chapter 2, one of the main advantages of digital signals is the ease with which they can be time division multiplexed. This advantage arises for communication between integrated circuits as well as for communication between switching offices.

Time Stages in General

Time switching stages inherently require some form of delay element to provide the desired time slot interchanges. Delays are most easily implemented using RAMs that are written into as data arrive and read from when data are to be transferred out. If one memory location is allocated for each time slot in the TDM frame format, the information from each TDM channel can be stored for up to one full frame time without being overwritten.

There are two basic ways in which the time stage memories can be controlled: written sequentially and read randomly or written randomly and read sequentially. Figure 5.18 depicts both modes of operation and indicates how the memories are accessed to translate information from time slot 3 to time slot 17. Notice that both modes of operation use a cyclic control store that is accessed in synchronism with the time slot counter.

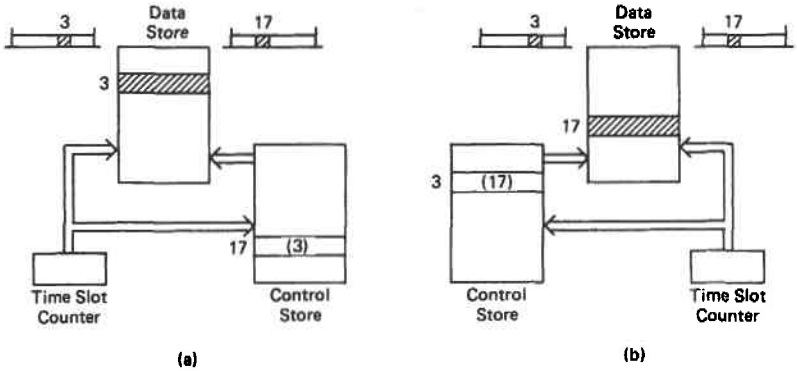


Figure 5.18 Time switch modules: (*a*) sequential writes/random reads; (*b*) random writes/sequential reads.

The first mode of operation in Figure 5.18 implies that specific memory locations are dedicated to respective channels of the incoming TDM link. Data for each incoming time slot are stored in sequential locations within the memory by incrementing a modulo-*c* counter with every time slot. As indicated, the data received during time slot 3 are automatically stored in the third location within the memory. On output, information retrieved from the control store specifies which address is to be accessed for that particular time slot. As indicated, the seventeenth word of the control store contains the number 3, implying that the contents of data store address 3 is transferred to the output link during outgoing time slot 17.

The second mode of operation depicted in Figure 5.18 is exactly the opposite of the first one. Incoming data are written into the memory locations as specified by the control store, but outgoing data are retrieved sequentially under control of an outgoing time slot counter. As indicated in the example, information received during time slot 3 is written directly into data store address 17, where it is automatically retrieved during outgoing TDM channel number 17. Notice that the two modes of time stage operation depicted in Figure 5.18 are forms of output-associated control and input-associated control, respectively. In a multiple-stage design example presented later, it is convenient to use one mode of operation in one time stage and the other mode of operation in another time stage.

5.4 TWO-DIMENSIONAL SWITCHING

Larger digital switches require switching operations in both a space dimension and a time dimension. There are a large variety of network configurations that can be used to accomplish these requirements. To begin with, consider the simple switching structure shown in Figure 5.19. This switch consists of only two stages: a time stage T followed by a space stage S. Thus this structure is referred to a time–space (TS) switch.

The basic function of the time stage is to delay information in arriving time slots until the desired output time slot occurs. At that time the delayed information is trans-

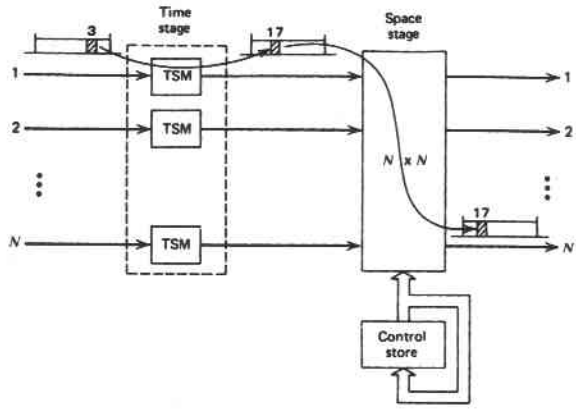


Figure 5.19 Time–space (TS) switching matrix.

ferred through the space stage to the appropriate output link. In the example shown the information in incoming time slot 3 of link 1 is delayed until outgoing time slot 17 occurs. The return path requires that information arriving in time slot 17 of link N be delayed for time slot 3 of the next outgoing frame. Notice that a time stage may have to provide delays ranging from one time slot to a full frame.

Associated with the space stage is a control store that contains the information needed to specify the space stage configuration for each individual time slot of a frame. This control information is accessed cyclically in the same manner as the control information in the analog time division switch. For example, during each outgoing time slot 3, control information is accessed that specifies interstage link number 1 is connected to output link N. During other time slots, the space switch is completely reconfigured to support other connections.

As indicated, a convenient means of representing a control store is a parallel end-around-shift register. The width of the shift register is equal to the number of bits required to specify the entire space switch configuration during a single time slot. The length of the shift register conforms to the number of time slots in a frame. Naturally, some means of changing the information in the control store is needed so that new connections can be established. In actual practice, the control stores may be implemented as RAMs with counters used to generate addresses in a cyclic fashion (as in the time stage control stores shown previously).

Implementation Complexity of Time Division Switches

In previous sections, alternative space division switching structures were compared in terms of the total number of crosspoints required to provide a given grade of service. In the case of solid-state electronic switching matrices, in general, and time division switching, in particular, the number of crosspoints alone is a less meaningful measure of implementation cost. Switching structures that utilize ICs with relatively large numbers of internal crosspoints are generally more cost effective than other structures

that may have fewer crosspoints but more packages. Hence a more relevant design parameter for solid-state switches would be the total number of IC packages. If alternate designs are implemented from a common set of ICs, the number of packages may closely reflect the number of crosspoints.

In addition to the number of crosspoints in space division stages, a digital time division switch uses significant amounts of memory that must be included in an estimate of the overall cost.* The memory count includes the time stage memory arrays and the control stores for both the time stages and the space stages. In large switches, the number of space stage crosspoints can be reduced at the expense of increasing the amount of memory required. Thus, a thorough analysis of implementation complexity requires knowing the relative cost of a crosspoint relative to the cost of a bit of memory. Because a crosspoint is closely associated with an external connection, it costs significantly more than a bit of memory. The use of standard medium-scale integration ICs leads to one crosspoint costing about the same as 100 bits of memory.† Use of custom- or application-specific ICs with integrated memory can change this factor, particularly because large-scale IC packages provide access to many more crosspoints per pin. For purposes of illustrating various design trade-offs, the following analyses of implementation complexity continue to consider crosspoint costs to be 100 times the cost of a memory bit. Depending on the implementation approach, this factor may not be accurate, but minimizing the cost of a matrix is no longer much of a concern, except in extremely large switching systems (e.g., in switches exceeding 100,000 voice channels).

The implementation complexity is expressed as follows:

$$\text{Complexity} = N_X + \frac{N_B}{100} \tag{5.13}$$

where N_X = number of space stage crosspoints
N_B = number of bits of memory

Example 5.2. Determine the implementation complexity of the TS switch shown in Figure 5.19 where the number of TDM input lines $N = 80$. Assume each input line contains a single DS1 signal (24 channels). Furthermore, assume a one-stage matrix is used for the space stage.

Solution. The number of crosspoints in the space stage is determined as

$$N_X = 80^2 = 6400$$

*It is worth noting that digital memories are inherently implemented with at least two crosspoints per bit. In this case the crosspoints are gates used to provide write and read access to the bits. These crosspoints, however, are much less expensive than message crosspoints that are accessed from external circuits.
†See the first or second edition of this book.

(The crosspoints on the main diagonal are included since two channels within a single TDM input may have to be connected to each other.) The total number of memory bits for the space stage control store is determined as

$$\begin{aligned} N_{BX} &= \text{(number of links)(number of control words)(number of bits per control word)} \\ &= (80)(24)(7) \\ &= 13{,}440 \end{aligned}$$

The number of memory bits in the time stage is determined as the sum of the time slot interchange and the control store bits:

$$\begin{aligned} N_{BT} &= \text{(number of links)(number of channels)(number of bits per channel)} \\ &\quad + \text{(number of links)(number of control words)(number of bits per control word)} \\ &= (80)(24)(8) + (80)(24)(5) \\ &= 24{,}960 \end{aligned}$$

Thus the implementation complexity is determined as

$$\text{Complexity} = N_X + \frac{N_{BX} + N_{BT}}{100} = 6784 \text{ equivalent crosspoints}$$

The implementation complexity determined in Example 5.2 is obviously dominated by the number of crosspoints in the space stage. A significantly lower complexity (and generally lower cost) can be achieved if groups of input links are combined into higher level multiplex signals before being switched. The cost of the front-end multiplexers is relatively small if the individual DS1 signals have already been synchronized for switching. In this manner, the complexity of the space stage is decreased appreciably while the overall complexity of the time stage increases only slightly. (See the problems at the end of this chapter.) The implementation costs are reduced proportionately, up to the point that higher speeds dictate the use of a more expensive technology.

A significant limitation of the TS structure of Figure 5.19 occurs when a connection has to be made to a specific channel of an outlet as opposed to any channel of an outlet. A connection to a specific channel is blocked whenever the desired time slot of the inlet TSM is already in use. For example, if time slot 17 of the first inlet is in use for some connection from the first inlet to some link other than link N, the connection from channel 3 of inlet 1 to channel 17 of inlet N cannot be made. Because of this limitation, the TS structure is useful only if the outlets represent trunk groups, which implies any channel of an outlet is suitable.* Applications that require connections to specific channels require additional stages for adequate blocking probability performance.

*A full-duplex connection requires the reverse connection to be established, which adds restrictions to which outlet channels can be used. See the problems at the end of the chapter.

Multiple-Stage Time and Space Switching

As discussed in the preceding section, an effective means of reducing the cost of a time division switch is to multiplex as many channels together as practical and perform as much switching in the time stages as possible. Time stage switching is generally less expensive than space stage switching—primarily because digital memory is much cheaper than digital crosspoints (AND gates). To repeat, the crosspoints themselves are not so expensive, it is the cost of accessing and selecting them from external pins that makes their use relatively costly.

Naturally, there are practical limits as to how many channels can be multiplexed into a common TDM link for time stage switching. When these practical limits are reached, further reductions in the implementation complexity can be achieved only by using multiple stages. Obviously, some cost savings result when a single space stage matrix of a TS or ST switch can be replaced by multiple stages.

A generally more effective approach involves separating the space stages by a time stage, or, conversely, separating two time stages by a space stage. The next two sections describe these two basic structures. The first structure, consisting of a time stage between two space stages, is referred to as a space–time–space (STS) switch. The second structure is referred to as a time–space–time (TST) switch.

5.4.1 STS Switching

A functional block diagram of an STS switch is shown in Figure 5.20. Each space switch is assumed to be a single-stage (nonblocking) switch. For very large switches, it may be desirable to implement the space switches with multiple stages. Establishing a path through an STS switch requires finding a time switch array with an available write access during the incoming time slot and an available read access during the desired outgoing time slot. When each individual stage (S, T, S) is nonblocking, the operation is functionally equivalent to the operation of a three-stage space switch. Hence a probability graph in Figure 5.21 of an STS switch is identical to the probability graph of Figure 5.8 for three-stage space switches. Correspondingly, the blocking probability of an STS switch is

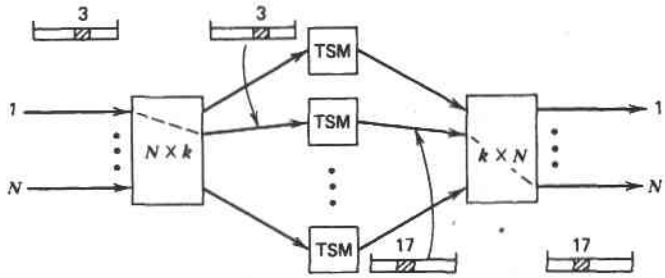


Figure 5.20 Space–time–space (STS) switching structure.

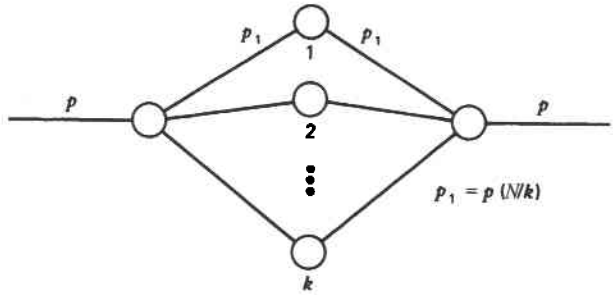


Figure 5.21 Probability graph of STS switch with nonblocking stages.

$$B = (1 - q'^2)^k \tag{5.14}$$

where $q' = 1 - p' = 1 - p/\beta \quad (\beta = k/N)$
k = number of center-stage time switch arrays

Assuming the space switches are single-stage arrays and that each TDM link has c message channels, we may determine the implementation complexity of an STS switch as*

$$\begin{aligned}\text{Complexity} &= \text{number of space stage crosspoints}\\ &\quad + (\text{number of space stage control bits}\\ &\quad + \text{number of time stage memory bits}\\ &\quad + \text{number of time stage control bits})/100\\ &= 2kN + \frac{2kc\log_2 N + kc(8) + kc\log_2 c}{100}\end{aligned} \tag{5.15}$$

Example 5.3. Determine the implementation complexity of a 2048-channel STS switch implemented for 16 TDM links with 128 channels on each link. The desired maximum blocking probability is 0.002 for channel occupancies of 0.1.

Solution. The minimum number of center-stage time switches to provide the desired grade of service can be determined from Equation 5.14 as $k = 7$. Using this value of k, the number of crosspoints is determined as (2)(7)(16) = 224. The number of bits of memory can be determined as (2)(7)(128)(4) + (7)(128)(8) + (7)(128)(7) =

*This derivation assumes output-associated control is used in the first stage and input-associated control is used in the third stage. A slightly different result occurs if the space stages are controlled in different manners.

20,608. Hence the composite implementation complexity is 430 equivalent crosspoints.

The value of implementation complexity obtained in Example 5.3 should be compared to the number of crosspoints obtained for an equivalent-sized three-stage switch listed in Table 5.2. The space switch design requires 81,920 crosspoints while the STS design requires only 430 equivalent crosspoints. The dramatic savings comes about as a result of the voice signals having already been digitized and multiplexed (for transmission purposes). When digital switches were first inserted into an analog environment, the dominant cost of the switch occurred in the line interface. Digital interface costs are much lower than analog interface costs, particularly on a per-channel basis.

5.4.2 TST Switching

The second form of multiple-stage time–space switch is shown in Figure 5.22— the TST switch. Information arriving in a TDM channel of an incoming link is delayed in the inlet time stage until an appropriate path through the space stage is available. At that time the information is transferred through the space stage to the appropriate outlet time stage where it is held until the desired outgoing time slot occurs. Assuming the time stages provide full availability (i.e., all incoming channels can be connected to all outgoing channels), any space stage time slot can be used to establish a connection. In a functional sense the space stage is replicated once for every internal time slot. This concept is reinforced by the TST probability graph of Figure 5.23.

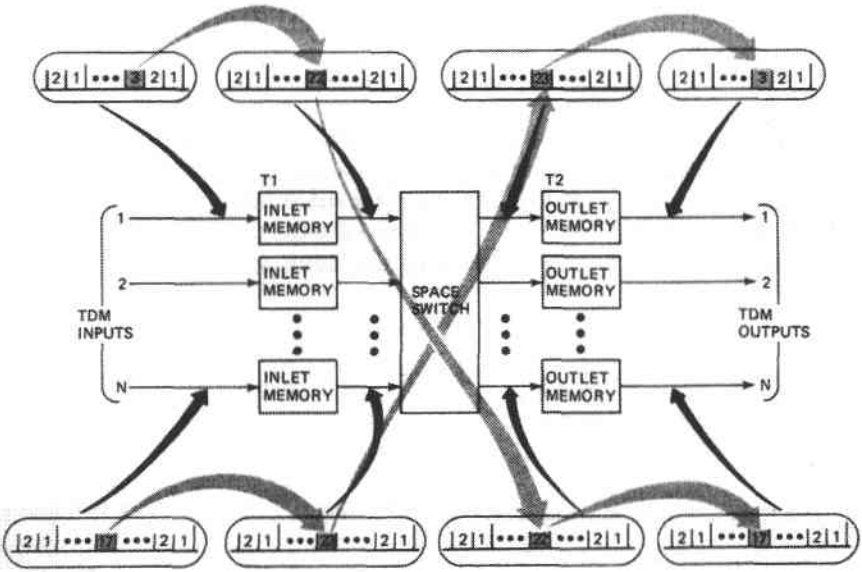


Figure 5.22 Time–space–time (TST) switching structure.

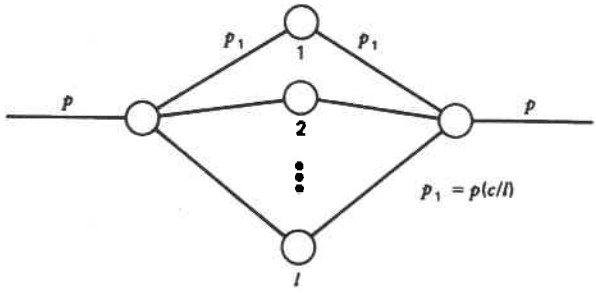


Figure 5.23 Probability graph of TST switch with nonblocking stages.

An important feature to notice about a TST switch is that the space stage operates in a time-divided fashion, independently of the external TDM links. In fact, the number of space stage time slots l does not have to coincide with the number of external TDM time slots c.

If the space stage is nonblocking, blocking in a TST switch occurs only if there is no internal space stage time slot during which the link from the inlet time stage and the link to the outlet time stage are both idle. Obviously, the blocking probability is minimized if the number of space stage time slots l is made to be large. In fact, as a direct analogy of three-stage space switches, the TST switch is strictly nonblocking if $l = 2c - 1$. The general expression of blocking probability for a TST switch with nonblocking individual stages (T, S, T) is

$$B = [1 - q_1^2]^l \tag{5.16}$$

where $q_1 = 1 - p_1 = 1 - p/\alpha$
α = time expansion (l/c)
l = number of space stage time slots

The implementation complexity of a TST switch can be derived as

$$\text{Complexity} = N^2 + \frac{Nl \log_2 N + 2Nc(8) + 2Nl \log_2 c}{100} \tag{5.17}$$

Example 5.4. Determine the implementation complexity of a 2048-channel TST switch with 16 TDM links and 128 channels per link. Assume the desired maximum blocking probability is 0.002 for incoming channel occupancies of 0.1

Solution. Using Equation 5.16, we can determine the number of internal time slots required for the desired grade of service as 25. Hence time concentration of $1/\alpha = 5.12$ is possible because of the light loading on the input channels. The implementation complexity can now be determined from Equation 5.17 as 656.

The results obtained in Examples 5.3 and 5.4 indicate that the TST architecture is more complex than the STS architecture. Notice, however, that the TST switch operates with time concentration whereas the STS switch operates with space concentration. As the utilization of the input links increase, less and less concentration is acceptable. If the input channel loading is high enough, time expansion in the TST switch and space expansion in the STS switch are required to maintain low blocking probabilities. Since time expansion can be achieved at less cost than space expansion, a TST switch becomes more cost effective than an STS switch for high channel utilization. The implementation complexities of these two systems are compared in Figure 5.24 as a function of the input utilization.

As can be seen in Figure 5.24, TST switches have a distinct implementation advantage over STS switches when large amounts of traffic are present. For small switches, the implementation complexities favor STS architectures. The choice of a particular architecture may be more dependent on other factors such as modularity, testability, and expandability. One consideration that generally favors an STS structure is its relatively simpler control requirements [8]. For very large switches with heavy traffic loads, the implementation advantage of a TST switch is dominant. Evidence of this fact is provided by the No. 4 ESS, a TST structure introduced in 1976 to service an excess of 100,000 voice channels.

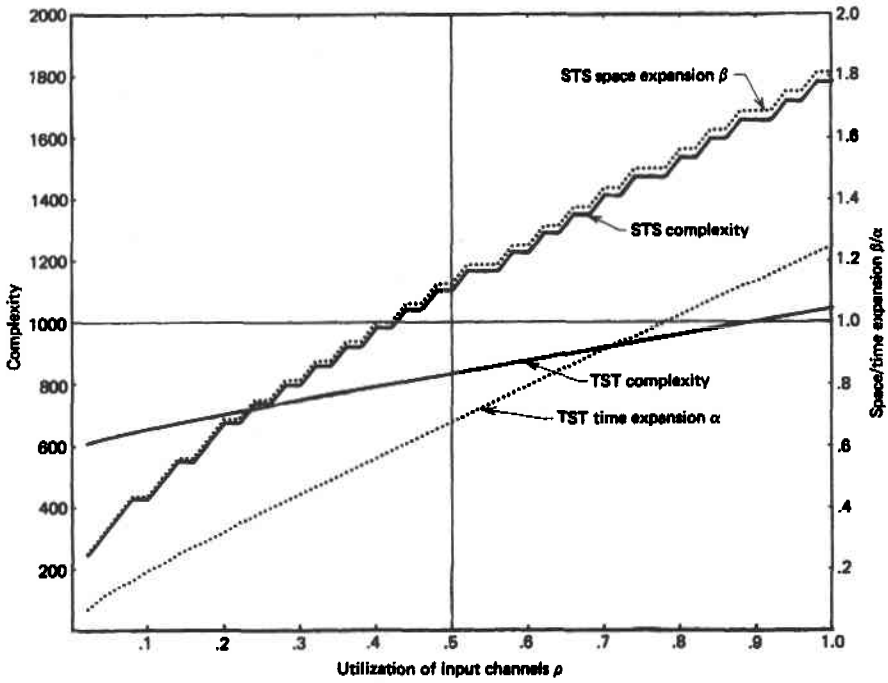


Figure 5.24 Complexity comparison of STS and TST switching structures for a blocking probability of 0.002.

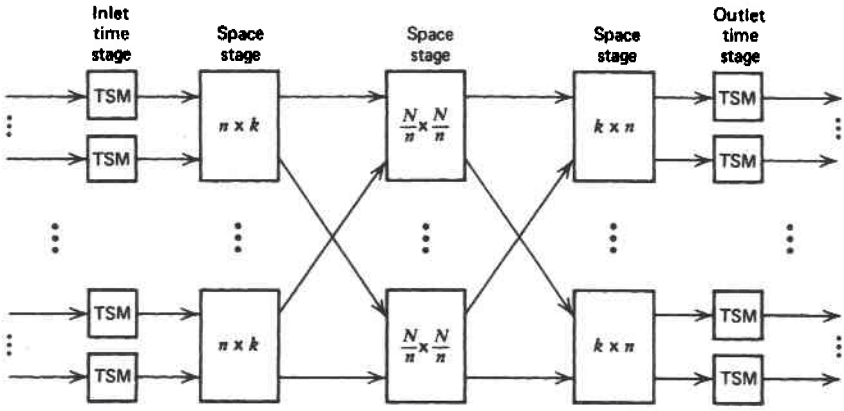


Figure 5.25 Time–space–space–space–time (TSSST) switching structure.

TSSST Switches

When the space stage of a TST switch is large enough to justify additional control complexity, multiple space stages can be used to reduce the total crosspoint count. Figure 5.25 depicts a TST architecture with a three-stage space switch. Because the three middle stages are all space stages, this structure is sometimes referred to as a TSSST switch. The EWSD switch of Siemens [9] uses a TSSST structure.

The implementation complexity of a TSSST switch can be determined as*

$$\text{Complexity} = N_X + \frac{N_{BX} + N_{BT} + N_{BTC}}{100} \tag{5.18}$$

where N_X = number of crosspoints, $= 2Nk + k\,(N/n)^2$
N_{BX} = number of space stage control store bits, $= 2k(N/n)l\,\log_2(n) + k(N/n)l\,\log_2(N/n)$
N_{BT} = number of bits in time stages, $= 2Nc\,(8)$
N_{BTC} = number of time stage control store bits, $= 2Nl\,\log_2(c)$

The probability graph of a TSSST switch is shown in Figure 5.26. Notice that this diagram is functionally identical to the probability graph of a five-stage space switch shown in Figure 5.10. Using the probability graph of Figure 5.26, we can determine the blocking probability of a TSSST switch as

$$B = \{1 - (q_1)^2[1 - (1 - q_2^2)^k]\}^l \tag{5.19}$$

*The assumed control orientations by stages are output, output, output, input, and input.

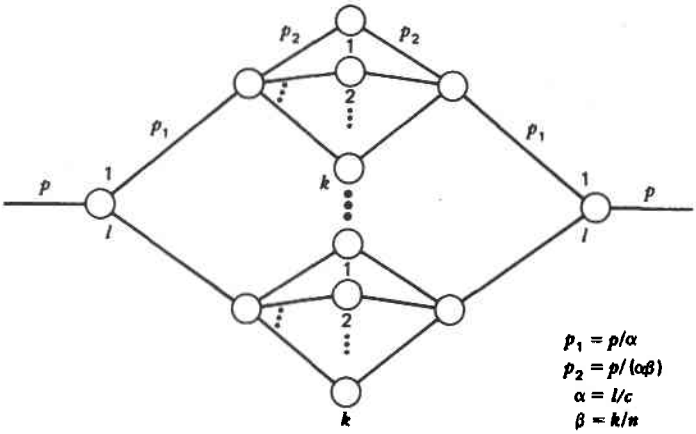


Figure 5.26 Probability graph of TSSST switch.

where $q_1 = 1 - p_1 = 1 - p/\alpha$
$q_2 = 1 - p_2 = 1 - p/\alpha\beta$

Example 5.5. Determine the implementation complexity of a 131,072-channel TSSST switch designed to provide a maximum blocking probability of 0.002 under channel occupancies of 0.7. Assume the switch services 1024 TDM input links with 128 channels on each link. Also assume that unity time expansion is used on the space stages.

Solution. The space switch can be designed in a variety of ways depending on how many links are assigned to each array in the first (and third) space stages. A value of 32 is chosen as a convenient binary number near the theoretical optimum presented earlier $(N/2)^{1/2}$. With this value for n, the only unknown in Equation 5.19 is the number of center-stage arrays k, which is determined to be 27 for a blocking probability of 0.0015. The complexity is determined from Equation 5.18 as

$$N_X = (2)(1024)(27) + (27)(32)^2 = 82{,}944$$

$$N_{BX} = 2(27)(32)(128)(5) + (27)(32)(128)(5) = 1{,}658{,}880$$

$$N_{BT} = 2(1024)(128)(8) = 2{,}097{,}152$$

$$N_{BTC} = 2(1024)(128)(7) = 1{,}835{,}008$$

$$82{,}944 + 5{,}591{,}040/100 = 138{,}854 \text{ equivalent crosspoints}$$

5.4.3 No. 4 ESS Toll Switch

As shown in Figure 5.27, the basic structure of a No. 4 ESS matrix is time–space–time with four stages in the space switch (i.e., a TSSSST) [10, 11]. The inputs to the matrix are 120-channel TDM links formed by multiplexing five DS1 signals together. Before these inputs are interfaced to the matrix, they are passed through a decorrelator to distribute the channels of incoming TDM trunk groups across multiple TDM links into the matrix. Decorrelation is used because the No. 4 ESS is a toll switch in which the incoming TDM links represent trunk groups from other switching machines. In contrast to TDM links of an end office switch formed by multiplexing independent subscriber channels, the channels of a TDM trunk group are not independently busy. In fact, the act of concentrating multiple, independent sources onto a trunk group causes high levels of correlation between the activity on the individual channels. If the channels of a trunk group were not decorrelated, they would experience much higher blocking probabilities because they would all be vying for the same paths through the matrix. Decorrelation shuffles the trunk groups so the alternate paths available for any particular connection are more likely to be statistically independent. Notice that, besides shuffling channels, the decorrelator provides space expansion (7 to 8) and time expansion (120 to 128).

The maximum size of the No. 4 ESS uses 128 decorrelators with seven 120-channel inputs each. Thus the maximum channel capacity is (128)(7)(120) = 107,520 channels. The space stage is a 1024×1024 matrix with four possible paths provided during each of the 128 space stage time slots. The probability graph of the No. 4 ESS switch is shown in Figure 5.28, from which the following blocking probability is derived:

$$B = [1 - (1 - p_1)(1 - p_2)(1 - p_1)]^{128} \tag{5.20}$$

where p is the external channel utilization, $p_1 = (7/8)(120/128)p$, and p_2 is the blocking probability of the four-stage space switch derived in Appendix A as

$$p_2 = 2p_1^2q_1^6 + 16p_1^3q_1^5 + 50p_1^4q_1^4 + 52p_1^5q_1^3 + 28p_1^6q_1^2 + 8p_1^7q_1 + p_1^8 \tag{5.21}$$

Example 5.6. Determine the blocking probability and number of paths that have to be searched in a No. 4 ESS with inlet channel occupancies of 0.7 and 0.9.

Solution. The occupancies of the interstage links (p_{21}) are determined to be 0.574 and 0.738, respectively. Using Equation 5.21, the corresponding values of p_2 are determined as 0.737 and 0.934. Equation 5.20 can now be used to determine the blocking probabilities as 0.002 and 0.56 for inlet occupancies of 0.7 and 0.9, respectively.

The probability that all four space paths are busy in a particular time slot is $1 - q_1q_2q_1$ (the base term in Equation 5.20), which evaluates to 0.952 for $p = 0.7$ and

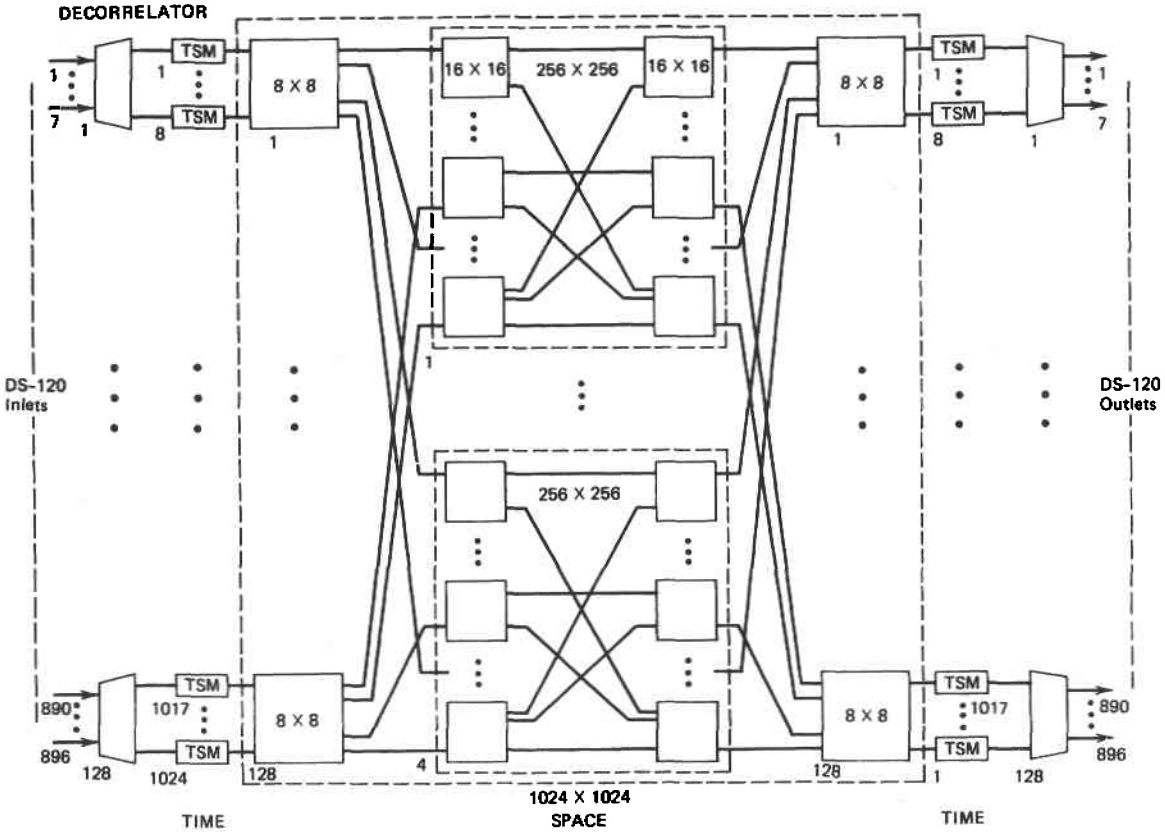


Figure 5.27 No. 4 ESS matrix architecture.

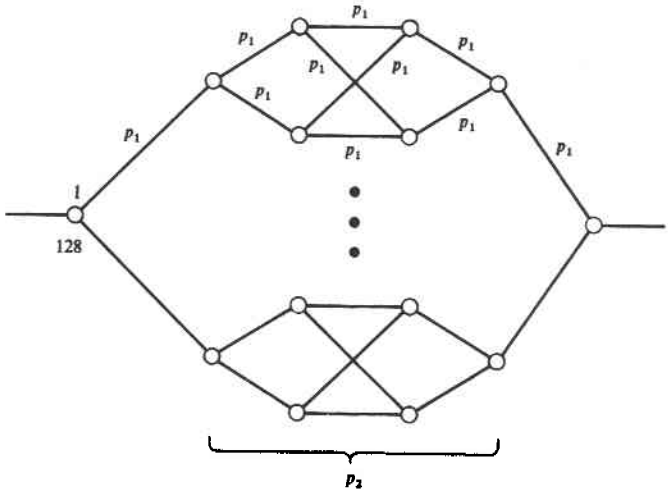


Figure 5.28 Probability graph of No. 4 ESS matrix.

0.9955 for $p = 0.9$. Equation 5.11 now determines the average number of time slots (four paths per time slot) that must be tested to find an idle path as 21 and 98, respectively.

Example 5.6 demonstrates that low blocking probabilities can be achieved even when individual paths are highly utilized if there are enough alternatives—a feature that is most practical with time division switching. This example also demonstrates the sensitivity of the call processing time as a function of load in that path search times more than quadruple when the node becomes overloaded with traffic. Reference [11] reports that the path search time is only 12% of the total processing load at $p = 0.7$, but this example implies that it would become more than 50% of the load at $p = 0.9$ if other aspects of call processing were linear functions of the traffic volume (a risky assumption).

5.4.4 System 75 Digital PBX

As originally released, the System 75 PBX of AT&T was a midrange PBX that can support 400 stations, 200 data instruments, and 200 trunks [12]. The switching matrix consists of two bytewide TDM buses as shown in Figure 5.29. Because the buses operate at 2.048 MHz, there are 256 time slots per bus per frame. Thus the total number of time slots on the buses is 512, which support 256 full-duplex connections.* All ports of the switch have access to all time slots on both buses. The dual-bus architecture allows the use of slower electronics and provides redundancy in the case of failure.

As long as there is an idle time slot on either bus, any source can be connected to any destination. Thus the blocking probability is zero from the standpoint of matching

*Because some of the time slots are used for overhead and continuous distribution of various tones, the maximum traffic capacity is 7200 CCS [12], which relates to 200 connections.

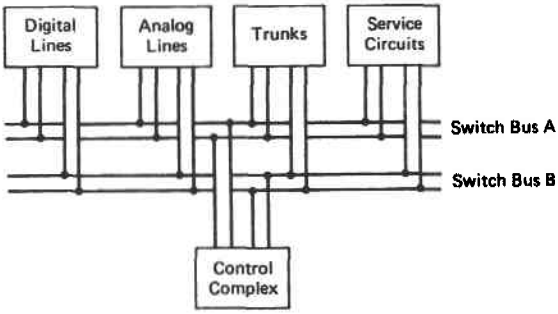


Figure 5.29 System 75 matrix architecture.

loss. Blocking can occur, however, as a result of congestion: if more connections are requested than can be supported by the number of time slots. Blocking analyses of this type require a different form of mathematics referred to as congestion theory. As presented in Chapter 12, congestion theory deals with the probability that the offered traffic load exceeds some value. Matching loss, on the other hand, is concerned with average or busy-hour traffic volumes.

As an example, congestion analyses in Chapter 12 show that if 800 stations are 37.5% busy, on average, the probability that a station requests service when 400 stations are already busy is 10^{-5}. Thus the system is virtually nonblocking for voice applications. Because data connections are often full-period connections, the blocking probability for voice connections may become more significant in applications involving intensive voiceband data switching. See the problem set in Chapter 12 for some examples.

5.5 DIGITAL CROSS-CONNECT SYSTEMS

A DCS is basically a digital switching matrix with an operations interface for setting up relatively static connections between input and output signals or channels. Instead of establishing connections in response to signaling information pertaining to call-by-call connection requests, DCS connections are established in response to network configuration needs. The most basic function of a DCS is to act as an electronic patch panel in lieu of a manual cross-connect facility. Manual cross-connect frames were typically installed in switching offices as demarcation points between transmission facilities and switching machines and in wire centers as demarcation points between feeder cables and distribution facilities. In both cases, the major purpose of the cross-connect frame was to allow rearrangement of transmission circuits and access for testing the circuits in both directions.

Figure 5.30 shows a typical manual cross-connect environment while Figure 5.31 shows the same function implemented with a DCS. Manual cross connects are implemented with bridging clips and punch-down wires. Cross connections within the DCS are established by entering commands at the management console (either locally or re-

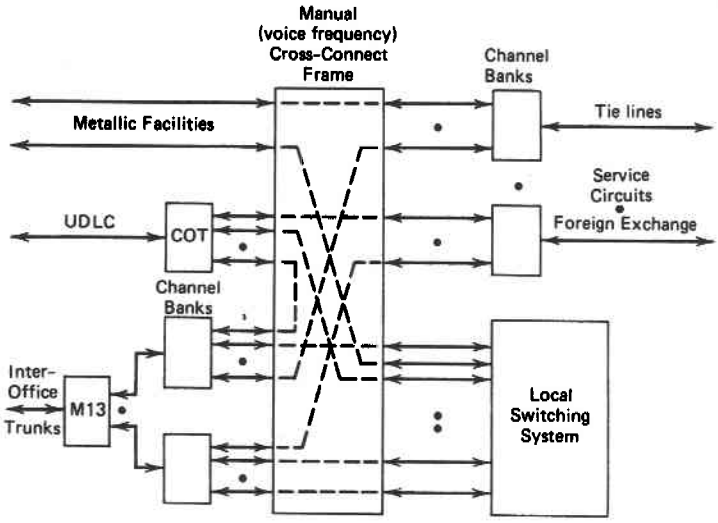


Figure 5.30 Manual cross-connect system.

motely). As indicated in the two figures, a major feature of a DCS is its elimination of back-to-back multiplexing functions when cross connecting individual channels within TDM transmission links. Additional advantages of a DCS system with respect to manual cross-connect systems are:

1. *Automatic Record Keeping.* Because the cross connects are under processor control, circuit connection reports are readily available through the management interface. In contrast, records in manual systems were inherently error prone and often out of date.

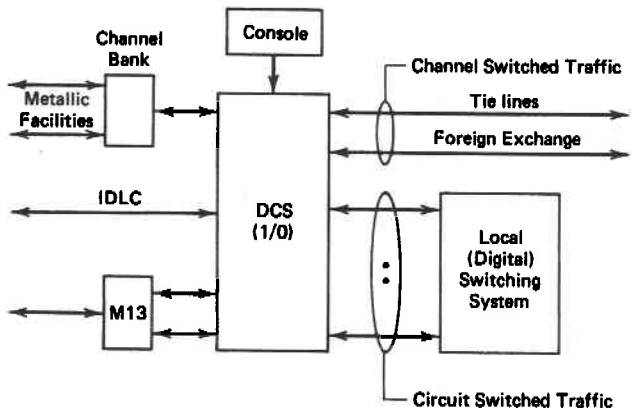


Figure 5.31 Electronic digital cross-connect system.

2. *Remote and Rapid Provisioning.* Provisioning is the basic process of providing (or discontinuing) service to a particular customer. The basic operations involved are outside-plant cross connections, inside-plant cross connections, configuration changes in switching system data base, and customer record updates in business (billing) systems. Obviously, the more these processes were automated, the faster and more accurately they could be performed.
3. *Automated Test Access.* Testing analog circuits at a manual cross-connect frame involves physically breaking the connection (by removing bridging clips) and attaching the test equipment to the side of the circuit to be tested. All manual operations are eliminated with an electronic patch panel by entering commands at the management console to connect the desired test tones and (DSP) measurement channels to the circuit under test.

Figure 5.31 depicts two types of network traffic: circuit-switched traffic and channel-switched traffic. Circuit-switched traffic represents locally switched traffic (DS0 circuits typically) and channel switched-traffic refers to leased line equivalents of digital channels. Channel-switched traffic might terminate at another public network switching office as in a foreign exchange (FX) circuit or at a distant customer premise as a tie line of a private network. In the latter case, more than one DS0 channel might be concatenated together to form a single higher rate channel referred to as a fractional T1 circuit or $M \times 64$-kbps channel. Channel-switched services typically account for over one-half of the transmission bandwidth between U.S. metropolitan offices [13]. The process of separating channel-switched services from circuit-switched services is often referred to as "grooming." Figure 5.30 also shows that universal digital loop carrier (UDLC) becomes integrated digital loop carrier (IDLC) in a digital environment (Figure 5.31). (Chapter 11 describes DLC systems in more detail.)

5.5.1 Consolidation and Segregation

Two basically different DCS grooming functions are depicted in Figures 5.32 and 5.33: consolidation and segregation. When multiple-access lines carrying traffic destined to a common distant node are partially filled, the per-channel costs of transport to the remote node can be reduced by consolidating the traffic. Conversely, when different types of traffic originate at a single location, it is desirable to allow a single fa-

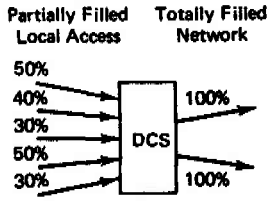


Figure 5.32 Consolidation.

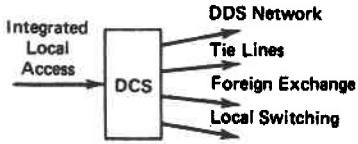


Figure 5.33 Distribution.

cility (e.g., a T1 line) to carry all types of traffic and segregate it at the DCS. Examples of such traffic are circuit-switched channels, DDS channels, tie lines, multidrop data circuits, FX circuits, or other special services that might be available only at another office.

5.5.2 DCS Hierarchy

The cross-connect system depicted in Figure 5.31 terminates DS1 signals and interchanges DS0 signals, which leads to the designation DCS 1/0. Similarly, a digital cross-connect system that terminates DS3 signals and rearranges DS1 signals within the DS3s is referred to as a DCS 3/1. If a DCS, such as a DCS 3/0, provides rearrangement of lower level signals, such as DS0s, it does not necessarily mean that it also provides cross-connect services for intermediate-level signals, such as DS1s. Cross connection of DS1s requires transparent mapping of the entire 1.544 Mbps, which includes the framing channel—a capability that may not be provided when 64-kbps DS0 channels are the primary focus. When a DCS provides multiple levels of cross connects, the intermediate levels are sometimes designated—as in DCS 3/1/0 for terminating DS3s and rearranging DS1s and DS0s. In most cases, it is not necessary to provide cross-connect services at all levels of the digital signal hierarchy for all of the terminations. Figure 5.34 depicts a DCS hierarchy that provides rearrangement of lower level signals on only subsets of higher level signals. The higher level (e.g., DS3) signals that are not cross connected to a lower level DCS may be unchannelized high-speed (44.736-Mbps) signals or transit DS3s that terminate on a DCS 3/1 or 3/0 at a

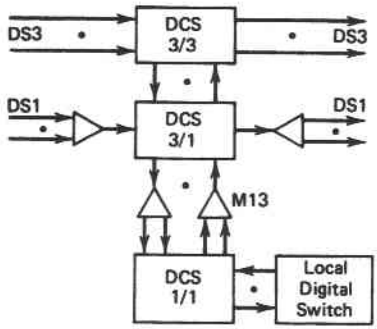


Figure 5.34 DCS hierarchy.

distant location. The primary purpose of the DCS 3/3 is to provide network restoration or protection switching of the DS3 signals and to possibly provide time-of-day rearrangement of trunk group assignments. SONET transmission links described in Chapter 8 are cross connected at much higher rates producing additional DCS layers.

5.5.3 Integrated Cross-Connect Equipment

Figure 5.31 depicts distinct multiplexing equipment and two distinct and colocated switching systems: the DCS and the local digital switch. Although DCS functions within the public network are traditionally implemented with separate equipment, private networks often utilize equipment that provides integrated multiplexing and cross-connect functions. This equipment has evolved from CPE-based T1 multiplexers because TSI circuits are insignificant hardware additions to T1 multiplexing hardware. Such equipment are variously referred to as intelligent multiplexers [14], nodal processors, or networking T1 multiplexers. Cross-connect functions are also incorporated into newer DLC systems because, again, the cost of the TSI function is virtually nil.

Integration of cross-connect functions with higher level digital signals (such as a DS3) has not occurred because the manner in which the higher level signals are multiplexed (described in Chapter 7) is not amenable for direct termination on a switching system. The newer form of (synchronous) multiplexing, as specified by the SONET standard described in Chapter 8, allows integration of higher level digital multiplexers with cross-connect systems for both public and private equipment.

Although cross-connect functions are basically nothing more than "pegged" or "nailed-up" circuit-switched connections, the two functions are traditionally implemented separately because of the following differences in the application requirements:

1. A DCS needs to be strictly nonblocking at the DS0 level, which is generally uneconomical in large public network switches.
2. Transparent cross-connection of a 1.544-Mbps signal requires transport of the framing bit that is not possible in typical digital circuit switches because DS1 interface equipment assumes the DS1 signals are channelized with framing in the 193rd bit.
3. DCS functionality does not involve processing of signaling bits so fully functional circuit switch interfaces have excess costs when used as interfaces to cross-connect systems.
4. Cross connecting multiple, concatenated DS0 channels (fractional T1 channels) requires maintaining order in the concatenated data stream. Although the order can always be maintained by careful mapping of the individual 64-kbps connections, it is a function that is typically not provided in DS0 circuit switching software. Figure 5.35 shows how the order of bytes in the concatenated data stream can be transposed by ill-chosen 64-kbps connections.

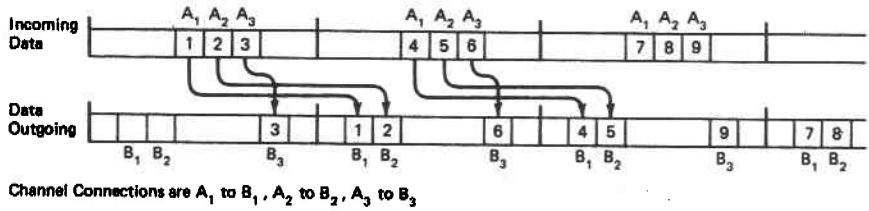


Figure 5.35 Transposition of data in concatenated time slots.

5.6 DIGITAL SWITCHING IN AN ANALOG ENVIRONMENT

When digital end office switches (or PBXs) are installed in an analog environment, the analog interfaces are necessarily unchanged. Although the digital switch may interface with digital subscriber carrier or digital fiber feeder systems, these systems merely extend the analog interface point closer to the subscriber. This section describes the basic considerations of using digital switching in such an analog environment. Chapter 11 describes digital end office switching with digital subscriber loops in relation to the integrated services digital network.

5.6.1 Zero-Loss Switching

As already mentioned, a well-designed digital transmission and switching system adds no appreciable degradation to the end-to-end quality of digitized voice. In particular, the derived analog signal coming out of a decoder can be adjusted to the same level as that presented to the encoder at the far end. Curiously, zero-loss transmission presents some significant problems when digital switching is used in the analog transmission environment of a class 5 central office.

Analog end office switches are two-wire switches designed to interconnect bidirectional two-wire customer loops. The voice digitization process, however, inherently requires separation of the go and return signal paths involved in a connection. Thus, a class 5 digital end office or digital PBX must be a four-wire switch. When inserted into a two-wire analog environment, hybrids are required to separate the two directions of transmission. As shown in Figure 5.36, hybrids at each end of the internal digital subnetwork produce a four-wire circuit with the potential for echoes and singing. (Amplifiers are shown in conjunction with the encoders to offset forward path loss inherent in the hybrids.) Instability arises as a result of impedance mismatches at the hybrids causing unwanted coupling from the receive portion to the transmit portion of the four-wire connection. Impedance mismatches occur because of the variability in the lengths and wire sizes of the subscriber loops. In particular, loaded and unloaded wire pairs have markedly different impedance characteristics.

The instability problems are compounded by certain amounts of artificial delay that are required in a digital time division switch. Although the delay through a digital switch (several hundred microseconds, typically) is basically unnoticeable to a user,

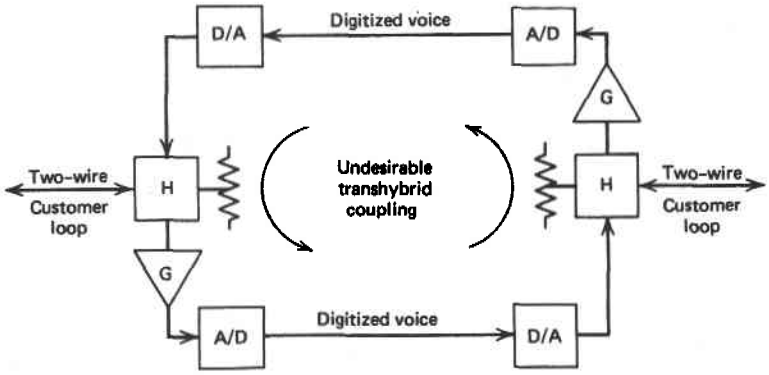


Figure 5.36 Four-wire digital switch with two-wire analog interfaces.

it represents the equivalent of as much as 30 to 40 miles of wire. This increased delay has the effect of lowering the oscillation frequencies that might otherwise be outside the voiceband and effectively be removed by encoder/decoder filters.

As mentioned in Chapter 1, hybrids were traditionally used at the interface to the toll network where the transmission medium is basically four-wire. In these instances, the instability of the four-wire circuit was controlled by designing a prescribed amount of net attenuation (by way of net loss) into the transmission path of shorter toll network circuits. On the longer circuits, the echoes and singing are eliminated by echo suppressors or echo cancelers. Class 5 digital switches could prevent instability in the same manner: by designing a certain amount of loss into the encoding/decoding process. While the necessary amount of signal attenuation (approximately 2–3 dB in each path [15]) could be tolerated on local connections, the added loss to some toll connections would be unacceptably large.

From the foregoing discussion it can be seen that the use of selectable attenuation is one solution to the instability problems. The necessary loss is inserted into the talking path for local connections but not used on long-distance calls, which already have loss designed into them. A second solution involves matching the impedance at the hybrids more closely. Before the advent of DSP nearly complete elimination of the unwanted coupling was prohibitively expensive. However, adequate isolation of the transmission paths could be accomplished with just two different matching networks: one for loaded loops and one for unloaded loops [16]. Impedance matching has been simplified by the use of advanced interface electronics (DSP) that contain trainable and automatic impedance matching circuitry. The impedance matching circuits essentially represent short-delay echo cancelers [17].

Notice that an all-digital network (with four-wire telephones) avoids instability problems because there are no two-wire analog lines. Voice is digitized at the telephone and uses a separate path from the receive signal all the way to the destination telephone. Thus an end-to-end four-wire circuit completely eliminates echoes and singing, allowing all connections to operate on a zero-loss basis.

5.6.2 BORSCHT

In Chapter 1 the basic functional requirements of the subscriber loop interface are described. These requirements are repeated here with two additional requirements for a digital switch interface: coding and hybrid. The complete list of interface requirements is unaffectionately known as BORSCHT [18]:

B: Battery feed
O: Overvoltage protection
R: Ringing
S: Supervision
C: Coding
H: Hybrid
T: Test

As mentioned in Chapter 1, the high-voltage, low-resistance, and current requirements of many of these functions are particularly burdensome to integrated circuit implementations. First-generation digital end office switches reduced the termination costs by using analog switching (concentrators) to common codecs. The DMS-100 [19] of Northern Telecom and the No. 5 ESS of AT&T [20] use analog concentration at the periphery. Integrated circuit manufacturers have worked diligently to implement the BORSCHT functions in what is called a subscriber loop interface circuit (SLIC). Per-line SLICs allow implementation of per-line BORSCHT functions. SLICs can be used in PBX applications with a minimum of other external components. In central office applications, where lightning protection and test access are more demanding, SLICs typically need other components for a complete interface.

5.6.3 Conferencing

In an analog network conference calls are established by merely adding individual signals together using a conference bridge. If two people talk at once, their speech is superposed. Furthermore, an active talker can hear if another conferee begins talking. Naturally, the same technique can be used in a digital switch if the signals are first converted to analog, added, and then converted back to digital.

As described in Chapter 3, μ255 and the *A*-law (ITU) code were designed with the specific property of being easily digitally linearizable (EDL). With this property, the addition function can be performed digitally by first converting all codes to linear formats, adding them, and then converting back to compressed formats. To the user, the operation is identical to the customary analog summation. For a conference involving N conferees, N separate summations must be performed, one for each conferee and containing all signals but his own. For a description of the conferencing algorithm in the System 75 PBX of AT&T, see reference [12]. For more general descriptions of conferencing implementations in digital switches, see reference [21].

Another conferencing technique involves monitoring the activity of all conferees and switching the digital signal of the loudest talker to all others. Although this technique is functionally different from a customary analog conference bridge, it is advantageous for large conferences because the idle channel noise of the inactive talkers does not get added into the output of the conference bridge. High-quality conference circuits also include echo cancelers so higher signal powers can be provided.

REFERENCES

1 M. R. Aaron, "Digital Communications—The Silent (R)evolution?" *IEEE Communications Magazine*, Jan. 1979, pp. 16–26.

2 C. Clos, "A Study of Non-Blocking Switching Networks," *Bell System Technical Journal*, Mar. 1953, pp. 406–424.

3 C. Y. Lee, "Analysis of Switching Networks," *Bell System Technical Journal*, Nov. 1955, pp. 1287–1315.

4 C. Iacobaeus, "A Study of Congestion in Link Systems," *Ericsson Techniques*, No. 48, Stockholm, 1950, pp. 1–70.

5 A. A. Collins and R. D. Pedersen, *Telecommunications, A Time For Innovation*, Merle Collins Foundation, Dallas, TX, 1973.

6 M. Schwartz, *Telecommunications Network Protocols, Modeling and Analysis*, Addison-Wesley, Reading MA, 1987.

7 T. E. Browne, D. J. Wadsworth, and R. K. York, "New Time Division Switching Units for No. 101 ESS," *Bell System Technical Journal*, Feb. 1969, pp. 443–476.

8 S. G. Pitroda, "Telephones Go Digital," *IEEE Spectrum*, Oct. 1979, pp. 51–60.

9 N. Skaperda, "EWSD: Where It Is," *Siemens Telecom Report*, Vol. 12, No. 2-3, 1989, pp. 56–59.

10 J. H. Huttenhoff, J. Janik, G. D. Johnson, W. R. Schleicher, M. F. Slana, and F. H. Tendick, "No. 4 ESS: Peripheral Systems," *Bell System Technical Journal*, Sept. 1977, pp. 1029–1042.

11 A. E. Ritchie and L. S. Tuomenoksa, "No. 4 ESS: System Objectives and Organization," *Bell System Technical Journal*, 1977, pp. 1017–1027.

12 L. A. Baxter, P. R. Berkowitz, C. A. Buzzard, J. J. Horenkamp, and F. E. Wyatt, "System 75: Communications and Control Architecture," *AT&T Technical Journal*, Jan. 1985, pp. 153–173.

13 R. K. Berman, R. W. Lawrence, and P. C. Whitehead, "Providing and Managing Channel Switched Services in an Intelligent ISDN Environment," *IEEE Clobecom*, 1987, pp. 4.6.1–4.6.5.

14 J. L. Melsa and H. R. Scull, "The Application of Intelligent T1 Multiplexers in Hybrid and ISDN Networks," *IEEE Globecom*, 1989, pp. 15.3.1–15.3.6.

15 J. L. Neigh, "Transmission Planning for an Evolving Local Switched Digital Network," *IEEE Transactions on Communications*, July 1979, pp. 1019–1024.

16 R. Bunker, F. Scida, and R. McCabe, "Zero Loss Considerations in Digital Class 5 Office," *IEEE Transactions on Communications*, July 1979, pp. 1013–1018.

17 D. L. Duttweiler, "A Twelve Channel Echo Cancellor," *IEEE Transactions on Communications*, May 1978, pp. 647–653.

18 F. D. Reese, "Memo to Management—You Must Appraise How New Technology Fits Customers," *Telephone Engineering and Management*, Oct. 1, 1975, pp. 116–121.

19 J. B. Terry, D. R. Younge, and R. T. Matsunaga, "A Subscriber Line Interface for the DMS-100 Digital Switch," *IEEE National Telecommunications Conference*, 1979, pp. 28.3.1–28.3.6.

20 D. L. Carney, J. I. Cochrane, L. J. Gitten, E. M. Prell, and R. Staehler, "The 5 ESS Switching System: Architectural Overview," *AT&T Technical Journal*, Aug. 1985.

21 R. J. D'Ortenzio, "Conferencing Fundamentals for Digital PABX Equipments," *IEEE International Conference on Communications*, 1979, pp. 2.5-29–2.5-36.

PROBLEMS

5.1 How many four-wire voice circuit connections can be provided by a bidirectional PAM switching bus if the minimum achievable pulse width is 250 nsec?

5.2 The TS switch of Figure 5.19 uses DS1 signals on each TDM link. What is the implementation complexity if groups of five DS1 inputs are first multiplexed to form 16 input links with 120 channels on each link?

5.3 Determine the Lee graph and Jacobaeus blocking probabilities of the first switch in Table 5.5 ($k = 6$) if two inlets out of 16 become connected to 1-erlang subscribers. (Possibly these lines could be dial-up ports to a computer.) (*Hint*: Two inlets are permanently busy but the remaining inlets continue to be 0.1-erlang ports.)

5.4 Repeat Problem 5.3 but instead of two inlets being busy assume that two of the output links of the first-stage module have failed.

5.5 How many crosspoints are needed in a 1024-line, three-stage space switch if the input loading is six common-channel signals per line and the maximum acceptable blocking probability (using a Lee graph analysis) is 0.005?

(a) If $n = 16$, $N/n = 64$
(b) If $n = 32$, $N/n = 32$
(c) If $n = 64$, $N/n = 16$

5.6 What is the (Lee graph) blocking probability of the TS switch in Figure 5.19 for channel loading of 0.2 erlang?

(a) Assume each TDM input is a 24-channel interoffice trunk group.
(b) Assume the TDM inputs are derived from 24-channel banks with each analog interface connected to a dedicated 0.2-erlang line.

5.7 Design an STS switch for 128 primary TDM signals of the CCITT hierarchy (30 voice channels per input). Blocking should be less than 0.002 and the loading is 0.2 erlang per channel. How many time slot interchange modules are needed? What is the complexity of the switch?

5.8 Repeat Problem 5.7 for a TST design.

5.9 Determine the number of crosspoints N_X and the total number of memory bits N_B required for a TST switch defined as follows: number of lines is 32, single-stage space switch, number of channels per frame is 30, and time expansion is 2.

5.10 What is the blocking probability of the switch in Problem 5.9 if the channel loading is 0.9 erlang per channel?

5.11 How many bits of memory are needed in a time slot interchange circuit for a 60-channel signal with 9 bits per time slot?

5.12 Derive Equation 5.15 assuming input-associated control in the first stage and output-associated control in the third stage.

5.13 Derive Equation 5.18 assuming all control orientations are output associated.

6

DIGITAL MODULATION AND RADIO SYSTEMS

Chapter 4 discusses various means of encoding digital information for transmission over wireline (or fiber) transmission links. In the context of this chapter the data encoding described in Chapter 4 is referred to as baseband coding. To impress the same information onto a carrier for radio transmission, a baseband signal is commonly used to modulate the carrier. Modulation shifts the dc-centered spectrum of a baseband signal to the carrier frequency. In contrast to Chapter 4, which emphasizes binary transmission, this chapter emphasizes multilevel digital modulation to achieve high data rates within the rigidly defined bandwidth of a radio channel or analog voiceband data circuit. As described later in this chapter, most multilevel modulation techniques can be implemented by directly modulating a carrier with one of the multilevel baseband signals described in Chapter 4.

Applications for the digital modulation techniques described in this chapter cover point-to-point microwave systems, digital cellular systems, and voiceband modems. Although point-to-point microwave systems are no longer used along heavy traffic routes of the U.S. public network, they are still utilized in thin-route applications in the U.S. and around the world where cost of a fiber right-of-way is prohibitive. Furthermore, digital point-to-point microwave is still commonly used in private networks (along with some old analog systems) and in special applications such as digital video transport from a studio to a transmitter or to and from a digital cellular base station. Digital cellular transmission between base stations and mobile units uses modulation techniques that are identical to that used in some original digital point-to-point systems.

When digital point-to-point radios were first being developed, the FCC established certain data rates [1] that must be achieved by these radios in the United States to ensure a minimum utilization efficiency of the radio frequency spectrum. Basically, these minimum rates (shown in Table 6.1) specify approximately the same number of voice circuits (at 64 kbps) as were available in the analog FM radios in use at the time (1974). Although individual voice channels in a FDM signal occupy 4 kHz of bandwidth, frequency modulation typically expands the signal bandwidth by a factor of

TABLE 6.1 Information Densities Required by FCC for Common-Carrier Microwave Channels Using 64 kbps per Voice Circuit

Band (GHz)	Channel BW (MHz)	Bit Rate (Mbps)	Density (bps/Hz)
2.110–2.130	3.5	6.144	1.8
3.700–4.200	20	73.7	3.7
5.925–6.425	30	73.7	2.5
6.525–6.875	30	73.7	2.5
10.7–11.7	40	73.7	1.8

about 4, depending on the amount of FM deviation applied. Thus the digital systems were required to compete with an equivalent voice channel bandwidth of about 16 kHz. If the digital radios were forced to compete with analog single-sideband radios [2, 3] introduced in the early 1980s, they could not have done so without using lower bit rate voice encoding.

Over and above the minimum bit rates established by the FCC, competition among digital microwave manufacturers and the economics of maximizing the number of voice circuits per radio stimulated the development of more advanced digital modulation techniques to achieve even greater transmission rates. As an example of signal processing advances, the 6G150 MB digital radio of NEC carries a 155-Mbps payload in 30 MHz of bandwidth—an information density of just over 5 bps/Hz. The first part of this chapter describes basic modulation techniques and the transmission efficiencies they provide. The latter sections describe radio system design considerations for point-to-point microwave relay systems.

Information Density

A useful parameter for characterizing the bandwidth efficiency of a digital modulation system is the information density, defined as

$$\delta = \frac{R}{\mathrm{BW}} \tag{6.1}$$

where R = data rate in bits per second
BW = bandwidth of digital signal in hertz

The units of information density are sometimes referred to loosely as bits per hertz. However, as defined in Equation 6.1, the units should be bits per second per hertz or simply bits per cycle. Since bits per second per hertz conveys the nature of information density more completely, it is the preferred unit.

The bandwidth factor in Equation 6.1 can be defined in a variety of ways. In theoretical studies the bandwidth of a signal is usually determined as the width of the ideal filter used to strictly bandlimit the signal (i.e., the Nyquist bandwidth). In practical systems, where the spectrum can never be strictly bandlimited, the bandwidth is more

difficult to define. In radio systems a channel is usually specified with a certain minimum signal attenuation at the band edges. In this case, the information density is easily determined as the bit rate of a signal meeting these requirements divided by the allotted bandwidth.

When no particular emission specifications are provided, however, the bandwidth required by a digital signal must be defined in another more or less arbitrary manner. Using a 3-dB bandwidth is usually inappropriate since the signal spectrum may drop rather slowly and thereby allow significant amounts of energy to spill over into adjacent channels. Using a greater attenuation level to define the band edges would be more appropriate but is still somewhat arbitrary. A generally more relevant criterion for practical systems defines the bandwidth of a signal as the channel spacing required to achieve a specified maximum level of interference into identical adjacent channels. Yet another definition of bandwidth, commonly used in Europe, is the 99% power bandwidth.

The theoretical maximum information density for binary signaling is 2 bps/Hz for a two-level line code or 1 bps/Hz for a modulated double-sideband signal. If a four-level line code is used to achieve 2 bits per signal interval, the theoretical information density is 4 bps/Hz for the line code, or 2 bps/Hz for a double-sideband carrier signal. The information density of amplitude-modulated signals can be doubled by using single-sideband transmission to effectively achieve the same efficiency as the line codes (baseband signals).

As a practical example, a representative data rate on a dial-up analog telephone line is 28,800 bps (usually with 12 bits per signal interval and a 2400-Hz signaling rate). Since the usable bandwidth of a telephone channel is approximately 3 kHz, a typical information density of a dial-up line is 9.6 bps/Hz.*

As a starting point for digital microwave radios, Table 6.1 includes the minimum bit rates and corresponding information densities required in common-carrier bands of the United States. Competition and compatibility with the digital hierarchy led to realizations of even greater transmission efficiencies.

6.1 DIGITAL MODULATION

Digital transmission on a radio link differs from wireline transmission in two important regards. First, the digital information must modulate a carrier in some manner to produce the radio frequency (RF) signal. In many cases the modulation process can be viewed as a special form of amplitude modulation by a nonreturn-to-zero (NRZ) line code signal. Thus the line code represents a baseband signal that amplitude modulates the carrier in the transmitter and is reproduced by demodulation in the receiver. Representing the modulation process in this manner has the advantage that the RF spectra can be determined by merely translating the line code (baseband) spectra to the selected carrier frequency.

*Slightly higher rates are sometimes achieved with V.34 modems. Rates approaching 64 kbps are also possible in special (V.90) applications described in Chapter 11.

Second, a radio link differs from wireline transmission due to the necessity of strictly bandlimiting the transmitted signals to prevent interference into other channels. Although wireline transmission links automatically filter the line signal to some extent, explicit filter requirements sometimes occur only in the receivers to reject as much noise as possible.* Since radio links are *bandlimited* in the transmitter and *noise filtered* in the receiver, the end-to-end filter function must be partitioned between the two ends. Figure 6.1 shows a block diagram of a radio link showing representative baseband and RF waveforms along with corresponding frequency spectrums. For modulation, Figure 6.1 shows multiplication of the carrier by the baseband waveform.

Modulation in a digital radio must ultimately be designed in conjunction with the filter functions. For ease of description, however, modulation techniques are considered first. Later on, the filtering requirements of each type of modulation are discussed. Figure 6.1 also shows the baseband encoding process to be independent of the modulation process. For most of the modulation techniques described in this chapter this view is appropriate. In the case of trellis-coded modulation (TCM) and carrierless amplitude and phase (CAP) modulation, described later, encoding is incorporated into the modulation process.

6.1.1 Amplitude Modulation

Historically, the simplest form of modulation to generate and detect is amplitude modulation (AM). A conceptual illustration of amplitude modulation is shown in Figure 6.2. The mathematical definition is

$$x(t) = [1 + am_n(t)]\cos \omega_c t \tag{6.2}$$

where a = modulation index ($0 < a \leq 1$)

$m_n(t)$ = n-level, symmetric NRZ baseband signal normalized to maximum amplitude of 1

ω_c = is the radian carrier frequency, $= 2\pi f_c$

Amplitude modulation is an example of a class of special modulation techniques referred to as "linear modulation." Linear modulation implies that the spectrum of the modulated signal is obtained by translating the baseband spectrum to the selected carrier frequency band. As shown in Figure 6.2, amplitude modulation by a two-level digital baseband signal essentially translates the $\sin(x)/x$ baseband spectrum up to the carrier frequency f_c. Other linear modulation techniques are double-sideband modulation, single-sideband modulation, and vestigial-sideband modulation.

Inspection of Equation 6.2 or Figure 6.2 indicates that if 100% modulation is used ($a = 1$), no carrier is produced for a logic 0. For obvious reasons, this form of amplitude modulation is often referred to as on–off keying, or *amplitude shift keying* (ASK). As

*Adherence to EMI compatibility specifications defined by the FCC (Part 15) often requires some amount of transmit filtering in on-premises wireline transmission systems.

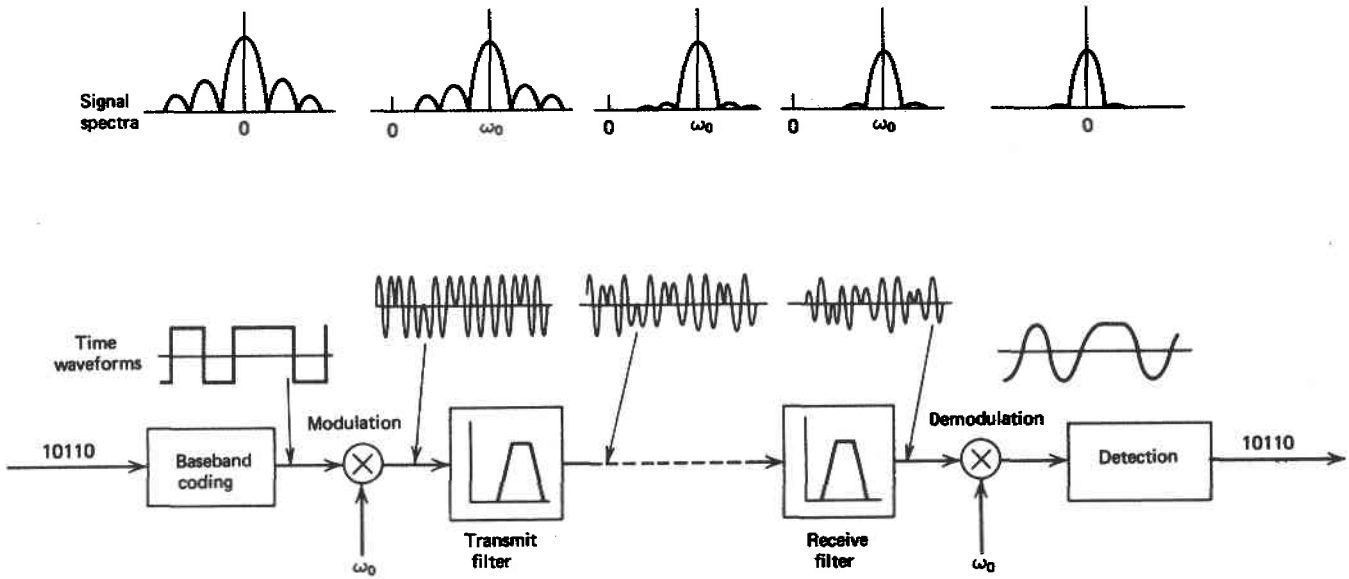


Figure 6.1 Simplified block diagram of end-to-end radio channel.

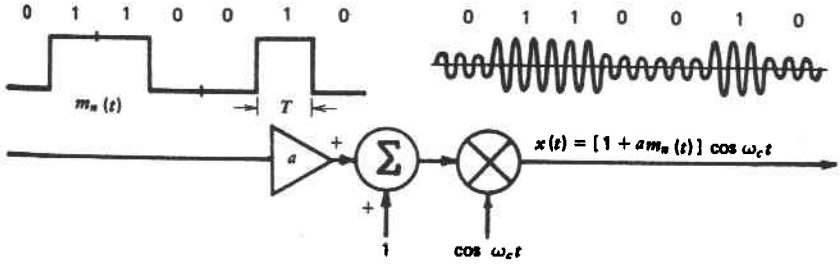


Figure 6.2 Digital binary amplitude modulation.

shown in Figure 6.3, on–off keying can be obtained by direct multiplication of a carrier with the two-level (unipolar) line code described in Chapter 4.

Amplitude-modulated signals are usually demodulated with a simple envelope detector. The cost effectiveness of this detector is the basic reason that commercial analog broadcasting uses amplitude modulation. Unfortunately, the error performance of digital amplitude modulation in general, and envelope detection in particular, is inferior to other forms of digital modulation and detection. For this reason, amplitude modulation is used only where the cost of the receiver is a significant consideration. Digital microwave links and digital cellular systems use other forms of modulation and demodulation to minimize the error rate for a given signal-to-noise ratio.

Conventional amplitude modulation provides suboptimum error performance for two basic reasons. First, if $a < 1$, a discrete (informationless) spectral line occurs at the carrier frequency. Although the existence of this spectral line simplifies carrier recovery, it increases the transmitter power without aiding discrimination between information signals.

With 100% modulation (on–off keying) no line spectra are produced, but the system is still inefficient in its use of transmitted power. As discussed in Chapter 4 for two-level line coding, the maximum use of transmitted power is achieved when one signal is the negative of the other. Thus, a second deficiency of amplitude modulation arises because a 0 signal is not the exact negative of a 1 signal. To achieve optimum performance, a symmetric two-level baseband signal should directly modulate (multiply) the carrier. As shown in Figure 6.4, this form of modulation produces two iden-

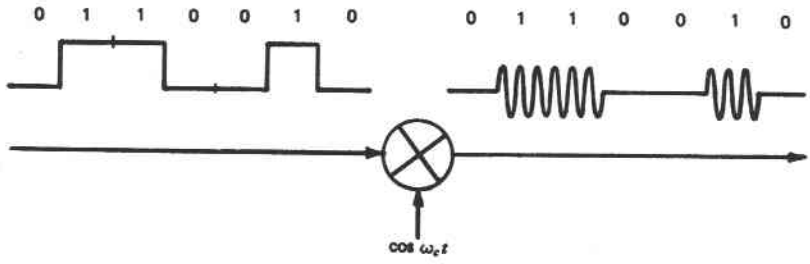


Figure 6.3 On–off keying.

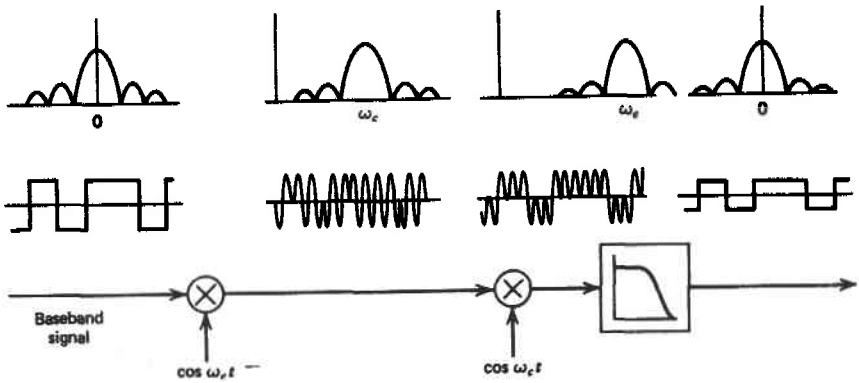


Figure 6.4 Phase-reversal keying.

tical signals except for a 180° phase reversal. Hence this modulation is sometimes referred to as phase reversal keying (PRK), or more often, two-level phase shift keying (2-PSK). Notice that PRK cannot be detected by envelope detection. Instead, a PRK signal must be detected by comparing it to a coherent carrier reference.

Coherent detection involves comparing the incoming signal to a local carrier synchronized in phase to that used at the transmitter. With conventional amplitude modulation there is no information in the phase. With PRK signaling, however, all of the information is in the phase. The use of a coherent reference for PRK signals allows optimum (antipodal) error rate performance. The basic equation defining a PRK signal is

$$x(t) = m_2(t) \cos \omega_c t \tag{6.3}$$

where $m_2(t)$ is the binary digital baseband signal [$m_2(t) = +1$ for a 1 and $m_2(t) = -1$ for a 0]. The demodulation equation is

$$\begin{aligned} y(t) &= x(t)[2 \cos \omega_c t] \\ &= [m_2(t) \cos \omega_c t]\,[2 \cos \omega_c t] \\ &= m_2(t) + m_2(t) \cos 2\omega_c t \end{aligned} \tag{6.4}$$

where the double-frequency term is removed with a low-pass filter.

As indicated in Figure 6.4 and Equation 6.4, the coherent demodulation process produces a symmetric two-level signal that can be processed further to detect the data. Since the demodulated baseband signal is, in essence, equivalent to a line code, it must include timing considerations as discussed in Chapter 4. In particular, there must be enough signal transitions in the baseband signal to allow recovery of a sample clock. However, the baseband signal need not preclude dc wander since dc levels

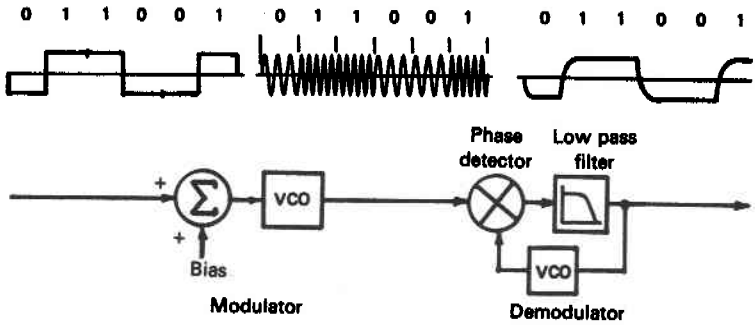


Figure 6.5 Binary frequency shift keying.

are translated to the carrier frequency and passed adequately by the double-sideband system.*

6.1.2 Frequency Shift Keying

In addition to an inefficient use of signal power, conventional amplitude-modulated signals (not PRK) have one other undesirable characteristic. By definition, an amplitude-modulated signal uses multiple signal levels, which implies AM is quite vulnerable to signal saturation that narrows the distance between amplitude levels and produces spectrum spreading. A common source of saturation in a radio system occurs in the output power amplifier of the transmitter. In most cases output amplifiers are operated at less than maximum power to eliminate saturation and other nonlinearities, so they can accommodate amplitude-modulated signals [4].

Angle-modulated systems FM or PM use constant-amplitude signals not adversely affected by signal saturation. Hence FM and PM can be transmitted at higher power levels than AM systems. The ability to use saturating power amplifiers is one of the reasons why FM was originally chosen for analog microwave radios. This section discusses digital frequency modulation, commonly referred to as frequency shift keying (FSK). The next section discusses digital phase modulation, commonly referred to as phase shift keying (PSK). Both systems provide a constant-amplitude signal. Systems using constant-amplitude carriers are also referred to as constant-envelope systems.

The general expression for an n-ary FSK signal is

$$x(t) = \cos\left[\left(\omega_c + \frac{m_n(t)\,\Delta\omega}{2}\right)t\right] \tag{6.5}$$

*A single-sideband system does not pass dc energy. Thus if single-sideband modulation is used, the baseband signal must exclude dc energy from its spectrum. Some double-sideband systems might also require the elimination of baseband dc energy so that a carrier tone can be inserted into the center of the passband without affecting the signal.

where ω_c = radian center frequency
m_n = n-level NRZ digital baseband signal
$\Delta\omega$ = radian difference frequency between signals

A typical binary FSK signal is shown in Figure 6.5 along with a simple (but low-performance) means of implementing the modulator and demodulator. The modulator is a voltage-controlled oscillator (VCO) that is biased to produce the center frequency when no modulation is applied. The amplitude a of the symmetric two-level baseband signal produces a frequency deviation of $+\Delta\omega/2$ for a 1 and $-\Delta\omega/2$ for a 0.

The demodulator is implemented as a phase-locked loop: a VCO, a phase detector, and a loop filter. The phase detector measures the difference in phase between the FSK signal and the VCO output. A positive voltage is produced when the receive signal leads the VCO, and a negative voltage is produced otherwise. After being filtered to minimize the effects of noise, the phase detector output drives the VCO in such a way as to reduce the phase difference. Ideally, the input control voltage of the demodulator VCO will be identical to the input of the modulator VCO. The loop filter, however, necessarily slows the demodulator response to minimize the effects of noise.

Frequency shift keying generally provides poorer error performance than PSK, particularly for multilevel signaling in a confined bandwidth. Older asynchronous modems for data transmission over the analog telephone network used FSK modulation. Newer systems do not use FSK because the simpler implementation compared to better performing techniques is no longer significant.

Minimum Shift Keying

One particular form of frequency shift keying that has received considerable attention in radio systems is minimum shift keying (MSK) [5]. Basically, MSK is binary FSK with the two signaling frequencies selected so that exactly 180° difference in phase shift exists between the two frequencies in one signal interval. In this manner MSK produces a maximum phase differential at the end of an interval using a minimum difference in signaling frequencies. Furthermore, an MSK signal maintains continuous phase at signaling transitions. For this reason MSK belongs to a class of FSK signals referred to as continuous-phase frequency shift keying (CPFSK). Figure 6.6 depicts a representative MSK waveform. Notice that there is exactly one-half cycle difference between a 1 signal and a 0 signal (2 cycles versus 1.5 cycles).

A mathematical expression for MSK signaling can be derived from Equation 6.5, where $m_n(t) = m_2(t)$ is a symmetric binary NRZ signal and $\Delta\omega = \pi/T$. The result for any particular signal interval is

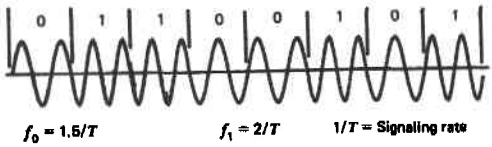


Figure 6.6 Minimum shift keying.

$$x(t) = \begin{cases} \cos\left(\omega_c t + \dfrac{\pi t}{2t} + \phi_0\right) & \text{(logic 1)} \\ \cos\left(\omega_c t - \dfrac{\pi t}{2t} + \phi_0\right) & \text{(logic 0)} \end{cases} \tag{6.6}$$

where $1/T$ is the signaling rate and ϕ_0 is the phase at the beginning of the signal interval ($\pm\pi$ in Figure 6.6).

The main attraction of MSK is its comparatively compact spectrum. Furthermore, with an appropriate means of detection, MSK can provide optimum error performance in terms of the energy-per-bit-to-noise-density ratio (E_b/N_0). The expression for the power spectral density of the MSK signal defined in Equation 6.6 is

$$S(\omega) = 16\pi^2 T\left(\frac{\cos z}{\pi^2 - 4z^2}\right)^2 \tag{6.7}$$

where $z = |\omega - \omega_c|T$. The frequency spectrum of an MSK signal is plotted in Figure 6.7, where it is compared to the frequency spectrum of PRK (2-PSK) signaling with the same data rate. Notice that the MSK spectrum is more compact and has its first spectral null at $3/4T$ instead of $1/T$ for PRK.*

Minimum shift keying is actually just one example of a general class of continuous-phase modulation (CPM) techniques that maintain a constant amplitude, narrow power spectrum, and good error performance. For a good overview and bibliography of this class of modulation schemes see reference [6]. CPM schemes have not been used in point-to-point microwave applications because they do not provide high information densities. Many versions of CPM have been used in satellite applications where nonlinear transponder amplifiers preclude the use of modulation techniques with multiple amplitude levels [7].

Gaussian MSK

Like MSK, Gaussian minimum shift keying (GMSK) produces a constant-amplitude and continuous-phase RF signal. GMSK differs from MSK through the use of a Gaussian baseband pulse shape in place of a square pulse shape for MSK. Because the Gaussian pulse rises and decays asymptotically with respect to a zero response level, it has a much more constrained bandwidth. Although it is conceivable that a GMSK signal could be generated by filtering the modulated signal, a typical implementation [8] utilizes baseband filtering, as shown in Figure 6.8, where it is contrasted to unfiltered MSK. The MSK signal is generated by direct FSK modulation of a carrier with a baseband signal that is scaled in amplitude to produce a modulation index or 0.5. A modulation index of this value produces the difference of 180° of phase shift for the two data values. One complication of the bandwidth-limiting filter of GMSK is the

*MSK is actually more closely related to 4-PSK and therefore is compared to it in a later section of this chapter.

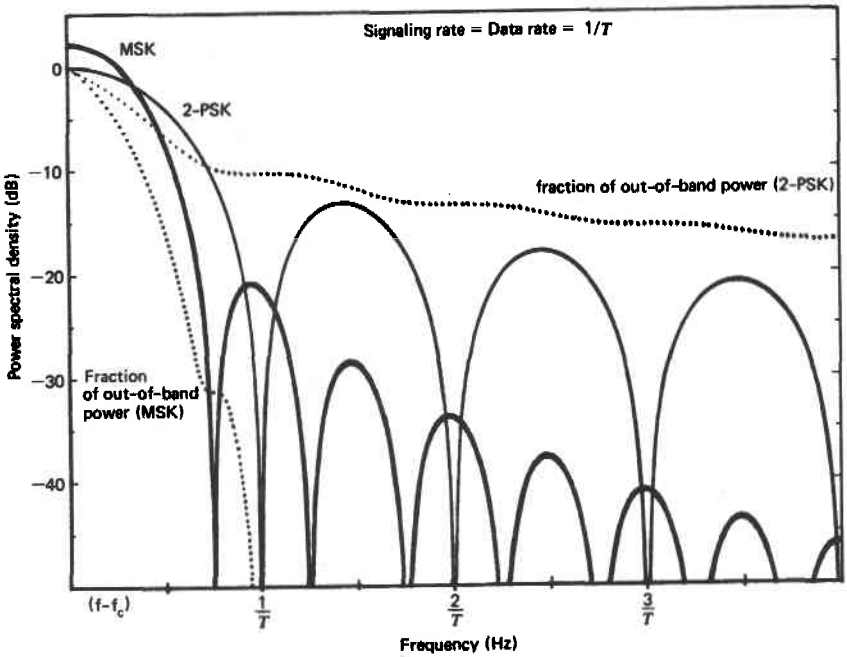


Figure 6.7 Power spectra of unfiltered MSK and 2-PSK signals.

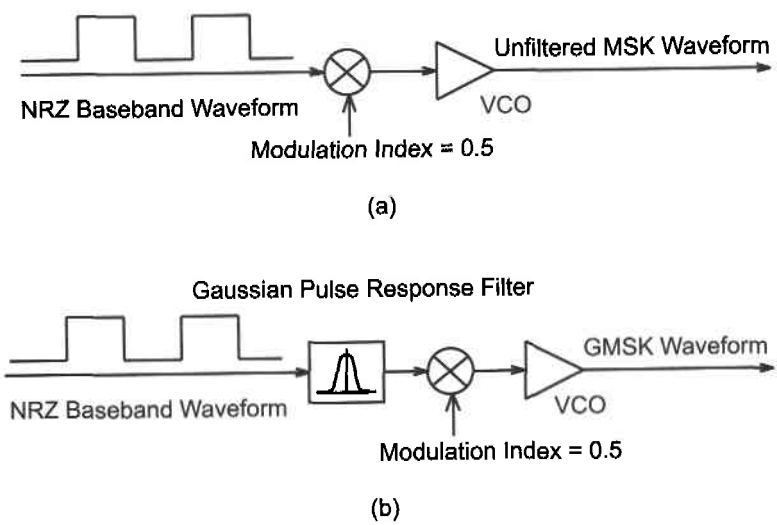


Figure 6.8 Comparison of (*a*) direct modulation MSK and (*b*) GMSK.

creation of intersymbol interference similar to partial-response systems discussed later in this chapter. In essence, GMSK trades a small amount of adjacent channel interference for a small amount of intersymbol interference. GMSK modulation is utilized in GSM digital cellular [9] discussed in Chapter 9 and in cellular digital packet data (CDPD) applications.

6.1.3 Phase Shift Keying

The second category of angle-modulated, constant-envelope signals is referred to as phase shift keying (PSK). Actually, one form of PSK has already been discussed as phase reversal keying (PRK), more commonly referred to as 2-PSK, indicating that each signal interval uses one of two phases that are 180° apart to encode binary data. Multiple-phase shift keying is also possible. Four-PSK (also called QPSK) and 8-PSK are the most common examples of multiple-phase PSK.

Phase shift keying (which includes 4-QAM described next) is the most popular modulation technique for intermediate information density high-performance applications. The popularity is primarily due to its constant envelope, insensitivity to level variations, and good error performance. Both 2-PSK and 4-PSK provide theoretical optimum error performance in terms of a signal-to-noise ratio (E_b/N_0).

A general expression for n-ary phase shift keying is provided in Equation 6.8. This expression assumes 100% modulation is employed. That is, the phase shift from one interval to the next can range anywhere from −180° to +180°. It is possible to devise PSK systems with lower modulation indices that allow only transitions to neighboring phases:

$$x(t) = \cos\left(\omega_c t + \frac{m_n(t)\,\Delta\phi}{2}\right) \tag{6.8}$$

where $\Delta\phi = 2\pi/n$ is the separation between adjacent signal phases and $m_n(t)$ is a symmetric n-level NRZ baseband signal with levels $\pm 1, \pm 3, \ldots$.

Examples of typical 2-PSK and 4-PSK waveforms are shown in Figure 6.9. The signaling rate for the 4-PSK system is shown to be exactly one-half the 2-PSK signaling rate so that equal data rates are provided. The same figure also shows corresponding phasor diagrams of the signaling phases of a cosine wave as defined in Equation 6.8. Other phase orientations are possible. The particular phases shown, however, are convenient for later discussions relating 4-PSK systems to other types of digital modulation.

Quadrature Signal Representations

Despite the somewhat exotic sounding name, quadrature signal representations are a very convenient and powerful means of describing PSK signals and many other digitally modulated signals. Quadrature signal representations involve expressing an arbitrary phase sinusoidal waveform as a linear combination of a cosine wave and a sine

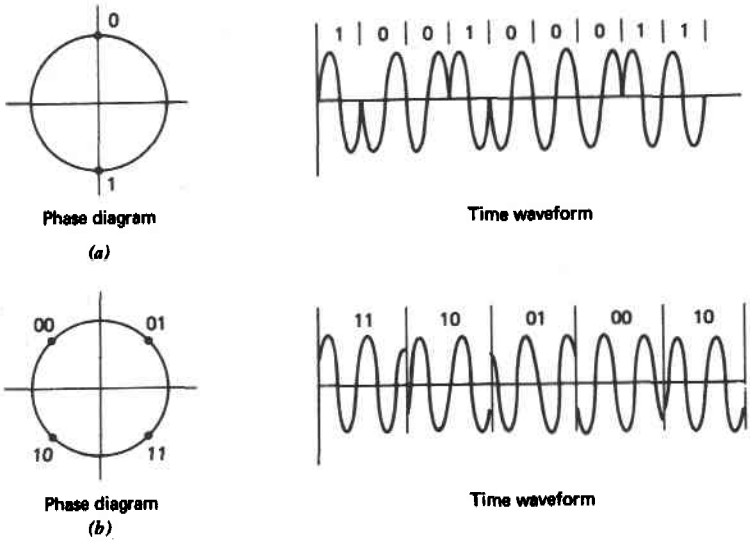


Figure 6.9 Phase shift keying: (*a*) 2-PSK; (*b*) 4-PSK.

wave with zero starting phases. The derivation of this representation is provided by the trigonometric identity:

$$\cos(\omega_c t + \phi) = \cos\phi \cos\omega_c t - \sin\phi \sin\omega_c t \tag{6.9}$$

Notice that $\cos\phi$ and $\sin\phi$ are constants over a signaling interval and hence represent coefficients for expressing $\cos(\omega_c t + \phi)$ as a linear combination of the signals $\cos\omega_c t$ and $\sin\omega_c t$. Since $\cos\omega_c t$ and $\sin\omega_c t$ are 90° out of phase with respect to each other, they are orthogonal in a phasor diagram and hence are said to be "in quadrature."

In essence, $\cos\omega_c t$ and $\sin\omega_c t$ represent basis vectors in a two-dimensional phasor diagram. The cosine signal is usually referred to as the in-phase or I signal, and the sine signal is referred to as the out-of-phase or Q signal. Table 6.2 provides an example

TABLE 6.2 Quadrature Signal Coefficients for 4-PSK Modulation

	Quadrature Coefficients		
Data Values	$\cos\omega_c t$	$\sin\omega_c t$	Composite Signal
01	0.707	−0.707	$\cos(\omega_c t + \pi/4)$
00	−0.707	−0.707	$\cos(\omega_c t + 3\pi/4)$
10	−0.707	0.707	$\cos(\omega_c t - 3\pi/4)$
11	0.707	0.707	$\cos(\omega_c t - \pi/4)$

TABLE 6.3 Quadrature Signal Coefficients for 8-PSK Modulation

	Quadrature Coefficients		
Data Values	$\cos \omega_c t$	$\sin \omega_c t$	Composite Signal
011	0.924	−0.383	$\cos(\omega_c t + \pi/8)$
010	0.383	−0.924	$\cos(\omega_c t + 3\pi/8)$
000	−0.383	−0.924	$\cos(\omega_c t + 5\pi/8)$
001	−0.924	−0.383	$\cos(\omega_c t + 7\pi/8)$
101	−0.924	0.383	$\cos(\omega_c t - 7\pi/8)$
100	−0.383	0.924	$\cos(\omega_c t - 5\pi/8)$
110	0.383	0.924	$\cos(\omega_c t - 3\pi/8)$
111	0.924	0.383	$\cos(\omega_c t - \pi/8)$

of quadrature signal representations for the 4-PSK signals presented in Figure 6.9. Table 6.3 provides a corresponding specification for an 8-PSK system using signal phases as provided in the phasor diagram of Figure 6.10. The phasor diagram assumes counterclockwise rotation, and hence the sine function lags the cosine function by 90°.

Most of the rest of this chapter relies heavily on quadrature signal representations to describe various modulation concepts, analyses, and implementations.

Modulator Implementations

A variety of techniques are possible for implementing PSK modulators. As mentioned when discussing PRK modulation, a 2-PSK modulator can be implemented by merely inverting the carrier (multiplying by −1) for a logic 0 and by not inverting for a logic 1. Some of the basic techniques used for generating multiple-phase PSK signals are the following:

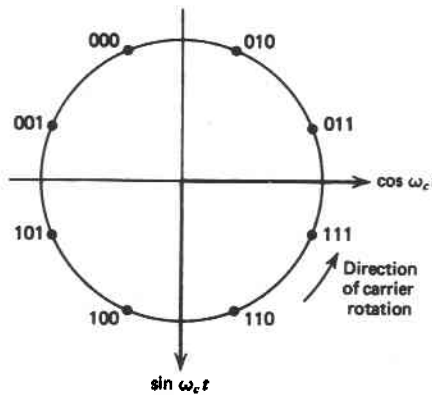


Figure 6.10 Phasor diagram of 8-PSK signal.

1. Synthesis of the desired waveforms using digital signal processing at suitably low carrier frequencies (as in voiceband modems).
2. Generating multiple phases of a single carrier and selecting between the phases depending on the data values.
3. Using controlled delays selected through a switching arrangement to provide the desired phase shifts. Delays are often used to generate the separate signals in method 2.
4. Generating the PSK signals as a linear combination of quadrature signals.

None of the foregoing techniques represent a direct implementation of multiphase PSK modulation as defined in Equation 6.8. To implement Equation 6.8 directly, a device is needed that produces carrier phase shifts in direct proportion to the levels in the baseband signal $m_n(t)$. Except for the special case of 2-PSK where a multiplier can be used to produce $\pm 180°$ phase shifts, such devices do not exist.

Direct modulation by baseband signals can produce PSK signals if a quadrature signal implementation is used, as indicated in method 4. Actually, two multilevel baseband signals need to be established: one for the in-phase (I) signal and one for the out-of-phase (Q) signal. These baseband signals are referred to as $m_I(t)$ and $m_Q(t)$ for the I and Q signals, respectively. The levels chosen for the two baseband signals correspond to the coefficients needed to represent a PSK signal as a linear combination of the I and Q signals. For example, Figure 6.11 shows how an 8-PSK signal, defined in Table 6.3, can be generated by adding two amplitude-modulated quadrature signals. A block diagram of the corresponding implementation is provided in Figure 6.12. This particular form of modulator is chosen, not so much as a recommendation for actual implementation, but because it demonstrates important modulation concepts and is useful in analyzing the spectrum requirements of PSK signaling.

Demodulator Implementation

Owing to the constant envelope,* all PSK systems must be detected with the aid of a local reference. For 2-PSK systems, the ideal reference is coherent with one of the two possible phases. When this reference is multiplied (mixed) with an in-phase signal, a maximum positive output is obtained. When multiplied by the opposite phase, a maximum negative output is obtained. In this manner, 2-PSK provides antipodal performance when a local coherent reference is established. The demodulation process for 2-PSK is presented in Figure 6.4 and Equation 6.4, where the mixing and filtering process effectively recovers the baseband signal $m_2(t)$.

The combination of a mixer and a low-pass filter shown in Figure 6.4 is generally referred to as a phase detector. The phase detection property is represented mathematically as

*Discussions to this point have not considered the effects of filtering on constant-envelope signals. A heavily filtered PSK signal does not have a constant envelope. However, as long as filtering occurs after channel nonlinearities (e.g., power amplifiers), the most harmful effect of a nonlinearity, spectrum spreading, is avoided.

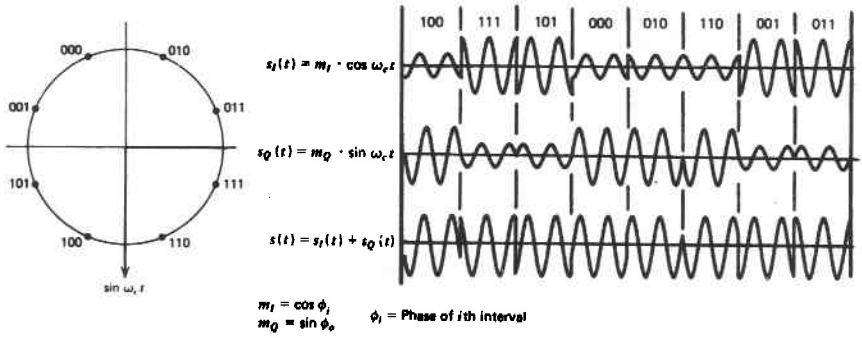


Figure 6.11 Generation of 8-PSK signals by superposition of quadrature-amplitude-modulated signals.

$$
\begin{aligned}
y_I(t) &= \text{lowpass}\{\cos(\omega_c t + \phi)\,(2)\cos\omega_c t\} \\
&= \text{lowpass}\{\cos\phi + \cos\phi\cos 2\omega_c t - \sin 2\omega_c t\} \\
&= \cos\phi
\end{aligned}
\tag{6.10}
$$

where lowpass$\{\cdot\}$ is a low-pass filter function designed to remove twice carrier terms.

When detecting 2-PSK modulation, a single-phase detector indicates whether the received phase is closer to 0° or to 180°. The desired information is directly available as the polarity of the phase detector output $\cos\phi$. In multiphase systems, however, the information provided by a single-phase detector is inadequate for two reasons. First,

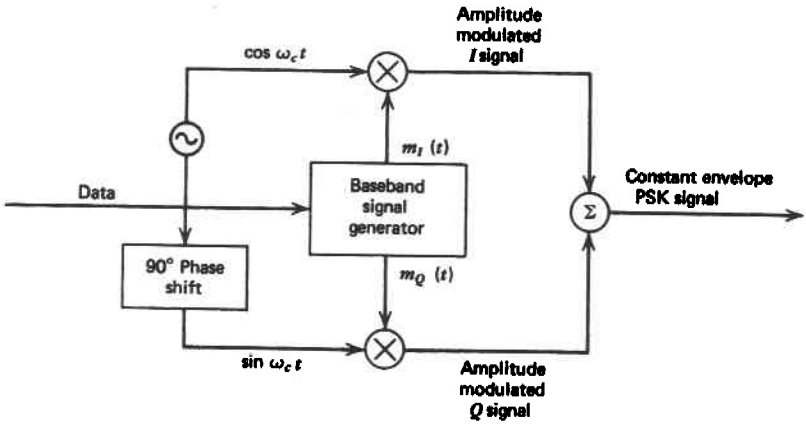


Figure 6.12 Generalized PSK modulator structure.

a measure of cos ϕ provides no information as to whether ϕ is positive or negative. Second, the output of the phase detector is proportional to the signal amplitude as well as to cos ϕ. Hence the magnitude of the phase detector output is meaningless unless referenced to the signal amplitude.

Both of the aforementioned problems are overcome if a second mixer and filter measure the phase with respect to a different reference. As expected, the best performance is obtained when the second reference is orthogonal to the first. If $y_I(t)$ is the output of the first phase detector defined in Equation 6.10, then the second phase detector output $y_Q(t)$ becomes

$$\begin{aligned} y_Q(t) &= \text{lowpass}\{\cos(\omega_c t + \phi)\,(2)\,\sin \omega_c t\} \\ &= -\sin \phi \end{aligned} \tag{6.11}$$

The second phase detector not only resolves the positive/negative phase ambiguity but also eliminates the need to establish an amplitude reference. All decisions can be based on the polarity of a phase detector output and not on the magnitude. As a first example, consider detection of the 4-PSK signals defined in Table 6.2. Notice that the first data bit in a pair is a 0 when the phase angle is positive ($\pi/4$ or $3\pi/4$) and a 1 otherwise. Hence the first data bit is completely specified by the polarity of sin ϕ: the output of the second phase detector $y_Q(t)$. Similarly, the second data bit is a 1 when the phase is $\pm\pi/4$, indicating that the polarity of $y_I(t)$ provides all information necessary to detect the second bit. The basic implementation of a 4-PSK (QPSK) demodulator/detector is shown in Figure 6.13. Also shown is a 4-PSK modulator to emphasize the relationships between the modulator and demodulator. A 4-PSK system is presented specifically because it is a popular system and because it is a useful foundation for describing other modulation techniques.

An important concept to notice in the 4-PSK system shown in Figure 6.13 is that there are, in essence, two separate data streams. The modulator divides the incoming bit stream so that bits are sent alternately to the in-phase modulator I and the out-of-phase modulator Q. These same bits appear at the output of the respective phase detectors in the demodulator, where they are interleaved back into a serial bit stream. In this manner two essentially independent binary PSK channels are established using the orthogonal carriers cos $\omega_c t$ and sin $\omega_c t$. The channels are usually referred to as the I and Q channels, respectively. This technique of establishing two channels within an existing bandwidth is sometimes referred to as *quadrature multiplexing*.

As long as the carriers in the modulator and the references in the demodulator are truly orthogonal (coherence is maintained for both channels in the receiver), the I and Q channels do not interfere with each other. Any amount of misalignment in these relationships causes crosstalk between the two quadrature channels. Crosstalk also arises if there is unequalized phase distortion in the transmission channel.

At first thought it might seem that quadrature multiplexing increases the capacity of a given bandwidth by a factor of 2. It must be remembered, however, that the binary PSK signal on each quadrature channel is a double-sideband signal. Hence the bandwidth of the channel, *without quadrature multiplexing*, is only 50% utilized in com-

parison to a single-sideband system. When quadrature channels are used, a single-sideband operation is no longer possible since the sideband separation process destroys the orthogonality of the two signals. In essence, quadrature multiplexing only recovers the loss of capacity incurred by the double-sideband spectrum. In fact, some single-sideband modulators [10] possess a remarkable resemblance to the QPSK modulator shown in Figure 6.13.

Demodulation and detection of higher level PSK systems are complicated by the fact that the use of only two references does not provide a simple means of detecting all data bits. There are two basic ways in which simple decisions (positive versus negative) can be established to detect all data. One method is to establish more references in the receiver and measure the phase of the received signal with respect to the additional references. The second method is to use only two references and related phase detectors and generate all additional measurements as linear combinations of the first two.

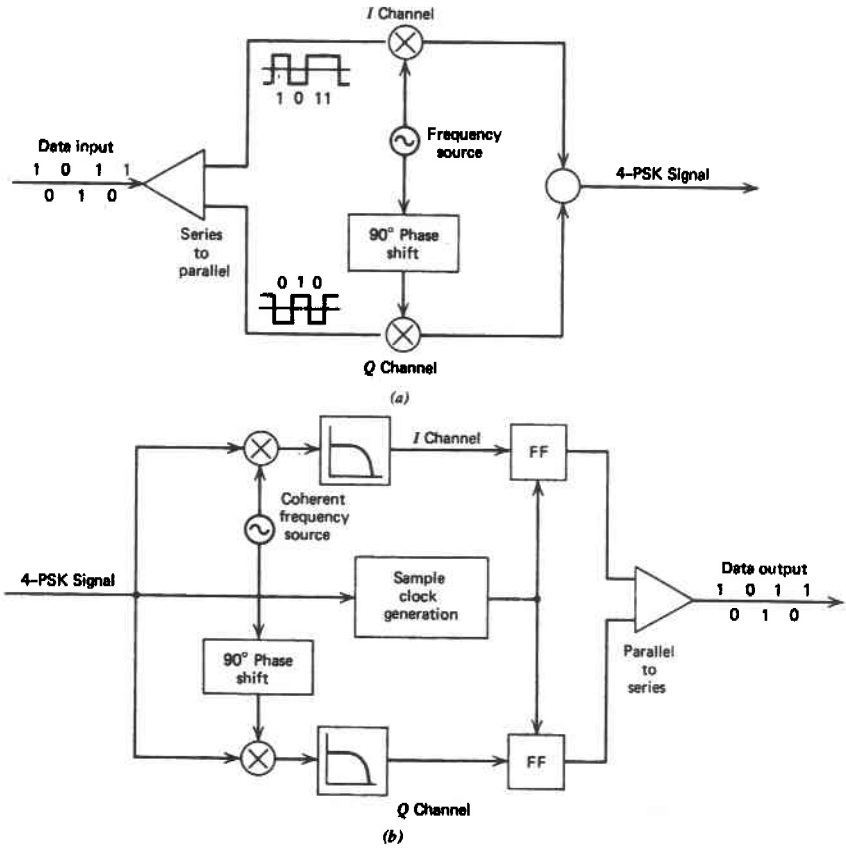


Figure 6.13 Four-PSK modulator–demodulator structure: (*a*) modulator; (*b*) demodulator–detector.

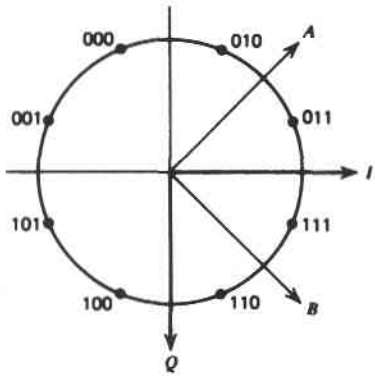


Figure 6.14 Receiver references for 8-PSK detection.

As an example of the first method, consider the 8-PSK system defined in Table 6.3 and shown in Figure 6.10. Optimum detection for this system can be obtained if two additional phase detector outputs are provided with respect to references at $+\pi/4$ and $-\pi/4$. The two new references are designated A and B as shown in Figure 6.14.

The outputs of four phase detectors corresponding to the four references are determined as follows:

$$
\begin{aligned}
y_Q &= \text{lowpass}\{\cos(\omega_c t + \phi)(2)\sin\omega_c t\} = -\sin\phi \\
y_B &= \text{lowpass}\{\cos(\omega_c t + \phi)(2)\sin(\omega_c t + \tfrac{1}{4}\pi)\} \\
&= 0.707\cos\phi - 0.707\sin\phi \\
y_I &= \text{lowpass}\{\cos(\omega_c t + \phi)(2)\cos\omega_c t\} = \cos\phi \\
y_A &= \text{lowpass}\{\cos(\omega_c t + \phi)(2)\cos(\omega_c t + \tfrac{1}{4}\pi)\} \\
&= 0.707\cos\phi + 0.707\sin\phi
\end{aligned}
\tag{6.12}
$$

As an aid in determining the appropriate decision logic, these equations have been evaluated for each of the eight possible signal phases and listed in Table 6.4. Examination of Table 6.4 indicates that the first data bit is 1 whenever y_Q is positive. Similarly, the second data bit is 1 whenever y_I is positive. The third data bit is a 1 whenever y_A, y_I, and y_B are all positive or when they are all negative. Hence the third data bit is determined as a logical combination of phase detector outputs. In summary,

$$D_1 = Q \qquad D_2 = I \qquad D_3 = AIB + \overline{A}\,\overline{I}\,\overline{B} \tag{6.13}$$

where D_i is the ith data bit and Q, I, A, and B are logic variables representing positive outputs from y_Q, y_I, y_A, and y_B, respectively.

TABLE 6.4 Eight-PSK Phase Detector Outputs

Data	Phase	Y_Q	Y_B	Y_I	Y_A
011	$\pi/8$	−0.383	0.383	0.924	0.924
010	$3\pi/8$	−0.924	−0.383	0.383	0.924
000	$5\pi/8$	−0.924	−0.924	−0.383	0.383
001	$7\pi/8$	−0.383	−0.924	−0.924	−0.383
101	$-7\pi/8$	0.383	−0.383	−0.924	−0.924
100	$-5\pi/8$	0.924	0.383	−0.383	−0.924
110	$-3\pi/8$	0.924	0.924	0.383	−0.383
111	$-\pi/8$	0.383	0.924	0.924	0.383

Another method of detecting 8-PSK signals, which avoids two extra references and phase detectors, is revealed in the phase detector Equation 6.12. The extra measurements y_A and y_B can be determined as

$$y_A = 0.707y_I - 0.707y_Q \qquad y_B = 0.707y_I + 0.707y_Q \tag{6.14}$$

Hence the y_A and y_B measurements can be obtained as linear combinations of the quadrature channel phase measurements y_I and y_Q, and no additional phase detectors are required. The resulting 8-PSK demodulator/detector is shown in Figure 6.15. Notice that implementation of the linear combinations in Equation 6.14 can ignore the magnitude of the 0.707 multipliers since only the sign of the result is needed.

The linear combinations in Equation 6.14 essentially represent a rotation of the quadrature channel basis vectors through an angle of $\pi/4$ radians. By changing the angle of rotation, the linear combinations needed for other phase measurements are eas-

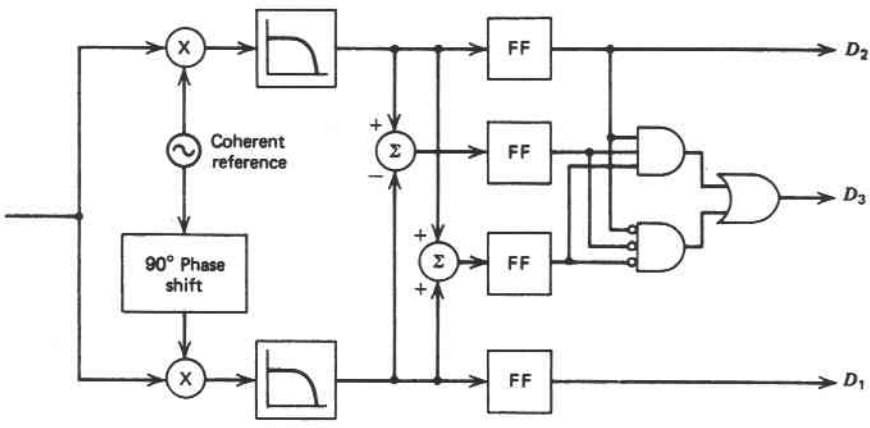


Figure 6.15 Eight-PSK demodulator–detector using only two references.

ily determined. Hence all phase-detector-based demodulators can be implemented with two phase detectors and as many linear combinations as needed. The general equation for the linear combinations produced by a rotation of α radians is

$$y_A = y_Q \cos\alpha - y_I \sin\alpha \qquad y_B = y_Q \sin\alpha + y_I \cos\alpha \tag{6.15}$$

Reference and Clock Recovery

The demodulators discussed in the preceding paragraphs all require a local, phase-coherent, carrier reference for their operation. Furthermore, multiple-phase systems require at least one more reference in quadrature with the first one. Recovery of any reference is complicated by the fact that phase shift keying is a double-sideband *suppressed* carrier modulation system. In other words, there is no discrete spectral line at the carrier frequency as there is with some other types of modulation. In fact, the FCC has ruled specifically that no spectral lines are allowed in the transmitted signal. The absence of a spectral line at the carrier frequency is overcome by using one of several nonlinear processing techniques [11]. After one coherent reference is established, a quadrature reference is obtained by delaying or differentiating the first.

Clock recovery is obtained in several ways. Once carrier recovery is accomplished and demodulation occurs, the clock can be obtained by locking onto transitions in the baseband signal. Baseband clock recovery occurs as though the signal has never been modulated and demodulated. In contrast, the clock can sometimes be recovered directly from the modulated signal. If a PSK signal is heavily filtered, the envelope is not constant and, in fact, contains amplitude modulation at the signaling frequency. Hence the clock can be recovered, independently of signal demodulation, by envelope detecting a heavily filtered PSK signal [12].

Clock transitions are often assured by purposely shifting the carrier reference halfway between signal points in every interval. Using 4-PSK, for example, the references would be shifted by 45°. During any one interval there are still only four possible signal states separated by 90°. However, the signal states alternate between being aligned with the I and Q axes and being aligned at 45°, as shown in Figure 6.9. Because a phase shift is ensured with every symbol interval, both the (filtered) carrier envelope and the recovered baseband signal have amplitude transitions. Without the reference shifts, proper clock recovery requires a minimum density of data transitions as described in Chapter 4.

Differential Detection

As an alternative to recovering a coherent reference, some systems merely compare the phase in the present interval to the phase in the previous interval. The signal received in the previous interval is delayed for one signal interval and is used as a reference to demodulate the signal in the next interval. Assuming that the data have been encoded in terms of phase shift, instead of absolute phase positions, the data can be decoded properly. Hence this technique, referred to as "differential detection," inherently requires differential encoding.

In general, PSK systems require differential encoding since the receivers normally have no means of determining whether a recovered reference is a sine reference or a cosine reference.* Furthermore, the polarity of the recovered reference is ambiguous. Thus error probabilities for PSK systems are doubled automatically because of the differential encoding process. Differential detection, on the other hand, implies an even greater loss of performance since a noisy reference is used in the demodulation process. Typically, differential detection imposes a penalty of 1 to 2 dB in signal-to-noise ratio [10].

PSK Spectra

By far the easiest way to determine the spectrum of a PSK signal is to analyze the baseband waveforms applied to the quadrature channels. Owing to the orthogonality of the two channels, the signals are uncorrelated, and the composite spectrum is merely the sum of the individual (identical) spectra.

In either 2-PSK or 4-PSK systems the baseband signal is a symmetric two-level NRZ waveform. The corresponding spectrum is the common $\sin(x)/x$ spectrum shown in Figure 4.2. High-level systems (8-PSK or greater) use symmetric multilevel NRZ baseband signals similar to that shown in Figure 4.16. The multiphase PSK baseband signal is somewhat different since unevenly spaced levels as defined in Table 6.3 are used.

As mentioned in Chapter 4, a multilevel NRZ signal has the same spectrum as a two-level signal. Hence all conventional PSK systems produce a spectrum that follows the $\sin(x)/x$ response defined in Equation 4.1 but translated to the carrier frequency.† Figure 6.16 shows the PSK spectrum for two-, four-, and eight-phase systems designed to provide the same data rate. Hence the higher level systems signal at lower rates and have proportionately narrower spectra.

PSK Error Performance

The error performance of any digital modulation system is fundamentally related to the distance between points in a signal space diagram. For example, a 2-PSK system, as represented in the phase diagram of Figure 6.9, is capable of optimum error performance since the two signal points have maximum separation for a given power level (radius of the circle). In other words, one 2-PSK signal is the exact negative of the other. Hence 2-PSK modulation provides antipodal error performance as defined in Chapter 4.

The error performance of a multiphase PSK system is easily compared to a 2-PSK system by determining the relative decrease in the error distance (voltage output of a properly referenced phase detector). In addition to the error distance, however, the relative values of the noise bandwidths must also be considered. (Recall that the noise bandwidth effectively determines the variance of the noise samples.)

*Absolute phase can be determined if a particular pattern in the data stream such as a framing pattern is unambiguously related to some particular phase of the carrier.

†The baseband levels must be uncorrelated to each other to produce a $\sin(x)/x$ spectrum. If phase transitions from one interval to the next are restricted in some manner, the baseband levels are correlated, and a different spectrum results.

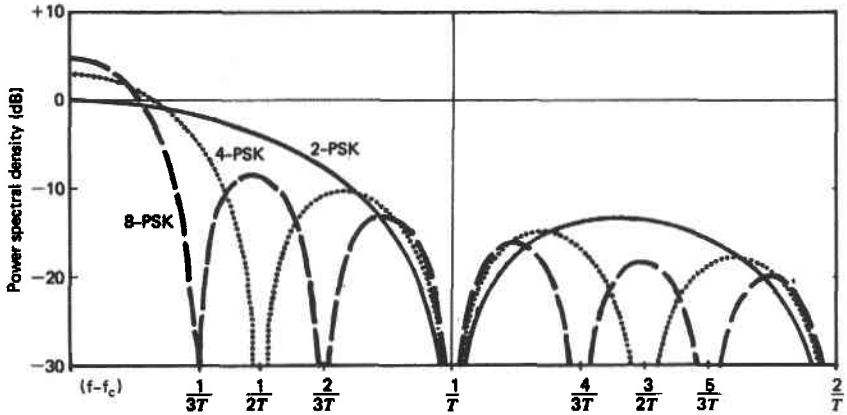


Figure 6.16 Spectrum of unfiltered PSK signals carrying equal data rates.

The general expression for the distance between adjacent points in a multiphase PSK system is

$$d = 2\sin\left(\frac{\pi}{N}\right) \tag{6.16}$$

where N is the number of phases.

A general expression of the bit error probability (or bit error rate) of an N-phase PSK system is derived in Appendix C as

$$P_E = \frac{1}{\log_2 N}\,\mathrm{erf}(z) \tag{6.17}$$

where

$$z = \sin\left(\frac{\pi}{N}\right)(\log_2 N)^{1/2}\left(\frac{E_b}{N_0}\right)^{1/2}$$

Equation 6.17 reveals that, with respect to E_b/N_0, 4-PSK provides the same error performance as does 2-PSK. Thus, as mentioned earlier, both systems provide optimum performance, but 4-PSK utilizes half as much bandwidth. The 4-PSK system has an error distance that is 3 dB smaller than the error distance of 2-PSK. However, the shorter error distance is offset by a 3-dB decrease in the noise bandwidth (indicating a 3-dB reduction in noise power at the detector). For the 4-PSK system to have the same noise power at the detector it would have to experience a 3-dB greater noise spectral density. Hence conventional signal-power-to-noise-power ratios (SNRs) can be misleading parameters for comparing digital modulation systems. However, as mentioned in Chapter 4, error rate performances in terms of SNRs are desired when

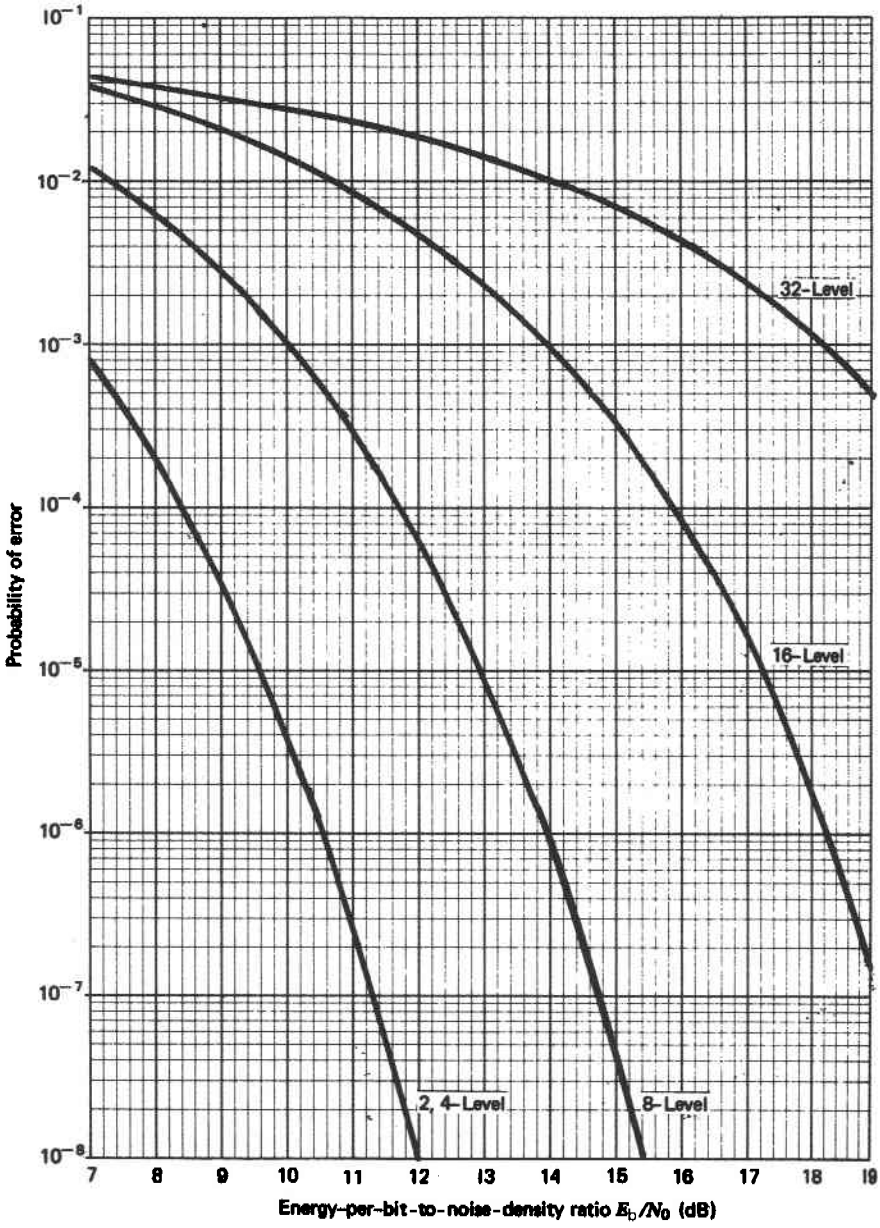


Figure 6.17 Error rates of PSK modulation systems.

determining the effects of interference or when specifying error rates with respect to measurable quantities. Figure 6.17 displays the error performance of 2-, 4-, 8-, 16-, and 32-PSK as a function of E_b/N_0. Appendix C provides the relationships needed for error rates in terms of SNR.

6.1.4 Quadrature Amplitude Modulation

As described previously, a convenient means of representing phase shift keying with four or more phases involves the use of quadrature signals. In the case of 4-PSK, the quadrature signals represent two separate channels by virtue of the independence of the baseband signals for each quadrature channel. In higher level PSK systems, the level of a baseband signal for the I channel is not independent of the baseband level for the Q channel (see Table 6.3 or Figure 6.11). After the baseband signals have been established, however, the quadrature channel modulation and demodulation processes are independent for all PSK systems.

Quadrature amplitude modulation (QAM) can be viewed as an extension of multiphase PSK modulation wherein the two baseband signals are generated independently of each other. Thus two completely independent (quadrature) channels are established including the baseband coding and detection processes. In the special case of two levels (± 1) on each channel, the system is identical to 4-PSK and is often referred to as such. Higher level QAM systems, however, are distinctly different from the higher level PSK systems. Figure 6.18 shows a signal constellation of a 16-QAM system obtained from four levels on each quadrature channel. The dots represent composite signal points while the hatch marks on the axes represent amplitude levels in each quadrature channel. Figure 6.19 shows a basic QAM modulator and demodulator structure along with a representative waveform for 16-QAM.

Notice that, in contrast to PSK signals, the QAM signal shown in Figure 6.18 does not have a constant envelope. A constant envelope is maintained with PSK modulation by restricting the combination of levels on the quadrature channels. A QAM system does not restrict the combinations since the levels on each channel are selected independently. Thus QAM modulation cannot be used with saturating amplifiers.

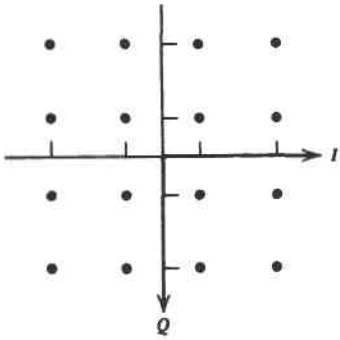


Figure 6.18 Signal constellation of 16-QAM modulation.

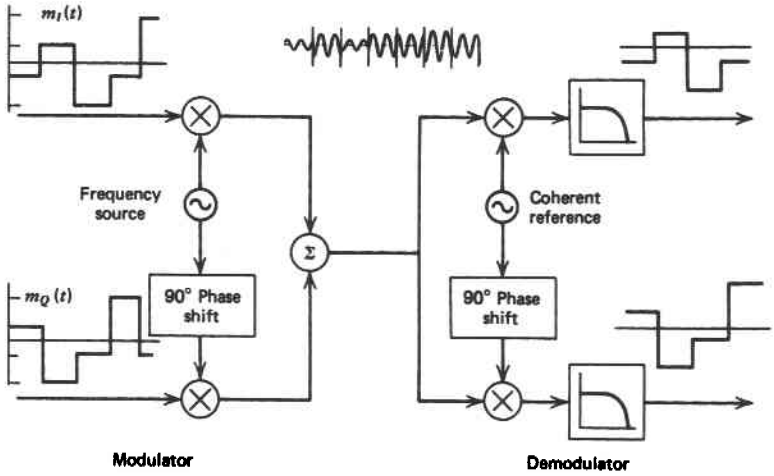


Figure 6.19 QAM modulator–demodulator.

The spectrum of a QAM system is determined by the spectrum of the baseband signals applied to the quadrature channels. Since these signals have the same basic structure as the baseband PSK signals, QAM spectrum shapes are identical to PSK spectrum shapes with equal numbers of signal points. Specifically, 16-QAM has a spectrum shape that is identical to 16-PSK, and 64-QAM has a spectrum shape identical to 64-PSK.

Even though the spectrum shapes are identical, the error performances of the two systems are quite different. With large numbers of signal points, QAM systems always outperform PSK systems. The basic reason is that the distance between signal points in a PSK system is smaller than the distance between points in a comparable QAM system. Figure 6.20 compares the signal points of a 16-QAM system with the signal set of a 16-PSK system using the same peak power.

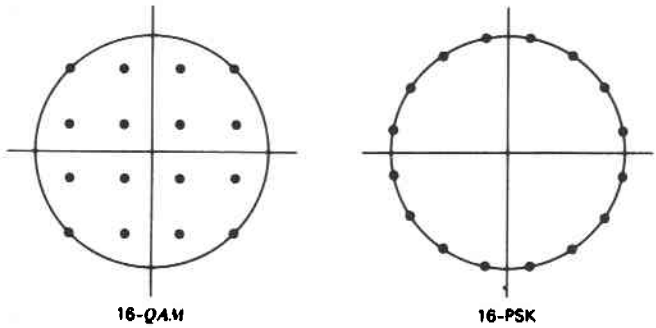


Figure 6.20 Comparison of 16-QAM and 16-PSK signal sets.

The general expression for the distance between adjacent signal points in a unit peak-amplitude QAM system with L levels on each axis is

$$d = \frac{\sqrt{2}}{L-1} \tag{6.18}$$

Equations 6.16 and 6.18 reveal that an n-ary QAM system has an advantage over an n-ary PSK system with the same peak power level. In terms of average power levels, the QAM system has an even greater advantage. The following equation, derived in Appendix C, provides a general expression for the peak-to-average ratio of a QAM system:

$$\frac{\text{Peak power}}{\text{Average power}} = \frac{L(L-1)^2}{2\sum_{i=1}^{L/2}(2i-1)^2} \tag{6.19}$$

Example 6.1. Determine the error performance of a 16-PSK system relative to a 16-QAM system with the same peak power level. Also determine the relative performance with respect to identical average powers.

Solution. Since the two systems provide identical numbers of signal points, they signal at the same rate and require the same bandwidth for a given data rate. Thus the relative error performance is completely determined by the relative distances between signal points. (When different signal rates are used, the effect of different noise bandwidths in the receivers must also be considered.) Evaluating Equations 6.16 and 6.18 indicates that 16-QAM has a 1.64-dB advantage over 16-PSK for a given peak power. Equation 6.19 indicates that 16-QAM has a peak-to-average power ratio of 2.55 dB. Since PSK systems have unity peak-to-average ratios, the advantage of a 16-QAM system over a 16-PSK is 4.19 dB for equal average powers.

The results of Example 6.1 show that 16-QAM is a significantly better modulation format than 16-PSK for applications like voiceband models that have no technology limits.* In digital microwave applications the performance advantage of 16-QAM is diminished by available amplifier technology, but, nevertheless, it is at 16 signal points that PSK modulation defers to QAM modulation for performance reasons. Some first-generation digital radios used 8-PSK modulation [13] and some used

*Voiceband modems are not restricted by technology but by signal power limits to prevent interference with other signals in the network. It is ironic that these restrictions arise primarily from old FDM analog radio systems in which a high-powered signal in one channel could create interference (crosstalk) into other channels. In a predominantly digital network the main signal limitation would be the saturation point of PCM encoders (and possibly crosstalk in subscriber pairs). It is also ironic that increasing the signal power of a voiceband modem (V.34 and earlier) to fall just short of the PCM saturation point would not improve performance in a mostly digital network. Performance of high-speed modems in this case is primarily determined by multiplicative noise created by PCM companders [18].

$= 21.96\,\text{dB}$

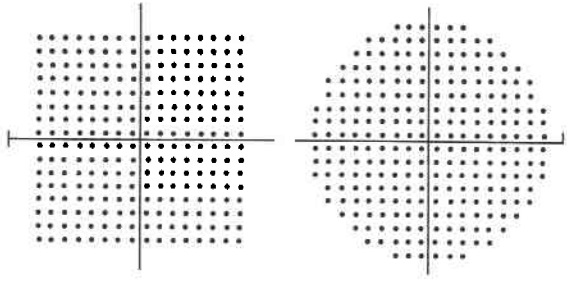


Figure 6.21 256 QAM and stepped square 256 QAM.

16-QAM [14] with comparable performance. Second-generation radios use 64-QAM [15] followed by third-generation radios with 256-QAM [16, 17].

Reliable operation of high-density modulation formats like 256-QAM requires extremely linear amplifiers to tolerate the wide range of signal amplitudes and extremely accurate adaptive equalizers to remove small percentages of intersymbol interference that arise when a high-amplitude pulse is transmitted adjacent to a low-amplitude pulse. Amplifier and equalizer performance requirements are basically a function of the dynamic range of the signal amplitudes*:

$$\text{Dynamic range} = 10 \log_{10}\left(\frac{P_{\max}}{P_{\min}}\right) \tag{6.20}$$

The dynamic range of a square QAM modulation format can be improved by modifying the signal set to eliminate corner points that produce high peak powers. One such technique referred to as 256-stepped square QAM (256-SSQAM) [19] is shown in Figure 6.21, where it can be compared with the signal set of conventional 256-QAM.

Example 6.2. Determine the dynamic range of both the conventional 256-QAM signal set and the 256-SSQAM signal set shown in Figure 6.21.

Solution. Using integral units along each quadrature axis of 1, 3, 5, 7, 9, 11, 13, and 15 for 256-QAM, the dynamic range (DR) is

$$\text{DR(256-QAM)} = 10 \log_{10}\left(\frac{15^2 + 15^2}{1^2 + 1^2}\right)$$

$$= 23.52\ \text{dB}$$

*If the quadrature carriers are perfectly recovered, intersymbol interference would only be a functio n of the dynamic range along each axis. To make an allowance for quadrature carrier phase error, which causes interference between I and Q channels, the composite signal power is used.

The highest amplitude signal point of 256-SSQAM is the (5,17) point. Thus the dynamic range of the 256-SSQAM signal is

$$\mathrm{DR(256\text{-}SSQAM)} = 10\log_{10}\left(\frac{5^2+17^2}{1^2+1^2}\right)$$

$$= 21.96 \text{ dB}$$

The result of Example 6.2 shows that the peak power of 256-SSQAM is 1.56 dB lower than the peak power of conventional 256-QAM for the same signal point separation. Thus in a peak-power limited application, 256-SSQAM can operate with 1.56 dB more signal separation than can conventional 256-QAM. When other, less significant factors are considered, reference [20] reports that a 140-Mbps 256-SSQAM radio at 6 GHz has a 2-dB performance advantage.

The theoretical error rate equation for QAM modulation, derived in Appendix C, is identical to the multilevel baseband error rate equation with the same number of levels. This equation is used to produce error rate curves for 4-, 16-, 64-, and 256-QAM modulation in Figure 6.22. Table 6.5 compares the most common forms of digital modulation used in point-to-point microwave radio systems.

Offset Keying

Because of the interdependence of the baseband signals in high-level PSK modulation, the signal transitions on the quadrature channels necessarily coincide. However, since QAM systems independently modulate the quadrature channels, they are not constrained to align the signal intervals on the two channels. When the signal intervals overlap each other by 50% (Figure 6.23), the mode of operation is referred to as offset keying. Offset keying is commonly used on 4-QAM systems (more commonly referred to as 4-PSK or QPSK systems). The main advantage of offset keying lies in the ability of the reference recovery circuitry to become synchronized to the incoming carrier at lower signal-to-noise ratios than conventional (aligned) QAM and QPSK systems [21, 22].

QAM Representation of Minimum Shift Keying

The preceding discussions of QAM systems have assumed the use of baseband signals with NRZ level encoding. A more general view of QAM allows arbitrary pulse shapes in generating the baseband signals. One particularly interesting pulse shape is a half-sinusoid pulse:

$$A(t) = \cos\frac{\pi t}{T} \quad \left(-\tfrac{1}{2}T \le t \le \tfrac{1}{2}T\right) \tag{6.21}$$

where T is the duration of a signal interval.

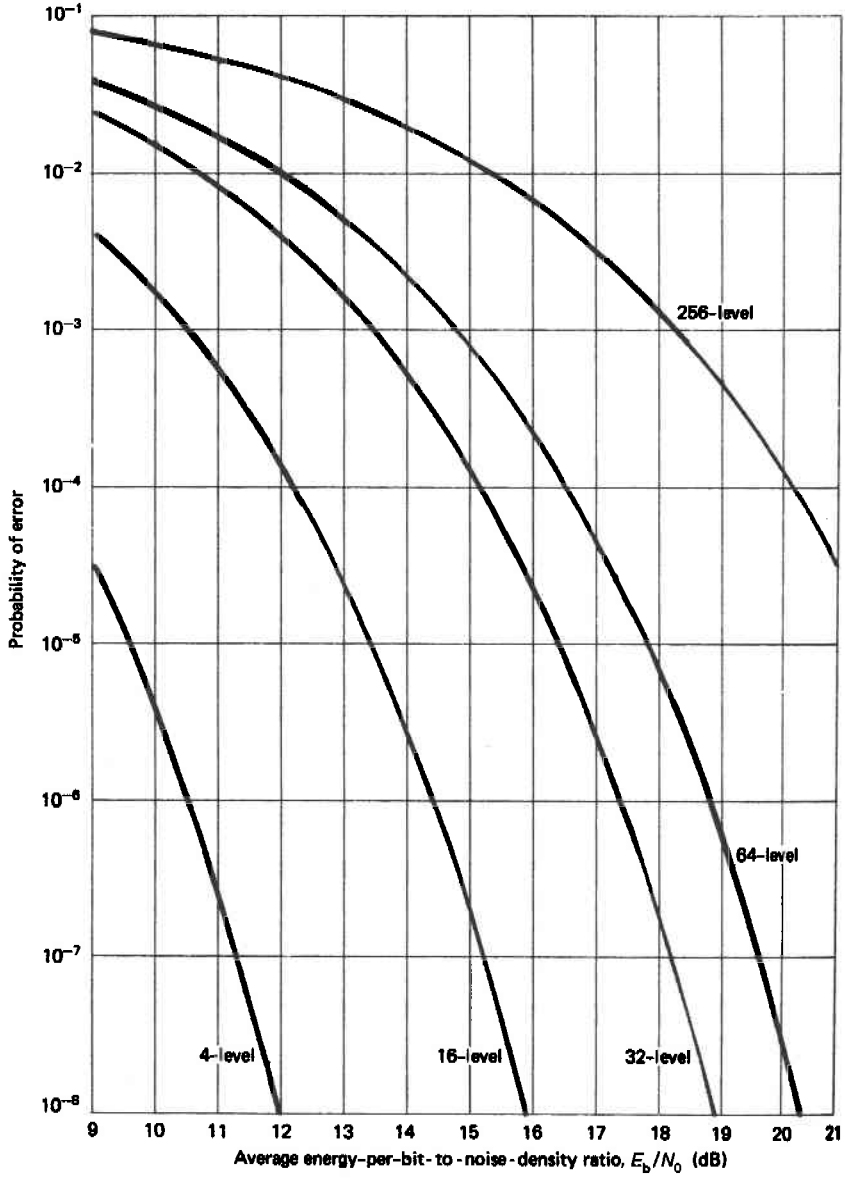


Figure 6.22 Error rates of QAM modulation systems.

TABLE 6.5 Comparison of Various Digital Modulation Techniques Based on Equal Data Rates

System Designation	Information Density (bps/Hz)	Signal–to-Noise Ratios for BER = 10^{-6} (dB): E_b/N_0 on the Channel	SNR at Decision Circuit	Peak-to-Average Ratio (dB)[a]
2-PSK	1	10.6	13.6	0.0
4-PSK, 4-QAM	2	10.6	13.6	0.0
QPR	2[b]	12.6	17.6	2.0
8-PSK	3	14.0	18.8	0.0
16-QAM	4	14.5	20.5	2.55
16-QPR	4[b]	16.5	24.5	4.55
16-PSK	4	18.3	24.3	0.0
32-QAM	5	17.4	24.4	2.3
64-QAM	6	18.8	26.6	3.68

[a]Ratio of maximum steady-state signal power to average signal power with random data. Measured on a channel with square-root filter partitioning.

[b]In a strict sense, the signal bandwidth of a partial-response system is no narrower than the theoretical (Nyquist) bandwidth of a corresponding full-response system. As a practical manner, however, the partial-response systems require about 17% less bandwidth [21].

If half-sinusoidal pulse shaping of duration T is used on both channels of an offset keyed QAM system, the quadrature channel signals $p(t)$ and $q(t)$ can be expressed as follows:

$$p(t) = a_i \cos\left(\frac{\pi t}{2T}\right)\cos \omega_c t \quad \left(-\tfrac{1}{2}T \le t \le \tfrac{1}{2}T\right)$$

$$q(t) = a_q \sin\left(\frac{\pi t}{2T}\right)\sin \omega_c t \quad (0 \le t \le T) \tag{6.22}$$

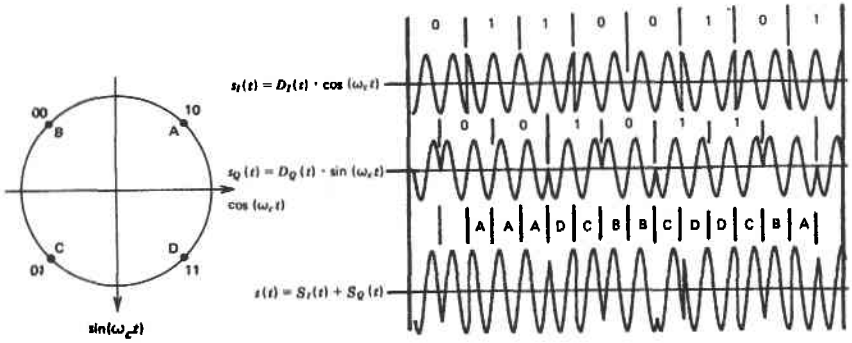


Figure 6.23 Offset keyed 4-PSK signaling.

where a_i and a_q are data values (± 1) for the I and Q channels, respectively, $\cos(\pi t/2T)$ is shaping on the I channel, and $\sin(\pi t/2T)$ is shaping on the Q channel. (The sine function is used to account for offset keying.) Adding the two quadrature signals together and applying some fundamental trigonometric identities produces the result

$$
\begin{aligned}
x(t) &= p(t) + q(t) \\
&= a_i \cos\left(\omega_c t + \frac{\pi t}{2T}\right) \quad (a_i = a_q) \\
&= a_i \cos\left(\omega_c t - \frac{\pi t}{2T}\right) \quad (a_i \neq a_q)
\end{aligned}
\tag{6.23}
$$

Equation 6.23 is essentially identical to Equation 6.6 defining minimum shift keying. Hence, except for a logic-level transformation of data values, offset QAM with sinusoidal pulse shaping is identical to MSK [5, 24]. The relationship is further demonstrated in Figure 6.24, which shows how an MSK signal is generated by offset keyed quadrature channel modulation with sinusoidal pulse shaping.

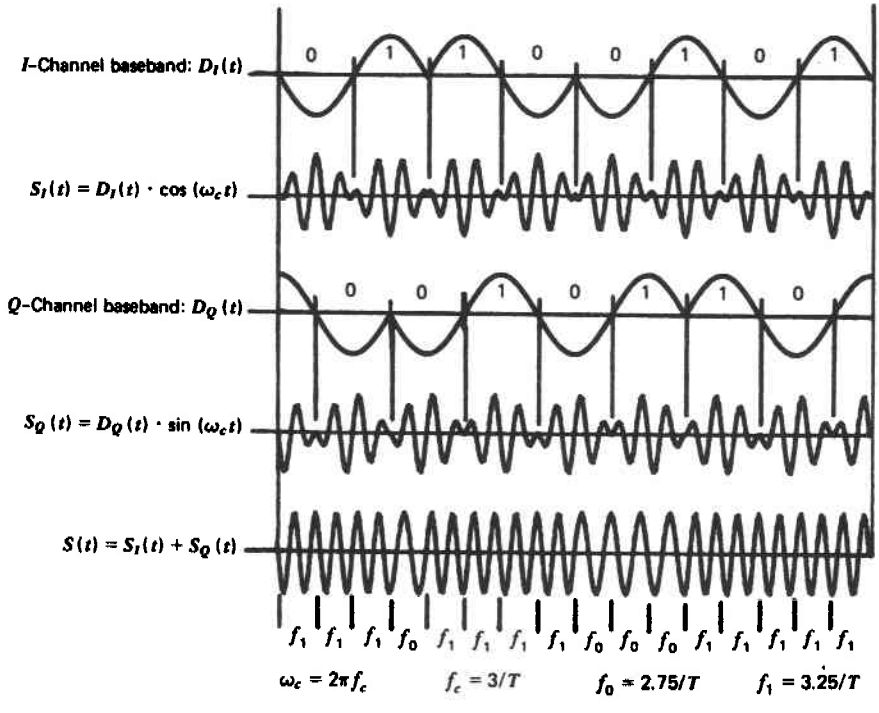


Figure 6.24 MSK signaling produced by offset keyed QAM with sinusoidal pulse shapes.

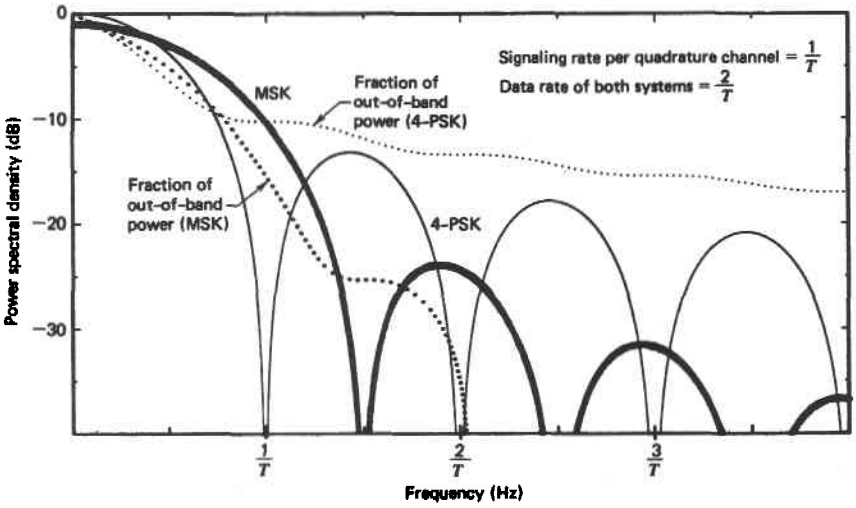


Figure 6.25 Power spectra of MSK and 4-PSK signals.

These foregoing results show that MSK is very closely related to offset-keyed 4-PSK. The only difference is in the use of half-sinusoidal baseband pulse shapes for MSK and square pulse shapes for 4-PSK. Because of this close relationship, it is interesting to compare the amplitude spectra of the two systems in Figure 6.25. As indicated, the MSK spectrum has its first spectral null at a 50% higher frequency than 4-PSK's first null. Other than this, the MSK spectrum is more compact than the 4-PSK spectrum and is significantly lower in amplitude at frequencies outside the main lobe of the spectrum. For this reason MSK is an attractive modulation technique where constant envelopes and little or no filtering are desired. One such application is on digital satellite links with frequency division subchannels [7]. In applications requiring transmitter filters, MSK has no particular advantage over 4-PSK and usually requires a more complicated modulator.

6.1.5 Carrierless Amplitude and Phase Modulation

Carrierless amplitude and phase (CAP) modulation [25] is a specialized form of QAM. As indicated in Figure 6.26, no explicit modulation of a carrier (or carriers) occurs. Instead, two separate modulated signals are directly generated with DSP filter functions: one for an in-phase channel and one for a quadrature channel. The in-phase filter converts I channel data directly into a filtered DSP cosine wave while the quadrature filter converts Q channel data into a filtered DSP sine wave. In this manner the symbol rate is locked to an implied carrier embedded into the filter algorithm. (Typically, the embedded carrier frequency is equal to the baud rate so there is one cycle of a "carrier" in each symbol interval.) After the in-phase and quadrature-phase signals

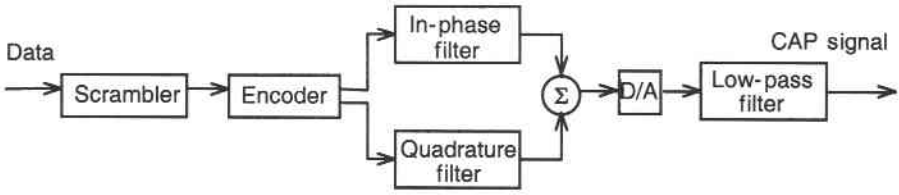


Figure 6.26 CAP modulator block diagram.

are added, the result is converted to an analog signal and filtered to smooth the DSP sampling frequency.

CAP modulation is used in one version of asymmetric digital subscriber line (ADSL) [26] and for a 51.84-Mbps unshielded twisted-pair ATM standard [27]. ADSL applications are described in Chapter 11.

6.1.6 Partial-Response QAM

Another popular modulation technique is quadrature partial-response signaling (QPRS). As commonly implemented [28], a QPRS modulator is nothing more than a QAM modulator followed by a narrow-bandpass filter that "over filters" the quadrature signals and produces controlled intersymbol interference in each channel. The most common application of QPRS involves two levels on each channel before filtering and three levels afterward (see Chapter 4).

This system is essentially a 4-PSK system with partial-response filtering to increase the information density. As shown in Figure 6.27, the effect of partial-response filtering is to produce nine signal points from the original four. In a similar manner, a 16-QAM partial-response system, with four levels on each channel before filtering, has seven levels afterward and 49 signal points in all [29].

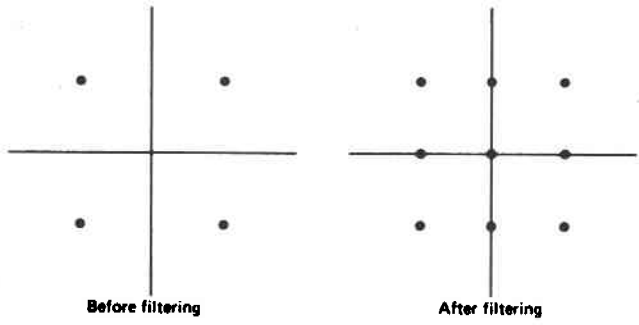


Figure 6.27 QPRS signal constellations.

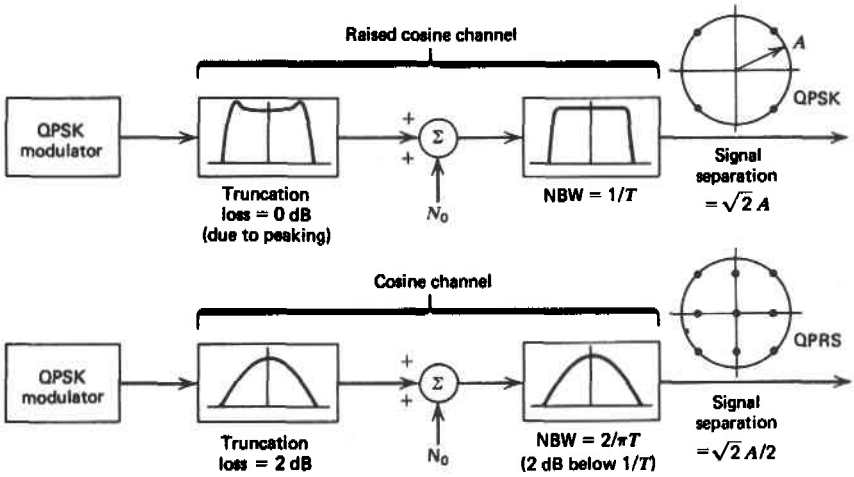


Figure 6.28 Comparison of QPSK and QPRS systems with equal data rates.

Although the modulators of a QPRS system can be conventional QAM modulators, the demodulator/detectors must be modified to account for the extra levels in the waveform. After the signal is demodulated, the detection processes for each channel are independent and identical to the baseband PRS detection procedures described in Chapter 4.

Figure 6.28 compares a QPRS system to a QPSK (4-PSK) system in terms of equal peak power out of the modulator. The average transmit powers are different, however, because the partial-response system overfilters the signal to reduce the transmitted bandwidth. In practice, the information density is increased by about 17% [23].

As discussed in Chapter 4, partial-response filtering cuts the distance between signal points in half, indicating a 6-dB reduction in error performance. However, the noise bandwidth of a PRS receive filter is lower than the noise bandwidth of the corresponding full-response system so that some of the error distance degradation is recovered. For the filter systems shown in Figure 6.28 the net performance loss of a QPRS system is 4 dB. Notice, however, that with respect to power on the channel the performance loss is only 2 dB.

6.1.7 Trellis-Coded Modulation

The 2-dB performance disadvantage of PRS mentioned in the previous section assumes that data are detected by making hard decisions one bit at a time. As described in Chapter 4, however, a PRS signal has redundancy that overlaps two bit intervals. If the detection circuitry processes the signal spanning two intervals before making hard decisions the performance penalty can be recovered. Processing a redundant signal in

this way is an example of a maximum-likelihood or Viterbi decoding algorithm [30] for redundant signals. If the redundancy is extended to more levels and more intervals and optimally processed, even greater coding gains are possible. As typically implemented, these extensions use extra signal points with restricted sequences (transitions between points). When the allowed sequences are represented with state transition diagrams, they form a trellis. Hence the term trellis-coded modulation (TCM) [31, 32].

As mentioned in Chapter 4, TCM is very similar to convolutional coding. The major difference is that convolutional coding adds redundant symbols (increases the bandwidth) but TCM adds redundant signal levels (increases the signal space). Both systems are detected in similar fashions using trellis diagrams. The decoding algorithms of both systems essentially determine the most likely *sequence of states* of the transmitted codes. The *transitions* between states then determine the data.

As a TCM example, consider the expansion of a 4-PSK signal set to an 8-PSK signal set as shown in Figure 6.29. Although there are eight signal points in the 8-PSK constellation, only four points can be freely chosen in any one interval. Which four points can be chosen is dependent of the signal points chosen in previous intervals. A TCM demodulator/decoder then determines which of the allowed signal sequences most closely matches a received sequence before making hard data decisions.

Figure 6.29 shows that if data decisions are made one interval at a time, without processing the redundancy, the error performance of the expanded signal set is 5.33 dB worse than the original signal set (a consequence of the noise power margin of adjacent signal points being reduced from 0.5 to 0.146). Also shown in Figure 6.29 are the noise margins of more widely separated signal points. TCM recovers the 5.33-dB penalty, and more, by ensuring all allowed sequences include some of the larger distances. Furthermore, noise in only one interval is not likely to produce a decision error. When four-state coded 8-PSK uses the transition trellis shown in Figure 6.30, the performance approaches a 3-dB improvement with respect to uncoded 4-PSK [31].

To understand Figure 6.30, it is important to realize that it represents state transitions in the encoding process, not signal waveforms. The use of the states in this man-

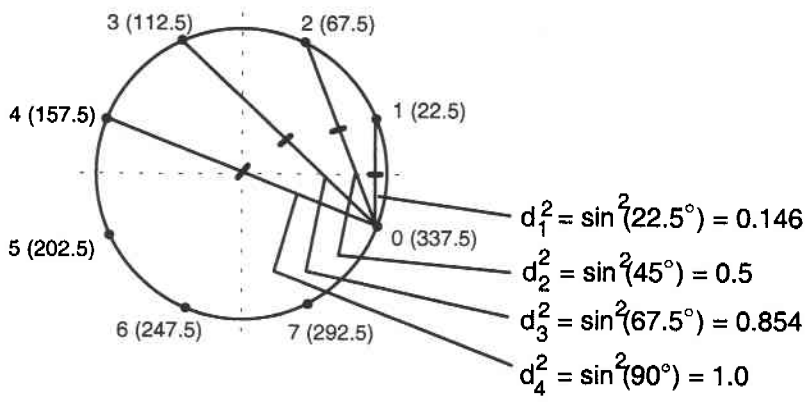


Figure 6.29 Eight-PSK signal points and error distances.

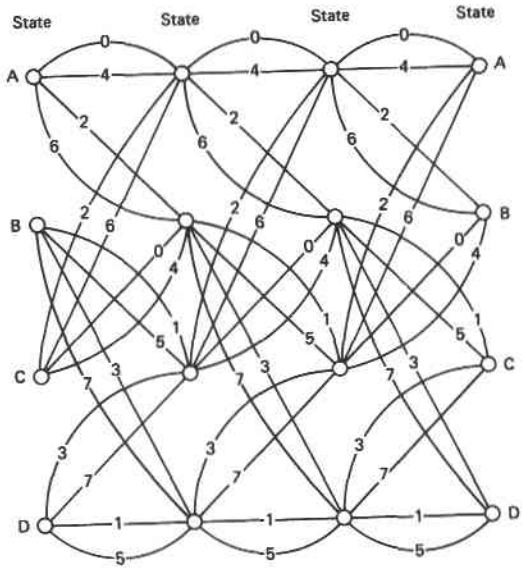


Figure 6.30 Four-state trellis for 8-PSK modulation.

ner is just a convenient means of representing past history. (All past histories are encapsulated into just four states.) Transitions from one state to another are labeled with the 8-PSK signal point chosen for encoding 2 bits of data. Notice that only four of the 8-PSK signal points can be freely chosen and that these points have the same separation as uncoded 4-PSK. The significant aspects of the decoding process are summarized as follows:

1. An allowed transition between two states can occur with either of two signals, which implies sequence information does not help discriminate between those two particular signals. However, the two signals in question are chosen to have maximal separation (noise margin 1) so redundancy is unnecessary.
2. Transitions that originate in different states and terminate in any particular state are encoded with signals having noise power margins of 0.5.
3. All sequences that begin and end in common states are at least three intervals long with minimum noise power margins of 0.5, 0.146, and 0.5. Thus the total noise margin between any two minimum-length sequences is 1.146.

It may be necessary to discriminate between sequences that are longer than three but in all cases the endpoints involve transitions with noise power margins of 0.5. Thus the noise margin is always greater than 1, which is 3 dB greater than the noise margin of uncoded 4-PSK. Determination of the exact error rate of a TCM system is much more involved than when interval-by-interval detection is used. Reference [31] shows

that the bit error rate of coded 8-PSK asymptotically approaches 3-dB improvement with respect to 4-PSK as the noise margin would indicate.

Example 6.3. A receiver for a trellis-coded modulation system as shown in Figure 6.30 detects a phase sequence of 20°, 220°, and 10°. Determine the allowable sequence of three signals with the closest distance to the given received sequence. Assume the starting state is state B.

Solution. To begin with we can determine the closest encoder signal phases as 22.5°, 202.5°, 22.5°, which correspond to signal points 1, 5, 1. Examination of Figure 6.30 indicates that this is a disallowed sequence because a 5 signal cannot follow a 1 signal. If data were decided on a symbol-by-symbol basis, an error would certainly be made. By tracing the trellis from state A, a list of allowable state sequences can be determined as provided in Table 6.6. For each transition, the most likely signal of each pair of signals that can produce a particular transition is indicated. The symbol errors (in degrees) for each of these signals is then determined followed by the total sequence error. As indicated, the most likely state sequence is CAA (signal sequence: 1,6,0).

Although the result of Example 6.3 indicates one particular sequence is more likely than any other, there are two other allowable sequences (141 and 162) that are fairly close to the most likely signal sequence. Notice that these two sequences have end states (C and B) that are different than end state A of the selected sequence. Thus, there is more information to come as to which is the most likely sequence. Because states A and B have no common allowed signals, the next symbol will provide additional discrimination between 160 and 162. The very next signal will not help discriminate between 160 and 141 but the signal following that will. A thorough determination of the most likely transmitted signal sequence needs to consider other possible starting phases and, consequently, previous signal values. (This is an exercise left to the student.)

Coded 8-PSK TCM has been used in satellite circuits [33] where nonlinearities dictate the use of a constant-envelope signal. TCM with higher density QAM signal sets

TABLE 6.6 Sequence and Error Determination for Example 6.3

States	Signals	Symbol Errors	Total Error
CBC	141	2.5, 62.5, 12.5	77.5
DDD	351	92.5, 17.5, 12.5	122.5
CAA	160	2.5, 27.5, 32.5	62.5
DCB	370	92.5, 72.5, 32.5	197.5
CBD	143	2.5, 62.5, 102.5	167.5
DDC	353	92.5, 17.5, 102.5	212.5
CAB	162	2.5, 27.5, 57.5	87.5
DCA	372	92.5, 72.5, 57.5	222.5

is used in high-speed voiceband QAM modems such as the 14.4-kbps V.33 [34] and 28.8-kbps V.34 [35]. TCM can also be used in conjunction with MSK modulation [36]. TCM has been utilized in point-to-point microwave radios such as the 155-Mbps 6G150 MB radio of NEC. The modulation format is 128-QAM. NEC refers to TCM as multilevel coded modulation (MLCM).

6.1.8 Multicarrier Modulation

In all of the modulation formats described to this point it has been implicitly assumed that modulation occurs on a single carrier, or possibly on two quadrature carriers operating at the same frequency. Multicarrier modulation (MCM) involves dividing the data stream into multiple, lower rate streams that are transmitted in parallel on multiple carrier frequencies. Although MCM has been utilized in some special applications in the past [37], it is the emergence of DSP technology that allows economic implementation of multiple modulators and demodulators for commercial applications. The most common means of implementing MCM utilizes fast Fourier transforms (FFTs), as shown in Figure 6.31. Source data are demultiplexed into separate data streams for each subchannel. The data for each subchannel are then independently encoded into discrete-level complex numbers representing the amplitudes of quadrature channel carriers. An inverse FFT transforms the complex spectrum components into a time-domain waveform that is converted to analog and transmitted.

The basic steps of the process are reversed in the receiver where the incoming time-domain waveform is sampled and fed to an FFT that recovers a complex spectrum. The spectrum components are then processed as individual QAM (or possibly PSK) signals to decode the data, which are then multiplexed back into a composite data stream. Not included in Figure 6.31 are ancillary functions of equalization and clock recovery. Normally, one frequency component is dedicated as a pilot signal that carries clock information for all subchannels. Amplitude equalization involves merely scaling the FFT component amplitudes at the output of the FFT. Amplitude equalization may not even be required if PSK modulation is utilized because, even though there may be amplitude distortion in the channel, within each narrow subchannel the

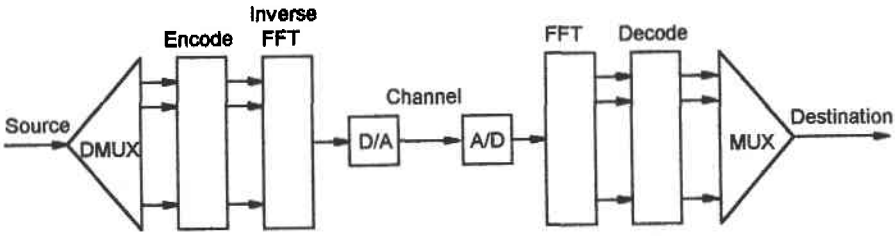


Figure 6.31 Multicarrier modulation utilizing fast Fourier transforms.

distortion is essentially flat, which implies that PSK data can be recovered by merely determining the phase of each complex frequency term.

Multicarrier modulation with an FFT implementation is commonly referred to as discrete multi-tone (DMT) in North America and as orthogonal frequency division multiplexing (OFDM) in Europe. (The term "orthogonal" occurs because the frequency components of an inverse FFT are harmonically related and therefore have zero-valued cross-correlation products.) The terms DMT and OFDM are interchangeable with the exception that in some OFDM applications it is understood that all subchannels utilize the same form of modulation with the same number of bits per channel. In DMT systems the modulation of the subchannels is more general so that different data rates can be carried on different subchannels depending on the transmission quality of the respective subchannels. A principal application of DMT is ADSL [38] standardized by ANSI committee T1E1.4. OFDM is specified by European Telecommunications Standards Institute (ETSI) as the modulation format for Digital Video Broadcasting (DVB) [39].

One of the most attractive features of DMT is the inherent ability to match an information signal spectrum to a channel response. An example of such a system is shown in Figure 6.32 that is representative of an ADSL application on subscriber wire pairs. A significant impairment of using existing wire pairs for high-bandwidth data is the possible presence of bridged taps. ISDN basic rate installations require removal of bridged taps. In an ADSL implementation bridged taps are accommodated by detecting their presence during channel characterization and then transmitting only as much information in the affected subchannel(s) as can be reliably supported. Notice that a conventional wideband data signal would experience significant distortion if transmitted on the channel of Figure 6.32. Thus, a significant amount of amplitude (and probably phase) equalization would be required.

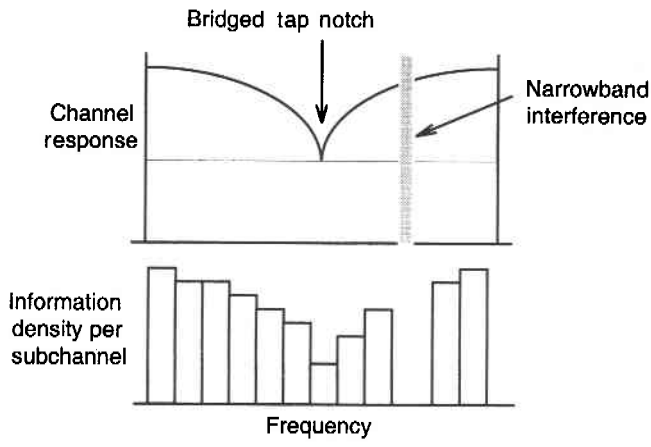


Figure 6.32 Use of DMT modulation to match subchannel data capacity to transmission channel.

Figure 6.32 shows that a DMT system assigns information densities to individual subchannels based on respective signal-to-noise ratios determined during channel characterization. Included in the process is detection of narrowband interference that causes a subchannel to be eliminated altogether. Notice that a conventional wideband system might be totally disabled by the narrowband interference.* For more details on DMT implementations see the ADSL discussion in Chapter 11.

6.2 FILTER PARTITIONING

The transmitting and receiving equipment of a digital radio system typically contains several filters that limit the signal spectrum to some degree or another. Since the end-to-end frequency response of the channel must conform to certain baseband pulse response objectives, the desired composite filter function must be partitioned among the individual filters. Normally, a number of the filters can be designed to provide their respective functions without significantly impacting the channel pulse response. For example, a mixing process produces a sum and a difference of the input frequencies. Only the sum is wanted when mixing upward, or only the difference is wanted when mixing downward. Usually the undesired terms can be eliminated by a filter that does not significantly affect the pulse shape of the underlying signal. The following discussion assumes that only two filters significantly influence the baseband pulse response: one in the transmitter and one in the receiver.

6.2.1 Adjacent-Channel Interference

One basic purpose of the radio channel receive filter is to minimize the amount of noise present at the detector. A second purpose of this filter is to reject energy in adjacent radio channels. Energy from an adjacent channel that does not get rejected is referred to as *adjacent-channel interference.* In this discussion, we assume that the spectrum shapes in the adjacent channels are identical to the desired spectra. This situation is shown in Figure 6.33, which depicts a number of frequency-division-multiplexed digital channels.

As indicated in Figure 6.33, adjacent-channel interference occurs as a result of two phenomena. First, the receive filter passes unwanted power P_1 because the adjacent signal is not completely truncated to prevent overlap into the desired channel. The second source of interference, P_2, occurs because the receive filter does not provide infinite attenuation of power properly belonging in the adjacent channel. Unwanted power P_1 is minimized by narrowing the transmit filter, while P_2 is minimized by nar-

*Totally avoiding the effects of high-energy narrowband interference in a DMT system is not as simple as it might seem. If the interference is present at the input to the A/D converter in the receiver, the interference may cause A/D saturation (or increased quantization noise if a companded converter is utilized). Thus, total avoidance of narrowband interference requires front-end notching of the signal. A front-end notch could also be used in a conventional wideband system followed by decision feedback equalization to accommodate the inserted distortion [40].

rowing the receive filter. Since channel pulse response considerations constrain the composite filter function to some minimum width, one component of the interference cannot be reduced without increasing the other. Hence the total filter function must be partitioned in some manner to minimize the sum of P_1 and P_2.

6.2.2 Optimum Partitioning

The optimum filter design for any particular application may depend on a number of factors including legislated emission specifications, available technology for power amplifiers, the availability of cross polarization for adjacent-channel isolation, and the relative effects of noise versus adjacent-channel interference. In the absence of external constraints, a classical result attributed to Sunde [41] and also presented in references [42] and [43] determines optimum partitioning as one matching the output of the transmit filter to the square root of the desired channel response. The desired output spectrum $Y(\omega)$ is obtained as

$$Y(\omega) = H_{RX}(\omega)H_{TX}(\omega)X(\omega) \tag{6.24}$$

where $X(\omega)$ = channel input spectrum
$H_{TX}(\omega)$ = transmit filter response
$H_{RX}(\omega)$ = receive filter response

Then the optimum partitioning is obtained as

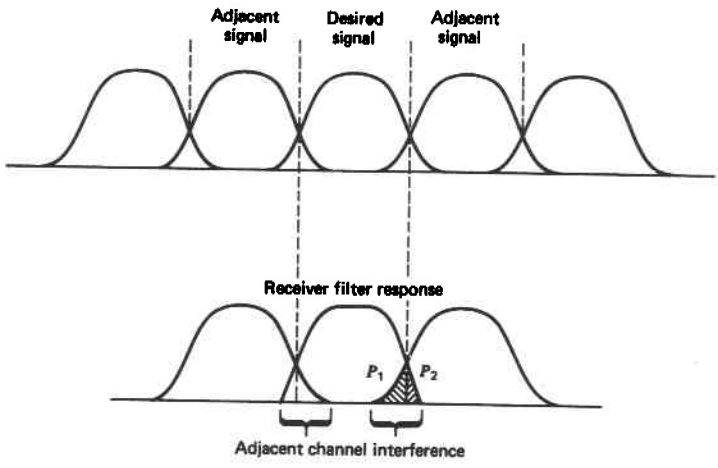


Figure 6.33 Adjacent-channel interference.

$$|H_{Rx}(\omega)| = |Y(\omega)|^{1/2}$$

$$|H_{Tx}(\omega)| = \frac{Y(\omega)^{1/2}}{X(\omega)} \tag{6.25}$$

Equation 6.25 defines the filter amplitude responses to minimize the adjacent-channel interference under the condition that the adjacent channels contain identical signals with identical power levels. In addition, if the transmitted spectrum $H_{TX}(\omega)X(\omega)$ is the complex conjugate of the receive filter response $H_{RX}(\omega)$, the receive filter is matched to the channel spectrum, and the best possible error performance is obtained with respect to signal power on the channel.

Partitioning as defined by Equation 6.25 is shown in Figure 6.34 for a pulse input and a raised cosine output (see Appendix C). Even though only baseband signal spectra are shown, the concept is extended easily to passband filter functions.

Although Equation 6.25 provides a sound theoretical basis for determining the optimum filter partitioning, practical considerations of a particular application may require deviation from the "optimum." One problem that arises in digital microwave radios is related to the peak at the band edges of the transmit filter in Figure 6.34. With a passive filter this peak can be obtained only by inserting loss into the midband response. If the transmit power can be increased as compensation, no ill effects result. However, the output level of many digital microwave radios is limited by the technology of microwave frequency power amplifiers (a few watts typically). Since these radios are device power limited, midband insertion loss cannot be overcome and therefore subtracts directly from the received signal level.

In device-power-limited applications the optimum transmit filter is one having a flat response in the passband. To achieve the desired channel response at the detector, the receive filter must then be peaked at the band edges. Channel insertion losses at this point do not degrade performance since both the signal and the noise are attenuated equally. However, the peaks have the undesirable effect of increasing the receiver noise bandwidth and the P_1 component of adjacent-channel interference. Although these increases represent a poorer performance than that obtained by theoretically optimum partitioning, the degradation is not as great as the insertion loss needed to achieve the "theoretical optimum." As one further note, it should be mentioned that

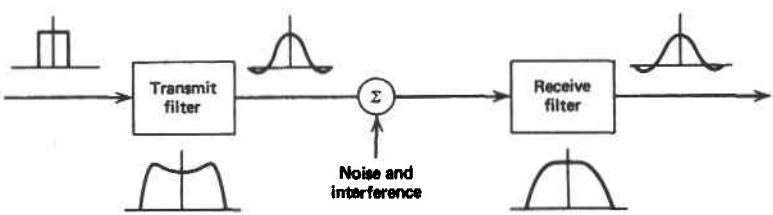


Figure 6.34 Theoretical optimum filtering for square-pulse excitation and raised-cosine output.

not all theoretically optimum transmit filters have a peak at the band edges. In particular, partial-response systems do not require band edge peaking (see Appendix C).

Another aspect of the optimum design to be kept in mind is that optimum error performance is achieved with respect to signal power on the channel, and not with respect to unfiltered transmitter power. It is this feature that allows any amount of attenuation to be inserted into the channel at the transmitter and not degrade the optimum performance. If performance is measured with respect to unfiltered transmit power, the best transmit filter may be different from that defined in Equation 6.25. Not only would less midband attenuation be desirable, as discussed, but it might also be desirable to widen the transmit filter response. This decreases the truncation loss in the transmit filter so the receive filter can be narrowed to decrease the receiver noise bandwidth.

As an extreme example of how widening the transmit filter can improve error rate performance, consider removing the transmit filter and incorporating it into the receiver. For a given output at the power amplifier, the signal at the detector is unchanged. The receiver noise bandwidth, however, is reduced because the composite filter is narrower than the original receive filter. Hence a higher signal-to-noise ratio is present at the detector.

The penalty incurred for removing the transmit filter, of course, is a greatly increased P_1 component of adjacent-channel interference. If adjacent channels do not exist or if they are adequately isolated by cross-polarization (i.e., if the system is noise limited), the performance can be improved by moving some of the transmitter truncation losses to the receiver until adjacent-channel interference matches the noise. Keep in mind, however, that if the system is adjacent-channel interference limited, the optimum partitioning is indeed defined by Equation 6.25.

6.3 EMISSION SPECIFICATIONS

One unavoidable consideration when determining the filter functions of a transmitter and receiver is the out-of-band emission specifications established by the FCC in the United States or the ITU-R in other parts of the world. In many cases the emission specifications dictate a narrower transmit filter than the theoretical optimum. Thus the partitioning aspect of the filter designs may be predetermined.

It is somewhat ironic that the FCC emission specifications were intended to control adjacent-channel interference but, in some cases, actually cause the interference to increase by forcing the use of a wider than optimum receive filter. The FCC specifications, however, were intended to protect adjacent analog radio channels from out-of-band digital emissions. They were not selected with adjacent digital channels in mind.

In January 1975, when the FCC established the out-of-band emission specifications, two separate specifications were established: for radios operating below 15 GHz and for radios operating above 15 GHz. The emission limitations for operation below 15 GHz are more restrictive than those for operation above 15 GHz because the lower frequencies were heavily used for analog FDM–FM radios, which are more sensitive to interference. The higher frequencies have not been used extensively because of vul-

nerability to rain attenuation. These frequencies are used for short-distance, special-purpose applications and are not as congested as the lower frequencies. Initially, the most popular bands for digital microwave were 11 and 6 GHz.

The emission limitations for operation below 15 GHz are

$$A = 35 + 10 \log_{10} B + 0.8(P - 50) \qquad (P > 50) \tag{6.26}$$

where A = power in 4-kHz band relative to mean output power (dB)
B = authorized bandwidth (MHz)
P = percent removed from carrier frequency (50 is the band edge)

In addition, the attenuation A must be at least 50 dB everywhere outside the band but does not need to exceed 80 dB at any point. Notice that the emission limitations are not specified in terms of absolute power levels, but only relative to the transmitted power. Thus these emission limitations do not constrain the output power level of the radio. Microwave output powers are often constrained by microwave amplifier technology. (A limit of 10 W of output power does exist when technology can reach it.)

Figure 6.35 displays the emission mask for 11-GHz radios using a 40 MHz of bandwidth. Also shown is the power spectrum of a 90-Mbps 8-PSK radio using a signaling rate of 30 MHz. To be compatible with the emission specification, the 8-PSK power spectrum is shown in terms of power per 4-kHz band. The difference between the signal spectrum and the FCC mask represents the minimum attenuation required of the transmit filter. The filtered 8-PSK signal thereby provides an information density of 90/40 = 2.25 bps/Hz. This is the basic modulation and filtering format used by Collins/Rockwell in their 90-Mbps digital radio, the MDR-11 [13].

The emission limitation for microwave bands above 15 GHz is defined as

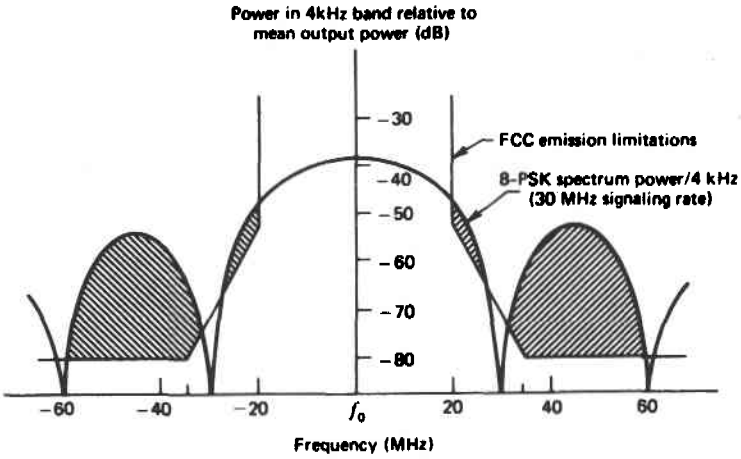


Figure 6.35 FCC emission mask at 11 GHz and a 90-Mbps 8-PSK spectrum.

$$A = 11 + 10 \log_{10} B + 0.4(P - 50) \qquad (P > 50) \tag{6.27}$$

where A = attenuation in 1-MHz band below mean output power (dB)
B = authorized bandwidth (MHz)
P = percent removed from carrier frequency

The attenuation A must be at least 11 dB but does not have to exceed 56 dB.

In conjunction with the emission limitations, the FCC has stipulated that no discrete lines exist in the transmitted spectrum [44]. Thus no carrier components can exist and no repetitive data patterns are allowed to occur. The repetitive data patterns are effectively eliminated by using a scrambler in the transmitter and a descrambler in the receiver. As mentioned already, PSK and QAM are forms of double-sideband *suppressed* carrier modulation so that the carrier terms are eliminated automatically as long as modulation is continuous.

6.4 RADIO SYSTEM DESIGN

The foremost design requirement of a point-to-point radio system for telephony is operational dependability, usually referred to as availability. Availability is expressed as the percentage of time that a system provides a specified minimum quality of service. With analog systems, minimum performance is determined by the noise power in the received signal. The performance of a digital system is determined by the bit error rate. Typical objectives for bit error rates range from 10^{-3} for voice traffic to 10^{-6} or 10^{-7} for data traffic. Recall that a bit error rate of 10^{-4} corresponds to the threshold of perceptibility for errors in a PCM voice signal. Typical design objectives for microwave radio systems specify availability on the order of 99.98% [45, 46]. Hence the maximum acceptable accumulation of outage, due to all causes, is on the order of 2 hr per year.

Radio system availability is dependent on equipment reliability and path propagation characteristics. The high-availability objectives of a typical radio system mandate redundant equipment and often require redundant paths. The need for redundant paths is determined by the likelihood of atmospheric-induced outages (rain attenuation or multipath interference). Rain is a dominant consideration at higher carrier frequencies (above 11 GHz), and multipath interference must be considered at all frequencies. Multipath fading is dependent on prevailing climate and terrain.

Redundant radio equipment typically operates in either a backup or a hot-standby mode with automatic protection switching. The transmission path is backed up with spare channels (frequency diversity) or spare paths (space diversity receivers). In extreme cases a backup route may even be utilized.

6.4.1 Fade Margins

The main technique used to circumvent atmospheric-induced outages is to provide an extra strong signal at the receiver during normal path conditions. The difference be-

tween the normal received power and the power required for minimum acceptable performance is referred to as the fade margin. Greater fade margins imply less frequent occurrences of minimum performance levels. Radios operating in higher frequency bands generally require greater fade margins because they are more susceptible to rain attenuation. A 50-dB fade margin is typical for a digital radio at 11 GHz, while a 40-dB fade margin is typical for lower microwave frequencies.

The amount of fade margin actually required for a particular route depends on the probability of multipath-induced fades and heavy rainfall occurrences. Thus drier climates permit lower fade margins, thereby allowing greater repeater spacing. In some mountain-based microwave links in the western United States, microwave hops can be 100 miles long. By comparison, the average hop in other parts of the country is less than 30 miles long.

When large fade margins are provided, the received signal power during unfaded conditions is so strong that bit errors are virtually nonexistent. Nevertheless, problems with extra strong signals do exist. Namely, automatic gain control in a receiver must operate over a wide dynamic range. If the maximum signal level into the demodulation and detection circuitry is not controlled, saturation is likely to degrade performance, especially in high-density modulation formats such as 64- or 256-QAM, where information is encoded into the signal amplitudes.

To minimize dynamic range requirements in a receiver and reduce interference between systems, adaptive transmitter power control (ATPC) is sometimes used [47]. ATPC uses a feedback data link from a receiving station to control the output power of a transmitting station. Thus when excess power is unnecessary, it is not used. ATPC is commonly used in digital mobile telephone systems where interference control is a primary concern.

6.4.2 System Gain

One of the most important parameters used to characterize digital microwave system performance is the system gain A_s. System gain is defined to be the difference, in decibels, of the transmitter output power and the minimum receive power for the specified error rate:

$$A_s = 10 \log_{10}\left(\frac{P_T}{P_R}\right) \tag{6.28}$$

where P_T = transmitter output power
P_R = receive power for specified error rate

The minimum acceptable receive power is sometimes referred to as the threshold power and is primarily dependent on the receiver noise level, the signal-to-noise ratio required by the modulation format, and various system degradations such as excess noise bandwidth, signal distortions, intersymbol interference, carrier recovery offsets,

timing jitter, and coupling and filter losses that either attenuate the signal or increase the noise level.

Noise power in a receiver is usually dominated by thermal noise generated in the front-end receiver amplifier. In this case, the noise power can be determined as follows:

$$P_N = (F)(N_0)(B)$$

$$= F(kT_0)B \tag{6.29}$$

where $F =$ the receiver noise figure
$N_0 =$ the power spectral density of the noise
$B =$ the receiver bandwidth*
$k =$ $1.38(10)^{-23}$ is Boltzmann's constant
$T_0 =$ the effective receiver temperatures in degrees Kelvin

Equation 6.29 essentially states that the receiver noise power is determined by the spectral noise density of the receiver input resistance and the additional noise introduced by the amplification process (noise figure F). Normally, a reference temperature of 290 K is assumed so that the thermal noise density (kT_0) is 4×10^{-21} W/Hz.

The noise figure of any device is defined as the ratio of the input signal-to-noise ratio to the output signal-to-noise ratio:

$$F = \frac{(S/N)_{in}}{(S/N)_{out}} \tag{6.30}$$

In effect, the noise figure specifies the increase in noise power relative to the increase in signal power. Since all physical devices introduce noise, the noise figure of any system is always greater than 1 and is usually expressed in decibels. If a system has no gain or attenuation, the noise figure is exactly equal to the ratio of output noise to input noise. Noise figures of low-noise microwave amplifiers typically range from 2 to 5 (3–7 dB). Radio receiver noise figures are typically 6–10 without a low-noise amplifier.

Combining Equations 6.28 and 6.29 and incorporating a term D for the degradation from ideal performance produces the following general expression for system gain:

$$A_s = 10 \log_{10}\left(\frac{P_T}{(\text{SNR})(F)(kT_0)(B)}\right) - D \tag{6.31}$$

where SNR is the theoretical signal-power-to-noise-power ratio required for the maximum acceptable error rate and D includes all degradations from ideal performance.

*Typically, B is assumed to be the minimum theoretical bandwidth for the particular modulation format in use. Excess bandwidth required by practical implementations is then incorporated into a system degradation factor.

Notice that the SNR term in Equation 6.31 refers to signal-power-to-noise-power ratios and not E_b/N_0. The relationship between SNR and E_b/N_0 is provided in Appendix C.

The system gain, in conjunction with antenna gains and path losses, determines the fade margin:

$$\text{Fade margin} = A_s + G_T + G_R + 20 \log_{10} \lambda - A_f - A_0 \tag{6.32}$$

where G_T = transmitter antenna gain (dB)
G_R = receive antenna gain (dB)
λ = transmitted wavelength
A_f = antenna feeder and branching loss (dB)
A_0 = free-space attenuation (distance d must be in the same units as λ)
$= 20 \log_{10} (4\pi d)$

The directivity, and consequently the gain of, an antenna is directly proportional to the size of its aperture and inversely proportional to the square of the transmitted wavelength. In determining the receive power, however, it is actually only the area of the antenna that is important and not the directivity or gain. Thus radio system designers conveniently consider transmit and receive antenna gains as contributing to signal power but include a wavelength normalization ($20 \log_{10} \lambda$) to relate the gain of the receiving antenna back to the size of its aperture.

In addition to providing increased antenna gain, greater directivities also reduce multipath problems. The longer secondary paths, which arise from greater emanation angles, have lower power levels when the directivity is increased. Unfortunately, antenna gains are limited by practical considerations in several regards: Economically sized towers can support only limited-sized antennas, mechanical alignment is difficult, and directional stability of both the antenna and the path is limited.

The feeder and branching losses A_f included in Equation 6.32 arise because single antenna systems typically carry several channels for separate radios. Furthermore, reliability considerations usually dictate that spare transmitters and receivers be available for protection switching. The process of combining signals for transmission or distributing them after reception inherently introduces various amounts of attenuation or splitting of signal power.

Example 6.4. Determine the system gain of a 10-Mbps, 2-GHz digital microwave repeater using 4-PSK modulation and an output power of 2.5 W. Assume the excess bandwidth of the receiver is 30% and that other departures from ideal performance amount to 3 dB degradation. Assume a noise figure of 7 dB for the receiver, and the desired maximum error rate is 10^{-6}. Also determine the fade margin assuming antenna gains of 30 dB each and a path length of 50 km. The branching and coupling losses are 5 dB.

Solution. From Figure 6.17, the required value of E_b/N_0 for 4-PSK modulation can be determined as 10.7 dB. Using Equation 3.42 (in Appendix C), it can be determined

that the signal-power-to-noise-power ratio at the detector is 3 dB higher than E_b/N_0. Thus, the required SNR is 13.7 dB.

Since 4-PSK modulation provides 2 bps/Hz, the signaling rate is 5 MHz, which is the theoretical minimum (Nyquist) bandwidth. Equation 6.31 can now be used to determine the system gain:

$$A_s = 10 \log_{10}\left(\frac{2.5}{4(10)^{-21}(5)10^6}\right) - 13.7 - 7 - 3 - 10 \log 1.3$$

$$= 116 \text{ dB}$$

At a carrier frequency of 2 GHz, the wavelength is $3 \times 10^8/2 \times 10^9 = 0.15$ m. Thus the fade margin can be determined from Equation 6.31:

$$\text{Fade margin} = 116 + 60 + 20 \log_{10}(0.15) - 5 - 20 \log_{10}(4\pi 5 \times 10^4)$$

$$= 38.5 \text{ dB}$$

6.4.3 Frequency Diversity

As mentioned previously, neither the transmitting and receiving equipment nor the path is normally reliable enough to provide an acceptable level of system availability. Frequency diversity is one means of providing backup facilities to overcome both types of outages. A deep multipath-induced fade occurs when a signal from a secondary path arrives out of phase with respect to the primary signal. Since the phase shift produced by a path is proportional to frequency, when one carrier fades, it is unlikely that another carrier fades simultaneously. Frequency diversity involves the use of a spare transmitter and receiver operating in a normally unused channel. Since separate hardware is used, frequency diversity also provides protection against hardware failures.

The simplest means of implementing frequency diversity is to use one-for-one (1 : 1) protection switching as indicated in Figure 6.36. One-for-one protection switching implies that one spare channel is provided for each assigned message channel. When high-spectrum efficiency is required, it is generally necessary to have only one spare for a group of N channels (1 : N protection switching). In fact, the FCC has stipulated that, in some frequency bands, a system must be implementable in one-for-N configurations. The main impact of a 1 : N protection system is the complexity of the switching unit and the need to switch back to the assigned channel in the first available hop so that a single spare channel can be reused repeatedly on a long route.

Frequency diversity normally does nothing to alleviate rain outages since all channels in a particular frequency band are simultaneously affected. When rain is a particular problem, it can be overcome only by using higher transmit powers or shorter hops.

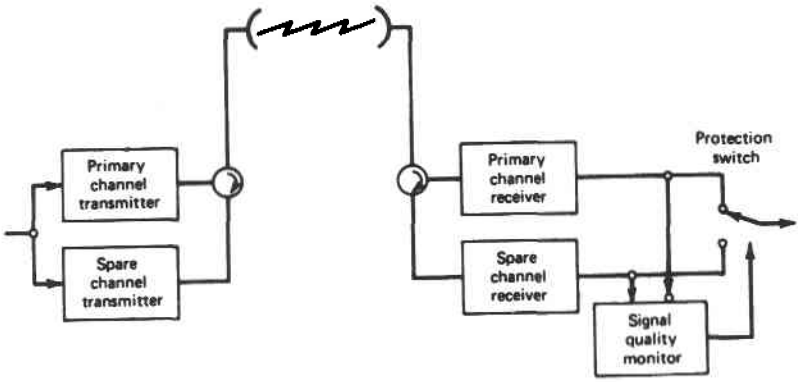


Figure 6.36 Protection switching with one-for-one frequency diversity.

6.4.4 Space Diversity

Space diversity is implemented, as shown in Figure 6.37, by vertically separating two receive antennas on a single tower. The resulting difference in the two paths is normally sufficient to provide independent fading at the two antennas. Space diversity is the most expensive means of improving the availability, particularly if separate receivers for multiple channels are used for each antenna. The cost can be minimized, however, by combining the two received signals in a phase-coherent manner for input to a common receiver [45]. This technique provides less hardware backup than when completely separate receivers are used for each antenna.

6.4.5 Angle Diversity

Because multipath fading is produced by multiple incident rays arriving at slightly different angles, protection from fading can be achieved by discriminating on the angle

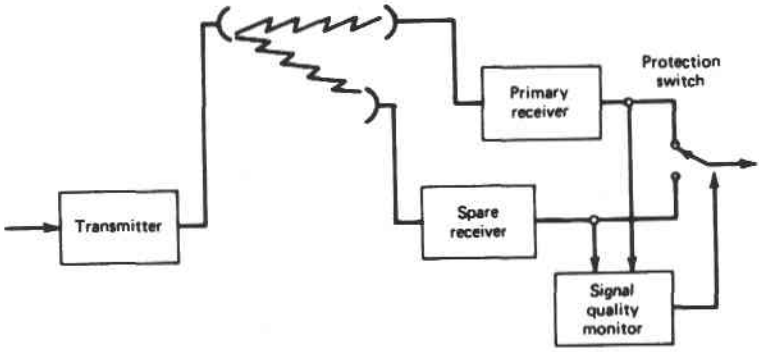


Figure 6.37 Protection switching with space diversity receivers.

of arrival. Angle diversity utilizes two side-by-side receiving antennas with slightly different angular elevations to provide the discrimination. Although angle diversity is generally not as effective as space diversity, it can improve the availability of digital signals [48, 49] in applications where the tower requirements of space diversity are impractical or disallowed.

6.4.6 Adaptive Equalization

Since multipath effects are frequency dependent, not all frequencies in a particular channel simultaneously experience the same amount of fading. Thus multipath fading can produce not only a general attenuation of the received signal but also the equivalent of in-band amplitude (and phase) distortion. Distortion of the spectrum amplitude produces a general degradation in the error performance over and above the penalty incurred by the general attenuation of the signal. In wideband digital radios, frequency-selective fading (as opposed to flat fading) has proved to be the dominant source of multipath outages. Fortunately, better performance can be achieved if the spectrum amplitude is equalized (adjusted to a uniform level).

Since atmospheric-induced multipath interferena varies with time, removal of multipath-induced amplitude distortion requires adaptive equalization. A common approach to adaptive amplitude equalization merely samples the energy at selected frequencies in the received signal spectrum. A compensating filter shape is then inserted into the signal path to adjust all energy samples to a common level. This basic technique is used in digital radios manufactured by Bell Northern [45] and Rockwell International [50]. Reference [50] reports a 7-dB improvement in effective fade margin can be achieved with an adaptive equalizer.

Because channel distortions produce intersymbol interference in the received baseband signal, time-domain equalization with adaptive transversal filters or decision feedback equalizers is also used. Reference [51] contrasts the relative merits of both approaches and provides an extensive list of references. The use of high-density modulation formats such as 256-QAM places very stringent requirements on the adaptive equalizers so these radios commonly use both types of equalizers [47].

Frequency-selective fading, which corrupts only a portion of a wide-bandwidth channel, is particularly difficult to equalize because it changes rapidly and requires complicated equalizer structures (in the frequency domain or the time domain). The problem can be greatly simplified if, instead of using a single carrier, multiple carriers (i.e., MCM) within the allotted bandwidth are used to create multiple subchannels. With this approach, fading within any subchannel is relatively flat and easy to equalize. The disadvantage of this approach is the cost of replicated circuitry for each subchannel. As reported in [52], however, atmospheric-induced outages are greatly reduced.

6.4.7 Route Design

The layout of a point-to-point microwave relay system involves numerous considerations of the local terrain, prevailing atmospheric conditions, radio frequency interference with other symbols, and interference from one hop to another within a single system.

Path Length

The foremost consideration in setting up a single hop of a microwave system is that line-of-sight transmission is required. Using antenna heights of about 60 m, the curvature of the earth limits the transmission distance to approximately 50 km. Longer line-of-sight distances are possible if taller antenna towers are used. However, practical considerations of mechanical stability limit the height of a tower, particularly since longer distances imply larger and more directive antennas.

When antennas are rigidly mounted on the top of a building or the side of a mountain, mechanical stability may not be a problem but the stability of the path itself may become a limiting factor. Under some atmospheric conditions, radio waves can be refracted to the point that a narrow beam completely misses the receiving antenna. Bending of the propagation path can also cause a fixed obstacle to intermittently obstruct transmission, implying that clearance between the normal path and nearby physical obstructions is required. Thus there is always a practical limit on how narrow the beam can be.

Local terrain also influences the prevalence of multipath fading. Nearby bodies of water contribute significantly to multipath conditions during late evening or early morning hours when there is little wind. Direct transmission over water is usually very difficult because of reflections off the water surface. Transmission over water often requires higher antennas and some means of blocking direct reflections.

Cochannel Interference

Since there are a limited number of channels available for point-to-point microwave systems, the same channels must be used over and over again. Reuse of microwave frequency bands is enhanced by the directivity of the antennas and the general need for line-of-sight reception. In many metropolitan areas, however, there is so much traffic converging into one particular area that it is impossible to completely isolate two systems using the same channel. This type of interference is referred to as cochannel interference.

Cochannel interference results from converging routes or from overreach of one hop into another hop of the same system reusing a channel. Sometimes reflections or atmospheric refractions can contribute to overreach, even when direct line of sight is not present. Overreach is usually controlled by zigzagging the hops so that a beam from one transmitter misses all subsequent receive antennas in a route. The reduced use of point-to-point microwave for long-distance traffic has reduced the cochannel interference in point-to-point systems. However, cochannel interference is a major consideration in the design and deployment of digital cellular systems, as discussed in Chapter 9.

REFERENCES

1 Federal Communications Commission, *Report and Order*, Docket No. 19311, FCC 74-985, released Sept. 27, 1974.

2 "Special Issue: AR6A Radio System," *Bell System Technical Journal*, Dec. 1983.

3 W. A. Conner, R. J. Girvin, and W. R. McClellan, "Frequency Control of the MSR-6 Single Sideband System," *IEEE International Conference on Communications*, 1983, pp. E2.5.1–E2.5.5.

4 Y. Yoshida, Y. Kitahara, and S. Yokoyama, "6G-90 Mbps Digital Radio System with 16-QAM Modulation," *IEEE International Conference on Communications*, 1980, pp. 52.4.1–52.4.5.

5 S. Pasupathy, "Minimum Shift Keying: A Spectrally Efficient Modulation," *IEEE Communications Magazine*, July 1979, pp. 14–22.

6 C.-E. Sundberg, "Continuous Phase Modulation," *IEEE Communications Magazine*, Apr. 1986, pp. 25–38.

7 Y. Morihiro, S. Nakajima, and N. Furuya, "A 100 Mbit/s Prototype MSK Modem for Satellite Communications," *IEEE Transactions on Communications*, Oct. 1979, pp. 1512–1518.

8 "Practical GMSK Data Transmission," MX-COM Application Note, 1998 (www.mxcom.com).

9 K. Murota and K. Hirade, "GMSK Modulation for Digital Mobile Radio Telephony," *IEEE Transactions on Communications*, July 1981.

10 R. E. Ziemer and W. H. Tranter, *Principles of Communications*, Houghton Mifflin, Boston, 1976.

11 E. Panayirci and E. Y. Bar-Ness, "A New Approach for Evaluating the Performance of a Symbol Timing Recovery System Employing a General Type of Nonlinearity," *IEEE Transactions on Communications*, Jan. 1996.

12 C. R. Hogge, Jr., "Carrier and Clock Recovery for 8-PSK Synchronous Demodulation," *IEEE Transactions on Communications*, May 1978, pp. 528–533.

13 P. R. Hartmann and J. A. Crosset, "A 90 MBS Digital Transmission System at 11 GHz Using 8 PSK Modulation," *IEEE International Conference on Communications*, 1976, pp. 18-8–18-13.

14 I. Horikawa, T. Murase, and Y. Saito, "Design and Performances of 200 Mbit/s 16 QAM Digital Radio System," *IEEE Transactions on Computers*, Dec. 1979, pp. 1953–1958.

15 J. A. Crosset and P. R. Hartmann, "64-QAM Digital Radio Transmission System Integration and Performance," *IEEE International Communications Conference*, 1984, pp. 636–641.

16 W. A. Conner, "Direct RF Modulation 256QAM Microwave System," *IEEE Globecom Proceedings*, 1988, pp. 52.7.1–52.7.6.

17 H. Matsue, T. Shirato, and K. Watanabe, "256 QAM 400 Mb/s Microwave Radio System with DSP Fading Countermeasures," *IEEE International Conference on Communications*, June 1988, pp. 41.5.1–41.5.6.

18 K. Pahlavan and J. L. Holsinger, "A Model for the Effects of PCM Compandors on the Performance of High Speed Modems," *IEEE Globecom Proceedings*, 1985, pp. 28.8.1–28.8.5.

19 T. Ryu, J. Uchibori, and Y. Yoshida, "A Stepped Square 256 QAM for Digital Radio System," *IEEE International Conference on Communications*, 1986, pp. 1477–1481.

20 M. Tahara, T. Deguchi, S. Mizoguchi, and Y. Yoshida, "6 GHz 140 Mbps Digital Radio System with 256-SSQAM Modulation," *IEEE Globecom Proceedings*, 1987, pp. 38.3.1–38.3.6.

21 S. A. Rhodes, "Performance of Offset-QPSK Communications with Partially Coherent Detection," *National Telecommunications Conference Record*, Nov. 1973, pp. 32A-1–32A-6.

22 M. K. Simon and J. G. Smith, "Offset Quadrature Communications with Decision-Feedback Carrier Synchronization," *IEEE Transactions on Communications*, Oct. 1974, pp. 1576–1584.

23 J. W. Bayless, R. D. Pedersen, and J. C. Bellamy, "High Density Digital Data Transmission," *National Telecommunications Conference Record*, 1976, pp. 51.3-1–51.3-6.

24 S. A. Gronemeyer and A. L. McBride, "MSK and Offset QPSK Modulation," *IEEE Transactions on Communications*, Aug. 1976, pp. 809–820.

25 R. D. Gitlin, S. K. Rao, J. Werner, and N. Zervos, "Method and Apparatus for Wideband Transmission of Digital Signals Between, for Example, a Telephone Central Office and Customer Premises," U.S. Patent No. 4,924,492, May 8, 1990.

26 B. Saltzberg, "Comparison of Single-Carrier and Multitone Digital Modulation for ADSL Applications," *IEEE Communications Magazine*, Nov. 1998, pp. 114–122.

27 "Mid-range Physical Layer Specification for Category 3 Unshielded Twisted-Pair," afphy-0018.000, *ATM Forum*, Sept. 1994.

28 C. W. Anderson and S. G. Barber, "Modulation Considerations for a 91 Mbit/s Digital Radio," *IEEE Transactions on Communications*, May 1978, pp. 523–528.

29 D. P. Taylor and M. Shafi, "Fade Margin and Outage Computation of 49-QPSR Radio Employing Decision Feedback Equalization," *IEEE International Conference on Communications*, 1983, pp. 1453–1458.

30 A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, Apr. 1967, pp. 260–269.

31 G. Ungerboeck, "Trellis-Coded Modulation with Redundant Signal Sets Part I: Introduction," *IEEE Communications Magazine*, Feb. 1987, pp. 5–11.

32 G. Ungerboeck, "Trellis-Coded Modulation with Redundant Signal Sets Part II: State of the Art," *IEEE Communications Magazine*, Feb. 1987, pp. 12–21.

33 R. J. F. Fang, "A Coded 8-PSK System for 140-Mbit/s Information Rate Transmission over 80-MHz Nonlinear Transponders," *Proceedings 7th International Conference on Digital Satellite Communications*, 1986, pp. 12–16.

34 T. Kamitake, K. Uehara, M. Abe, and S. Kawamura, "Trellis Coding 14.4kb/s Data Modem Implemented with a Single-Chip High-Speed Digital Signal Processor," *IEEE Globecom Proceedings*, 1987, pp. 12.8.1–12.8.6.

35 "A Modem Operating at Data Signaling Rates Up to 28,800 bps for Use on the General Switched Telephone Network and on Leased Point-to-Point 2-Wire Telephone-Type Circuits," ITU-T Rec. V.34, Sept. 1994.

36 Ramin Sadr and J. K. Omura, "Generalized Minimum Shift-Keying Modulation Techniques," *IEEE Transactions of Communications*, Jan. 1988, pp. 32–40.

37 J. A. C. Bingham, "Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come," *IEEE Communications Magazine*, May 1990, pp. 5–14.

38 J. M. Cioffi, "Asymmetric Digital Subscriber Lines," in *The Communications Handbook*, J. Gibson, Ed., CRC Press, Boca Raton, FL, 1998, pp. 450–479.

39 Digital Video Broadcasting (DVB): Framing structure, channel coding and modulation for digital terrestrial television, EN 300 744, v1.2.1, ETSI, 1999-01.

40 B. Saltzberg, "Comparison of Single-Carrier and Multitone Digital Modulation for ADSL Applications," *IEEE Communications Magazine*, Nov. 1998, pp. 114–122.

41 E. D. Sunde, "Ideal Binary Pulse Transmission by AM and FM," *Bell System Technical Journal*, Nov. 1959, pp. 1357–1426.

42 W. R. Bennet and J. R. Davey, *Data Transmission*, McGraw-Hill, New York, 1965.

43 R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communications*, McGraw-Hill, New York, 1968.

44 Federal Communications Commission, Memorandum Opinion and Order modifying FCC *Report and Order* of Docket 19311, released January 29, 1975.

45 G. H. M. de Witte, "DRS-8: System Design of a Long Haul 91 Mb/s Digital Radio," *IEEE National Telecommunications Conference*, 1978, pp. 38.1.1–38.1.6.

46 H. Kostal, D. R. Jeske, and V. K. Prabu, "Advanced Engineering Methods for Digital Radio Route Design," *IEEE International Communications Conference*, 1987, pp. 19B.6.1–19B.6.6.

47 C. P. Bates, P. L. Penney, and K. L. Seastrand, "Digital Radio Technology in the AT&T Network," *IEEE International Communications Conference*, 1987, pp. 19B.5.1–19B.5.7.

48 E. A. Sweedyk and P. Balaban, "A Multipath Fading Model for Terrestrial Microwave Radio," *IEEE Globecom Proceedings*, 1988, pp. 52.6.1–52.6.7.

49 S. H. Lin, "Opening New Vistas for Microwave Radio," *Bellcore Exchange*, July/Aug. 1988, pp. 9–12.

50 P. R. Hartmann and E. W. Allen, "An Adaptive Equalizer for Correction of Multipath Distortion in a 90 MB/S 8-PSK System," *IEEE International Conference on Communications*, 1979, pp. 5.6.1–5.6.4.

51 J. K. Chamberlain, F. M. Clayton, H. Sari, and P. Vandamme, "Receiver Techniques for Microwave Digital Radio," *IEEE Communications Magazine*, Nov. 1986, pp. 43–54.

52 Y. Nakamura, H. Ohtsuka, S. Aikawa, and H. Takanashi, "Advanced Techniques for Super Multi-Carrier Digital Microwave Radio with Trellis-Coded 256 QAM Modulation," *IEEE Globecom Proceedings*, 1989, pp. 11.3.1–11.3.6.

PROBLEMS

6.1 To prevent the transmission of line spectra, digital radio terminals use data scramblers to randomize the data patterns. Furthermore, differential encoding is normally required for proper data detection. Both functions introduce error multiplication. If the combined effect of these operations causes an average of five

decoded errors for every channel error, what is the effective penalty in transmit power for a 4-PSK system at BER = 10^{-6}? At BER = 10^{-3}?

6.2 If a digital radio receiver is experiencing thermal-noise-induced errors at a rate of 1 per 10^6 bits, what is the new error rate if the path length is decreased from 30 to 25 miles?

6.3 Derive Equation 6.18.

6.4 What is the minimum theoretical bandwidth of an 8-PSK signal carrying 4800 bps?

6.5 A 32-QAM signal set is implemented by eliminating the four corner points of a 36-QAM signal. What is the minimum error distance in terms of the peak signal power? How does this answer compare to the error distance for antipodal signaling? (Express the answers in decibels.)

6.6 What is the peak-to-average ratio of a 32-QAM signal (Problem 6.5) assuming transmission of random data?

6.7 How does 32-QAM (Problem 6.5) error performance compare to 32-PSK error performance?
(a) In terms of average signal powers?
(b) In terms of peak signal powers?

6.8 A carrier transmission system using 4-PSK modulation provides an error rate of 10^{-6}. If the modulation is changed to 16-PSK to narrow the transmission bandwidth (data rate unchanged), how much must the transmit power be increased to maintain the same error rate?

6.9 Repeat Problem 6.8 assuming that the bandwidth is unchanged and the data rate is doubled.

6.10 What is the error rate of an ideal 16-QAM signal with a signal-to-Gaussian-noise ratio (SNR) of 18 dB?

6.11 What is the error rate of Problem 6.10 if interference is also present at a level that is 21 dB below the received signal? (Assume the effect of the interference is identical to Gaussian noise at the same rms power.)

6.12 How much must the transmit power in Problem 6.11 be increased to offset the effect of the interference? (Assume the interference level is fixed.)

6.13 Repeat Problem 6.12, but assume that the interference increases in proportion to the increase in transmit power. (Signal powers in the adjacent channels are increased along with the desired signal power.)

6.14 Repeat Problem 6.3 assuming a starting state of D.

7

NETWORK SYNCHRONIZATION CONTROL AND MANAGEMENT

In Chapters 4 and 6 some synchronization requirements of transmission systems are discussed. These requirements involve carrier recovery for coherent detection of modulated signals, clock recovery for sampling incoming data, and framing procedures for identifying individual channels in a TDM signal format. All of these considerations are inherent in digital transmission systems and, for the most part, operate independently of other equipment in a network. An instance of one subsystem's dependency on another was noted for T1 lines. T1 source data must include minimum densities of 1's to maintain timing on the original transmission link. In contrast, other line codes are described that maintain clock synchronization independently of the source data.

This chapter discusses network-related synchronization considerations for interconnecting various digital transmission and switching equipment. Foremost among these considerations is synchronization of switching equipment. When digital switching equipment was first installed in the public telephone network, the interfaces were analog transmission systems. Hence, each switching machine could operate with an autonomous frequency source (clock) that converted all voice signals into digital signals with precisely the same data rate (nominally 64 kbps). These switching matrices were designed to carry one channel rate and one channel rate only. The advent of subsequent digital switch interconnection required switches to carry digital channels originating someplace else in the network—from a different frequency source. Thus, network synchronization requirements arose.

When individual synchronous transmission and switching equipments are interconnected to form a network, certain procedures need to be established that either synchronize the clocks to each other or provide for their interoperability when each subsystem uses independent clocks. Following the discussion of network clock synchronization, the concept of synchronization is extended to other aspects of network control.

7.1 TIMING

All digital systems inherently require a frequency source, or "clock," as a means of timing internal and external operations. Operations timed from a single frequency source do not require particularly stable sources since all commonly clocked elements experience timing variations in common. A different situation occurs when transfers are made from one synchronous equipment to another (as from a transmitter to a receiver). Even if the clock of the receiving terminal is "synchronized" to the transmitting terminal on a long-term or average basis, short-term variations in either clock may jeopardize the integrity of the data transfer. Thus it is generally necessary to use frequency sources (oscillators) in both the transmitter and the receiver that are as stable as is economically feasible.

7.1.1 Timing Recovery: Phase-Locked Loop

A common means of synchronizing a receiver clock to a transmitter clock uses a phase-locked loop (PLL) as shown in Figure 7.1. A phase detector continuously measures the phase difference between the incoming clock and a locally generated clock. The phase detector in Figure 7.1 merely measures the difference in the zero crossings between the two signals. When the zero crossing of the line clock precedes the zero crossing of the local clock, a positive voltage is generated; otherwise, a negative voltage is produced. The output of the phase detector is filtered to eliminate as much receive noise as possible, and then the phase measurement adjusts the frequency of the voltage-controlled oscillator (VCO) to reduce the phase difference. Some amount of noise or interference inevitably passes through the phase detector and the filter, causing erroneous adjustments in the VCO frequency. As time passes, however, a frequency offset produces ever-increasing phase shifts. When the phase difference builds up, it is easier to detect, and the appropriate changes in the VCO occur. Hence the local

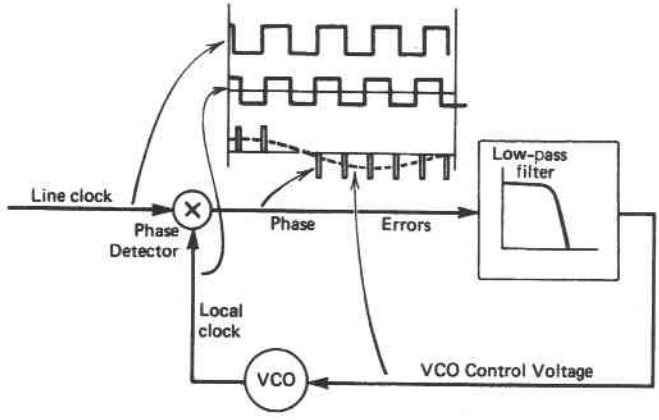


Figure 7.1 Phase-locked loop clock recovery circuit.

clock maintains the desired average frequency but inherently produces short-term variations of phase and frequency as it "hunts" the underlying frequency of the line clock.

The line clock in Figure 7.1 is shown to have a transition in every clock interval, a situation that does occur with Manchester or diphase-type line codes. With some line codes (bipolar or AMI in particular) there are no clock transitions during intervals of binary 0's. In these cases either the clock extraction circuitry inserts artificial transitions extrapolated from previous intervals or the phase detector is disabled during intervals when no pulse is detected.

7.1.2 Clock Instability

The variation in the output frequency of the VCO described above is an example of clock instability. All clocks have a certain amount of instability—even free-running oscillators. An important aspect of clock instability is its frequency: the rate at which the clock frequency changes from being too high to being too low. The frequency of the instability can be directly observed as the frequency spectrum of the VCO control voltage in a PLL clock recovery circuit. When the VCO control voltage varies slowly, the variations are referred to as clock wander. When the variations are more rapid, the instability of the clock is referred to as jitter. The most generally accepted dividing point between wander and jitter is 10 Hz. Pure wander on a T1 line therefore produces positive phase errors for more than 77,200 bit intervals (0.05 sec) followed by more than 77,200 bit intervals of only negative phase errors.* The main sources of clock instability (both wander and jitter) in a network are:

1. Noise and interference
2. Changes in the length of transmission media
3. Changes in velocity of propagation
4. Doppler shifts from mobile terminals
5. Irregular timing information

Noise and Interference

If the low-pass filter of a PLL is designed with a very low cutoff frequency, it can filter out almost all of the noise and interference on a transmission link that would otherwise corrupt the timing recovery. There are three main reasons why arbitrarily low pass filters cannot be used. First, the ability of the PLL to acquire synchronization (e.g., lock on to the underlying clock) is inversely related to the PLL bandwidth. If the VCO begins oscillating at the wrong frequency and the bandwidth is too narrow, the PLL may never pull the oscillator to the frequency of the line clock. Sometimes, this problem is accommodated by using two bandwidths: a wide one to acquire synchronization and a narrow one that is selected after lock is achieved.

*As discussed later, observing wander in this way requires extremely stable clocks as a reference. The normal PLL clock recovery circuitry cannot be used to observe wander because the VCO tracks the relatively long-term phase offsets. Thus the clock recovery circuit does not filter out the wander but passes it on.

A second consideration that generally precludes the use of very narrow filters, even after acquisition, is that the source may vary in frequency that cannot be tracked by a slowly responding PLL. In this case, the recovered clock does not track the ideal sample times leading to high error rates or, worse yet, the PLL loses synchronization altogether and has to reacquire lock.

The third limiting factor for low-bandwidth PLLs is the instability of the VCO itself. If the VCO begins to drift in frequency, very low bandwidth filters preclude adjusting the VCO input voltage soon enough to prevent bit errors or possible loss of synchronization.

Because operational considerations dictate certain minimum PLL bandwidths, noise and interference on the transmission link always cause the recovered clock to be more impaired than the source clock. The PLL does, however, eliminate that portion of a disturbance with frequency content above the bandwidth of the PLL. Thus disturbances with low-frequency content are the most difficult to deal with. Systematic jitter, as produced by particular patterns of intersymbol interference, can have an arbitrarily low frequency content and is discussed more fully in a later section.

An important consideration in the design of a digital transmission link is the accumulation of jitter in tandem clock recovery circuits. If a recovered clock is used to time the transmission of outgoing data, as in a regenerative repeater, some amount of incoming jitter is imbedded into the outgoing clock. The clock recovery circuit in the next receiver tracks its incoming clock but introduces even more jitter due to noise and interference on the second section. Thus jitter accumulates at every regenerative repeater using its received line clock as its transmit clock. If there is a large number of regenerative repeaters, the jitter can accumulate to a point where subsequent clock recovery circuits have difficulty tracking the receive clock, produce sampling errors, and possibly lose lock.

Changes in Length of Transmission Media

Path length changes occur as a result of thermal expansion or contraction of guided transmission media or of atmospheric bending of a radio path. While a path is increasing in length, the effective bit rate at the receiver is reduced because more and more bits are being "stored" in the medium. Similarly, as the path shortens, the bit rate at the receiver increases because the number of bits stored in the transmission link is decreasing. After the path has stabilized, the receive signal returns to the nominal data rate. The most significant changes in path length occur with communications satellites. Geostationary satellites produce path length variations of approximately 200 miles, which cause propagation time variations of approximately 1 msec [1]. Path length changes also occur in guided transmission media such as copper wire and optical fibers. These changes are referred to as diurnal changes because they occur once a day.

Changes in the Velocity of Propagation

Temperature changes not only cause expansion and contraction of wireline transmission media but also can change those propagation constants of the media that deter-

mine the velocity of propagation. The resulting change in received clocked stability, however, is much less than that produced by the change in path length [2].

The propagation velocity of radio waves in the atmosphere also changes with temperature and humidity. Although these velocity changes are more significant than those occurring in wirelines, they are still smaller than the path-length-induced variations. Notice that a change in propagation velocity is effectively equivalent to a change in path length since the number of bits "stored" in the transmission path is changed.

Doppler Shifts

The most significant source of potential timing instability in a received clock occurs as a result of Doppler shifts from airplanes or satellites. For example, a Doppler shift induced by a 350-mph airplane amounts to an equivalent clock instability of 5×10^{-7}. Digital mobile telephone receivers must accommodate Doppler shifts equivalent to clock instabilities of about one part in 10^7.* Again, Doppler shifts occur, in essence, as a result of path changes.

Irregular Timing Information

As discussed in Chapter 4, a fundamental requirement of a digital line code is that it provide sufficient timing information to establish and maintain a receiver line clock. If the timing information is data dependent, jitter in the recovered clock increases during periods of relatively low density timing marks. The magnitude of the jitter is dependent not solely on the density of timing marks but also on the timing (data) patterns. In an ideal repeater, only the density would matter. In practice, however, various imperfections lead to pattern-dependent jitter [3].

As discussed later in this chapter, higher level digital multiplexers insert overhead bits into a composite data stream for various purposes. When the higher rate data stream is demultiplexed, the arrival rate of data within individual channels is irregular. This irregularity produces timing jitter when generating new line clocks for the lower rate signals. This source of jitter (waiting time jitter) is often the most troublesome and is discussed in more detail later.

7.1.3 Elastic Stores

The timing instabilities described in the preceding paragraphs essentially represent changes in the number of bits stored in a transmission link. In the case of noise- and interference-induced jitter, the change in "bits stored" occurs because data are sampled a little earlier or a little later than nominal. Since the outgoing data of a regenerative repeater are transmitted according to the recovered clock, a phase offset in the clock means the delay through the repeater is different from when there is no misalignment in timing.

*If you are paranoid about cooperation between cellular operators and law enforcement authorities, you may not want to use your cell phone while speeding.

If phase offsets in successive regenerative repeaters coincide, a net change of several bits of storage in a long repeated transmission link occurs. Since these extra bits enter or leave the transmission link over relatively short periods of time, the accumulated jitter may represent a relatively large, but short lived, instability in the receive clock.

Because regenerative repeaters use incoming sample clocks as output clocks, sustained timing differences between inputs and outputs do not exist. The endpoints of a transmission link, however, may interface to a local clock. In this case a difference between a received and a relatively fixed local clock must be reconciled with an elastic store. An elastic store is a data buffer that is written into by one clock and read from by another. If short-term instabilities exist in either clock, the elastic store absorbs the differences in the amount of data transmitted and the amount of data received. An elastic store can compensate only for short-term instabilities that produce a limited difference in the amounts of data transmitted and received. If sustained clock offsets exist, as with highly accurate but unsynchronized clocks, an elastic store will eventually underflow or overflow.

TDM–Switch Interface

A typical need for an elastic store occurs when a digital transmission link is interfaced to a digital time division switch. As shown in Figure 7.2, the elastic store is placed between the incoming digital transmission link and the inlet side of the switch. In most instances the digital switch provides timing for all outgoing TDM links so that no timing discrepancies exist between these links and the switch. For the time being, assume that the far end of the digital link derives its clock from the receive signal and uses that clock to time digital transmissions returning to the switch. This is the situation that arises when a remote channel bank is connected to a digital switch through T1 lines and is commonly referred to as channel bank "loop timing." When loop timing is used, the line clock on the incoming link of the switch is nominally synchronized to the switch clock. However, for reasons discussed previously, a certain amount of instability in the incoming clock necessarily exists. The elastic store absorbs these instabilities so that purely synchronized data are available for the switch.

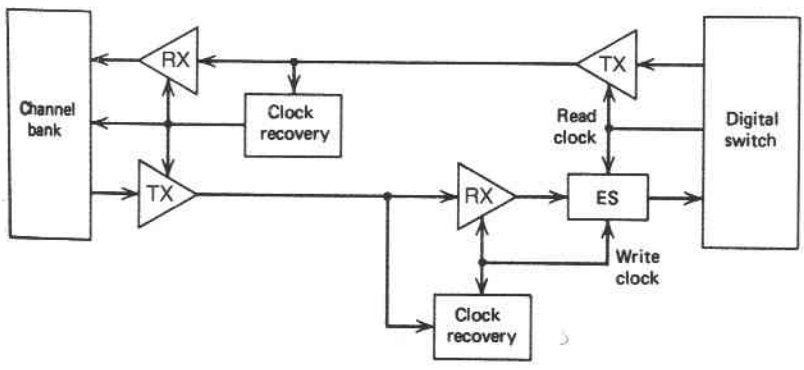


Figure 7.2 Interface between TDM transmission link and a digital switch using an elastic store.

In essence, the loop formed by the transmission links and the elastic store maintains a constant and integral number of clock intervals between the inlet and outlet of the switch. Thus, from a timing point of view, the inlets and outlets operate as though directly connected to each other using a common source of timing.

Removal of Accumulated Jitter

Another application for an elastic store is shown in Figure 7.3, where it is used to remove transmission-induced timing jitter in a regenerative repeater. Normally, a regenerative repeater establishes the transmit timing directly from the locally derived sample clock. In Figure 7.3, however, the transmit timing is defined by a separate local clock. The elastic store absorbs the short-term instabilities in the receive clock, but the long-term frequency of the transmit clock is controlled by maintaining a certain "average level of storage" in the elastic store. Thus the transmit clock is synchronized to the line clock on a long-term basis, but not on a short-term basis. If the elastic store is large enough to accommodate all transient variations in the data rate, high-frequency instability of the input clock is removed.

All regenerative repeaters, regardless of the mechanism used to recover timing, derive their output clocks by averaging the incoming timing information over a period of time. Tuned circuits average the incoming clock for relatively few signal intervals; phase-locked loops do so for many intervals. In all cases a certain amount of storage or delay is implied. An elastic store is merely a mechanism to increase the available delay so that output timing adjustments can be made more gradually. Notice that an elastic store always inserts significant amounts of artificial delay into the data path. To remove arbitrarily low frequency instabilities (wander), arbitrarily large elastic stores (and arbitrarily stable VCOs) are required. Thus a jitter-removing elastic store should be inserted into a chain of repeaters only when jitter accumulation threatens the ability of a regular regenerative repeater to maintain synchronization.

Jitter is not just a phenomenon of a digital transmission link but also occurs in digital storage systems. For example, jitter removal like that shown in Figure 7.3 is used in laser disc players to eliminate blurring of the visual images [4].

Elastic Store Implementations

The required size of an elastic store varies from a few bits to several hundred bits for high-speed, long-distance communications links. Figure 7.4 shows one means of implementing a small elastic store utilizing a series-to-parallel converter, a register, and

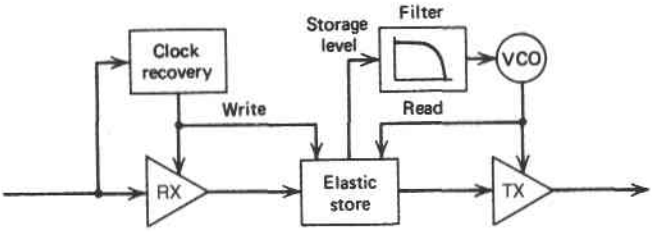


Figure 7.3 Jitter-removing regenerative repeater.

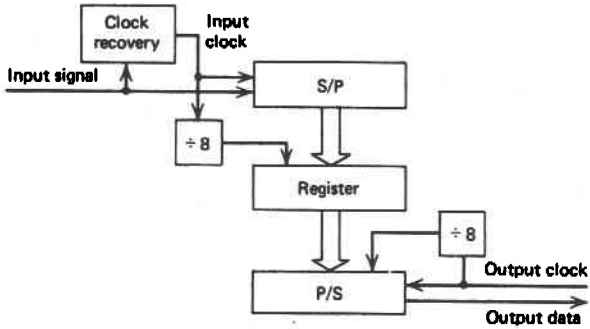


Figure 7.4 Basic implementation of an elastic store.

a parallel-to-series converter. As indicated, incoming data are transferred into the register as soon as each word is accumulated in the series-to-parallel converter. Some time later, data in the register are transferred to the output parallel-to-series converter as a complete word is shifted out. Notice that transfers to the parallel-to-series converter are independent of the incoming clock. As long as output transfers occur between input transfers, no data are lost, and short-term jitter is absorbed by varying delays through the elastic store.

Normally, some control circuitry (not shown) is needed to initialize the elastic store so that the first transfer into the register occurs midway between output transfers. This process means some incoming data are initially discarded by the series-to-parallel register until the desired transfer time occurs.

The relative times of the parallel transfers into and out of the holding register provide a direct indication of the relative phase of the input and output clocks. Thus the parallel transfer clocks contain the information needed to generate VCO control voltages if the elastic store is being used to remove accumulated transmission jitter.

The basic structure shown in Figure 7.4 can be extended to implement larger elastic stores, as shown in Figure 7.5. The only change involves the substitution of a first-in, first-out (FIFO) buffer for the holding register of Figure 7.4. This data buffer is designed specifically to allow input transfers under the control of one clock while outputs are controlled by a different clock. Normally, the FIFO buffer is initialized by inhibiting output transfers until it is half full. In fact, some commercially available FIFO buffers have an output signal specifically indicating when it is half full or greater.

7.1.4 Jitter Measurements

A simple circuit for measuring timing jitter is shown in Figure 7.6. As indicated, it is nothing more than a phase-locked loop (PLL) with the output of the phase comparator providing the measurement of the timing jitter. Normally, the bandwidth of the low-pass filter (LPF) is very small so the VCO ignores short-term jitter in the timing signal. If there is no jitter at all, the output of the phase comparator is constant and no signal is passed by the high-pass selection filter (HPF).

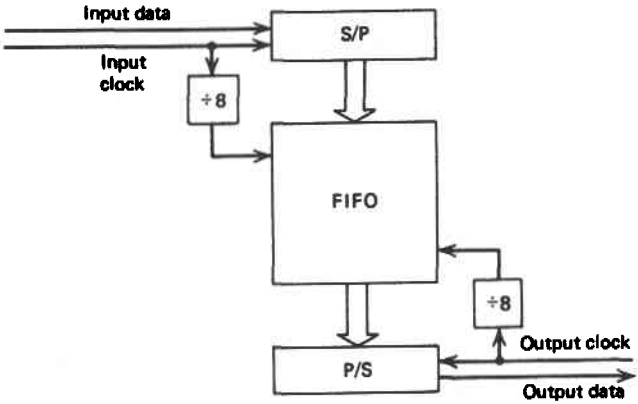


Figure 7.5 FIFO implementation of an elastic store.

Very low frequency jitter cannot be measured by the circuit in Figure 7.6 because the VCO tracks slowly changing phase shifts. Low-frequency jitter may be of no concern, however, because it can be tracked by a PLL. Higher frequency jitter, on the other hand, is more apt to cause sampling errors or a loss of lock in the clock recovery of a repeater. Thus the spectral content of the jitter as well as its magnitude is of interest. Besides not being able to measure low-frequency jitter, the circuit of Figure 7.6 cannot operate if it cannot lock onto the fundamental clock signal. Section 7.4.7 describes another way to measure jitter and other timing impairments within a network.

Phase jitter is commonly specified by communications theorists as a power measurement in units of radians squared or cycles squared. (One cycle = 2π radians.) As indicated in Figure 7.7, phase jitter power is then a measure of the variance of the number of clock cycles or unit intervals (UIs) stored in the transmission link. In a physical sense jitter "power" has little meaning because it represents timing variations, not power. Some physical justification for expressing jitter as a power can be obtained by observing that the rms power σ_v^2 of the phase detector output signal is proportional to the rms phase jitter σ_ϕ^2:

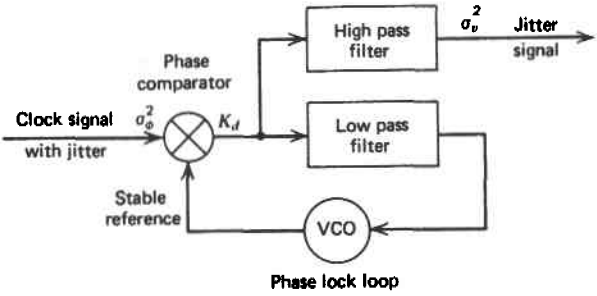


Figure 7.6 Circuit for measuring timing jitter.

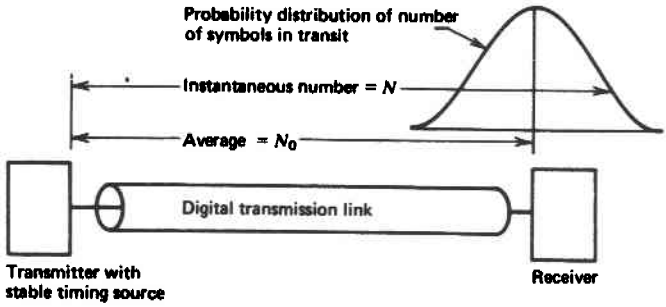


Figure 7.7 Phase jitter modeled as the variance in the number of symbols "stored" in the transmission link.

$$\sigma_\nu^2 = K_d^2 \sigma_\phi^2 \quad (\mathrm{V}^2) \tag{7.1}$$

where K_d is the phase detector gain factor in volts per radian.

Example 7.1. Given an rms phase jitter of 10.7 dB relative to one UI, what is the standard deviation of the phase offset?

Solution. The variance of the signal phase is determined as $\sigma_\phi^2 = 10 \exp(10.7/10) = 11.76$ UI squared. Hence, the standard deviation is $(11.76)^{1/2} = 3.43$ UI (symbol intervals). Since 68% of a normal probability distribution lies within one standard deviation, the phase of this signal is within ±3.43 symbol intervals for 68% of the time. One percent of the time the signal phase will be outside 2.6 standard deviations, or ±8.9 symbol intervals.

If phase jitter arises as a result of additive Gaussian noise on a stable signal, the phase noise can be approximated as

$$\sigma_\phi^2 = \frac{\sigma_n^2}{2P_s} \quad (\mathrm{rad}^2) \tag{7.2}$$

where σ_n^2 = additive noise power
P_s = signal power

Equation 7.2 is the basic equation of phase jitter produced by additive noise on a continuous sinusoid [5]. When timing is extracted from a data signal, the timing information is usually not continuous. The distinction is not important because jitter produced by additive noise is normally insignificant compared to other sources [6]. For an analysis of phase jitter produced by regenerative repeaters operating on randomly occurring timing transitions, see reference [7].

TABLE 7.1 Maximum Instability of DS1 Customer Interface

Frequency Band	Maximum Peak-to-Peak Variation
< 10 Hz	28 UIs in 24 hr
< 10 Hz	5 UIs in 15 min
10 Hz–40 kHz	5 UIs
8 kHz–40 kHz	0.1 UIs

As an example of a particular instability specification, Table 7.1 lists the wander and jitter specifications of a DS1 digital carrier (T1) interface to the public network [8]. Notice that the higher frequency jitter specification allows deviations in pulse centers of only 5%, which effectively specifies the accuracy of the data sample clock. Larger variations are allowed for lower frequency instabilities because the clock recovery circuits can track the changes to maintain a good sample clock.

7.1.5 Systematic Jitter

An original analysis of jitter in a chain of digital regenerators was reported by Byrne, Karafin, and Robinson [3]. Figure 7.8 shows the basic model of their analysis. Each of the regenerative repeaters in a T-carrier line extracts timing from the received waveform and passes that timing on to the next regenerator as a transmit clock. Because of implementation imperfections (primarily intersymbol interference) in the timing recovery circuits, jitter produced by repeaters is dependent on the data patterns. One worst-case pattern produces an extreme phase lag. Another pattern produces an extreme phase lead. When the data pattern shifts from one worst case to the other, a phase ramp occurs. Because every repeater has the same basic implementation, the jitter produced by individual repeaters tends to be coherent. The systematic nature of this jitter makes it the most significant source of accumulated line clock jitter at the end of a chain of repeaters.

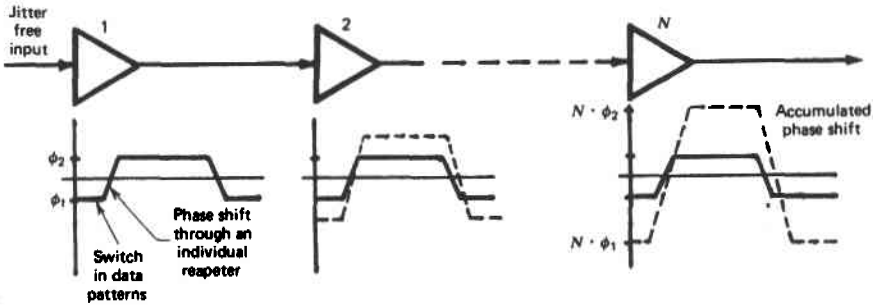


Figure 7.8 Model of systematic jitter in a string of regenerative repeaters: Φ_2 = phase shift produced by worst-case data pattern for phase lead; Φ_1 = phase shift produced by worst-case data pattern for phase lag.

As indicated in Figure 7.8, the last repeater in the chain experiences a large phase ramp equal to the number of repeaters times the phase shift of each individual repeater. This phase ramp represents an abrupt change in the clock frequency that may cause bit errors or a complete loss of synchronization. Thus there is a limit to the number of repeaters that can be used without jitter removal. For more analyses of jitter accumulation including the combined effects of systematic jitter and random perturbations, see references [9], [10], and [11].

7.2 TIMING INACCURACIES

In the preceding section the nature of certain instabilities or transient variations in timing was discussed. Although these variations represent shifts in the frequency of a line clock, the shifts are only temporary and can be absorbed by elastic stores. In some instances digital communications equipment using autonomous frequency sources must be interconnected. When this happens, the clock rates of the two systems are never exactly the same, no matter how much accuracy is designed into the frequency sources. An offset in the two clocks, no matter how small, cannot be reconciled by elastic stores alone.

In the preceding section, channel bank loop timing was mentioned as an example of how remote terminals are synchronized to a digital switch. When the remote terminal is another digital switch using its own frequency source as a reference, a different situation results. As shown in Figure 7.9, the outgoing clock for each direction of transmission is defined by the local switch clock. Thus the incoming clock at each switch interface contains not only transmission-line-induced jitter but also a small and unavoidable frequency offset.

7.2.1 Slips

As indicated in Figure 7.9, the interface of each incoming digital link necessarily contains an elastic store to remove transmission link timing jitter. The elastic store at the first digital switch is written into by the recovered line clock but *read from* the local rate R_1. If the average rate of the recovered line clock R_2 is different from R_1, the elastic

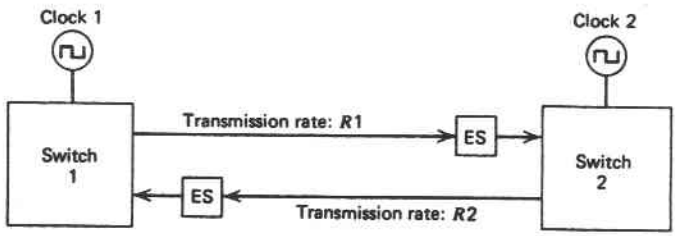


Figure 7.9 Connections between autonomously timed digital switches.

store will eventually underflow or overflow, depending on which rate is larger. When R_2 is greater than R_1, the elastic store at the first digital switch overflows, causing a loss of data. If R_2 is less than R_1, the same elastic store underflows, causing extraneous data to be inserted into the bit stream entering the switch. Normally, the extraneous data are a repetition of data bits already transferred into the switch. Disruptions in the data stream caused by underflows or overflows of an elastic store are referred to as "slips."

Uncontrolled slips represent very significant impairments to a digital network because they generally cause a loss of frame synchronization. Therefore, slips are allowed to occur only in prescribed manners that do not upset framing. One general approach to controlling the slips is to ensure that they occur only in the form of a repetition or deletion of an entire frame. Thus the time slot counters and framing logic associated with the multiplex group remain synchronized. Controlled slips comprising entire frames can be assured by using elastic stores with at least one frame of storage. As a slip occurs, the storage level in the elastic store is effectively increased or decreased by a full frame. Rather than actually inserting or deleting a frame of information, the desired effect is achieved more easily by indexing address pointers in a random-access memory. Such a system is shown in Figure 7.10.

The elastic store in Figure 7.10 operates by sequentially writing input information into memory addresses corresponding to individual TDM channels. Data for individual output channels are obtained by reading the same addresses in the same sequential manner. Ideally, if there is no offset between the clock rates, the read times of each channel occur midway between write times for corresponding channels. The elastic store then has the capability of absorbing transmission delay variations up to one-half of a frame time.

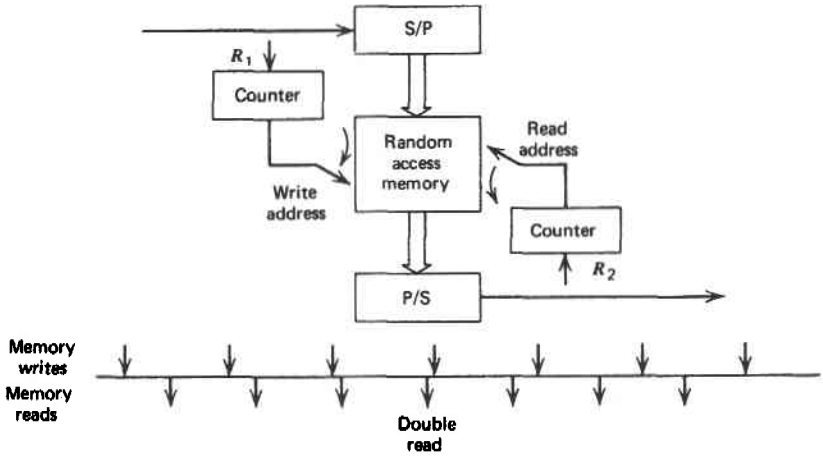


Figure 7.10 Elastic store operation with a one-frame memory.

The timing diagram in Figure 7.10 depicts an exaggerated timing offset in which the switch clock R_2 is greater than the incoming clock R_1. As indicated, the read times catch up gradually with the write times until a "double read" occurs. At that time the information retrieved for each channel is a repetition of the information retrieved for the previous outgoing frame. Although write and read times for only one channel are shown, the corresponding times for all other channels have the same relationship. Thus all channels slip together. Notice that R_1 is greater than R_2, a slip occurs when a "double write" on all channels causes the information in the previous incoming frame to be overwritten.

The elastic store operation depicted in Figure 7.10 is very similar to the operation of a time slot interchange memory described in Chapter 5. This relationship is exploitable in a TST switch where the inlet memory can provide both the elastic store function and the time switching function. When the two functions are combined, slips generally occur at different times for different channels. Nevertheless, individual channels maintain proper frame alignment since each channel is transferred through the inlet memory using dedicated memory addresses.

One attractive feature of using the inlet memory as an elastic store is that, when setting up a new connection, an internal switching time slot can be chosen so that the inlet memory read is halfway between inlet memory writes for the particular channel. Thus a slip in that connection will not occur for a long time, probably not until long after the connection is released. (With a clock inaccuracy of one part in 10^8, the time between slips in any one channel is 3.5 hr.)

One potential problem with the elastic store in Figure 7.10 occurs when write and read times nearly coincide. When both accesses to a single channel occur one after the other, transient timing instabilities can cause the two accesses to cross back and forth with respect to each other. Thus slips caused by double reads may follow slips caused by double writes and vice versa. To remedy this situation, some amount of hysteresis is needed in the counter adjustment process. The hysteresis, in turn, implies that additional storage is needed to defer the occurrence of one type of slip after a slip of the other type has recently occurred.

One means of implementing an elastic store with the desired hysteresis is to use two frames of storage as shown in Figure 7.11. For convenience, the elastic store is divided into an A-frame memory and a B-frame memory. The counter logic again accesses the memories in sequential fashion except that frames are written alternately into the A and B memories. Under normal operation, the memories are accessed in the same way for output data. When a slip is imminent, however, control logic causes the output channel counter to be reset so that the A memory is read twice in a row. This situation is depicted in the timing diagram of Figure 7.11, which again assumes that R_2 is greater than R_1. The important point to be noticed in the timing diagram is that after the counter adjustment produces a double read of memory A, the write and read times of each individual memory are approximately one frame time apart. Thus another adjustment can be deferred until the write and read accesses again drift one full frame time with respect to each other. The structure and mode of operation shown in Figure 7.11 describe the elastic store used for DS1 signal interfaces to AT&T's (now Lucent's) No. 4 ESS [12].

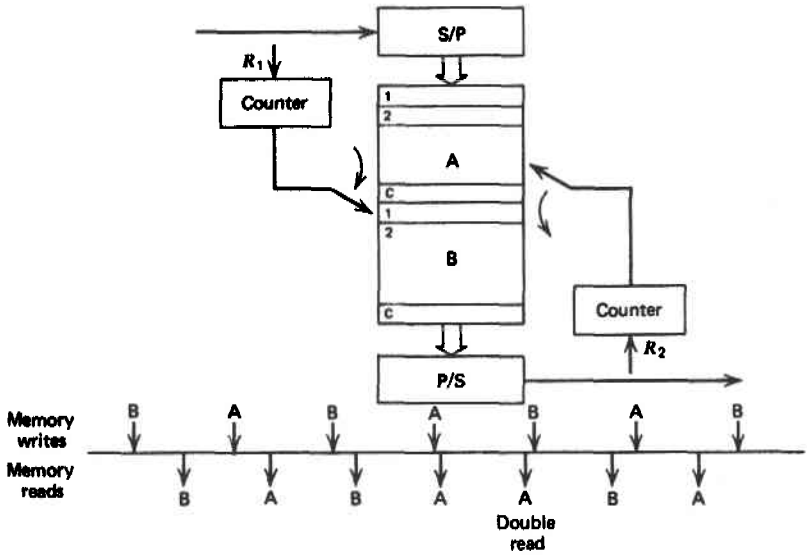


Figure 7.11 Elastic store with a two-frame memory.

Slip Rate Objectives

If the difference between an elastic store's input data rate and its output data rate is ΔR,* the time between slips is

$$\Delta T = \frac{N}{\Delta R} \tag{7.3}$$

where N is the number of bits that get dropped or repeated whenever a slip occurs. Normally, a slip involves a full frame of data, in which case the time between slips is determined as

$$\Delta T = \frac{1}{\Delta F} \tag{7.4}$$

where ΔF is the difference in frame rates.

As long as slips are controlled so they do not disrupt framing,† their only effect is an infrequent repetition or deletion of the information within affected TDM channels. The audible effect of slips on a digitized voice signal is an occasional "click." Only one slip in 25 produces an audible click in PCM voice [1]. Voice signals can therefore tolerate several slips per minute [13].

*Typically, clock offsets are specified in relative terms (e.g., one part in 10^6). A clock that is accurate to P parts per million (ppm) has a maximum offset of $\Delta R = R \cdot P/10^6$.

†With single-frame slips, framing is disrupted in the sense that framing pattern is shifted. Thus, the frame sequence has to be reacquired, but this process is simplified by the fact that the location of the framing bits and the integrity of the message channels are maintained.

A more subtle and troublesome aspect of slips occurs when a digitized channel carries voiceband data. High-speed data modems for the analog telephone network use QAM modulation with coherent detection in the receiver. Since these modems are particularly sensitive to phase shifts, they are particularly vulnerable to slips. An 8-bit slip in a digitized modem signal using a carrier of 1800 Hz causes an instantaneous phase shift of 81°. Obviously, a phase shift of this size causes a data error, but more importantly, it upsets the carrier recovery circuitry in the receiver and causes multiple errors. A single slip can upset the operation of some voiceband modems for several seconds [14, 15].

Characterizations of the effects of slips on Group 3 facsimile equipment [16] reveal that a single slip can cause the loss of four to eight scan lines without an error report. Sometimes the loss of the lines is not immediately evident to the reconstructed image. Diagonal lines, however, readily reveal missing vertical space.

Encrypted traffic (voice or data) is more susceptible to slips since the encryption–decryption process usually relies on bit-synchronous scramblers and unscramblers. When the bit count is altered by insertion or deletion of bits in a time slot, counters in the source and destination become unsynchronized. At best, the decryption process causes every slip to be audible. At worst, unintelligible speech or data result until the unscrambler is resynchronized. Another important aspect of encrypted communication is transmission of encryption keys or indexes to encryption keys. If synchronization is lost for some reason, it may be necessary to resynchronize the encryption keys, thereby compromising security.

When a digital transmission link is being used to transmit data directly, the effect of a slip may not be any more significant than a single channel error. Most data communications receiving equipment requests a complete retransmission of any block of data not satisfying certain redundancy checks. Thus one error is as bad as many errors or a complete loss of data. The effect of the slip will be more significant, however, if the communications protocol [e.g., the DDCMP of Digital Equipment Corporation (Compaq)] relies on byte count procedures to delimit message blocks. Insertion or deletion of data by the network causes the receive counter to become unsynchronized, and the normal exchange of information is disrupted until the loss of synchronization is recognized.

From the foregoing considerations for data transmission, the slip rate objective for the AT&T network and adopted by Bellcore and ANSI for North America was set at one slip in 5 hr for an end-to-end connection [17–19]. Since slips can occur at multiple points within a network, the objective for slips at individual trunk and switching interfaces was set at one slip every 20 hr.

Example 7.2. Determine the relative accuracy requirements of two independent clocks to maintain a mutual slip rate objective of one slip in 20 hr. Assume a frame rate of 8 kHz as in PCM voice signals.

Solution. The slip rate objective implies that the frame rate produced by one clock can be different than the frame rate produced by the other clock by no more than

$$\Delta F = \frac{1}{20 \times 60 \times 60} = 1.39 \times 10^{-5} \text{ slips per second}$$

Since there are 8000 frames per second, the relative accuracy is determined as

$$\frac{1.39 \times 10^{-5}}{8000} = 1.7 \times 10^{-9} \text{ slips/frame}$$

Hence the clocks must be accurate to 1.7 parts in 10^9.

Because Example 7.2 determines only a maximum relative inaccuracy, the absolute inaccuracy of each individual clock must be less than $(1.7 \times 10^{-9})/2$, or 0.85 parts in 10^9.

7.2.2 Asynchronous Multiplexing

In the preceding section certain aspects of network synchronization were discussed that implied the need for clock synchronization to prevent a loss of data by way of slips. In this section a procedure referred to as "pulse stuffing" is discussed that avoids both slips and clock synchronization. The term pulse stuffing can be somewhat misleading since it implies that pulses are inserted into the line code to make timing adjustments. Actually, pulse stuffing involves only the data stream and is independent of the line code or modulation system in use. Pulse stuffing is a term commonly used in North America while the same concept is referred to as "justification" in Europe.*

The basic concept of pulse stuffing involves the use of an output channel whose rate is purposely higher than the input rate. Thus the output channel can carry all input data plus some variable number of "null," or "stuff," bits. The null bits are not part of the incoming data. They are inserted, in a prescribed manner, to pad the input data stream to the higher output rate. Naturally, the extraneous null, or stuff, bits must be identifiable so that "destuffing" can recover the original data stream.

The practice of pulse stuffing arose when the initial digital TDM hierarchy (Table 1.10) was defined. At this time there were only isolated digital transmission links within the network that precluded synchronizing them to a common clock. When it came time to combine lower rate tributaries (e.g., DS1s) into higher level signals (e.g., DS2s or DS3s), the multiplexing procedure necessarily had to accommodate tributaries operating at slightly different rates. The generic term for combining these unsynchronized signals is asynchronous multiplexing. In this context, "asynchronous" refers to multiplexing of unsynchronized tributaries into a higher level signal (using pulse stuffing). It does not refer, in any way, to a means of transmission. The higher level signal is always carried on a synchronous transmission link.

As the amount of digital equipment in the network grew and more and more of it became interconnected, both the means and the necessity for a different form of mul-

*Justification is the printing industry practice that aligns the right side of lines of text by inserting a variable amount of space within the line. As you will see in the following discussion, aligning individual tributaries to the rate of a higher level multiplexer is conceptually the same process.

tiplexing arose. The chosen approach is referred to as synchronous multiplexing (SONET) in North America and Synchronous Digital Hierarchy (SDH) in the rest of the world. The principles of SONET and SDH are described in conjunction with fiber systems in the next chapter.

Pulse-Stuffing Concepts

As a starting point for understanding the need for pulse stuffing, consider the simple two-channel, bit-interleaved multiplexer in Figure 7.12. As indicated, within any string of even-numbered bits in the multiplexer output, the number of bits carried in each subchannel is necessarily identical. Thus the rates of the subchannels are also identical. If the two input data streams are running at different rates, the output can be synchronized to one of the channels but not both. Thus slips would necessarily occur in at least one of the tributaries.

As a simplified example of pulse stuffing, Figure 7.13 shows a two-channel, bit-interleaved format as before but with the additional detail needed to allow adjustments of the information flow within each subchannel. As indicated, the multiplexed output is formatted into 10-bit master frames with 5 bits assigned to each subchannel. The first 3 bits in each subchannel of each master frame always carry data from the respective tributaries. The fourth bit in each subchannel (C_1 and C_2) specify whether the last bits (S_1 and S_2) carry data or are stuff bits. When C_1 is a "1," a bit is stuffed; otherwise S_1 carries tributary data. Hence each master frame can carry 3 or 4 bits from each tributary. If, on average, each tributary sends 3.5 bits during a master frame, variations of ±14% in the tributary clock rates can be accommodated.

An important point to notice about an asynchronous multiplexer is that the output frame structure is unrelated to the frame structure of the lower level inputs. As far as the higher level multiplexer is concerned, each input signal is merely a serial bit stream with no particular structure assumed. Framing bits in the lower level multiplex signals are transmitted right along with the information bits. After the higher level signal is demultiplexed and the tributaries are unstuffed, framing of the lower level signals must be established for further demultiplexing.

Although pulse stuffing can be implemented with a variety of higher level framing formats, the generally most desirable features of a pulse-stuffing format are identified as follows:

1. The use of fixed-length master frames with each channel allowed to stuff or not to stuff a single bit in the master frame

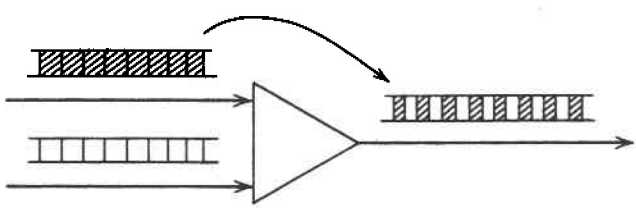


Figure 7.12 Two-channel multiplexer showing equal output data rates for each input.

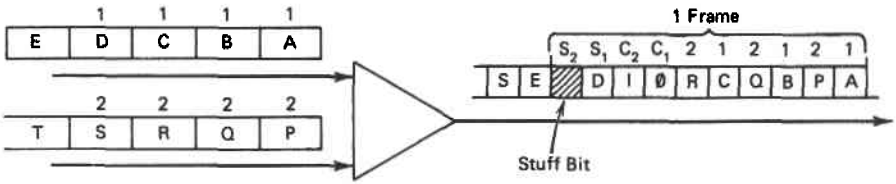


Figure 7.13 Simplified pulse-stuffing example.

2. Redundant stuffing specifications
3. Noninformation bits distributed across a master frame

Timing offsets are generally quite small, so only small adjustments from a single occasional stuff bit are required. Thus large numbers of tributary bits can be combined into a master frame with one specific bit position identified as the S bit. Nominally, approximately one-half of the master frames contain the maximum number of information bits N_M, and the other half contain $N_M - 1$ information bits.

The purpose of a pulse-stuffing operation is to prevent a loss of data when two interconnected digital transmission links are unsynchronized with respect to each other. If single bit errors can cause a stuff bit to be interpreted as information (or vice versa), the basic objective is lost. Furthermore, notice that an erroneous interpretation of a stuff code causes embedded, lower level multiplex signals to lose framing. For these reasons, the interpretation of C bits must be encoded redundantly. Under the assumption that channel errors are random, the probability of misinterpreting a stuff code is

$$P_F = \sum \binom{2n+1}{n+1} p^{n+1}(1-p)^n \tag{7.5}$$

where p = probability of a channel error
n = number of correctable stuff code errors ($2n + 1$ bits in a stuff code)

Information bits should be distributed across a master frame for several reasons. First, by separating these bits as much as possible, errors in redundant stuffing specification bits (C bits) are more likely to be independent. If the specification bits are too close together and burst errors are prevalent, the redundant encoding is of little use. Second, by distributing noninformation bits, the irregularity of information flow is minimized. When higher level multiplex signals are demultiplexed, a clock for each individual lower level signal needs to be derived from the irregular information rate in each channel. Generation of a suitably stable clock synchronized to the information rate is simplified if information bursts or gaps are minimized. Furthermore, elastic stores needed to smooth out the information rates are smaller when the length of information gaps is minimized.

M12 Multiplexer

An example of a higher level multiplexing format is provided in Figure 7.14. This is the format used for 6.312-Mbps DS2 signals in the North American digital hierarchy.* A DS2 signal is derived by bit interleaving four DS1 signals and adding the appropriate overhead bits.

A DS2 master frame is 1176 bits long. Of these there are 1148 information bits (287 per channel), 11 framing bits (M_0, M_1, F_0, F_1), 12 stuffing control bits (C_1, C_2, C_3, C_4), 4 S bits (S_1, S_2, S_3, S_4), and an alarm bit X. Since an S bit can be a null bit or an information bit, each channel can send 287 or 288 bits in a master frame. An S bit is designated as an information bit if all three of the corresponding C bits are 0. The S bit is a null (stuff) bit if all three corresponding C bits are 1. Obviously, this encoding procedure allows for single error correction in the stuffing control bits.

The first level of framing is established by the alternating F_0, F_1, F_0, . . . pattern. Notice that exactly 146 bits separate the F_0 and F_1 bits. Another level of framing for identifying the C and S bits is established by the M_0 and M_1 bits. A fourth M bit (X) is not used for framing and therefore can be used as an alarm service digit. Similar frame structures exist for other higher level digital signals. Figures 7.15–7.18 show the structures for DS3, DS4, DS1C, and the E2 second-level digital signal of the ITU, respectively.

Example 7.3. Determine the minimum and maximum input channel rates accommodated by an M12 multiplexer. Also determine the rate of DS1 misframes caused by an erroneous interpretation of a stuffed bit. Assume the bit error rate is 10^{-6}.

Solution. The maximum information rate per channel is determined as

$$\frac{6.312 \times 288}{1176} = 1.5458 \text{ Mbps}$$

The minimum information rate per channel is determined as

$$\frac{6.312 \times 287}{1176} = 1.5404 \text{ Mbps}$$

Since there are three possible combinations of two errors in the C bits, the probability of misinterpreting an S bit is closely approximated by $3 \times (10^{-6})^2 = 3 \times 10^{-12}$. The duration of each master frame is 1176/6.312 = 186 μsec. Thus the rate of misframes per DS1 signal is

$$\frac{3 \times 10^{-12}}{186 \times 10^{-6}} = 0.016 \times 10^{-6} \text{ misframes per second}$$

which is equivalent to one misframe every 2 years.

*DS2 signals are no longer transmitted as individual signals. They only exist as an intermediate level between 28 DS1s and a DS3 signal.

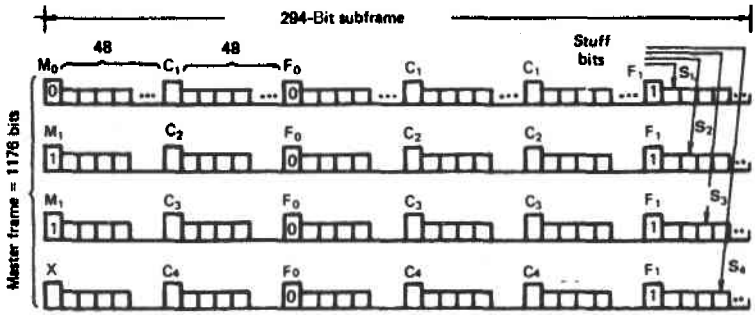


Figure 7.14 Frame format of DS2 digital signal. Stuffing occurs in channel i when the previous C_i bits = 111; X is an alarm bit that equals 1 for no-alarm condition. Framing is established by the $F_0F_1F_0$ . . . sequence with 146 intervening bits.

Example 7.3 demonstrates that the tolerance of a 1.544-MHz DS1 clock is −3.572 to +1.796 kHz. Thus the relative accuracy between the DS1 and DS2 clocks must be 1.796/1544, or only 1 part in 860. This relatively large timing tolerance is much greater than what is required for reasonable clock and line instabilities. The timing adjustment capabilities and unsymmetric tolerance range were chosen out of a desire to (1) minimize DS2 reframe times, (2) provide a line clock that is a multiple of 8 kHz, and (3) minimize waiting time jitter [20].

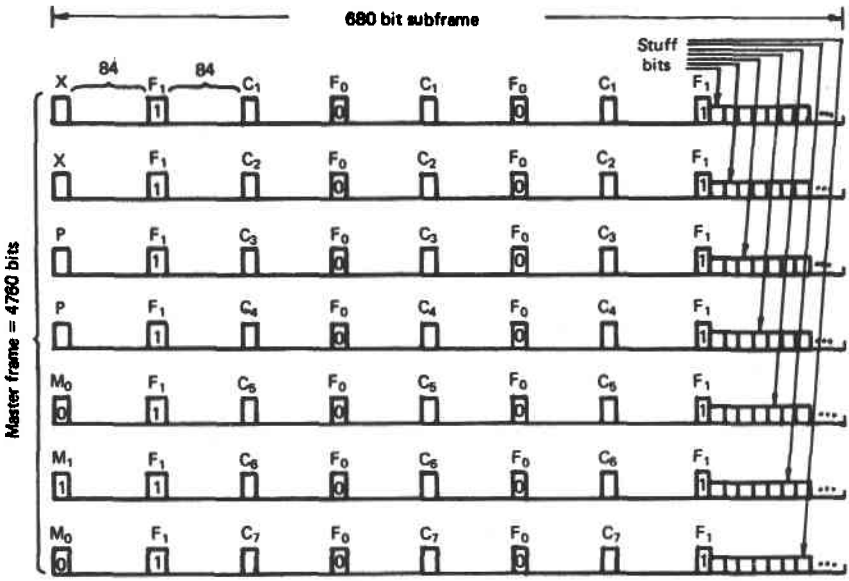


Figure 7.15 Frame format of DS3 digital signal; P is even parity over all message bits in the previous master frame. Stuffing occurs in channel i when the previous C_i bits = 111. The X (alarm) bits and the P bits must be 11 or 00 so the $M_0M_1M_0$ sequence can identify the end of the master frame.

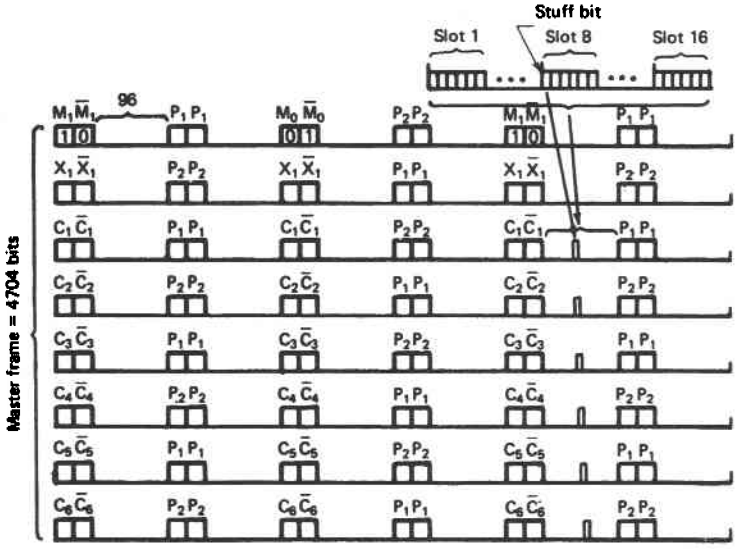


Figure 7.16 Frame format of DS4: C_i bits = 111 implies stuff the eighth message bit position for channel i following the last C_i; P_1 is even parity over the 192 previous odd-numbered message bits; P_0 is even parity over the 192 previous even-numbered message bits.

Elastic Store Size Requirements

A functional model of an M12 multiplexer is shown in Figure 7.19. Associated with each lower level (DS1) input is an elastic store to hold incoming data until it is transferred to the higher level (DS2) output. The elastic stores serve two purposes: to remove the arrival jitter of the incoming data and to hold data for the proper time slots. In addition to generating framing, the control logic of the multiplexer monitors the

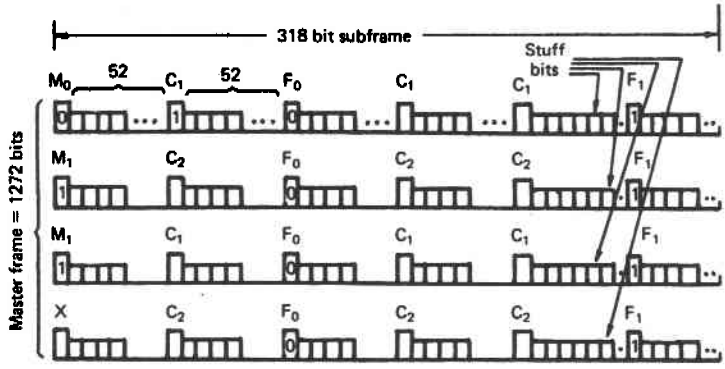


Figure 7.17 Frame format of DS1C digital signal. Stuffing occurs in channel i when the previous C_i bits = 111; X is an alarm bit that equals 1 for no-alarm condition. Framing is established by the $F_0F_1F_0$ . . . sequence with 158 intervening bits.

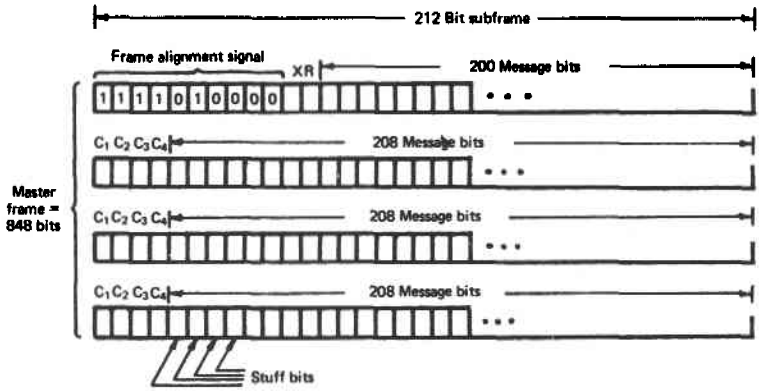


Figure 7.18 Frame format of second-level digital signal of ITU-T (E2). Stuffing occurs in channel i when the previous C_i bits are 111.

storage level (which serves as a phase comparator) of each elastic store and initiates a stuffing operation whenever the elastic store is less than half full. Conversely, no stuffing occurs when the elastic store is more than half full.

Since stuffing can occur at only certain times and only at a certain maximum rate, the elastic store must be at least as large as the peak jitter (peak phase offset) of the incoming signal. As discussed at the beginning of this chapter, jitter accumulates along the entire length of a repeatered transmission link. Thus longer links require larger elastic stores if slips are to be prevented.

As an example of the relationship between line length (number of repeaters) and number of bits of elastic storage needed by an M12 multiplexer, refer to Figure 7.20. This figure was submitted by AT&T to the CCITT special study group on jitter [21]. The analysis is an extension of the systematic jitter analysis of reference [3].

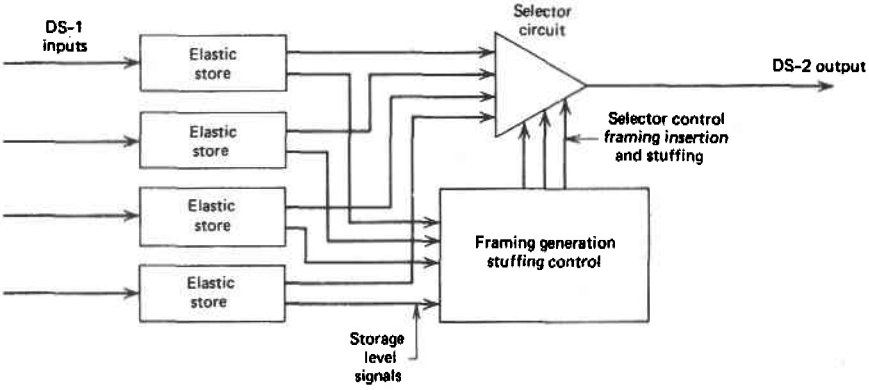


Figure 7.19 Functional diagram of a M12 multiplexer.

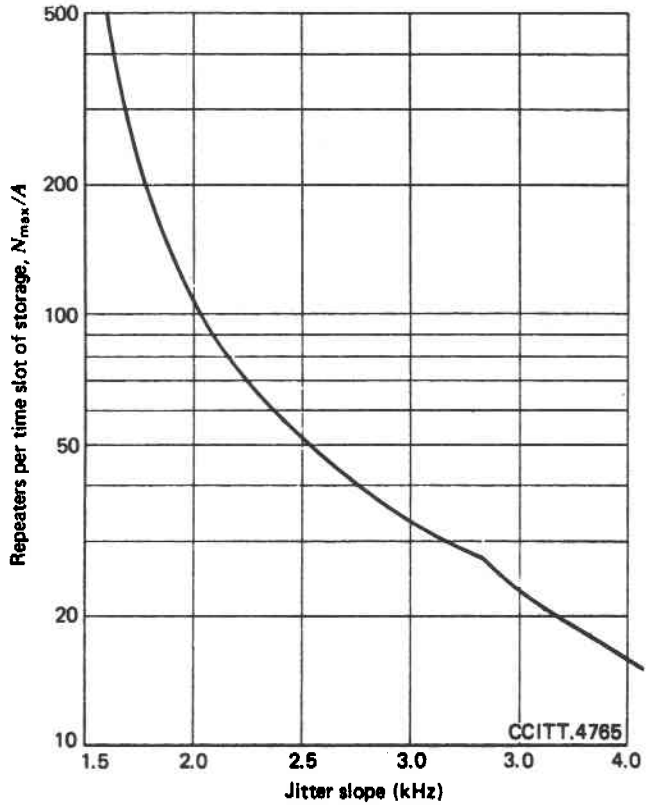


Figure 7.20 Maximum number of regenerative repeaters as a function of elastic store size and jitter slope; A = number of cells in elastic store.

The abscissa of Figure 7.20 is the maximum-phase slope produced by a clock recovery circuit in the presence of a worst-case shift in data patterns (worst-case systematic jitter). Since phase slope is nothing more than frequency offset, the required elastic store size can be determined as the maximum phase slope times its maximum duration. Since the total phase slope is proportional to the number of repeaters, Figure 7.20 displays the maximum number of repeaters per storage cell in an elastic store versus the jitter slope of an individual repeater when making worst-case timing transition.

As an example, first-generation T1 repeaters produce a worst-case slope of 2.4 kHz. M12 multiplexers allocate 5 bits of storage to input phase jitter (3 more bits are included for implementation ease and waiting time jitter). From Figure 7.20 it can be seen that the ratio of N_{max} to A is 56, which implies that $N_{max} = 56 \times 5 = 280$ repeaters.*

*The performance of long T1 lines is not as much of a concern as it once was because most long-distance DS1 circuits are now embedded in fiber links, which have much greater repeater spacing.

7.2.3 Waiting Time Jitter

When demultiplexing a higher level TDM data stream, it is necessary to generate a clock for each derived subchannel. Because the subchannels are transferred (or transmitted) as a synchronous data stream, the derived clock must be continuous. Derivation of subchannel clocks is complicated by the insertion into TDM data streams of overhead bits that create gaps in the bit arrival times. Irregularity in the data arrival rate caused by these gaps is referred to as *waiting time jitter*.

Most of the overhead bits (e.g., framing bits, parity bits, stuffing control C bits) occur on a regular and predictable basis. The waiting time jitter caused by these gaps (sometimes referred to as mapping jitter*) can be eliminated easily with an elastic store and an output clock derived from the incoming line clock. For example, a single PCM channel clock at 64 kbps can be derived from a 1.544-Mbps T1 line clock by multiplying by 8 and dividing by 193. Mapping jitter as it occurs in a E1 signal mapped into an E2 signal is depicted in Figure 7.21. In this figure the phase of the tributary data mapped into the higher level signal is shown relative to an unmapped (continuous) data clock at the same average rate. Notice that the phase of the tributary falls behind the reference during periods of data gaps but catches up during data fields because the average tributary data rate during these fields is slightly higher than the reference. Notice further that the fill level of an elastic store does not get perfectly reconciled at the end of a single E2 master frame. This result occurs because, on average, a nonintegral number of bits of a tributary are carried in a master frame.

In contrast to mapping data rate variations, waiting time jitter produced by pulse stuffing is significantly more difficult to deal with. The difficulty arises because waiting times produced by stuffed pulses are irregular and unpredictable. For this reason the subchannel output clocks derived from a pulse-stuffed TDM line must be derived independently and only from the average arrival rate of each channel's data—not from the higher level TDM rate!

Output clocks from M12 demultiplexers are generated using jitter-removing elastic stores, as shown in Figure 7.22. If large elastic stores and very slowly adjusted output clocks are used, most of the jitter can be removed. Unfortunately, waiting time jitter has frequency components down to zero frequency so the jitter (e.g., wander) can never be eliminated entirely. However, the jitter can be confined to as low a band of frequencies as desired by using a large enough elastic store. Figure 7.23 depicts time interval errors resulting from the pulse stuffing–destuffing process. The key point to notice is a full bit of offset between the source data clock and the mapped data can occur at any time but an adjustment can only be made when the next stuff opportunity occurs—hence the term waiting time jitter.

Waiting time jitter is basically a function of how often pulses are stuffed, but it is also dependent on the ratio of actual stuffs to stuffing opportunities. If the input clock

*Some references include mapping jitter to include waiting time jitter. Here, mapping jitter is used to represent repetitive data rate variations that occur when both the tributary and the higher level transport signals are respectively at their precise rates. Thus, waiting time jitter only occurs as a result of frequency adjustments.

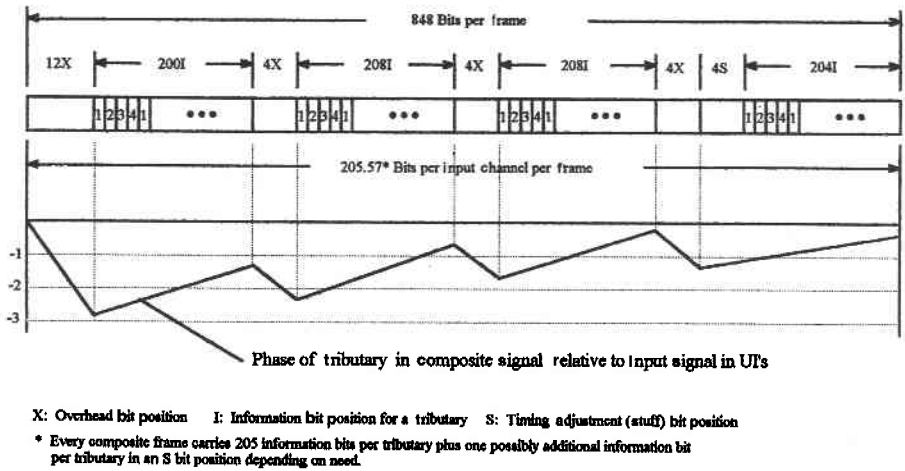


Figure 7.21 Mapping Jitter of a CCITT E1 in a CCITT E2 signal.

is jitter free, the output jitter peaks when one-half of the opportunities are used. From the point of view of maximum tolerance for clock offsets, a stuffing ratio of $\frac{1}{2}$ is ideal. To reduce the waiting time jitter, however, stuffing ratios of approximately $\frac{1}{3}$ are often used. For a thorough analysis of waiting time jitter, see reference [22]. As an example of waiting time jitter dependence on stuffing (justification) ratios, see Figure 7.24 obtained from reference [21]. The abscissa of Figure 7.24 represents the ratio of stuffs to opportunities while the ordinate is jitter power produced by a single pulse-stuffing process. The jitter power is expressed in decibels relative to one slot squared (a slot is an older term for a unit interval, U1). Curve *A* shows the output jitter pro-

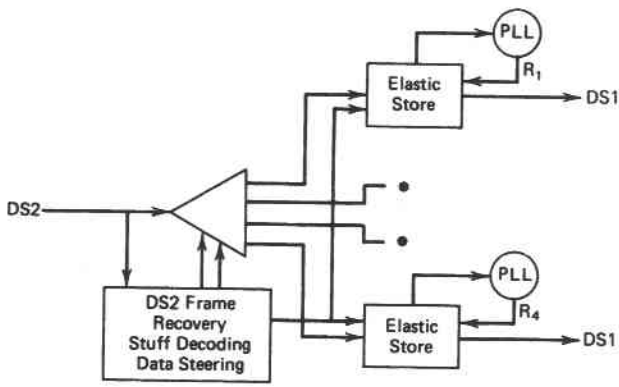


Figure 7.22 Functional diagram of M12 demultiplexer/desynchronizer.

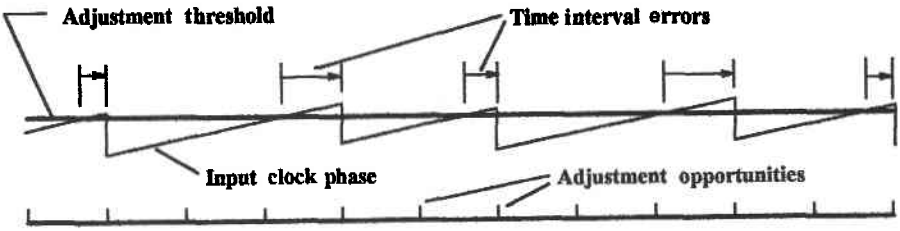


Figure 7.23 Time interval errors produced by adjustment waiting times.

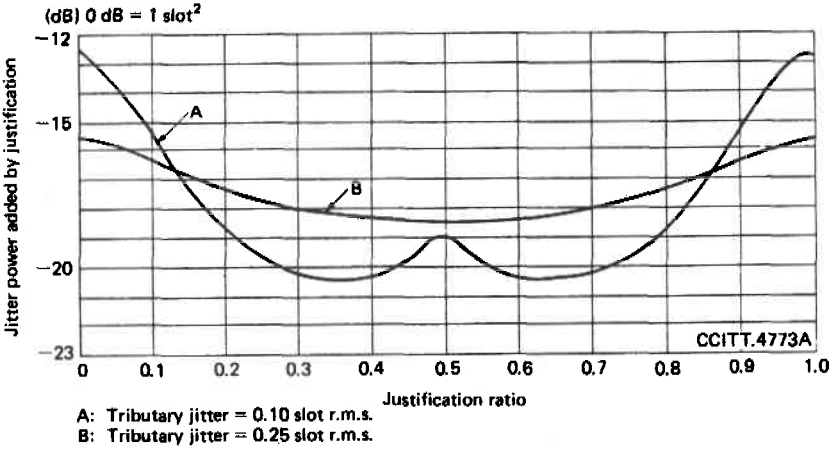


Figure 7.24 Waiting time jitter dependence on justification ratio.

duced when the input jitter is −20 dB (0.1 UI rms). Curve *B* shows the output jitter when the input jitter is −12 dB (0.25 UI rms). Figure 7.24 shows jitter produced by a single multiplexer. From measured data [22], a good order-of-magnitude estimate of the waiting time jitter accumulated by N_m tandem M_{12} stuffing–destuffing operations is

$$\sigma_w^2 = \frac{N_m}{100}(\text{UI})^2 \tag{7.6}$$

7.3 NETWORK SYNCHRONIZATION

As discussed in the preceding section, whenever a digital transmission link is connected to a digital switch, it is desirable to synchronize the two systems by having the

transmission link obtain its timing from the switch. An obvious exception to this mode of operation occurs when a digital transmission link is connected to a digital switch on both ends. Generally, a transmission link in an all-digital network derives its timing from just one of the switches to which it is connected. If the other switch is not synchronized to the first in some manner, an unsynchronized interface necessarily results. This section is concerned with network synchronization as a whole, not simply the synchronization of a single interface. Basically, network synchronization involves synchronizing the switches of the network. The transmission links can then be synchronized automatically by deriving timing directly from a switching node.

There are two basic reasons for paying assiduous attention to the timing requirements of a digital network. First, the network must prevent uncontrolled slips that could produce misframes, inadvertent disconnects, and cross connects. It is generally very difficult or very expensive to prevent slips altogether. Thus a second aspect of a network timing plan requires establishing a maximum rate of controlled slips as part of the end-to-end circuit quality objectives.

Synchronizing private networks is sometimes difficult because the network topologies are not designed with network synchronization in mind, and the switching equipment (PBXs) are not designed to provide synchronization to other nodes. Furthermore, the private networks often interface to multiple carriers in multiple locations. Determining which signals to synchronize to, particularly on a dynamic basis when a reference signal becomes unavailable, is exceptionally difficult.

There are six basic approaches used, or considered for use, in synchronizing a digital network:

1. Plesiochronous
2. Networkwide pulse stuffing
3. Mutual synchronization
4. Network master
5. Master–slave clocking
6. Packetization

7.3.1 Plesiochronous

A plesiochronous network does not synchronize the switches but merely uses highly accurate clocks at all switching nodes so the slip rate between the nodes is acceptably low. This mode of operation is the simplest to implement since it avoids distributing timing throughout the network. A plesiochronous network, however, implies that the smaller switching nodes carry the cost burden of highly accurate and redundant timing sources. As a compromise, large networks can be divided into subnetworks for timing purposes and use plesiochronous operations for inter-subnetwork synchronization and some other, more cost-effective, means of providing intra-subnetwork synchronization. As described in Section 7.5, the public telephone network in the United States uses plesiochronous synchronization at the upper levels.

Plesiochronous timing is also used to synchronize international digital network interconnections. In recommendation G.811 [23], the ITU has established the stability objectives for clocks of all international gateway digital switches. The stability objective of one part in 10^{11} implies that slips between international gateway switches will occur at a rate of one per 70 days. (This assumes one clock is positive one part in 10^{11} and another clock is negative one part in 10^{11}.)

7.3.2 Networkwide Pulse Stuffing

If all internal links and switches of a network were designed to run at nominal rates slightly higher than the nominal rates of the voice digitization processes, all voice signals could propagate through the network without slips by stuffing the information rate up to the local channel rate. None of the clocks would have to be synchronized to each other, and relatively coarse clock accuracies could be tolerated. At every interface between systems running under different clocks, however, the individual channels would have to be unstuffed from the incoming rate and stuffed up to the local or outgoing rate. In essence, the TDM links of the network would provide TDM channels through which user data flows at lower and variable rates, the differences being absorbed by internal pulse stuffing.

In contrast to pulse-stuffing operations of higher level multiplexers where all channels in a lower level digital signal are stuffed as a group, switching operations imply that each channel must be stuffed independently. The need for separate pulse-stuffing operations is illustrated in Figure 7.25, which depicts two voice signals being switched into a common TDM outlet link. Obviously, the bit rate R_3 of both output channels is identical. If the two channels originate in portions of the network running under different clocks R_1 and R_2, the pulse-stuffing adjustments must be made separately for each channel. The complexity of stuffing and unstuffing every 64-kbps channel at every network element would have been extremely expensive when the digital network began to take shape.

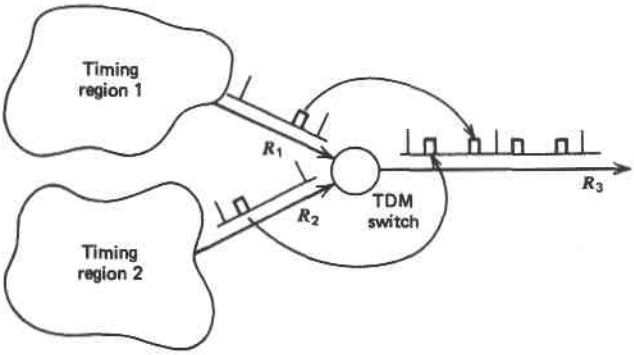


Figure 7.25 Switching two channels with different rates onto a common TDM output.

Although the availability of low-cost logic could minimize the cost aspect of networkwide pulse stuffing, other problems would occur. First, the 64-kbps clocks for speech reconstruction would be required for each channel and would contain relatively large amounts of waiting time jitter. Second, the network would no longer provide byte framing so the channel recovery process would have to also include byte framing logic. Byte boundaries of PCM data can be readily determined from statistical data patterns in the bit positions (e.g., the polarity bit), but other applications for the channels may require explicit byte boundary identifiers.

7.3.3 Mutual Synchronization

The two preceding sections discuss modes of operation for the network that do not involve synchronization of individual clocks. This section and the next two describe network timing plans that synchronize each individual clock to a common frequency. The first method, mutual synchronization, establishes a common network clock frequency by having all nodes in the network exchange frequency references as shown in Figure 7.26. Each node averages the incoming references and uses this for its local and transmitted clock. After an initialization period, the network clock normally converges to a single stable frequency. Under certain conditions, however, the averaging process can become unstable [24].

The main attractiveness of a mutually synchronized network is its ability to remain operational in spite of a clock failure in any node. The main disadvantages are the uncertainties of the exact average frequency and unknown transient behavior. Mutual synchronization has not been considered for the North American telephone network. In Great Britain, however, a hierarchical timing structure was once considered that utilized mutual synchronization within some portions of the network [25].

7.3.4 Network Master

Another method of synchronizing the network is shown in Figure 7.27. With this method a single master clock is transmitted to all nodes enabling them to lock onto a common frequency. As indicated, all network nodes are directly connected to the network master, implying the need for a separate transmission network dedicated to the

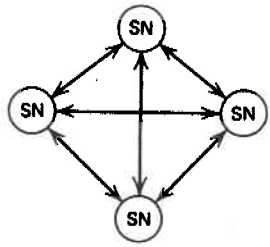


Figure 7.26 Mutual synchronization: SN, switching node.

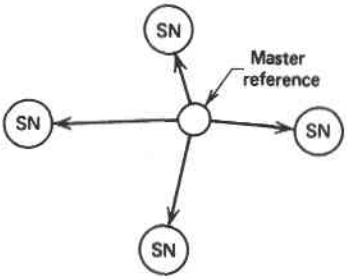


Figure 7.27 Network master synchronization: SN, switching node.

distribution of the reference. Reliability considerations also imply that alternate paths be provided to each node. Because of cost considerations for the separate timing network and reliability problems with reference distribution, a network master with direct transmission to each node is undesirable.

Something similar to a master synchronized network is evolving through the use of Global Positioning System (GPS) satellites for timing distribution to network nodes. Switching systems of telecommunications networks around the world are using GPS and other satellite systems to synchronize their switching office clocks. As the costs of the GPS receivers and suitably stable oscillators drop in cost, more and more network nodes are being timed from this master (highly accurate) source. Because CDMA digital mobile systems also lock to GPS, lower cost synchronization systems have also been developed that obtain GPS timing from the CDMA base stations in lieu of directly receiving GPS signals that often require outside antennas to receive from multiple GPS satellites simultaneously.

7.3.5 Master–Slave Synchronization

The main drawback to network master synchronization as described in the preceding section is its need for separate and reliable transmission facilities to every node. Figure 7.28 shows a network configuration that disseminates a master reference by way of the message links themselves. A network reference frequency is transmitted to a few selected higher level switching nodes. After these nodes synchronize their clocks to the reference and remove transmission link-induced timing jitter, the reference is passed on to lower level switches by way of existing digital links. The next lower level switches, in turn, synchronize to an incoming link from the higher level and pass timing on to another level of switches by way of their outgoing digital links. The process of passing the reference downward from one level to the next is referred to as "master–slave synchronization."

Since all switching nodes in the network are synchronized either directly or indirectly to the same reference, they all run at the same nominal clock rate. Thus slips should not occur. However, because of the different paths through which timing is disseminated, short-term frequency differences can occur between some nodes. If these

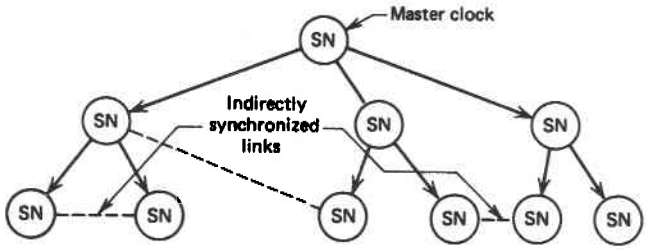


Figure 7.28 Master–slave synchronization.

nodes are synchronized indirectly, as shown in Figure 7.28, infrequent slips might occur. Furthermore, reliability considerations imply that backup clocks must be provided in all switches should the clock distribution system fail. When this happens, slips become more likely, but only after relatively stable backup clocks have had enough time to drift from the common reference frequency.

AT&T and the United States Independent Telephone Association (USITA) originally selected master–slave synchronization for the switched digital network in the United States [26]. The reference frequency was located in Hillsboro, Missouri, from which selected No. 4 ESS switching centers received their timing by way of dedicated transmission facilities. Synchronization of all other switches occurred by way of existing digital transmission links. As discussed in Section 7.4, the original plan has been changed to use plesiochronous synchronization at the highest level. Furthermore, as mentioned in the previous section, the availability of GPS timing sources is leading to more and more nodes at the higher level and, consequently, fewer and fewer slave nodes.

7.3.6 Packetization

The synchronization discussions of the five preceding sections have assumed implicitly that a synchronous, circuit-switched network was being considered, since prevailing digital voice networks operate in that manner. For completeness, however, another form of network must be mentioned—a packet-switched network.

As discussed in Chapter 10, packet-switched networks break up messages into identifiable blocks (packets or cells) of data. In between the blocks, the transmission links carry either idle codes or control messages. If all messages (control and data) are separated by a nominal interval of idle transmission, elastic stores can be reset in preparation for the next block. As long as each block is limited in length, the elastic stores can absorb clock differences and avoid losses of data. (In essence, slips occur in the idle codes.)

7.3.7 Network Timing Performance Measurements

After choosing a synchronization architecture for a network, it is necessary to be able to measure the quality of the timing signals within the network to qualify equipment,

determine performance margins, isolate faulty equipment or transmission links, and possibly evaluate design alternatives. ANSI and ITU standards use the concept of time interval error (TIE) and maximum time interval error (MTIE) for these purposes. An additional measure of clock performance is a time variance (TVAR).

Maximum Time Interval Error

A TIE is the difference in delay between a timing signal and an ideal timing signal measured at the end of a particular time period (i.e., the error at the end of a time interval). An MTIE is the maximum peak-to-peak variation in TIE values that occur within a specified time interval. These concepts are illustrated in Figure 7.29. Figure 7.29*a* displays the TIE that occurs when measuring a perfectly stable but inaccurate timing signal. Because the frequency of the signal under test differs from the reference frequency by a constant value (Δf), the TIE is directly proportional to the measurement interval S. The error is determined by counting clock cycles in both signals and expressing the difference as the time required by the test signal to catch up (or fall back) to the current reference count. Thus the TIE produced by a constant-frequency offset is

$$\begin{aligned} \text{TIE} &= \frac{\Delta n}{f} \\ &= \frac{S(f+\Delta f) - S(f)}{f} \\ &= S\left(\frac{\Delta f}{f}\right) \end{aligned} \tag{7.7}$$

where Δn is the accumulated difference in clock cycles, Δf is the frequency offset, and S is the measurement interval in seconds.

Because the TIE in Figure 7.29*a* increases monotonically with S, the MTIE and the TIE are identical. In contrast, Figure 7.29*b* depicts the TIE produced by a timing signal with no long-term frequency offset but some short-term instabilities (jitter and wander) as would occur if the reference was relayed through a network. As indicated, the TIE varies as a function of S but is bounded. The MTIE is the largest peak-to-peak difference in TIE values in a measurement interval S. The determination of MTIE with no long-term frequency offsets is also shown in Figure 7.30, where periodic samples of time interval errors are depicted.

Figure 7.29*c* displays the more general case where the signal under test contains both instability and a frequency offset. In this case the choice of the value for S is critical. If S is too small, jitter will mask the presence of the offset. If S is too large, the frequency offset dominates the TIE and MTIE measurements. Thus periodic measurements need to be recorded for a time history to fully characterize the timing imperfections.

The first three examples in Figure 7.29 assume the use of a "perfect" reference with which to make the measurements. If an "imperfect" reference is used, meaningful measurements are obtained as long as the TIE variations are significantly greater than

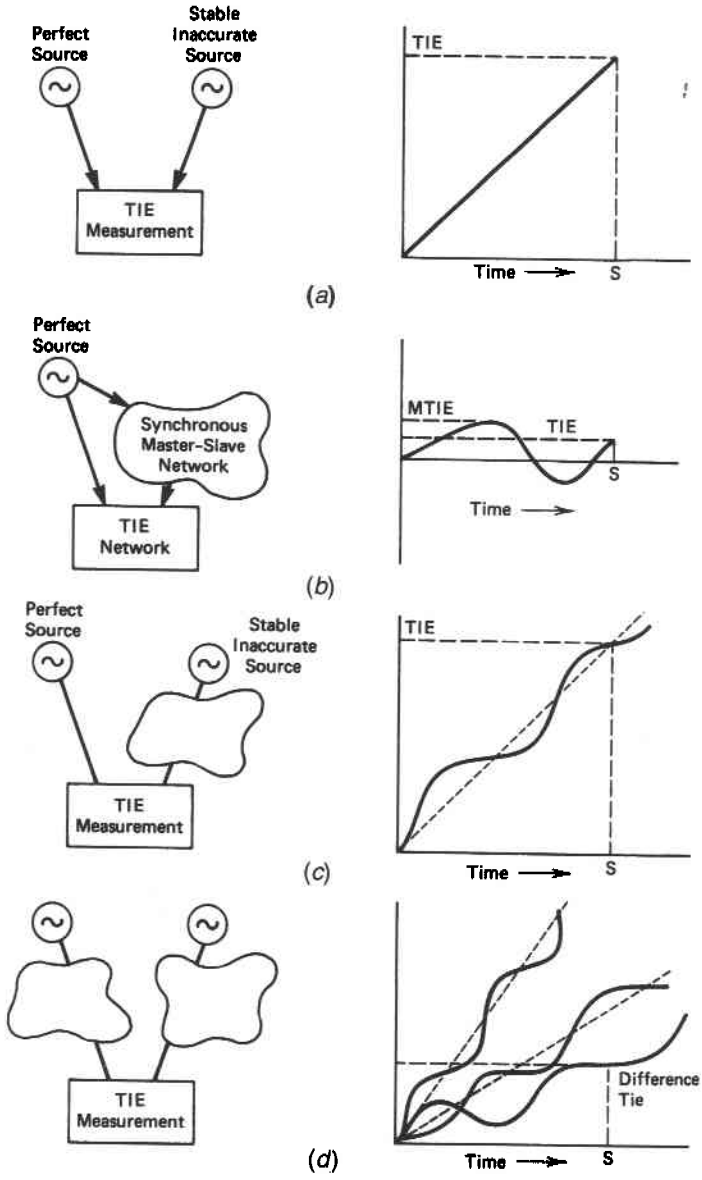


Figure 7.29 TIE and MTIE for various clock imperfections: (*a*) pure offset; (*b*) pure jitter; (*c*) offset and jitter; (*d*) TIE difference measurement.

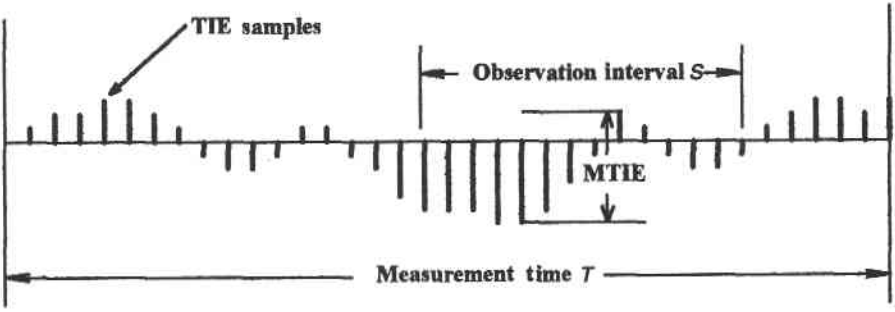


Figure 7.30 Maximum time interval error example.

the imperfections of the reference. Sometimes the relative performance between two timing signals is of interest. For example, a PBX with T1 connections to two different places (or carriers) is more dependent on the timing differences between the two transmission links than on the absolute performances. Figure 7.29*d* depicts measuring the relative performance of two such links. In this case, all measurements are meaningful, even arbitrarily low frequency wander, which can cause slips but is difficult or impossible to measure in an absolute sense. If relative measurements indicate timing problems exist, absolute measurements may be needed to isolate the problem source.

Time Variance

Neither TIE nor MTIE measurements convey any information regarding the frequency content of the jitter (other than that conveyed by the measurement interval S). A more general statistical characterization of the jitter requires representing the jitter magnitude as a function of frequency, or equivalently, as a function of time between TIE samples. The time variance (TVAR) is such a measure. TVAR values are determined as the expected variance of second-order differences between TIE samples that are separated by a time τ, where τ varies from zero (or some fraction of a second) to some maximum observation period. TVAR values are customarily measured in units of time squared (e.g., nanoseconds squared). The formula for calculating TVAR values from TIE samples $x_i (i = 1, \ldots, N)$ is

$$\begin{aligned}\mathrm{TVAR}(\tau) &= \sigma_x^2(\tau)\\ &= \tfrac{1}{6}E[(\Delta^2 x)^2]\\ &= \frac{1}{6n^2(N-3n+1)}\sum_{j=1}^{N-3n+1}\left[\sum_{k=0}^{n-1}(x_{j+2n+k} - 2x_{j+n+k} + x_{j+k})\right]^2 \qquad (7.8)\end{aligned}$$

where $\tau = n\tau_0$ (τ_0 is the sampling interval) and the observation period is $N\tau_0$.

The use of second-order differences removes the effect of a dc offset or of a linear phase ramp in the TVAR samples. Thus, there is no need to use a synchronized refer-

ence to determine the TIE samples before determining the TVAR values. However, if TVAR values for long observation periods τ are to be determined, a stable reference is needed to preclude reference wander from influencing the measurement values.

Time Deviation

Time deviation (TDEV) measurements are merely the square root of the TVAR measurements. Thus, TDEV and TVAR values have the same relationship as the classical standard deviation and variance of a probability distribution. TDEV is customarily measured in nanoseconds.

7.4 U.S. NETWORK SYNCHRONIZATION

The original plan for synchronization of the U.S. network formulated by the Bell System and the U.S. Independent Telephone Association was to use master–slave synchronization with a single master clock [26]. Due to the breakup of the network into multiple independent companies and to difficulties in reliably distributing a highly accurate master, the synchronization architecture for the United States was changed to a plesiochronous/hierarchical design in the late 1980s [13, 18, 19]. Although the hierarchical design is still in use, it is gradually changing to a "flatter" design by incorporating top-level functionality in more and more nodes.

7.4.1 Synchronization Regions

As shown in Figure 7.31, the public network is partitioned into synchronization regions that are internally synchronized with a master–slave timing hierarchy that establishes different levels of timing quality: stratum 1 to stratum 4. Stratum 1 clocks have the highest quality while stratum 4 clocks have the lowest. Timing for each region is established by a primary reference source (PRS) at stratum 1. Stratum 1 clocks are free-running clocks with inaccuracies no greater than one part in 10^{11}. Some regions may have their own PRS while others may use a synchronization signal from another party (e.g., an interexchange carrier such as AT&T).

For the most part, the synchronization regions correspond to LATAs. Every region must have at least one stratum 2 clock, which is typically associated with an access tandem switch. Toll offices within LATAs may also have stratum 2 clocks. All toll switches within AT&T network contain stratum 2 clocks [13]. Except when they use a common PRS, the synchronization regions are independently synchronized. Thus connections between the regions (using interexchange carriers like AT&T) are typically plesiochronously timed. Within a single region nodes are synchronized in a master–slave hierarchy as indicated.

The accuracy requirements of the four levels of stratum clocks are provided in Table 7.2. These accuracies pertain only to situations in which the nodes are operating in a free-running mode. Normally the nodes are synchronized to higher level clocks so the long-term accuracy is traceable back to the respective PRS. In addition to listing free-running accuracies, Table 7.2 lists accuracies that must be met during holdover

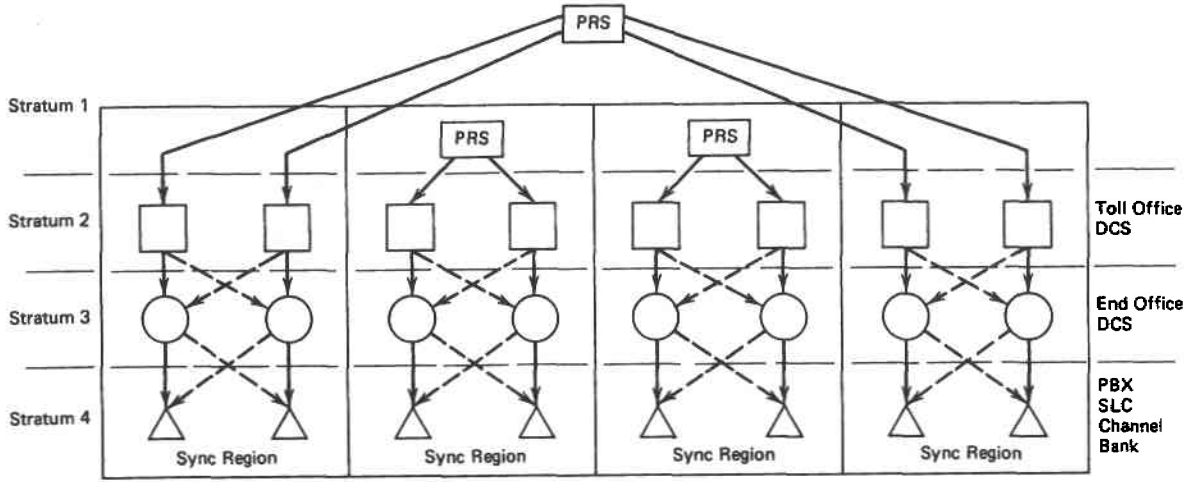


Figure 7.31 North American network synchronization.

TABLE 7.2 Stratum Clock Accuracy Requirements

Level	Free-Run Accuracy	24-Hr Holdover Accuracy
Stratum 1	$\pm 1 \times 10^{-11}$	Not applicable
Stratum 2	$\pm 1.6 \times 10^{-8}$	$\pm 1 \times 10^{-10}$ per day
Stratum 3E	$\pm 4.6 \times 10^{-6}$	$\pm 1 \times 10^{-8}$ (day 1)
Stratum 3	$\pm 4.6 \times 10^{-6}$	< 255 slips in day 1
Stratum 4	$\pm 32 \times 10^{-6}$	Not required

states. The holdover mode for stratum 2 and 3 clocks requires these nodes to gradually transition to the free-running state when they lose their reference signals.* A stratum 2 clock, for example, loses the previously established frequency at the rate of 1×10^{-10} per day, implying it would take 160 days to reach the worst-case free-running accuracy. Stratum 4 clocks, however, have no holdover requirements so they enter the free-running state immediately upon losing their reference(s).

Network synchronization requirements also specify how the various nodes respond to degradation or complete failure of their references. Stratum 2 and 3 clocks must bridge short interruptions in the reference with minimum specified time-keeping errors [19]. Stratum 3 clocks, which typically have a primary reference and a secondary reference, must perform a very gradual switchover to the secondary reference when the primary fails. Abrupt switchovers, as often occur in stratum 4 nodes (PBXs), can cause phase transients in the output clock, which can in turn disrupt synchronization in all downstream devices (e.g., within a private network).

7.4.2 Primary Reference Sources

The ANSI MTIE specification for the accuracy of a primary reference source is shown in Figure 7.32. Notice that jitter (instability that can be observed in less than 0.05 sec) is not specified. Wander, as measured over 500-sec time intervals, is limited to 3000 nsec. The asymptote for long-term timing errors corresponds to inaccuracies of 1×10^{-11}.

In addition to being designed with highly accurate (cesium beam or rubidium) clocks, all PRSs must be continuously verified with universal coordinated time (UTC). Such verification can be achieved by using a UTC-based navigation system such as Loran-C or GPS. MCI uses Loran-C and GPS to directly synchronize its PRS. AT&T uses GPS to monitor (verify) the long-term accuracy of each PRS node established in North America.† The typical accuracy of these nodes is much better than the ANSI or ITU requirement [13].

*Holdover operations are typically implemented by storing the last valid correction value (e.g., filtered phase detector output) and maintaining that value as an offset correction to a highly stable (temperature-compensated) oscillator.

†Originally, AT&T had 16 PRS nodes. Subsequent large-scale usage of GPS has led to the establishment of hundreds of PRS in the AT&T network.

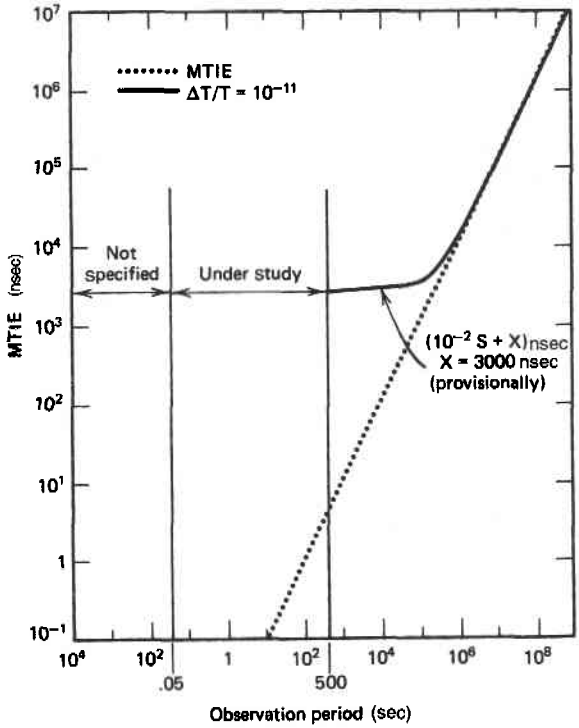


Figure 7.32 Permissible MTIE versus observation period at the output of a primary reference source.

7.4.3 1996 AT&T Synchronization Architecture

Beginning in 1996 AT&T began a move away from a two-tiered synchronization architecture to a single-level architecture with a stratum 2 clock in every office synchronized to GPS [27]. Significant features of this architecture are:

1. The synchronization distribution network has been eliminated.
2. Synchronization is independent of the network topology so the traffic-carrying network can be changed without affecting synchronization.
3. Each node is monitored by two adjacent nodes and each node monitors two adjacent nodes for performance verification.
4. Performance verification involves the use of both MTIE and TDEV.
5. DS1 timing signals are derived from SONET optical signals.

7.5 NETWORK CONTROL

The synchronization procedures described in the preceding section represent methods for controlling the timing between transmission and switching systems. In this section

synchronization is discussed in a more general sense. Instead of just time or frequency control, the synchronization concept is extended to higher level functions of connection control and network control as a whole. Fundamental to the control concept is the interaction between two processes (e.g., the exchange of information from one switching machine to another to set up or monitor a connection).

A particularly useful means of defining the interaction of two processes is a state transition diagram. The main purpose of the state transition diagram is to abstract the operational states of a process and indicate what events cause transitions from one state to another. When these events are messages (signaling tones or CCS messages) from another process, the state transition diagram effectively defines how two communicating processes interact.

7.5.1 Hierarchical Synchronization Processes

As an example of a more generalized concept of synchronization, Figure 7.33 has been included to demonstrate three distinct levels of control for a conventional telephone connection using digital transmission and multiplexing. The lowest level process shown in Figure 7.33*a* depicts nothing more than the clock synchronization process required to transmit and receive digital information. There are only two states to the process in both the transmitter and the receiver. The purpose of clock synchronization is to cause transitions between the two states in the receiver to coincide with transitions in the clocking process of the transmitter. To accomplish this, a certain amount of transmission capacity is required in the form of line code transitions.

Figure 7.33*b* depicts a higher level synchronization process involving the framing of a time division multiplexer. Both processes represent a modulo-N counter, where N is the number of channels in the TDM frame. The two processes are synchronized (framed) by utilizing some of the transmission capacity to send framing patterns. Once the receiver acquires framing, the counter in the receiver counts in synchronism with the counter in the transmitter so that individual TDM channels are properly identified.

Figure 7.33*c* provides state transition diagrams of a somewhat more complicated but easily understood process. The figure depicts the connection control of a conventional telephone call. The state transition diagram of the first process represents a subscriber placing a call (going off-hook). The second state transition diagram represents the sequence of states the control element in the local switch goes through to set up the connection.

As indicated, the process begins by the originating subscriber going off-hook and waiting for the dial tone. When the switch recognizes the off-hook signal (current flow in the line), it connects the subscriber line to a digit receiver that returns a dial tone. The subscriber then dials the address of the desired telephone and enters another wait state. Upon receiving the last digit of the address, the switch control processes the request. Once the status of the called party is determined, a busy tone or a ringback tone is returned to the originating subscriber. A busy tone prompts the subscriber to hang up (go on-hook) while a ringback signal causes the subscriber to stay in the wait state until the called party answers or until the caller "times out" and abandons the call.

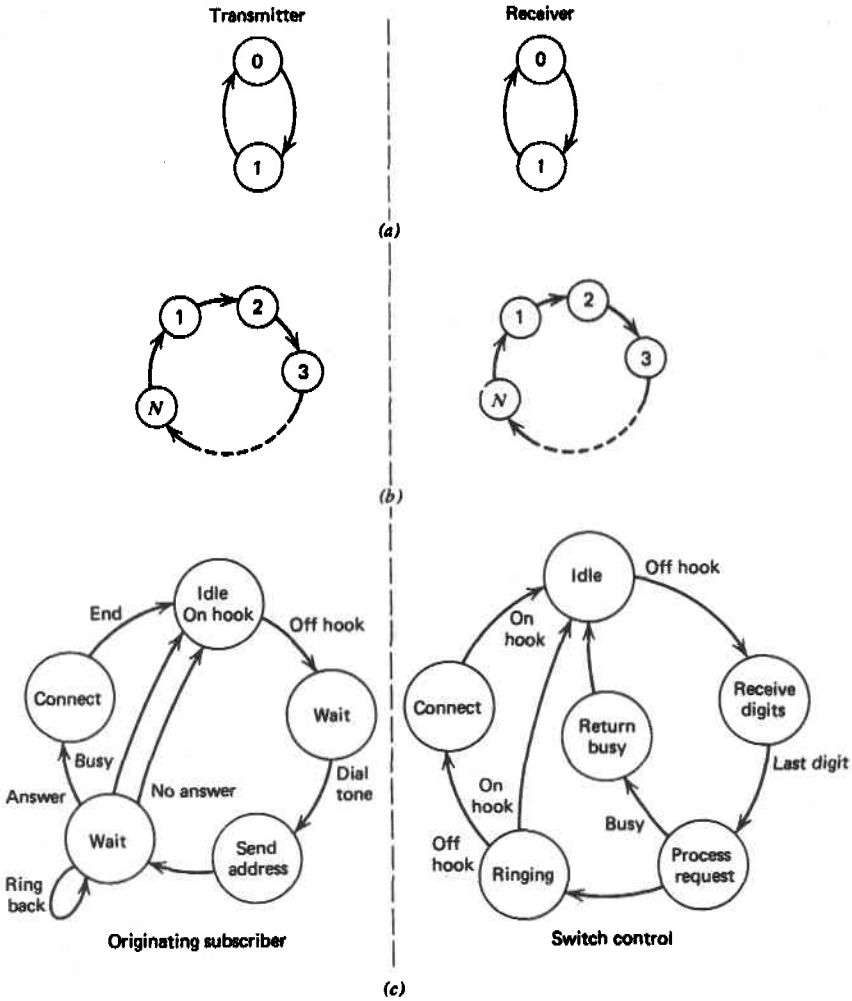


Figure 7.33 State transition diagram of synchronization processes: (*a*) timing process; (*b*) framing process; (*c*) telephone connection process.

When the called party answers, both processes enter the connected state and communication between the end users begins. The end users then get involved in yet another level of "synchronization." Voice telephone users begin by exchanging greetings and identities to establish a "connection" between their thought processes to communicate on a mutually understood subject. The message exchange process also requires synchronization so that only one person talks at a time. Hence various forms of control signals are needed to "turn the line around." Although being somewhat subtle in nature, these control signals represent transmission overhead in the same sense as control signals within the network. A talker may indicate his end of transmission

by asking a question, by inflections in the voice, by the message itself, or more commonly by a pause.

Data communications equipment goes through the same basic procedures in order to establish connections and exchange information. In this case, the procedures are defined more formally and, consequently, are more restrictive. The formal rules of communication between data communications equipment are usually referred to as a "protocol." Data communications protocols typically include a definition of certain control codes, code interpretations, message framing, turn-around procedures for half-duplex lines, error control, message sequencing, fault control, and recovery.

Automated fault control and recovery procedures for communications networks can become quite involved and difficult to implement reliably. When individual voice circuits malfunction (e.g., become noisy or disconnected), the recovery procedures are left to the users. They merely redial the connection and take up where they left off. However, large trunk groups or switching systems must be designed for higher levels of dependability and maintainability. The dependability criterion ensures that failures or malfunctions rarely occur or that they are circumvented automatically by protection switching. High levels of maintainability ensure that failures are repaired quickly when they occur. Within switching systems, most of the instructions and memory words of the processor are dedicated to hardware and software performance monitoring, recovery procedures, and maintenance diagnostics.

7.6 NETWORK MANAGEMENT

In addition to controlling individual connections and equipment, a communications network must also manage its facilities on more macroscopic levels. The basic goal of network management is to maintain efficient operations during equipment failures and traffic overloads. The main considerations are routing control and flow control.

7.6.1 Routing Control

Routing control refers to procedures that determine which paths in a network are assigned to particular connections. If possible, connections should use the most direct routes at the lowest levels of the network. The direct routes are obviously desirable because they use fewer network facilities and generally provide better transmission quality. However, economic considerations often limit the capacities of the direct routes so that alternate routes are needed to maintain suitably low blocking probabilities between one switching machine and another.

If a trunk group between two switching machines contains enough circuits to provide an acceptably low blocking probability, a significant number of the circuits in the group are idle during average traffic loads. A more economical design allocates a limited number of heavily utilized trunks in the direct route and provides alternate routes for overflow (alternately routed) traffic. In this manner the users are able to share larger portions of the network. Chapter 12 presents basic examples of how a network

can be engineered to minimize the transmission facilities while providing a given grade of service (blocking probability). As discussed in Chapter 1, the use of centralized control for the network, with common-channel signaling, provides significant efficiencies of operation in congested networks.

7.6.2 Flow Control

In the preceding section, alternate routing is discussed as one aspect of managing traffic in a communications network. Routing algorithms are concerned only with the utilization of paths or directions of travel within a network. Another requirement of network management is to control the amount of traffic in a network. Managing the rate at which traffic enters a network is referred to as flow control. A network without effective flow control procedures becomes very inefficient or ceases to function entirely when presented with excessively heavy traffic conditions.

The generalized performance of a large, uncontrolled network is shown in Figure 7.34 as a function of the offered traffic. As indicated, when light traffic conditions exist, the network carries all traffic requests presented. As the load increases, however, some of the offered traffic is rejected because no appropriate circuits are available for particular connections; that is, blocking exists. As the input load increases even further, a network with no flow control eventually begins to carry less traffic than it does when lighter loads are presented. If the offered load increases even more, the network may even cease to carry any traffic at all.

The reason that the volume of carried traffic decreases when the offered traffic exceeds some critical value is that partially completed requests tie up network resources while trying to acquire other resources tied up by other partially completed requests. Thus a form of dynamic deadlock occurs. Prior to the development of centralized network control, this situation would often arise on busy calling days (e.g., Mother's Day). In a network with distributed control all sources of traffic are serviced by successively seizing trunks to intermediate switching nodes until the destination is reached. If heavy traffic exists, requests emanating from all sides of the network en-

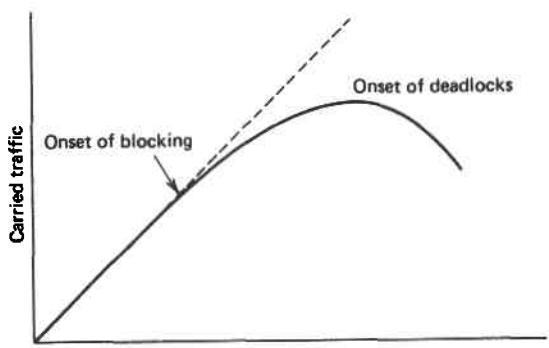


Figure 7.34 Traffic carried versus traffic offered for a network with no flow control.

counter congestion somewhere near the middle. At that time, the partially completed connections are tying up facilities needed by other requests. In the limit, when extremely heavy traffic exists, all network resources are held by partially completed requests and no complete connections can be established!

Another example of the need for flow control is automobile traffic at a metropolitan intersection. Have you ever encountered an intersection in which your lane of traffic was blocked by cross traffic backed up into the intersection? In essence, the driver blocking your direction of travel seized a common resource (the middle of the intersection) without being able to obtain the next resource required (the other side of the intersection). Bumper-to-bumper traffic in one direction can significantly degrade the throughput in other directions. With heavy traffic in all directions, total throughput can grind to a halt until the congestion is relieved from the periphery inward.

The fundamental principle demonstrated by these examples is that efficient use of the common resources of a heavily loaded network requires some form of flow control. In the automobile example, smooth operation of an intersection depends on each driver looking ahead and not entering an intersection unless he can get all the way across. A telecommunications network must use the same basic principle (it is hoped with more discipline). The control elements at the periphery of the network must be aware of the internal status of the network and control the flow of traffic from its sources.

More than one level of flow control is sometimes implemented within a network. In a data communications link, some form of flow control is required to keep a source terminal from overloading the terminal at the other end of the link. The receiving terminal uses a reverse channel (sometimes with a lower bit rate) to inform the source when to cease and when to begin transmissions. This level of flow control in a circuit-switched network involves the terminals themselves and is of no concern to the network since the traffic flows within an established connection. Of more concern to a circuit-switched network is how to control the flow of connection requests into the interior of the network. In setting up a long-distance connection, the first few circuits required should not be seized unless there is a reasonable chance that all of the circuits necessary to complete the connection can be obtained. Partially completed circuits only degrade the network capacity by increasing congestion without satisfying a service request. Network flow control is greatly simplified with common-channel signaling support for centralized network control. The following paragraphs describe basic, uncentralized flow control techniques and how centralized control simplifies their implementation.

Trunk Directionalization

The operation of trunk circuits can be classified according to two different ways of controlling seizures for particular calls. Two-way trunks can be seized at either end. One-way trunks, on the other hand, can be seized only at one end. (Notice that this has nothing to do with the direction of message transfer on established connections, which is always in both directions.) When one-way trunking is used, the trunk group is usually partitioned into one group that can be seized at one end and one group that can be

seized at the other end. Two-way trunk groups are obviously more flexible in servicing fluctuating traffic patterns, but they are more difficult to control since the possibility of simultaneous seizures (called glare) at both ends must be resolved.

A useful feature to incorporate into two-way trunks is the ability to directionalize them by marking them busy at one end and effectively creating one-way trunks. With this mechanism, a distant, overloaded switching node can be relieved of additional incoming traffic while providing it sole access to the trunk group for outgoing traffic. Thus the overloaded node relieves its congestion while inhibiting new arrivals.

When the network as a whole experiences heavy traffic loads, trunk directionalization can be used to reduce the flow of connect requests into the interior of the network while establishing one-way trunks from the interior to the periphery. Thus connect requests that manage to get to the interior have a much better chance of obtaining all facilities required for the complete connection.

Cancellation of Alternate Routing

Alternate routing of traffic accommodates localized overloads by transferring traffic to underutilized routes. During networkwide overloads, however, alternate routing is undesirable for two reasons. First, alternate routes imply that a greater number of transmission and switching facilities are needed for a connection. If these same facilities could be assigned to two or more direct connections, the total number of links per call could be reduced and the network could carry more traffic.

Second, the probability that an alternately routed call can acquire all the necessary resources is relatively low. Trying to set up a connection with a large number of facilities is undesirable if the probability of getting all of the facilities is low (particularly so, if some facilities are fruitlessly tied up while less demanding requests are pending).

Code Blocking

Code blocking refers to artificially blocking calls intended for specific destination codes. If the calls are blocked at originating end offices before they acquire internal network facilities, the destinations are relieved of incoming traffic without tying up facilities that may be needed for outgoing requests from the specified areas.

The method of flow control is particularly useful in times of natural disasters, which typically stimulate large numbers of calls both into and out of the area of the disaster. In these events a network control center can initiate code blocking for all, or a large percentage, of the calls into the area. The principle of giving preference to the outgoing calls serves two purposes. First, no network facilities are seized unless there is a reasonable chance of obtaining all facilities necessary. It is the trunks into or out of the disaster area that are the focal point of network congestion. Once one of these trunks is seized, the rest of the connection can probably be established. Second, code blocking is useful because outgoing calls are probably more important than incoming calls.

Centralized Connection Control

All of the flow control procedures described previously are designed to eliminate seizures of common resources if the desired connection has a low completion probability.

Because of the distributed nature of network control implied by these operations, these control procedures are necessarily probabilistic. To maintain a certain amount of network efficiency, the network is purposely operated at less than maximum capacity.

A more desirable mode of operation, from a throughput point of view, is to allocate network facilities from a single centralized control node. Since this central node has access to the status of all network resources, no facilities are assigned to a particular request unless all facilities needed for the complete connection are available. Network transmission links are assigned in a manner that is analogous to the assignment of internal links of common control switches.

Complete centralized control of a network as large as the public telephone network is obviously infeasible from the point of view of maintaining status of all interstage links within end office switches and from the point of view of survivability of the network when the control node fails. However, many aspects of centralized control have been implemented in North America and around the world with common-channel signaling (CCS). For example, INWATS call requests are routed to a central node that determines if the destination is busy or not. If the destination is busy, the originating end office is instructed to return the busy tone without any of the internal transmission links ever being seized. This mode of operation is particularly useful for 800 numbers (INWATS) that occasionally experience very heavy traffic flow because of national television announcements.

Without CCS, the previous mode of operation was to return a busy tone all the way through the network from the place at which the busy circuit or subscriber is located. Thus the path through the network was tied up during the time the busy tone was being returned. CCS allows the originating office to return the busy tone so internal network facilities can be released and reassigned immediately upon detecting the busy condition.

REFERENCES

1 M. Decina and U. deJulio, "International Activities on Network Synchronization for Digital Communication," *IEEE International Communications Conference*, 1979.

2 J. R. Pierce, "Synchronizing Digital Networks," *Bell System Technical Journal*, Mar. 1969, pp. 615–636.

3 C. J. Byrne, B. J. Karafin, and D. B. Robinson, "Systematic Jitter in a Chain of Digital Repeaters," *Bell System Technical Journal*, Nov. 1963, pp. 2679–2714.

4 "Digitally Dejittering Laser Disc Players," *IEEE Spectrum*, Feb. 1990, p. 14.

5 F. M. Gardner, *Phaselock Techniques*, 2nd ed., Wiley, New York, 1979.

6 E. D. Sunde, "Self-Timing Regenerative Repeaters," *Bell System Technical Journal*, July 1957, pp. 891–938.

7 D. L. Duttweiler, "The Jitter Performance of Phase-Locked Loops Extracting Timing from Baseband Data Waveforms," *Bell System Technical Journal*, Jan. 1976, pp. 37–58.

8 "Carrier-to-Customer Installation-DSI Metallic Interface," ANSI T1. 403–1989, American National Standards Institute, New York, 1989.

9 H. Meyer, L. Popken, and H. R. Mueller, "Synchronization Failures in a Chain of PLL Synchronizers," *IEEE Transactions on Communications*, May 1986, pp. 436–445.

10 P. R. Trischitta and E. L. Varma, *Jitter in Digital Transmission Systems*, Artech House, Norwood, MA, 1989.

11 P. R. Trischitta and P. Sannuti, "The Accumulation of Pattern-Dependent Jitter for a Chain of Fiber Optic Regenerators," *IEEE Transactions on Communications*, June 1988, pp. 761–765.

12 J. F. Boyle, J. R. Colton, C. L. Dammann, B. J. Karafin, and H. Mann, "No. 4ESS: Transmission/Switching Interfaces and Toll Terminal Equipment," *Bell System Technical Journal*, Sept. 1977, pp. 1057–1097.

13 J. E. Abate, E. W. Butterline, R. A. Carley, P. Greendyk, A. M. Montenegro, C. D. Near, S. H. Richman, and G. P. Zampetti, "AT&T's New Approach to the Synchronization of Telecommunication Networks," *IEEE Communications Magazine*, April 1989, pp. 35–45.

14 H. Drucker and A. C. Morton, "The Effects of Slips on Data Modems," *IEEE International Conference on Communications*, 1987, pp. 12.4.1–12.4.3.

15 J. F. Ingle, "Identification of Digital Impairments in a Voiceband Channel," *IEEE International Conference on Communications*, 1989, pp. 12.3.1–12.3.5.

16 J. E. Abate and H. Drucker, "The Effect of Slips on Facsimile Transmission," *IEEE International Conference on Communications*, 1988, pp. 32.3.1–32.3.4.

17 J. E. Abate, L. H. Brandenburg, J. C. Lawson, and W. L. Ross, "The Switched Digital Network Plan," *Bell System Technical Journal*, Sept. 1977, pp. 1297–1320.

18 "Digital Synchronization Network Plan," Bellcore TA-NPL000436, Issue 1, Bell Communications Research, Morristown, NJ, Nov. 1986.

19 "Synchronization Interface Standards for Digital Networks," ANSI T1. 101–1987, American National Standards Institute, New York, 1987.

20 Technical Staff, Bell Telephone Laboratories, *Transmission Systems for Communications*, Bell Telephone Laboratories, Winston-Salem, NC, 1971.

21 "Impact of Jitter on the Second Order Digital Multiplex at 6312 kbit/s," AT&T Submittal to CCITT study group on jitter, *Green Book*, Vol. 3, pp. 861–869.

22 D. L. Duttweiler, "Waiting Time Jitter," *Bell System Technical Journal*, Jan. 1972, pp. 165–207.

23 "Plesiochronous Operation of International Digital Links," CCITT Recommendation G.811, *Orange Book*, Geneva, Switzerland, 1976.

24 J. P. Moreland, "Performance of a System of Mutually Synchronized Clocks," *Bell System Technical Journal*, Sept. 1971, pp. 2449–2464.

25 P. A. Mitchell and R. A. Boutler, "Synchronization of the Digital Network in the United Kingdom," *IEEE International Conferences on Communications*, 1979, pp.11.2.1–11.2.4.

26 C. A. Cooper, "Synchronization for Telecommunications in a Switched Digital Network," *IEEE Transactions on Communications*, July 1979, pp. 1028–1033.

27 C. Olszewski, "AT&T Synchronization Network Architecture and Operations," Presentation to 1999 NIST-TIXI Workshop on Synchronization in Telecommunications Systems, Boulder, CO, March 9–11, 1999.

PROBLEMS

7.1 Determine the size of an elastic store needed to accommodate a velocity shift of ±1000 km/hr that lasts for 10 sec if the data rate is 10 Mbps. (The speed of light is 3×10^8 m/sec.)

7.2 How many bits are needed in an elastic store designed to interface the E1 digital signal of ITU-T to a digital switch? What is the maximum slip rate if the line clock and switch clock differ by +50 to −50 ppm (the maximum recommended offsets)?

7.3 What is the maximum (99% probability) phase offset (in signal intervals) produced by a jitter power of +10 dB relative to 1 rad^2?

7.4 Determine the rate at which DS1 signals in a DS2 multiplex lose framing because stuff codes of a DS2 signal are incorrectly interpreted. Assume the channel BER is 10^{-3}. Assume the BER is 10^{-6}.

7.5 What would the rate of incorrect DS2 stuff code occurrences be if a 5-bit stuff code were used instead of the 3-bit stuff code? Assume the channel BER is 10^{-3}.

7.6 A digital transmission link is to be used to transmit blocks (packets) of data without slips. If the transmission link is autonomously timed with respect to the receiving terminal, what is the maximum allowable block length if the clocks vary by ±50 ppm each and an elastic store of 16 bits exists in the receiving terminal? Assume the elastic store is initialized to half full between each block.

7.7 Assume the systematic jitter from a single repeater produces a symmetric, worst case phase slope of 300 rad/sec for 1 msec. What is the peak-to-peak jitter in decibels relative to a unit interval at the end of a line with 200 such repeaters?

7.8 Repeat Example 7.3 for DS2 signals in a DS3.

7.9 A jitter power of 20 dB relative to 1 rad^2 is observed at the receiving end of a digital microwave link. What is the probability that the phase offset will exceed ±4.0 symbol intervals?

7.10 Determine the amount of phase shift injected into a 2400-Hz carrier signal by a slip of one PCM sample.

7.11 Determine the TIE and the MTIE at the end of a 10-sec interval produced by a DS3 signal that has a constant offset of one part in 10^6.

7.12 Repeat Problem 7.11 but assume a 1-MHz jitter component is added that has a peak-to-peak amplitude of 8 UIs. Assume the starting phase of the jitter component is 0°.

7.13 Compare the slip rate of stratum 3E with stratum 3 clocks in holdover conditions.

7.14 What is the oscillator accuracy (in ppm) implied by the holdover requirement of a stratum 3 clock?

8

FIBER OPTIC TRANSMISSION SYSTEMS

The evolution of the common equipment portions of the public telephone network from analog technology to digital technology began and ended with digital transmission systems: T1 systems for short-haul trunks and fiber optic systems for long-haul trunks. If fiber optic technology had not emerged, extensive digitization of the long-distance network would not have occurred as rapidly. High-bandwidth coaxial cable transmission systems are too expensive to universally replace radio systems, and analog radios are more efficient than digital radios in terms of voice circuits per previously available bandwidth. Even though high-density digital modulation techniques and sophisticated voice compression algorithms can make up for the bandwidth inefficiencies of digitized voice, any ultimate advantage of digital radios over analog radios is not nearly great enough to warrant wholesale replacement of the analog radio backbone. Fiber transmission systems, on the other hand, provide such dramatic savings in equipment and operational costs that wholesale deployment of high-density routes with fiber systems was economically justified.

As a note of historical interest, Figure 8.1 shows a diagram of a "photophone," patented by Alexander Graham Bell in 1880. Bell developed several techniques for modulating a light source (the sun) so as to directly vary the resistance of a selenium cell detector and thereby create an analog electrical current for the speaker. The modulation mechanism shown in Figure 8.1 is a mirror that is vibrated by acoustic energy to deflect more or less light to the receiver. It was almost exactly 100 years later that a suitable transmission medium and suitable sources made optical transmission practical.

The particular characteristics of optical fibers that make them so useful for transmission systems are low loss, high bandwidth, small physical cross section, EMI immunity, and security.

Fiber Attenuation

As a result of the invention of the laser in 1960, materials scientists began searching for optical transmission media that could be usefully applied as a communication system. The first instance of a practical optical fiber was announced 10 years later in 1970

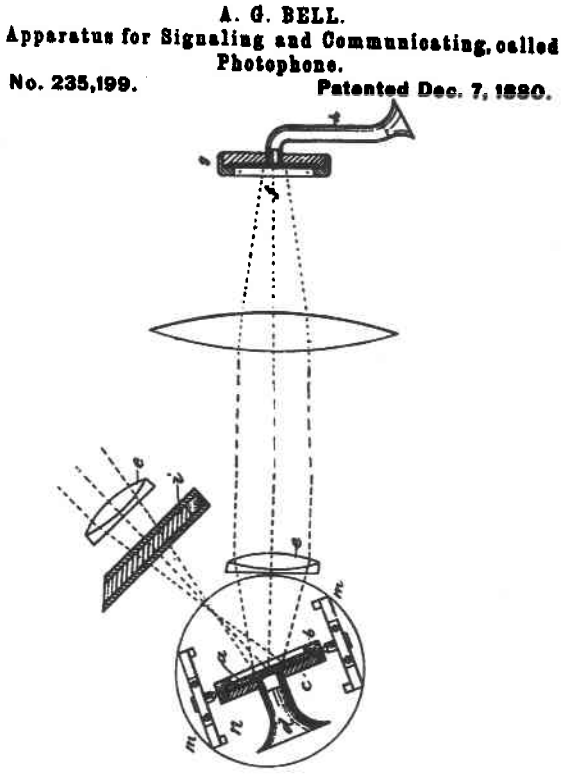


Figure 8.1 Alexander Graham Bell's photophone.

[1]. This announcement described a silica-based fiber with "only" 20 dB/km attenuation. In just a little less than 10 more years, commercially viable optical fibers with 0.2 dB/km attenuation had been developed [2]. Such remarkably low attenuation attracted immediate attention because it meant intercity transmission links could be traversed with very few repeaters, implying a dramatic savings in equipment and maintenance.

As shown in Figure 8.2, the attenuation of an optical fiber is very dependent on the wavelength of the light signal in use. Two wavelengths that experience particularly low attenuation in contemporary fibers are 1300 and 1550 nm. Representative attenuations at these wavelengths are 0.35 and 0.2 dB/km, respectively. With an easily achievable net loss of 20 dB the latter attenuation value allows 100 km between repeaters or amplifiers. Because 0.2 dB/km is close to the minimum theoretical attenuation at 1550 nm, additional reductions in attenuation cannot be realized without going to higher wavelengths. If fibers and associated electronics can be developed to operate at higher wavelengths, much lower levels of attenuation may be achieved in the future. The main application of such systems would be in submarine cables where the avoidance of repeaters is most desirable.

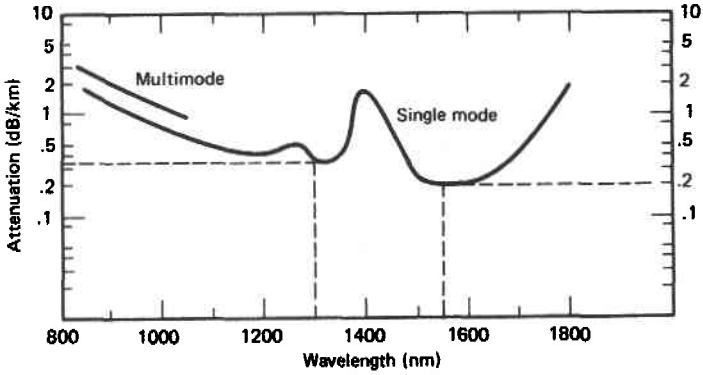


Figure 8.2 Fiber attenuation versus wavelength.

Bandwidth

When compared to electromagnetic transmission media, the bandwidth of an optical fiber is mind boggling: a single optical fiber operating at 1300 or 1550 nm of wavelength has a potential bandwidth of 20 THz (20×10^{12} Hz), which is enough for 312 million 64-kbps channels. Bandwidth limits of fiber optic transmission systems are mostly determined by the electro-optic drivers and receivers or the electronic interfaces to these devices. As described in Section 8.1.1, multimode fibers have an inherent distance-dependent bandwidth limitation, but this particular limitation is avoided in single-mode fibers. Single-mode fiber systems also have a distance-dependent bandwidth limitation, but this limitation is as much a limitation of the optical sources as it is a limitation of the fiber.

Small Physical Cross Section

One of the most beneficial features of T1 systems is their ability to relieve overcrowded cable ducts with a 12-to-1 savings of copper pairs compared to two-wire analog transmission. Fiber systems provide the same advantage with respect to T1 systems, only to a greater degree. A 25-pair copper cable has approximately the same cross section as a fiber cable with 24 fibers and some copper pairs for line-powered repeaters or amplifiers. The latter can easily carry 100,000 voice circuits. Furthermore, fiber systems can sometimes be upgraded for more capacity by merely installing higher speed electronics or wavelength division multiplexing (WDM). Thus, after copper cable is replaced by a fiber cable, duct congestion is unlikely to reappear unless extensive use of broadband services greatly expands the need for fiber facilities.

Electromagnetic Interference Immunity

Because glass has a very high resistance to electricity, the fiber itself cannot pick up interfering signals or noise or propagate harmful or damaging transients to personnel and equipment at the endpoints. The immunity to interference also means crosstalk is

not a problem between multiple-fiber cables. In systems utilizing line-powered repeaters, however, some amount of copper is included in the cable to carry power, implying that the immunity, particularly with respect to hazardous voltages, is compromised.

The immunity of fibers to external interference such as noise and crosstalk implies there is no crosstalk-induced limit from high powered transmitters or more sensitive receivers. Receiver sensitivity is ultimately limited by internal noise in the photodetectors, and transmit power has technology limits due to spectrum spreading in the sources and nonlinearities in the fiber. Until these limits are reached, however, the absence of crosstalk between fibers allows the capacity or transmission distance of a fiber system to be increased by merely upgrading the electronics, not the fiber. Although crosstalk from one fiber to another is nonexistent, crosstalk can arise between separate channels within WDM channels of a single fiber.

Security

Because optical fibers radiate no energy, noninvasive eavesdropping of the signal is impossible. Furthermore, invasive taps are more difficult to implement than are wireline taps, which merely require bridging across the conductors with a sufficiently high impedance to remove a usable but unnoticeable amount of signal energy. A similar process is possible with optical fibers, but it requires bending the fiber a very precise amount to allow a small amount of energy to escape and be amplified by a tap. Not only does this process allow passively tapping a fiber, but also signals can be injected into the fiber through such a bend. This technique has been used as a means of locally testing the effectiveness of a fiber splice and has been considered as a means of implementing distributed, passive taps for fiber distribution to the home [3].

8.1 FIBER OPTIC TRANSMISSION SYSTEM ELEMENTS

As shown in Figure 8.3, the basic elements of a fiber optic transmission system are the electrical-to-optical transducer in the transmitting terminal, the optical fiber itself, the optical-to-electrical transducer in the receiver, and signal processing circuitry for amplification, clock recovery, and data detection. Regenerative repeaters require optical-to-electrical conversion for the signal processing functions and conversion back to optical for transmission. Direct optical amplification with erbium-doped fiber amplifiers (EDFAs) are used in lieu of repeaters, particularly for WDM transmission links.

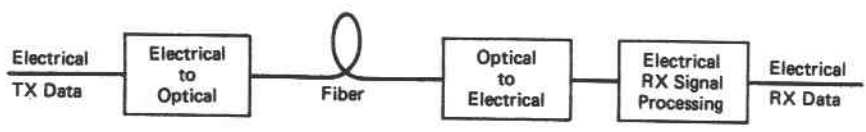


Figure 8.3 Fiber optic transmission system elements.

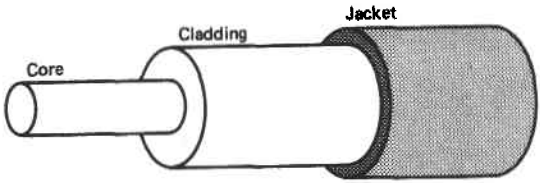


Figure 8.4 Optical fiber construction.

The system depicted in Figure 8.3 is not necessarily a digital one. Although all of the fiber applications for telephony have utilized digital signals, analog transmission is possible. Analog FM modulation of optical signals has been successfully applied to feeder applications of CATV systems [4]. Furthermore, systems that carry 80 channels of AM video on fiber to residences have been developed, a capability that allows analog television sets to receive fiber-based cable TV without a digital video decoder.

8.1.1 Optical Fiber Fundamentals

The basic construction of an optical fiber is shown in Figure 8.4. Both the core and the cladding are transparent to the desired light signal but the cladding is designed with a lower index of refraction, which causes most light waves in the core to be reflected back into the core. As shown in Figure 8.5, whether or not a ray is reflected back into the core is dependent on the angle at which it strikes the core–cladding boundary. If the angle is too sharp, the ray is not reflected but passes through the cladding and is absorbed by the opaque, protective jacket. The sharp angles can occur at two places: (1) near the source where all of the source's output is not focused into the center of the fiber and (2) at bends, splices, or other imperfections in the fiber.

Multimode Fibers

Although it is usually desirable to deliver as much signal energy as possible to the receiver, waves that reflect back and forth before reaching the receiver may be undesirable if they experience too much delay with respect to the primary ray traveling down the center of the fiber. An optical fiber that allows rays to arrive at the receiver via multiple paths is referred to as a multimode fiber. Multimode fibers have core diameters

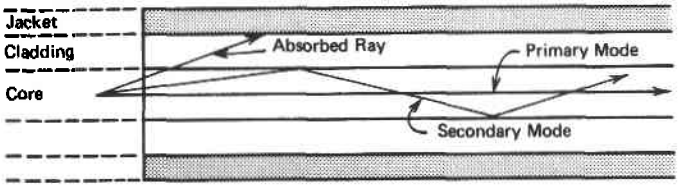


Figure 8.5 Fiber with multimode propagation.

that are large compared to the wavelength of the signal. A typical multimode fiber will have a 50-μm core diameter and a 125-μm cladding diameter (such a fiber is designated as a 50/125 fiber). The delayed rays cause pulse spreading referred to as multimode dispersion. The significance of the spreading depends on the width of the pulse or, conversely, the data rate being used. Multimode dispersion therefore creates an inherent operational limit defined as a bandwidth–distance product (BDP). The BDP of a typical step-index multimode fiber is 13 Mbps-km [5].

The previously described multimode fiber is referred to as a step-index fiber because the index of refraction in the core is constant with a step change in the index occurring at the core–cladding boundary. Multimode dispersion can be significantly reduced by varying the index of refraction within the core so that a high value occurs in the center and a low value occurs at the edge. Because the speed of propagation of light is higher in lower indices of refraction, rays that reflect back and forth within the core travel at an average speed that is greater than a primary ray that remains entirely within the center. Thus, if the index of refraction is carefully graded within the core, all rays can be made to arrive at the receiver with the same amount of delay. Such a fiber is referred to as a graded-index fiber. A representative BDP of a graded-index fiber is 2 Gbps-km [6]. Notice that this is an improvement of more than two orders of magnitude over typical step-index multimode fibers.

Example 8.1. Determine the loss limit and the multimode dispersion limit of a graded-index multimode fiber system operating at 820 nm and providing a bandwidth of 90 Mbps (enough to carry two DS3 signals). Assume that the difference between the available output power from the source and the input power required by the receiver for an acceptable maximum error rate is 42 dB.

Solution. From Figure 8.2, the attenuation of a multimode fiber operating at 820 nm is approximately 3 dB/km. Thus,

$$\text{Loss limit} = \frac{42}{3.0} = 14 \text{ km}$$

Using 2 Gbps-km as a typical BDP of a graded-index multimode fiber, the multimode dispersion distance limit is determined as

$$\text{Dispersion limit} = \frac{2000}{90} = 22.2 \text{ km}$$

The results of Example 8.1 are representative capabilities of first-generation fiber systems such as the FT3C system of AT&T [7]. In actual practice, repeater spacing would be less than the 14-km loss limit to allow margin for, for example, component tolerances, splicing, and aging. The first FT3C system, which began service between New York City and Washington, DC in February 1983, had repeater spac-

ing of 7 km—the location of repeaters in a coaxial cable system replaced by the fiber system.

Single-Mode Fibers

As the first-generation fiber systems were being brought into service in the long-distance network, optical technology had already advanced to the point that deployment of second-generation systems was underway. One key technology of the second generation was the use of single-mode fibers (SMFs) that provide two distinct advantages. First, SMFs have smaller diameter cores (8 μm typically) that restrict propagation to a fundamental mode and therefore eliminate multimode dispersion. Second, SMFs have less internal (Rayleigh) scattering, which reduces the attenuation by about 50% [6, 8]. For example, the attenuation of an SMF operating at 820 nm is about 2 dB/km, as opposed to 3 dB/km for a multimode fiber.

Another key development of second-generation systems involved the use of longer wavelengths (1300 nm), which, as shown in Figure 8.2, experience significantly less attenuation than the 800–900-nm wavelengths used in first-generation systems. Use of these wavelengths required new technology for sources and receivers, as described in Sections 8.1.2 and 8.1.3.

Chromatic Dispersion

The combination of eliminating multimode dispersion and using a wavelength with much lower signal attenuation revealed another transmission limitation referred to as chromatic dispersion. Chromatic dispersion arises when a photonic signal contains more than one wavelength and the individual wavelengths propagate at different speeds. Thus, chromatic dispersion is the photonic equivalent of phase distortion (also called delay distortion) in electromagnetic (wireline) propagation. Dispersion levels are quantified by a dispersion coefficient of picoseconds per nanometer per kilometer. A representative value for the SMF at 1550 nanometers is 16 psec/nm km.

The effect of chromatic dispersion is minimized by using one or more techniques. First, operation can be centered about a wavelength that exhibits a small amount of chromatic dispersion. Typical silica fibers, for example, produce approximately 15 times less chromatic dispersion at 1300 nm than at 1550 nm. Second, an optical source should be chosen that is as spectrally pure as possible (e.g., has a narrow spread of wavelengths). Third, narrow return-to-zero (RZ) pulses can be used to prevent intersymbol interference. Fourth, a dispersion-compensating fiber can be included in the transmission path that has a dispersion slope in opposition to the main fiber. When just the first two techniques are used, BDPs of 250 GHz-km are possible [6].

Example 8.2. Determine the loss limit and the chromatic dispersion limit of a high-performance SMF optical transmission system operating at 1300 nm and providing a bandwidth of 417 Mbps (enough to carry nine DS3 signals). Assume a narrowband source is used with an output power that is 42 dB greater than the minimum acceptable receive power (i.e., assume the system gain to be 42 dB).

Solution. As indicated in Figure 8.2, the attenuation of a single-mode fiber operating at 1300 nm is approximately 0.35 dB/km. Thus,

$$\text{Loss limit} = \frac{42}{0.35} = 120 \text{ km}$$

Using 250 Gbps-km as the BDP of a silica, single-mode fiber,

$$\text{Chromatic dispersion limit} = \frac{250{,}000}{417} = 600 \text{ km}$$

When the dispersion limit is much higher than the loss limit, as is the case in Example 8.2, it indicates that the optical sources are providing a narrower spectrum than is necessary. Thus, the system design could be changed to use less expensive sources or the power output of the source could be increased to get a longer loss limit. (Increasing the power output of the source may widen its spectrum and thereby reduce the BDP.)

The system parameters of Example 8.2 are representative of the second-generation FT series G systems deployed by AT&T [9]. The FT series G systems had maximum repeater spacing of 48 km (29 miles), indicating that the initial systems could be installed with significant margin for reliability and future upgrades. Many of the routes were, in fact, upgraded to 1.7 Gbps without changing the fiber [10, 11]. Even at the higher data rates, the FT series G routes had significant loss margin for splicing and the insertion of wavelength division multiplexers as another means of upgrading the system. Wavelength division multiplexing is described in Section 8.3.

Example 8.2 shows that attenuation is the limiting distance factor in 1300-nm systems operating at low and intermediate data rates. To operate with even longer distances, Figure 8.2 indicates that 1550 nm should be used. With typical silica fibers, the limiting factor at 1550 nm would be chromatic dispersion. To overcome this limitation, two other types of fiber have been developed: a dispersion-shifted single-mode fiber (DS-SMF) that has minimal dispersion at 1550 nm and a dispersion-flattened single-mode fiber (DF-SMF) that has less dispersion across a range of wavelengths. Figure 8.6 contrasts the dispersion of both of these types of fibers with a conventional SMF. Even though Figure 8.6 indicates zero dispersion exists at 1300 nm on the SMF and at 1550 nm on the DS-SMF, some amount of dispersion has to be assumed due to fiber manufacturing variations and uncertainty of the actual center wavelength produced by a source. Values of 1–3.5 psec/km nm are typically used as the dispersion coefficient at the "zero" dispersion points.

8.1.2 Electrical-to-Optical Transducers

Two basic types of semiconductor devices convert electrical signals into optical signals and can have the output coupled into an optical fiber: laser diodes (LDs) and light-emitting diodes (LEDs). LDs generally provide better performance in terms of higher output power, greater bandwidth, and narrower signal spectrum. LEDs, on the other

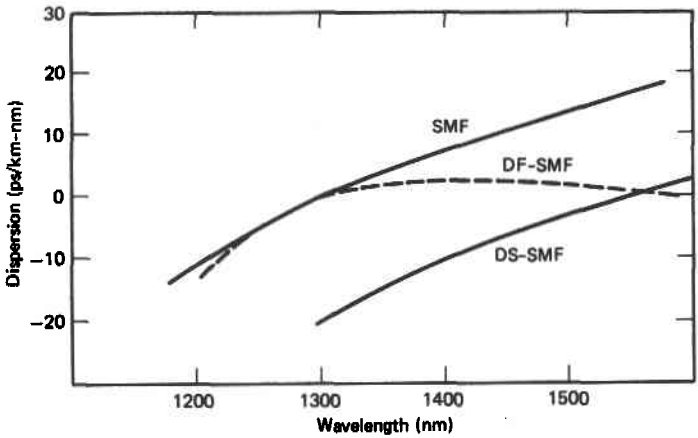


Figure 8.6 Dispersion of SMF and DS-SMF fibers.

hand, are less expensive, require only simple interface circuitry, are more tolerant of environmental conditions, and are generally more reliable. Thus LDs are used for long transmission distances and LEDs are used when ruggedness and interface costs are more important than performance.

To achieve maximum distances between repeaters, it is necessary to minimize chromatic dispersion by using narrowband sources. Various forms of LDs have been developed that come close to achieving a spectrally pure output. Figure 8.7 shows the typical output spectrum of one the most popular types of "single-frequency" devices: a Fabry-Perot laser diode. Another type of diode with a very narrow spectrum is a distributed-feedback LD (DFB-LD). As indicated, the dominant mode of a Fabry-Perot diode is typically 10 dB stronger than the adjacent modes. To maintain the spectrum indicated, it is generally necessary to tightly control the operational environment (bias points, modulation level, and temperature). Otherwise, modulation-dependent effects broaden the spectrum or even cause very short duration modes shifts that pro-

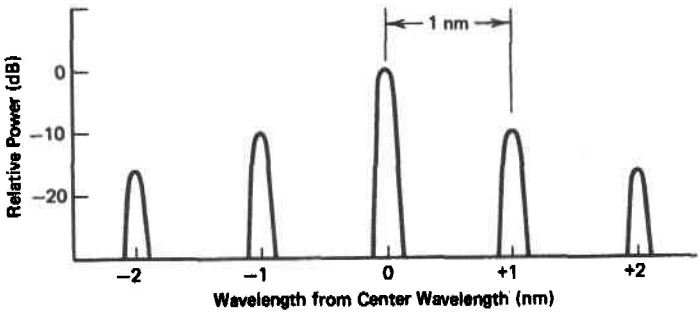


Figure 8.7 Representative laser output spectrum.

TABLE 8.1 Representative Parameters of Various Optical Sources

Device Type	Wavelength (nm)	Launched Output Power (dBm)	FWHM Spectrum Width (nm)
Si LED	850	−16	50
Ge LED	1300	−19	70
InGaAsP LED	1300	−10	120
DFB LD	1300	−5	1.0
DFB LD	1550	−5	0.4
IL/DFB LD	1550	+2	0.8

duce a dominant mode at one of the adjacent wavelengths. This latter phenomenon is referred to as mode partition noise [12], which can cause a halving of the BDP of a system [6]. The spectral width of an optical source is commonly specified as the full-width at half-maximum (FWHM), which represents the spread between the wavelengths at which the spectrum is at half the magnitude of the peak.

To achieve greater distances in loss-limited systems, either a more sensitive receiver or a higher powered source is necessary.* Laser diodes provide the highest available powers, as indicated in Table 8. 1, which lists representative parameters of various LD and LED sources. Systems that are dispersion limited cannot be improved by higher output powers because the intersymbol interference dominates the imperfections in the received signal.

The effective transmit power depends on the coupling efficiency between the source and the fiber. The coupling efficiency may be as low as 1% in the case of an unfocused LED connected to a single-mode fiber. High-efficiency couplings of 60% are possible with focused sources. The launched power values provided in Table 8.1 include the coupling efficiencies.

Bandwidth–Distance Products

The bandwidth–distance factor of a fiber system resulting from chromatic dispersion is determined from the fiber dispersion coefficient and the spectral width of the source. The maximum acceptable amount of pulse spreading with an NRZ line code is typically specified to be one-fourth of a bit interval T. Thus,

$$DL\,\Delta\lambda \leq 0.25T \tag{8.1}$$

where D = dispersion coefficient of fiber (psec/km nm)
L = distance (km)
$\Delta\lambda$ = spectrum width of source (nm)
T = duration of a bit interval (psec)

*Transmit power cannot be increased arbitrarily without encountering nonlinear phenomena in the fiber itself. The optical power threshold where nonlinear effects begin to occur may be as low as 10 mW [6].

Using the relationship that the bandwidth (B) is the reciprocal of the bit interval, the BDP is determined as

$$\text{NRZ-BDP} = BL$$

$$\leq \frac{250}{D\Delta\lambda} \quad \text{Gbps-km} \tag{8.2}$$

A 50% RZ line code allows twice as much pulse spreading because energy in one interval is half a bit period from energy in the next interval. Thus

$$\text{RZ-BDP} < \frac{500}{D\,\Delta\lambda} \quad \text{Gbps-km} \tag{8.3}$$

The RZ line codes are obviously preferred in dispersion-limited systems but are inferior to NRZ line codes in loss-limited systems because the receiver must detect pulse with half as much energy (assuming the peak transmit power is held constant).

Example 8.3. Determine the BDP of a 1550-nm SMF system and a 1550-nm DS-SMF system using the 1550-nm DFB LD source identified in Table 8.1. Assume an NRZ line code and a "zero" dispersion value of 3.5 psec/km nm to allow relaxed tolerances for the fiber and the source operating wavelength.

Solution. From Figure 8.6, the dispersion coefficient of an SMF fiber at 1550 nm is 16 psec/km nm. From Table 8.1, the spectral width of the source is 0.4 nm. Thus,

$$\text{NRZ-BDP} = \frac{250}{16 \times 0.4} = 39 \text{ Gbps-km}$$

Using the given "zero" dispersion value, the BDP of the DS-SMF system is determined as

$$\text{NRZ-BDP} = \frac{250}{3.5 \times 0.4} = 179 \text{ Gbps-km}$$

8.1.3 Optical-to-Electrical Transducers

Two basic types of photodetectors are available as transducers to convert the optical energy at the receiver to electrical energy for amplification and other processing such as clock recovery and data detection. Silicon based *p-i-n* diodes were the first photodetectors to be used, operating in systems using wavelengths of 800–900 nm. These diodes are inexpensive and reliable and provide good performance. Their major drawback is that they do not operate at higher wavelengths where fiber loss is mini-

TABLE 8.2 Representative Parameters of Optical Detectors (BER = 10^{-11})[a]

Device Type	Wavelength (nm)	Receive Power (dBm)	Data Rate (Mbps)
Si *p-i-n*	850	–48.0	50
Si APD	850	–58.0	50
InGaAs *p-i-n*	1310	–35.0	420
InGaAs APD	1310	–43.0	420
InGaAs APD	1310	–26.0	8000
InGaAs APD	1550	–37.5	678
InGaAs *p-i-n*	1550	–37.0	1200
InGaAs APD	1550	–33.0	4000

[a]The detector sensitivities assume an NRZ line code. The sensitivities decrease by 3 dB for (50%) RZ codes.

mized. Germanium devices function at the higher wavelengths and are more sensitive but are more temperature dependent and less reliable.

The second basic type of photodetector is an avalanche photo diode (APD), which enhances the receiver sensitivity because it operates with internal gain. (A *p-i-n* diode does not have internal gain and therefore requires all amplification to be externally applied, which raises the noise level.) The amplification inherent in the APD's conversion from an optical signal to an electrical signal is useful because it means the APD can be 10–15 dB more sensitive in detecting low-level signals at a given error rate. A major drawback of an APD is that it needs a high supply voltage to operate and is very sensitive to temperature. APDs tend to have somewhat lower GBPs than *p-i-n* diodes, which restricts their use in very high data rate systems. Table 8.2 lists various types of photodetectors and provides representative performance parameters for each.

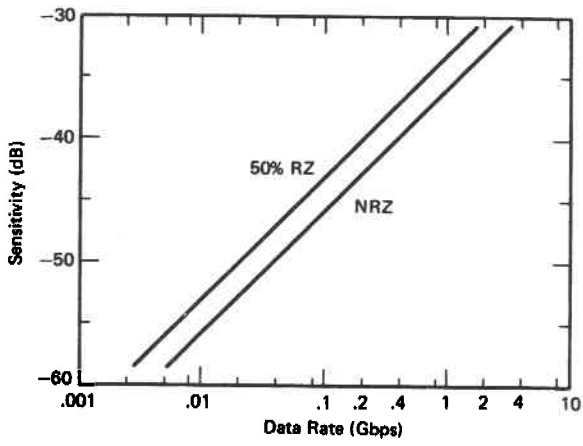


Figure 8.8 Receiver sensitivity versus data rate.

Both types of receivers utilize direct detection, which means that they merely measure the presence or absence of optical energy to detect data. Because higher data rates imply less energy per bit, the sensitivity decreases with data rate. As a first approximation, a doubling of the data rate implies the receiver becomes 3 dB less sensitive. The general formula for determining the change of sensitivity (in decibels) as a function of a change in the data rate is

$$\text{Sensitivity } R = \text{Sensitivity } R_0 + 10 \log_{10}\left(\frac{R}{R_0}\right) \tag{8.4}$$

where $R =$ new data rate
$R_0 =$ data rate at given sensitivity

Figure 8.8 plots the receiver sensitivity of representative *p-i-n* diode detectors and APD detectors as a function of the data rate for both NRZ and RZ line codes.

8.2 LINE CODES FOR FIBER OPTIC TRANSMISSION

Line codes for fiber optic transmission (FOT) systems must satisfy the same basic requirements as the line codes in electrical (wireline) transmission systems described in Chapter 4. Some fiber systems use the line codes described previously. A few line codes developed specifically for fiber applications are described here. Optical sources and detectors are primarily used in nonlinear modes of operation with significant gain and threshold variations, which implies they are best suited to operating in only two states: on or off. Hence, a two-level line code (on–off keying) is most natural. Considering the wide bandwidth available, multilevel line codes are usually unnecessary and extra timing transitions can be included in the line code without significant penalty. An exception to the availability of "free" bandwidth occurs in submarine systems where maximum repeater spacing is achieved by minimizing the bandwidth of the optical signal. These systems are typically dispersion limited because the need for ultrahigh reliability precludes the latest single-frequency lasers with unproven longevity.

A particularly useful property of direct detection optical receivers is the absence of any polarity ambiguity, which means the transmitter and receiver are inherently dc coupled so differential encoding is unnecessary. Even though the optical signal is inherently dc coupled, the dc balance of the line code may still be a consideration because the gain of some photodiodes (particularly APDs) is sometimes dependent on the dc level of the line code. Furthermore, it is usually desirable to ac couple amplifier stages in the receivers [13], which means electrical pulse amplitudes at the detecting comparators are dependent on optical pulse densities. In all, these constraints are typically easier to deal with than their counterparts in wireline systems, so line codes for fiber systems generally tolerate greater amounts of dc variation. Notice that, in fiber systems, both a Manchester and a CMI line code have unvarying dc levels equal to one-half of the pulse amplitude. A number of other line codes used in fiber systems allow some variation in the average dc level. Because these variations are controlled,

the line codes are sometimes referred to as "dc-constrained" line codes as opposed to being perfect or near-perfect "dc-balanced" line codes. The dc variations that do exist in these line codes either are too small to cause problems or are accommodated by providing some form of dc restoration [14].

FOT applications that do not require maximum repeater spacing or maximum data rates typically use Manchester (diphase) or CMI line codes because both of these codes provide strong timing content and a completely constrained dc level (50% on). An early example of the use of CMI is a high-speed intraoffice link developed by NTT of Japan [15].

Manchester and CMI line codes experience the worst case of both the loss limit and the dispersion limit. The worst-case loss limit arises because the receiver must detect optical pulses with duration equal to half the bit interval (i.e., with a sensitivity of an RZ code). The dispersion limit of a Manchester or CMI line code is actually half of the NRZ dispersion limit because the symbol rate is essentially doubled. (Each bit interval is divided in half, with the need to detect a pulse or no pulse in both halves.) Thus, the Manchester or CMI BDP can be determined as

$$\text{MC-BDP} = \frac{125}{D\,\Delta\lambda} \quad \text{Gbps} \tag{8.5}$$

From the previous discussion it can be seen that when performance is more important than low cost, some other line code is needed. Three basic line codes are commonly used: scrambled NRZ, scrambled RZ, or some variation of a code referred to as an mBnB line code, described below. The choice of an NRZ or RZ line code is primarily based on whether the system is loss limited or dispersion limited. Loss-limited systems use NRZ line codes whereas dispersion-limited systems use RZ codes. A scrambler is typically used to enhance timing recovery and control dc wander in a statistical sense (but not in an absolute sense). Often, overhead data patterns in the data stream provides some assurance that the data stream cannot exactly match the scrambler sequence so a certain minimum number of transitions are assured. This is the case in SONET transmission systems described in Section 8.5.

8.2.1 mBnB Line Codes

As an alternative to relying on scramblers for ensuring timing transitions, a class of line codes has been developed that encodes m binary data bits into blocks of n binary line bits (mBnB), where $m < n$. Because only 2^m data codes must be selected from 2^n codes in each block, there is coding flexibility for controlling timing and dc wander. For example, if $m = 4$ and $n = 5$, sixteen data codes must be chosen from the set of thirty-two 5-bit line codes. A specific choice of these codes is provided in Table 8.3, which depicts the data codes and control codes selected for the fiber-distributed data interface (FDDI) standard [16].

The significant aspects of the coding assignments in Table 8.3 are:

TABLE 8.3 FDDI 4B5B Line Codes

Line Code				
Decimal	Binary	Symbol	Name	Functional Assignment
0	00000	Q	Quiet	Line state symbol
1	00001	V	Violation	Disallowed
2	00010	V	Violation	Disallowed
3	00011	V	Violation	Disallowed
4	00100	H	Halt	Line state symbol
5	00101	L		Start delimiter
6	00110	V	Violation	Disallowed
7	00111	R	Reset	Control indicator
8	01000	V	Violation	Disallowed
9	01001	1		Data symbol 0001
10	01010	4		Data symbol 0100
11	01011	5		Data symbol 0101
12	01100	V	Violation	Disallowed
13	01101	T		End delimiter
14	01110	6		Data symbol 0110
15	01111	7		Data symbol 0111
16	10000	V	Violation	Disallowed
17	10001	K		Start delimiter
18	10010	8		Data symbol 1000
19	10011	9		Data symbol 1001
20	10100	2		Data symbol 0010
21	10101	3		Data symbol 0011
22	10110	A		Data symbol 1010
23	10111	B		Data symbol 1011
24	11000	J	Set	Start delimiter
25	11001	S	Set	Control indicator
26	11010	C		Data symbol 1100
27	11011	D		Data symbol 1101
28	11100	E		Data symbol 1110
29	11101	F		Data symbol 1111
30	11110	O		Data symbol 0000
31	11111	I	Idle	Line state symbol

1. The code space is used for control as well as data.
2. When transmitting data, the minimum pulse density is 40% and there can never be more than three intervals without a pulse.
3. The dc component is constrained to fall between a minimum of 40% pulses and a maximum of 60% pulses, a range that is one-fifth of an unconstrained random data stream.

TABLE 8.4 Digital Biphase (Manchester) 1B2B Line Code

Line Code		
Decimal	Binary	Functional Assignment
0	00	Disallowed
1	01	Data symbol 0
2	10	Data symbol 1
3	11	Disallowed

The features of the 4B5B line code described above are achieved at the expense of a 25% increase in the line data rate. By way of comparison, a digital biphase (Manchester) line code and the CMI line code described in Chapter 4 can both be represented as 1B2B line codes with the coding assignments provided in Tables 8.4 and 8.5, respectively. Notice that both of these codes come at the expense of a 100% increase in the line data rate.

Because the 4B5B line code defined for FDDI is intended for a specific application, it contains codes for control as well as for data. The 5B6B codes given in Table 8.6 are intended for transmission only and therefore do not allocate code space for control. Notice that the 5B6B code assignments are made in such a way that the dc level is fixed at 50% pulses but the maximum run length of no pulses is 6. Thus, tighter control of the dc level comes at the expense of increasing the worst-case duration between pulses. 5B6B and 7B8B line codes have been used extensively in Europe. Examples of 7B8B usage are a 565-Mbps terrestrial system developed by British Telecom [17] and a 280-Mbps NL1 submarine system developed by STC of Great Britain [18].

The 4B5B and 5B6B examples use output blocks that are only 1 bit greater than the input blocks. Considerably more coding flexibility is achieved when the output block is more than 1 bit longer than the input block. As an example, a 6B8B code allows all 64 input codes to be encoded with an output code containing exactly four 1's in every code. (The number of combinations of four 1's in 8 bits is 70.) Thus dc wander can be maintained without having to alternate between low-density and high-density codes but coding efficiency is sacrificed.

TABLE 8.5 Coded Mark Inversion (CMI) 1B2B Line Code

Line Code		
Decimal	Binary	Functional Assignment
0	00	Data symbol 1 (if 11 previously sent)
1	01	Data symbol 0
2	10	Disallowed
3	11	Data symbol 1 (if 00 previously sent)

TABLE 8.6 5B6B Line Code Assignments[a]

5B Input	6B Output	5B Input	6B Output
00000	011101/100010	10000	111010/000101
00001	101110/010001	10001	100011
00010	010111/101000	10010	100101
00011	000111	10011	100110
00100	101011/010100	10100	101001
00101	001011	10101	101010
00110	001101	10110	101100
00111	001110	10111	011011/100100
01000	110101/001010	11000	110001
01001	010011	11001	110010
01010	010101	11010	110100
01011	010110	11011	101101/010010
01100	011001	11100	111000
01101	011010	11101	110011/001100
01110	011100	11110	101101/010010
01111	110110/001001	11111	100111/011000

[a]A code 1/code 0 output must alternate between the code with four 1's and the code with two 1's to maintain dc balance.

An 8B10B coding arrangement [19, 20] has been defined as the fiber channel transmission standard (ANSI X3.230-1994). As indicated in Table 8.7, each input byte is separated into a 5-bit field and a 3-bit field that are respectively encoded using a 5B6B and 3B4B algorithm. In some cases, two codewords are provided for encoding particular input data. The choice of one code in a pair (which are complements of each other) is made to maintain dc balance. If the previously transmitted unbalanced code had more 1's than 0's, the code with fewer 1's is chosen. If the previously transmitted unbalanced code had more 0's than 1's, the code with fewer 0's is chosen.*

8.2.2 Bit Insertion Codes

The mBnB block codes described in the previous section have one significant disadvantage: They are difficult to implement on very high speed data streams. (In lower speed applications, decode logic or table lookup in a small read-only memory is trivial.) Very high speed links therefore use another type of code referred to as bit insertion codes. These codes are actually special cases of mBnB codes with $n = m + 1$ and the codes selected so the coding/decoding logic is greatly simplified.

The first bit insertion code to be considered is the MB1P code, which merely inserts an odd parity bit after every m bits of source data. Odd parity ensures that at least one

*Two special cases exist that may alter the rule when 00111 is 5B6B encoded or 011 is 3B4B encoded.

1 is included in the parity field of $m + 1$ bits. As a point of reference, notice that the digital biphase (Manchester) line code is a 1B1P line code. An example of the use of a 24B1P line code is the Trans Pacific Submarine Cable system (TPC-3) put into service in December 1988 by AT&T and KDD of Japan [21].

An even simpler bit insertion code described by engineers of NTT in Japan [22] is the mB1C code, which merely inserts a bit after every m source bits—the value of which is the complement of the immediately preceding bit. Thus, the added bit always forces a data transition and is very easy to implement. An obvious disadvantage of the simplicity is the loss of performance monitorability over the first $m - 1$ bits of a block. Again notice that a diphase line code is a degenerate case of an mB1C line code with $m = 1$. An example of the use of a 10B1C line code is the F-1.6G system of NTT in Japan [23].

TABLE 8.7 8B10B Fiber Channel Coding

5B Input	6B Output	5B Input	6B Output
0 (00000)	100111/011000	16 (10000)	011011/100100
1 (00001)	011101/100010	17 (10001)	100011
2 (00010)	101101/010010	18 (10010)	010011
3 (00011)	110001	19 (10011)	110010
4 (00100)	110101/001010	20 (10100)	001011
5 (00101)	101001	21 (10101)	101010
6 (00110)	011001	22 (10110)	011010
7 (00111)	111000/000111	23 (10111)	111010/000101
8 (01000)	111001/000110	24 (11000)	110011/001100
9 (01001)	100101	25 (11001)	100110
10 (01010)	010101	26 (11010)	010110
11 (01011)	110100	27 (11011)	110110/001001
12 (01100)	001101	28 (11100)	001110
13 (01101)	101100	29 (11101)	101110/010001
14 (01110)	011100	30 (11110)	011110/100001
15 (01111)	010111/101000	31 (11111)	101011/010100

3B Input	4B Output
0 (000)	0100/1011
1 (001)	1001
2 (010)	0101
3 (011)	1100/0011
4 (100)	0010/1101
5 (101)	1010
6 (110)	0110
7 (111)	0001/1110

8.3 WAVELENGTH DIVISION MULTIPLEXING

Wavelength division multiplexing (WDM) is the basic technique used to establish multiple, independent optical channels on a single fiber. The concept of WDM is illustrated in Figure 8.9 showing the coupling (multiplexing) of separate sources into a fiber and the separation (demultiplexing) of the signals out of the fiber. Prismatic refraction is depicted as the basic coupling mechanism at both ends of the link. Refraction is usable only for channels with relatively large wavelength separations. Diffractive grating devices, on the other hand, can operate with channel spacing on the order of 1–2 nm [24]. All WDM devices share the common characteristics of being purely passive and being reversible so any particular device can perform either multiplexing or demultiplexing functions. Other than some optical insertion loss and some crosstalk from imperfect separation, the multiplexing/demultiplexing operations are transparent to the individual channel signals. The insertion losses of the diffraction grating devices can vary from 1 to 7 dB depending on fiber size (multimode vs. single mode) and number of channels [24].

When the first fiber routes were installed, WDM was not used because a single high-speed transmitter and receiver was generally less expensive than separate pairs of lower speed transmitters and receivers and accompanying WDM devices. Thus, WDM would typically be installed as a subsequent upgrade to increase the capacity of a system without laying new fiber. For example, the original 90-Mbps, 820-nm FT3C system installed in 1983 by AT&T in the northeast corridor was upgraded in 1984 by adding an additional 180 Mbps on a 1300-nm carrier [25]. In addition, references [9] and [10] describe how the FT Series G can have its capacity increased by multiplexing a 1550-nm system onto an existing 1300-nm system and/or adding higher speed electronics without taking the existing system out of service.

The passive property of WDM devices is a dominant attraction for some applications such as fiber to the home. Allocating individual fibers from a central office to each home is expensive, implying the need for some form of fiber sharing. The outside plant environment (−40–70°C) and the need for remote power for active components imply that passive WDM devices are preferred. The use of WDM in this application is sometimes referred to as the passive photonic loop (PPL) [26].

WDM is fundamentally identical to FDM as used on electrical (copper) or electromagnetic (radio) transmission systems. In fact, using the following relationship allows wavelengths to be related to frequency:

$$v = f\lambda \tag{8.6}$$

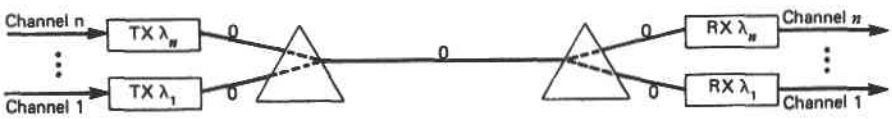


Figure 8.9 Wavelength division multiplexing.

where $v =$ velocity of light, $= 3 \times 10^8$ m/sec
$f =$ frequency (Hz)
$\lambda =$ wavelength (m)

Optical systems are defined in terms of wavelength as opposed to frequency because optical sources are traditionally specified in terms of the wavelengths of their outputs. Nevertheless, systems with extremely close WDM channel spacing (on the order of 0.04 nm) are sometimes referred to as FDM systems [27].

Systems that use electronic multiplexing to produced electrical FDM signals for modulation of a single optical carrier are referred to as subcarrier multiplexing systems. Because the individual channels of these systems are typically close together in frequency, and hence in wavelength, passive separation is usually infeasible.

Example 8.4. Determine the difference in wavelength of two optical signals separated by 2 GHz and centered at 1500 nm.

Solution. The frequency (in free space) of a 1500-nm signal is

$$f = \frac{3 \times 10^8}{1500 \times 10^{-9}} = 200{,}000 \text{ GHz}$$

Thus the upper and lower frequencies are determined as 200,001 and 199,999 GHz, respectively. The corresponding wavelengths are

$$\lambda_1 = \frac{3 \times 10^8}{199{,}999 \times 10^9} = 1500.007 \text{ nm}$$

$$\lambda_2 = \frac{3 \times 10^8}{200{,}001 \times 10^9} = 1499.992 \text{ nm}$$

and the difference in wavelengths is 0.015 nm.

Example 8.4 shows that even when subcarriers are separated by 2 GHz, the resulting wavelength difference is small—too small for passive demultiplexing.

Wavelength demultiplexing with passive diffraction does not inherently cause a loss of signal power in the individual channels. An alternative approach to demultiplexing is shown in Figure 8.10, which involves power splitting of the received signal followed by wavelength filtering to extract individual channels. This approach is passive and functionally identical to diffractive separation but is wasteful of optical power. Its main advantage is that power splitting can be implemented as passive taps distributed along the fiber route. Thus, it is a useful technique in applications (like local area networks) where distances are less important than flexibility in network topology.

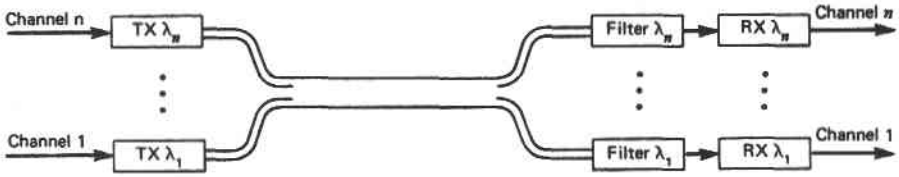


Figure 8.10 WDM with power splitting/filtering.

Dense Wavelength Division Multiplexing

Major advances in WDM technology occurred in the late 1990s, primarily in response to the needs of common carriers to dramatically expand the bandwidth of their fiber facilities for high-bandwidth data networking. In contrast to previous systems, which carried a small number of WDM channels, the newer systems carried a large number of closely spaced wavelengths. These systems are generally referred to as dense wavelength division multiplexing (DWDM) systems. An example of a first-generation (1996) DWDM system is the MultiWave 1600 system from Ciena Corporation. This system provided 16 channels spaced 0.8 nm apart in the region of 1550 nm. DWDM systems are enabled by optical amplifiers (EDFAs) that transparently amplify all wavelengths in the band and by fiber Bragg gratings fabricated into glass fiber for demultiplexing and filtering in a receiver.

DWDM usefulness is not confined to retrofitting of existing plant. DWDM may be the only possible way to achieve extremely high bandwidths. A 40-channel system operating at 2.488 Gbps (OC-48) per channel provides an aggregate bandwidth of 100 Gbps—a difficult speed for a single TDM channel, particularly on dispersion-limited fibers and with other high-speed limitations such as polarization mode dispersion. An additional advantages of DWDM is the inherent transparency of individual wavelengths, which allows transmission and interoperability of mixed types of services. Yet another advantage of DWDM is the inherent reliability provided by separate electronics for each wavelength interface.

8.4 FIBER SYSTEM DESIGN

The examples in the previous sections highlight only the most basic aspects of fiber optic technology. This section provides system-level considerations for the design of practical FOT systems. As is the case with any transmission technology, the system design is strongly dependent on system availability requirements and cost of maintenance. At the low end of these requirements are easily accessible fiber connections to individual workstations within a single building. At the other end of the spectrum of reliability requirements are submarine cables that are obviously very expensive to repair. Repairs are minimized by using only ultrareliable components. Even though submarine cables are immune to backhoes* and gophers, they must contend with oceanic

*Some outside plant engineers of the telephone companies would probably not be too surprised if a backhoe could somehow get to a submarine cable.

counterparts: fishing trawlers and sharks (which seem attracted to electromagnetic radiation). AT&T developed a "seaplow" for burying submarine cables in areas of large-scale fishing activity [28]. About the only advantage that submarine applications have in their favor is the constant, relatively low temperature environment for repeaters/amplifiers.

8.4.1 Fiber Connectors and Splices

Three basic techniques of joining fibers are connectors, mechanical splices, and fusion splices [29]. Connectors are used for terminal connections, patch panels, or other instances where loss (0.4–1 dB) is less important than ease of maintenance and reconfiguration.

Mechanical splices involve mechanical alignment and clamping of two adjoining fibers. There are many different types of mechanical splicing equipment that provide various levels of loss depending on the fiber size (multimode or single mode) and installation time. Splices can be made on individual fibers or in bulk on fiber bundles or ribbons. Two basic splicing procedures are possible: passive or active. Passive splices are faster because the craftsperson merely places the fibers in the alignment sleeves and clamps the fibers in place. Active alignment involves passing a signal through the splice and making fine mechanical adjustments to minimize the loss before the final clamp is applied. A single fiber splice described in reference [30] provides an average of 0.2 dB loss when passively installed and 0.05 dB loss when actively installed. The installation times are 5 and 7 min, respectively.

Fusion splices provide the lowest losses because they essentially form one continuous fiber that is almost as good as a single-drawn fiber. In fact, submarine fibers are pretested at the factory for tensile strength and fused at the weak points where breaks occur [28]. This process ensures that the fibers can withstand the tensile stress encountered when laying and recovering the fiber cable but introduces slightly higher levels of average attenuation in the fiber. For example, the 104-km repeaterless system installed between Taiwan and the island of Peng Hu has 0.24 dB/km at 1550 nm including all splices [28]. Fusion splicing equipment is available that is completely automated with high-resolution TV used to align the fibers and estimate the splice loss by measuring the misalignment after fusion [29]. Fusion losses can vary between 0.01 and 0.1 dB. For undersea applications, only values close to 0.01 dB are accepted.

8.4.2 Protection Switching

Protection switching within an FOT system is basically the same as in other high-capacity transmission systems wherein one spare line, including transmitters, repeaters, and receivers, can be switched into service if one of N main lines fails ($1 : N$ protection). There are, however, a few unique considerations for protecting fiber systems, arising primarily because of the high capacity of the systems. First, a system can be installed with lower cost (lower performance) optoelectronic devices with an upgrade for higher performance anticipated at a later date. In this case, the protection

switching procedures and packaging concepts should allow one-at-a-time upgrading of electronics without disrupting service [11].

A more significant and difficult aspect of protection switching arises because fiber routes tend to carry large-traffic volumes representing aggregations of traffic that would otherwise be carried by separate facilities on separate routes. One-for-N protection of most fiber systems is of no use when a backhoe cuts an entire cable. Protection from the loss of a complete facility requires network-level routing diversity, a feature of a network that tends to disappear when using fiber and large switching systems to their maximum economic advantage—large hubs fed by large trunk groups. A network architecture that provides 1 : 1 route diversity and no hub dependence is a SONET ring described later in this chapter.

8.4.3 System Gain

The first step in determining the repeater spacing of an FOT section is to determine the difference between the launched output power of a transmitter and the receive power required for a designated minimum BER. In analogous fashion to the same term introduced for radio systems, this parameter is referred to as the *system gain*. Notice that the system gain includes coupling losses at the optical source and optical detector. The system gain must be greater than the sum of all the losses in the path. Sources of loss include inherent attenuation of the fiber, splicing/connector losses, bending losses, WDM device losses, and photonic switch losses if any. Splicing losses that occur at regular intervals during the installation of the fiber are often added into the fiber attenuation so there is only one distance-dependent parameter involved in determining repeater spacing. The difference between the system gain and the sum of all the nominal losses is the loss margin. In any particular installation, the loss margin allows for equipment manufacturing tolerances, power penalties due to dispersion, unanticipated splices, component aging, and possible WDM upgrades.

Example 8.5. Determine the system gain, the BDP, the dispersion-limited repeater spacing, and the loss margin for an FOT system with the following parameters: data rate = 565 Mbps, line code = 5B6B RZ, wavelength = 1550 nm, source = –5 dBm DFB-LD with 0.4 nm FWHM, fiber = SMF, detector = InGaAs APD, repeater spacing = 65 km, and splicing losses = 0.2 dB/km.

Solution. The use of the 5B6B line code implies the line data rate of 565(6/5) = 678 Mbps. The receiver sensitivity for 678 Mbps is determined from Figure 8.8 or Table 8.2 as –34.5 dBm. Thus,

$$\text{System gain} = -5 - (-34.5) = 29.5 \text{ dB}$$

Because an RZ line code is in use,

$$\text{BDP} = \frac{500}{17 \times 0.4} = 73.6 \text{ Gbps-km} \qquad \text{BDP spacing} = \frac{73.6}{0.678} = 109 \text{ km}$$

$$\text{Path loss} = (0.2 + 0.2)(65) = 26 \text{ dB} \qquad \text{Loss margin} = 29.5 - 26 = 3.5 \text{ dB}$$

Example 8.5 is representative of the 565-Mbps system developed by Fujitsu of Japan for carrying an E5 digital signal (565.148 Mbps) [31]. The fact that the repeater spacing is 60% of the BDP limit implies that dispersion penalties are small (estimated to be 1.2 dB).

8.5 SONET/SDH

The first generations of fiber optic systems in the public telephone network used proprietary architectures, equipment, line codes, multiplexing formats, and maintenance procedures. Some commonality with other systems in the network came from suppliers who also supplied digital radio systems. In these cases, the multiplexing formats and maintenance protocols emulated counterparts in the radio systems, which also had proprietary architectures. The only thing in common with all of the radio and fiber systems from all of the suppliers was that the interface to the network was some number of DS3 cross-connect signals. Proprietary multiplexing formats for multiple DS3 signals evolved because there was no higher level standard compatible with the applications. A DS4 signal, which is composed of six DS3 signals, requires too much bandwidth for radio systems and carries a larger cross section of channels (4032) than needed in many applications.

The Regional Bell Operating Companies and interexchange carriers (IXCs), the users of the equipment, naturally wanted standards so they could mix and match equipment from different suppliers. This became particularly important as a result of competition among the IXCs who desired fiber interfaces to the local exchange carriers (LECs) but did not want to necessarily buy from the same suppliers as the LECs. (It might be necessary for an IXC to interface with a different supplier at each LEC.) To solve these problems, and others, Bellcore initiated an effort that was later taken up by the T1X1 committee of the Exchange Carriers Standards Association (ECSA) to establish a standard for connecting one fiber system to another at the optical level (i.e., "in the glass"). This standard is referred to as the synchronous optical network (SONET) [32, 33]. In the late stages of the development of this standard, CCITT became involved so that a single international standard exists for fiber interconnect between telephone networks of different counties. Internationally, the standard is known as the synchronous digital hierarchy (SDH) [34]. The SONET standard addresses the following specific issues:

1. Establishes a standard multiplexing format using some number of 51.84-Mbps (STS-1) signals as building blocks.
2. Establishes an optical signal standard for interconnecting equipment from different suppliers.

TABLE 8.8 SONET Signal Hierarchy

North American Designation			
Electrical Signal	Optical Signal	Data Rate (Mbps)	ITU-T Designation
STS-1	OC-1	51.84	
STS-3	OC-3	155.52	STM-1
STS-12	OC-12	622.08	STM-4
STS-24	OC-24	1244.16	STM-8
STS-48	OC-48	2488.32	STM-16
STS-96	OC-96	4976.64	STM-32
STS-192	OC-192	9953.28	STM-64

3. Establishes extensive operations, administrations, maintenance, and provisioning (OAM&P) capabilities as part of the standard.
4. Defines multiplexing formats for carrying existing digital signals of the asynchronous multiplexing hierarchy (DS1, DS1C, DS2, DS3).
5. Supports CCITT (ITU-T) digital signal hierarchy (E1, E2, E3, E4).
6. Defines a DS0 identifiable mapping format for DS1 signals.
7. Establishes a flexible architecture capable of accommodating other applications such as broadband ISDN with a variety of transmission rates. Wide-bandwidth signals (greater than 51.84 Mbps) are accommodated by concatenating multiple STS-1 signals. A STS-3c signal, for example, is an 155.52-Mbps signal that is treated by the network as a single entity.

The data rates and signal designations of the SONET hierarchy are provided in Table 8.8. At the lowest level is the basic SONET signal referred to as the synchronous transport signal level 1 (STS-1). Higher level signals are referred to as STS-N signals. An STS-N signal is composed of *N* byte-interleaved STS-1 signals. The optical counterpart of each STS-N signal is an optical carrier level *N* signal (OC-N). Table 8.8 also includes ITU nomenclature for the SDH, which refers to signals as synchronous transport modules *N* (STM-N). Because common applications of the ITU signal hierarchy cannot efficiently use a 51.84-Mbps signal, the lowest level STM signal is a 155.52-Mbps (STS-3c) signal.*

Although the SONET specification is primarily concerned with OC-N interconnect standards, STS-1 and STS-3 electrical signals† within the SONET hierarchy are useful within a switching office for interconnecting network elements (e.g., multiplexers, switching machines, and cross-connect systems) [35, 36].

*A 51.84-Mbps STS-1 SONET signal is sometimes referred to as a STM-0 SDH signal.
†An STS-1 electrical signal uses a B3Zs line code and a STS-3 electrical signal uses a CMI line code.

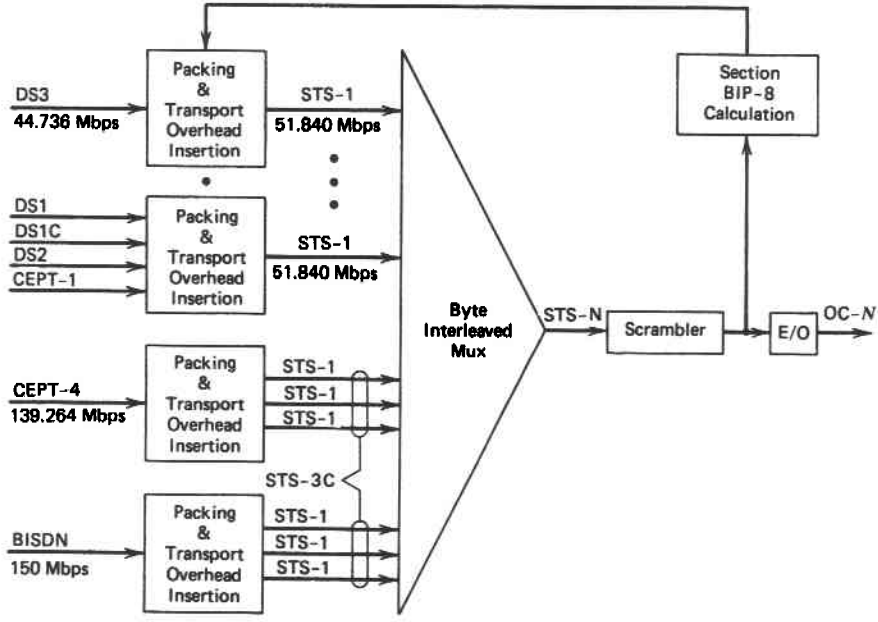


Figure 8.11 Functional block diagram of SONET multiplexing.

8.5.1 SONET Multiplexing Overview

The first step in the SONET multiplexing process (shown in Figure 8.11) involves generation of a 51.840-Mbps STS-1 signal for each tributary. The STS-1 signal contains the tributary (payload) traffic plus transport overhead. As indicated in the figure, a variety of tributary types are accommodated:

1. A single DS3 per STS-1 that can be a standard asynchronous DS3 signal generated by an M13 or M23 multiplexer. Asynchronous DS3 inputs are passed transparently through the system to a DS3 output. Because this transparent option exists, any 44.736-Mbps signal can be carried within the payload envelope.
2. A group of lower rate tributaries such as DS1, DS1C, DS2, or E1 signals can be packed into the STS-1 payload.
3. A higher rate (wideband) signal can be packed into a multiple number of concatenated STS-1 signals. Prevalent examples of higher rate signals are 139.264-Mbps fourth-level multiplexes of ITU or a broadband ISDN signal at 150 Mbps. Each of these applications requires three STS-1 signals concatenated together to form an STS-3c signal. Higher levels of concatenation (to form STS-Nc signals) are possible for higher rate tributaries. Concatenated STS-1 signals contain internal control bytes that identify the signal as a component of

a higher speed channel so the integrity of the concatenated data can be maintained as it passes through a network.

An STS-N signal is created by interleaving bytes from *N* STS-1 signals that are mutually synchronized. All timing (frequency) adjustment is done in generating each of the individual STS-1 signals. STS-1 signals that originate in another SONET node with a possibly different frequency are rate adjusted with the equivalent of byte stuffing (described later) to become synchronized to the clock of the local node. No matter what the nature of the tributary traffic is, all STS-1 tributaries in a STS-N signal have the same high-level format and data rate.

Optical carrier level-*N* signals are generated by first scrambling the STS-N signal (except for framing bytes and STS-ID bytes) and then converting the electrical signal to an optical signal. Other than scrambling, the OC-N signal is generated with direct conversion to an optical signal. Thus, the data rates, formats, and framing of the STS-N and OC-N signals are identical.

A SONET system is defined as a hierarchy of three levels—sections, lines, and paths—as indicated in Figure 8.12. Each of these levels has overhead bandwidth dedicated to administering and maintaining the respective level. As indicated in Figure 8.11, one of the overhead functions provided within an STS-N signal involves calculation and transmission of a parity byte for the entire STS-N signal. Parity is also defined for the other levels of the architecture as described in the following section.

8.5.2 SONET Frame Formats

The frame format of an STS-1 signal is shown in Figure 8.13. As indicated, each frame consists of 9 rows of 90 bytes each. The first 3 bytes of each row are allocated to transport overhead with the balance available for path overhead and payload mapping. The

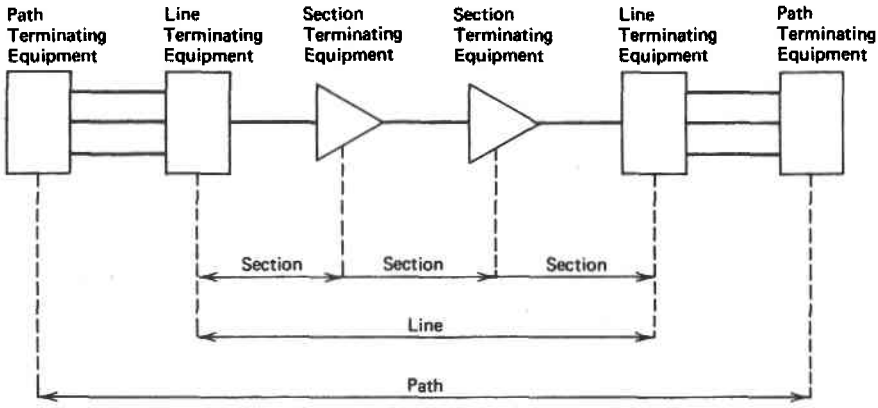


Figure 8.12 SONET system hierarchy.

	Transport Overhead			STS-1 Information Payload		
	A1	A2	C1	J1		
Section	B1	E1	F1	B3		
Overhead	D1	D2	D3	C2		
	H1	H2	H3	G1	Path	9 Rows
	B2	K1	K2	F2	Overhead	
Line	D4	D5	D6	H4		
Overhead	D7	D8	D9	Z3		
	D10	D11	D12	Z4		
	Z1	Z2	E2	Z5		
	3 Columns			87 Columns		

Figure 8.13 STS-1 frame format.

transport overhead is itself composed of section overhead and line overhead. Path overhead is contained within the information payload as indicated.

The 9 rows of 87 bytes (783 bytes in all) in the information payload block is referred to as the envelope capacity. Because the frame rate is 8 kHz, the composite data rate of each STS-1 signal can be represented as the sum of the transport overhead rate and the information envelope capacity:

$$\begin{aligned}\text{STS-1 rate} &= \text{overhead rate} + \text{information envelope rate}\\ &= 9 \times 3 \times 8 \times 8000 + 9 \times 87 \times 8 \times 8000\\ &= 1.728 \times 10^6 + 50.112 \times 10^6\\ &= 51.840 \text{ Mbps} \qquad (8.7)\end{aligned}$$

The internal format of the envelope capacity is dependent on the type of tributary traffic being carried. One aspect of the envelope format that is common to all types of traffic is the 9 bytes of path overhead indicated in Figure 8.13. The actual location and purpose of this overhead are described in the next two sections.

As a specific example of a higher level (STS-N) signal, Figure 8.14 depicts the details of an STS-3 signal that also represents the STM-1 signal format in ITU terminology. Transmission of the bytes occurs row by row and left to right. Thus, the first 3 bytes of an STS-3 frame are the three framing bytes A1, A1, A1. Most of the section and line overhead functions within an STS-3 signal are carried in the STS-1 number 1 overhead. Thus many of the corresponding bytes of the other STS-1 signals are unused and are so designated with an asterisk. Notice, however, that path overhead is included in the information envelope for each of the STS-1 signals.

After a frame of an STS-N signal is scrambled, a parity byte (BIP-8) is generated that provides even parity over corresponding bits in all bytes of the STS-N frame. This parity byte is inserted into the section overhead of the first STS-1 signal of the next STS-N frame.

Transport Overhead									STS-3 Information Payload						
A1	A1	A1	A2	A2	A2	C1	C1	C1			J1		J1		J1
B1	*	*	E1	*	*	F1	*	*			B3		B3		B3
D1	*	*	D2	*	*	D3	*	*			C2		C2		C2
H1	H1	H1	H2	H2	H2	H3	H3	H3			G1		G1		G1
B2	B2	B2	K1	*	*	K2	*	*			F2		F2		F2
D4	*	*	D5	*	*	D6	*	*			H4		H4		H4
D7	*	*	D8	*	*	D9	*	*			Z3		Z3		Z3
D10	*	*	D11	*	*	D12	*	*			Z4		Z4		Z4
Z1	Z1	Z1	Z2	Z2	Z2	E2	*	*			Z5		Z5		Z5
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	...	(1)	...	(2)
9 Columns									261 Columns						

Figure 8.14 STS-3 frame format.

8.5.3 SONET Operations, Administration, and Maintenance

The SONET standard places significant emphasis on the need for operations, administration, and maintenance (OAM) of an end-to-end system. As shown in Figure 8.15, the OAM architecture is based on the section, line, and path layers described previously. OAM standardization is a requirement for mixing equipment from multiple vendors and ease of management of all levels of a system (an individual repeater section or an end-to-end path).

Section Overhead

The functional allocation of the 9 bytes of section overhead in each STS-1 frame shown in Figure 8.13 are:

- A1 Framing byte = F6 hex (11110110)
- A2 Framing byte = 28 hex (00101000)
- C1 STS-1 ID identifies the STS-1 number $(1, \ldots, N)$ for each STS-1 within an STS-N multiplex
- B1 Bit-interleaved parity byte providing even parity over previous STS-N frame after scrambling
- E1 Section-level 64-kbps PCM orderwire (local orderwire)
- F1 A 64-kbps channel set aside for user purposes
- D1–D3 An 192-kbps data communications channel for alarms, maintenance, control, and administration between sections

The fact that there is such a richness of maintenance support at the section level (from one repeater to another) is indicative of the recognized need for extensive OAM facilities and the availability of economical technology to provide it.

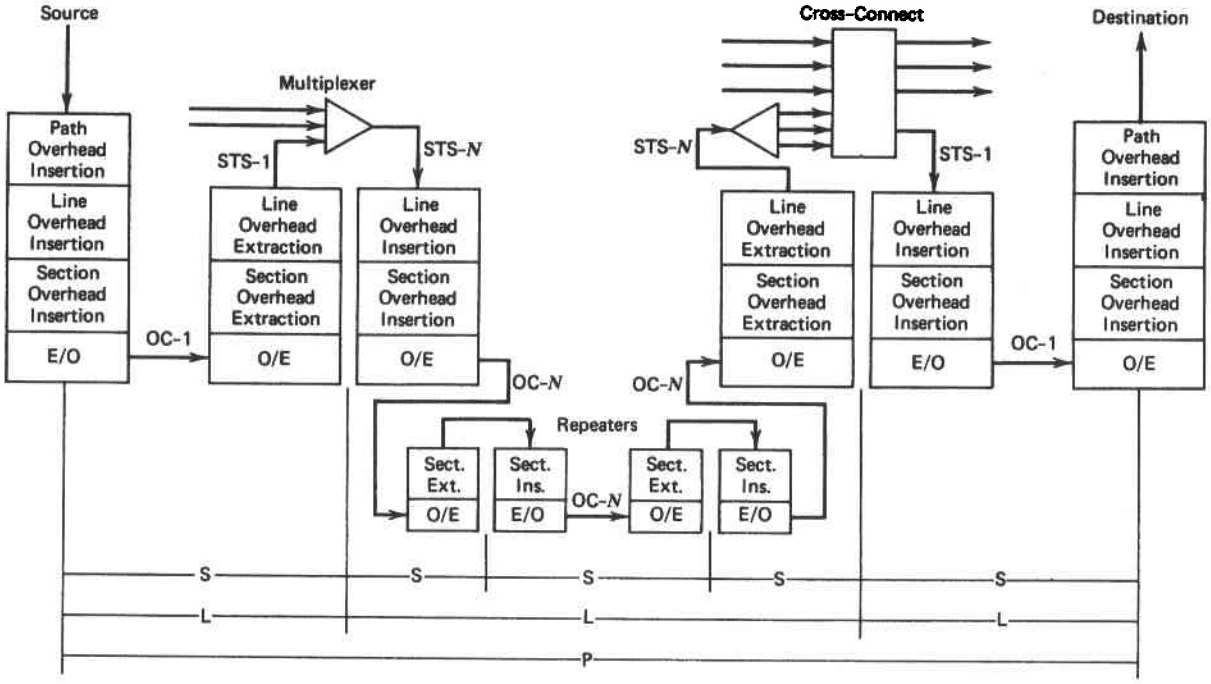


Figure 8.15 SONET overhead layers.

Line Overhead

The functional allocation of the 18 bytes of line overhead in each STS-1 frame shown in Figure 8.13 are as follows:

- H1–H3 Pointer bytes used in frame alignment and frequency adjustment of payload data; the functions of these bytes are defined in detail in Section 8.5.4
- B2 Bit-interleaved parity for line-level error monitoring
- K1, K2 Two bytes allocated for signaling between line-level automatic protection switching equipment
- D4–D12 A 576-kbps data communications channel for alarms, maintenance, control, monitoring, and administration at the line level
- Z1, Z2 Reserved for future use
- E2 A 64-kbps PCM voice channel for line-level orderwire

Notice that the line-level OAM facilities are similar to those available at the section level with the addition of the protection switching signaling channel and H1, H2, and H3 pointer bytes use for payload framing and frequency adjustment.

Path Overhead

As indicated in Figure 8.13, there are 9 bytes of path overhead included in every block (9×87 bytes) of information payload. The important aspect of this overhead is that it is inserted when the tributary data are packed into the synchronous payload envelope (SPE) and not removed (processed) until the tributary data are unpacked. Thus, it provides end-to-end OAM support independent of the path through the synchronous network, which may involve numerous intermediate multiplexers, cross-connect switches, or add–drop multiplexers. The exact location of these 9 bytes within the payload envelope is dependent on pointer values defined in the next section. The functions of the path overhead bytes are:

- J1 A 64-kbps channel used to repetitively send a 64-byte fixed-length string so a receiving terminal can continuously verify the integrity of a path; the contents of the message are user programmable
- B3 Bit-interleaved parity at the path level
- C2 STS path signal label to designate equipped versus unequipped STS signals and, for equipped signals, the specific STS payload mapping that might be needed in receiving terminals to interpret the payloads
- G1 Status byte sent from path-terminating equipment back to path-originating equipment to convey status of terminating equipment and path error performance (received BIP error counts)
- F2 A 64-kbps channel for path user
- H4 Multiframe indicator for payloads needing frames that are longer than a single STS frame; multiframe indicators are used when packing lower rate channels (virtual tributaries) into the SPE
- Z3–Z5 Reserved for future use

8.5.4 Payload Framing and Frequency Justification

Payload Framing

The location of the 9 bytes of path overhead in the STS-1 envelope is not defined in terms of the STS-1 transport framing. Instead, the path overhead is considered to be the first column of a frame of data referred to as the SPE, which can begin in any byte position within the STS-1 payload envelope (see Figure 8.16). The exact location of the beginning of the SPE (byte J1 of the path overhead) is specified by a pointer in bytes H1 and H2 of the STS line overhead. Notice that this means that an SPE typically overlaps two STS-1 frames.

The use of a pointer to define the location of the SPE frame location provides two significant features. First, SPE frames do not have to be aligned with higher level multiplex frames. It may be that when first generated, an SPE is aligned with the line overhead at the originating node (i.e., the pointer value is 0). As the frame is carried through a network, however, it arrives at intermediate nodes (e.g., multiplexers or cross connects) having an arbitrary phase with respect to the outgoing transport framing. If the SPE had to be frame aligned with the outgoing signal, a full SPE frame of storage and delay would be necessary. Thus, the avoidance of frame alignment allows SPEs on incoming links to be immediately relayed to outgoing links without artificial delay. The location of the SPE in the outgoing payload envelope is specified by setting the H1, H2 pointer to the proper value (0–782).

The second advantage of the pointer approach to framing SPE signals is realized when direct access to subchannels such as DS1s is desired. Because the pointer provides immediate access to the start of an SPE frame, any other position or time slot within the SPE is also immediately accessible. If the tributary uses a byte-synchronous mapping format, individual channel bytes have fixed positions with respect to the start of the SPE. This capability should be compared to the procedures required to demultiplex a DS3 signal. In a DS3 signal there is no relationship between the higher level framing and the lower level DS2 and DS1 framing positions. In essence, two more frame recovery processes are needed to identify a DS0 time slot. The use of pointers

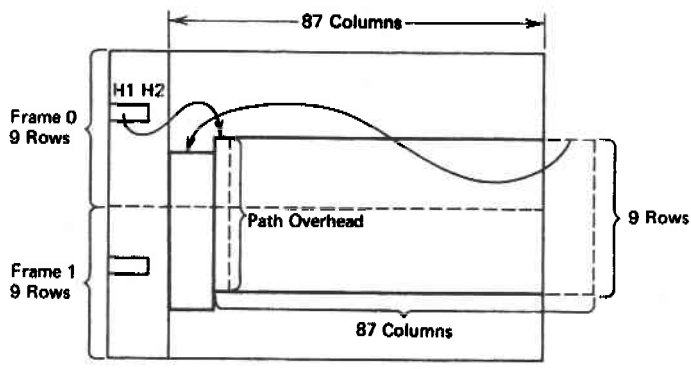


Figure 8.16 Representative location of SPE.

in the SONET architecture eliminates the need for more than one frame recovery process when accessing byte-synchronous lower level signals.

Frequency Justification

Although it is generally intended that SONET equipment be synchronized to each other or to a common clock, allowances must be made for the interworking of SONET equipment that operates with slightly different clocks. Frequency offsets imply that an SPE may be generated with one clock rate but be carried by a SONET transport running at a different rate. The means of accommodating a frequency offset is to accept variable SPE frame rates using dynamic adjustments in the SPE pointers. Pointer adjustments allow SPE frames to float with respect to the transport overhead to maintain a nominal level of storage in interface elastic stores. Figure 8.17 shows the basic means of accommodating a slow incoming SPE. If the elastic store begins to empty, positive byte stuffing is invoked to skip one information time slot (the slot immediately following the H3 byte) and simultaneously incrementing the pointer to delay the SPE frame by one byte.

Negative byte stuffing, to accommodate a fast SPE clock, requires sending an extra SPE byte whenever the elastic store begins to fill. As indicated in Figure 8.18, the H3 slot carries the extra byte of data, which requires the pointer to be simultaneously decremented, thereby advancing the SPE frame by 1 byte. To protect against errors in misinterpreting byte-stuffing operations, the need to increment or decrement a pointer is redundantly encoded,* and the new pointer value is transmitted for a minimum of three frames following the frame in which the pointer adjustment occurs. This implies that a 1-byte adjustment can be made once every four frames (once every 500 μsec). For an analysis of the effects of channel errors on SONET payload pointers, see reference [37].

Example 8.6. Determine the range of SPE data rates that can be accommodated by the byte-stuffing operation described above.

Solution. Four SPE frames nominally contain $4 \times 9 \times 87 = 3132$ bytes of data. Thus, the nominal SPE rate is $8 \times 3132 \times 2000 = 50.112$ Mbps. When positive byte stuffing is used to accommodate a slow incoming SPE rate, 3131 bytes of data are transmitted in four frames. Thus, the lowest, slip-free rate is

$$\text{Minimum SPE rate} = 8 \times 3131 \times 2000 = 50.096 \text{ Mbps}$$

When negative byte stuffing is used to accommodate a fast SPE rate, 3133 bytes of data are transmitted in four frames. Thus, the highest, slip-free rate is

$$\text{Maximum SPE rate} = 8 \times 3133 \times 2000 = 50.128 \text{ Mbps}$$

*During the adjustment frame only, 5 even-numbered bits of the pointer value are inverted to indicate a negative stuff (data in byte H3). When a positive stuff occurs, 5 odd-numbered bits of the pointer are inverted.

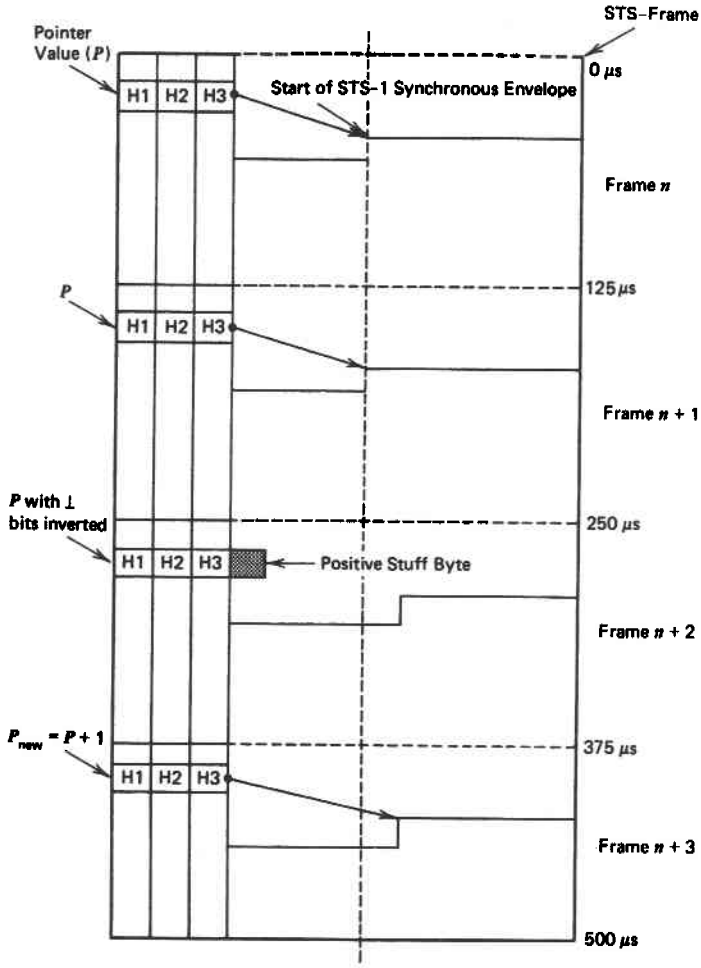


Figure 8.17 Positive STS-1 pointer justification operation.

Example 8.6 demonstrates that the SONET clock accuracy required for maintaining SPE data is 50.112 ± 0.016 Mbps—a very wide tolerance of ±320 ppm. In comparison, a SONET node is specified to maintain a minimum timing accuracy of 20 ppm if it loses its reference. Thus, the frequency of timing adjustments was chosen more from a desire to simplify the process than from just assuring a margin for slip-free operations.

The use of byte stuffing to accommodate timing differences introduces waiting time jitter into SPE payloads, just as bit stuffing introduces waiting time jitter into DS1 signals being asynchronously multiplexed into DS2 or DS3 signals. If the SPE is carrying DS1 traffic, the effect of byte stuffing at the SPE rate is about one-fourth of a bit interval at the DS1 rate. (Because the SPE rate is 32 times the DS1 rate, the duration of an SPE byte is one-fourth the duration of a DS1 bit.) Notice that this amount of

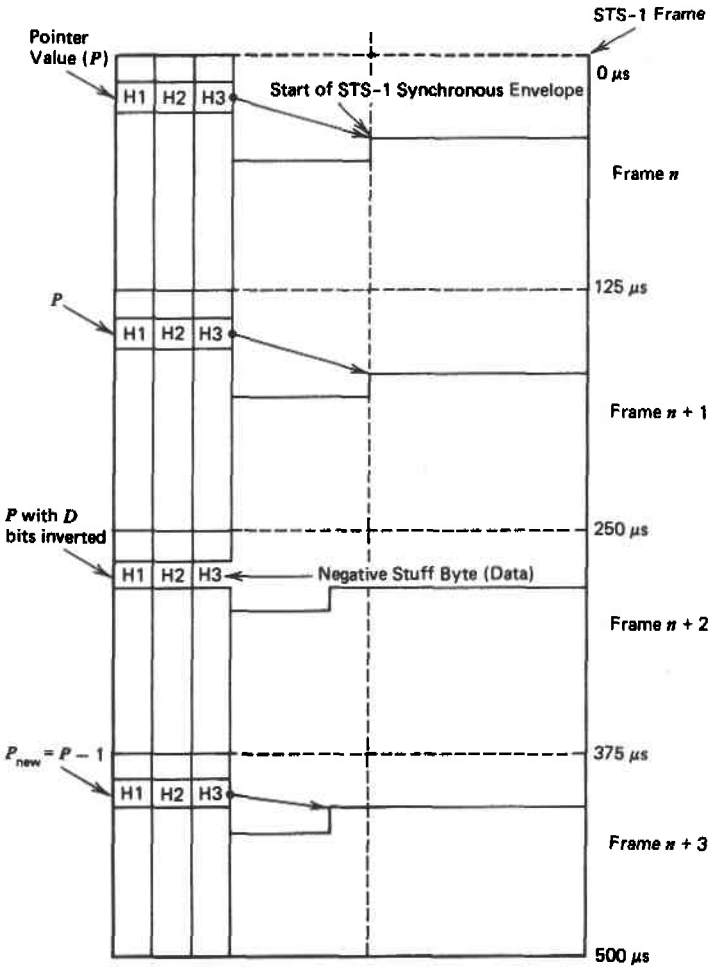


Figure 8.18 Negative STS-1 pointer justification operation.

phase jitter is comparable to the amount introduced by bit stuffing a DS2 signal into a DS3 signal.

8.5.5 Virtual Tributaries

To facilitate the transport of lower rate digital signals, the SONET standard uses sub-STS-1 payload mappings referred to as virtual tributary (VT) structures, as shown in Figure 8.19. This mapping divides the SPE frame into seven equal-sized subframes or VT blocks with 12 columns (108 bytes) in each. Thus, the subframes account for $7 \times 12 = 84$ columns with the path overhead and two unused columns (reserved bytes R) accounting for the remainder of the 87 columns in an SPE. The rate of each VT structure is determined as $108 \times 8 \times 8000 = 6.912$ Mbps.

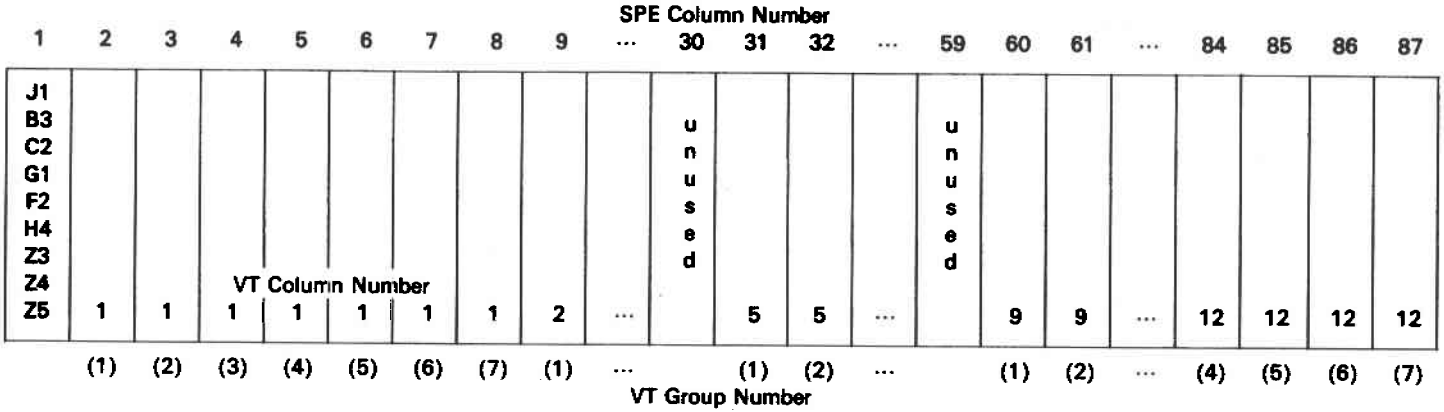


Figure 8.19 SPE mapping for virtual tributaries.

TABLE 8.9 SONET Virtual Tributaries

Tributary Type	VT Designation	Number of Columns per VT	Number in VT Group	Maximum Number in SPE
DS1	VT1.5	3	4	28
E1	VT2.0	4	3	21
DS1C	VT3.0	6	2	14
DS2	VT6.0	12	1	7

The VT structures can be individually assigned to carry one of four types of signals. Depending on the data rate of a particular signal, more than one signal may be carried within a VT structure as a VT group. All signals within a VT group must be of the same type, but VT groups within a single SPE can be different types. The particular lower rate signals accommodated as VTs are listed in Table 8.9. The last column indicates how many of the lower rate signals are carried in a single SPE if all seven VT groups are the same type.

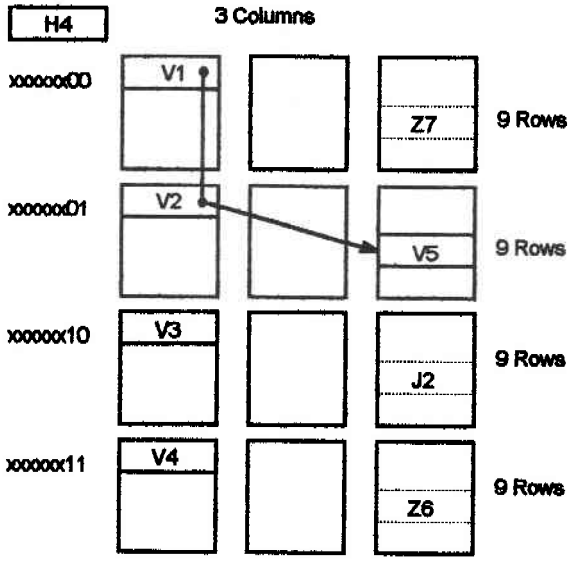
VT-SPE payloads are allowed to float within an STS-1 SPE in the same fashion as pointers to SPE payloads are allowed to float at the STS-1 level. Thus, a second level of pointer logic is defined for VT payloads. Again, a floating VT-SPE allows for minimal framing delays at intermediate nodes and for frequency justification of VT-SPEs undergoing transitions between timing boundaries. High-rate VT-SPEs are accommodated by inserting an information byte into V3 while slow-rate VT-SPEs are accommodated by stuffing into the information byte immediately following V3 when necessary.

The mapping format for a VT1.5 is shown in Figure 8.20. Each VT1.5 uses three columns of data to establish 108 bytes in a VT1.5 payload. There are four such payloads in a 12-column VT group. The V1, V2, V3, V4 bytes of the payload have fixed positions within the STS-1 payload. The remaining 104 bytes of the VT1.5 signal constitute the VT1.5 payload, the start of which is the V5 byte pointed to by V1 and V2. Figure 8.21 depicts two different mappings for a VT1.5 payload: an asynchronous mapping and a byte-synchronous mapping.

Asynchronous Mapping

The asynchronous operation is identical in concept to the bit-stuffing operation described in Chapter 7. The DS1 bit stream is inserted into the information bits (I) with no relationship to the VT-SPE frame or byte boundaries. As indicated, there are two stuffing opportunities (S_1 and S_2) available in every four-frame superframe. Thus, the VT1.5 superframe carries 771, 772, or 773 information bits depending on the value of the stuff control bits C_1 and C_2. The nominal number of information bits in each frame is $193 \times 4 = 772$. Nominal frames carry information in S_2 while stuffing in S_1.

Because the asynchronous operation is compatible with the asynchronous network, it is the format used in most SONET applications. The major advantage of the asynchronous mode of operation is that it provides for totally transparent transmission of



Notes:
1) A VT1.5 uses three columns of an SPE for 108 bytes in a 500 usec. superframe.
2) V1, V2, V3, V4 bytes have fixed locations in a SPE identified by the last two bits of H4.
3) V1 and V2 point to V5 which is the first byte of the floating VT1.5 SPE.
4) SPE overhead bytes V5, J2, Z6, Z7 occur in identical relative positions of an SPE.

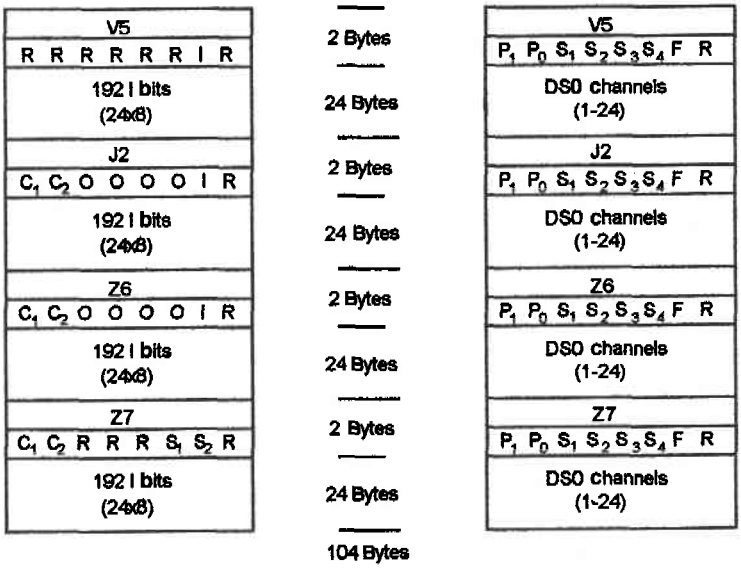
Figure 8.20 Superframe structure for VTI.5 tributaries.

the tributary signal in terms of information and in terms of information rate. The major disadvantage of the asynchronous mode is that 64-kbps DS0 channels and signaling bits are not readily extracted.

Byte-Synchronous Multiplexing

In contrast to the asynchronous mapping, the byte-synchronous payload mapping shown in Figure 8.21*b* allocates specific bytes of the payload to specific bytes (channels) of the DS1 tributary signal. Hence, this mode of operation overcomes the main drawback of the asynchronous mode in that 64-kbps DS0 channels and signaling bits within the payload are easily identified. In fact, when the DS1 tributary arises from legacy applications, the signaling bits of a DS1 are moved from the least significant bit (LSB) of every sixth frame of respective channels and placed in dedicated signaling bit positions within the VT-SPE. Thus byte-synchronous multiplexing offers an additional feature of converting from in-slot signaling to out-slot signaling for DS1 signals.

An important aspect of the byte-synchronous format shown in Figure 8.21*b* is the absence of timing adjustments for the source DS1 signal. Thus, the DS1 interface necessarily requires a slip buffer to accommodate a DS1 source that may be unsynchronized to the local SONET clock. Although slips in byte synchronously mapped DS1



No. of fixed information bits = 4(192) + 3(1) = 771
S_1, S_2 = Stuff/information bits
C_1, C_2 = Stuffing control bits
R bits are fixed stuff bits (unspecified)
O bits are reserved for future purposes

(*a*)

S_1, S_2, S_3, S_4 = Channel-associated signaling bits
P_1, P_0 = Signaling phase indicator bits
F = DS1 framing bit
R bits are fixed stuff bits (unspecified)

(*b*)

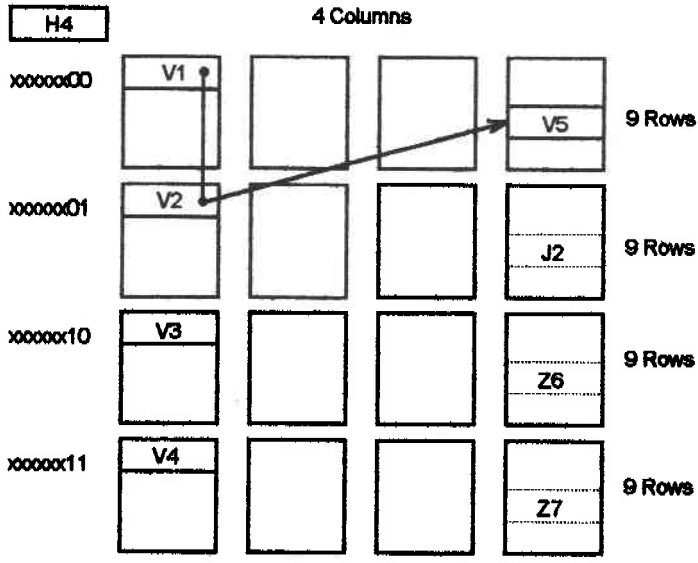
Figure 8.21 DS1 mappings in VT1.5 SPE: (*a*) asynchronous; (*b*) byte synchronous.

signals may occur at the SONET network interface (e.g., SONET gateway), slips cannot occur within the SONET network because internal nodes rate adjust the VT1.5 payloads with pointer adjustments.*

E1 Mappings

E1 signals are mapped into VT2 signals with the same basic procedures used for DS1s. As shown in Figure 8.22, the VT2 signal is composed of four columns of bytes in an STS-1 that produce a total of 144 bytes. After removing the V1, V2, V3, and V4 bytes, the VT2 payload has 140 bytes. Formats for asynchronously mapped EIs and byte synchronously mapped E1s are shown in Figure 8.23. Notice that the byte-synchronous mapping for a 30-channel E1 carries channel-associated signaling in slot 16—the form of out-slot signaling designed into E1 signals at their inception. The same basic format supports common-channel signaling, which is sometimes referred to as a

* The original specifications for SONET included a locked timing format for VT-SPEs that eliminated the VT pointers so DS0 channels could be identified directly within the STS-1. This mode of operation has since been abandoned.



Notes:
1) A VT2 uses four columns of an SPE for 144 bytes in a 500 usec. superframe.
2) V1, V2, V3, V4 bytes have fixed locations in a SPE identified by the last two bits of H4.
3) V1 and V2 point to V5 which is the first byte of the floating VT2 SPE.
4) SPE overhead bytes V5, J2, Z6, Z7 occur in identical relative positions of an SPE.

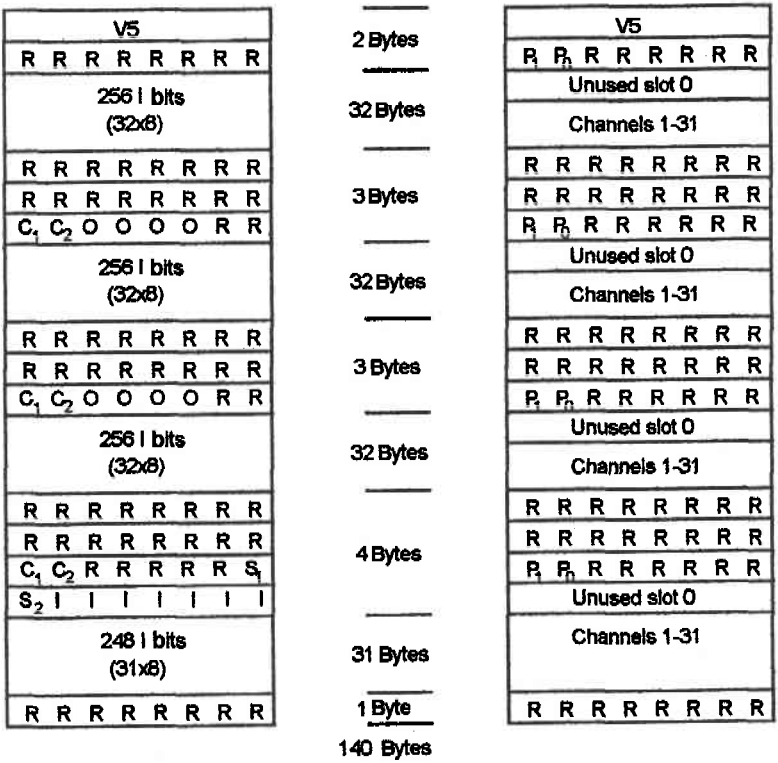
Figure 8.22 Superframe structure for VT2 tributaries.

31-channel E1 format. In this case channel 16 is the CCS channel and channels 1–15 and 17–31 are the bearer channels. Thus, the multiplex mapping is not changed, just the nomenclature of the channels and the SPE type designation in the VT path overhead byte V5.

8.5.6 DS3 Payload Mapping

The previous section describes several alternatives for packing virtual tributaries into an STS-1 envelope. When all seven VTs in an envelope are VT1.5s, a total capacity of 28 DS1s is provided—the same as a DS3 signal. Thus one method of carrying a DS3 signal involves demultiplexing it into its constituent DS1 (or DS2 signals) and packing the constituents as virtual tributaries. This approach is attractive in that the virtual tributaries are individually accessible for cross-connect or add–drop multiplexer systems. If the application does not need to access the individual tributaries, it is simpler to pack the DS3 signal directly into an STS-1, as indicated in Figure 8.24. The payload mapping in Figure 8.24 treats the DS3 signal simply as a 44.736-Mbps data stream with no implied internal structure. Thus, this mapping provides transparent transport of DS3-rate data streams.

Each row of a nine-row SPE envelope contains $87 \times 8 = 696$ bits, which can carry 621 or 622 DS3 data bits depending on the value of the C bits. Notice that this format



No. of fixed information bits = 4(256) -1 = 1023
S_1, S_2 = Stuff/information bits
C_1, C_2 = Stuffing control bits
R bits are fixed stuff bits (unspecified)
O bits are reserved for future purposes

(*a*)

CAS bits in channel 16
P_1, P_0 = Signaling phase indicator bits
R bits are fixed stuff bits (unspecified)

(*b*)

Figure 8.23 E1 mappings in VT2 SPE: (*a*) synchronous; (*b*) byte synchronous.

has five C bits, which allows for single and double bit error correction. The path overhead (POH) bytes carry the 9 bytes of POH as defined in Section 8.5.3.

Determination of the minimum and maximum DS3 rates is left as an exercise at the end of the chapter.

8.5.7 E4 Payload Mapping

One example of a SONET superrate mapping is shown in Figure 8.25 for a 139.264-Mbps fourth-level ITU-T signal (E4). This signal is packed into a 155.52-Mbps STS-3c (or STM-1) signal. Figure 8.25 shows only the synchronous payload envelope

87(8) = 696 Bits

	8	18	1	205	16	11111111	208	16	11111111	208
1	POH	R	C	I	R	CCRRRRRR	I	R	CCRROORS	I
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	POH	R	C	I	R	CCRRRRRR	I	R	CCRROORS	I

POH:	Path overhead byte	8
R:	Fixed stuff bit (reserved)	59
C:	Stuff control bit	5
I:	Information (payload) bit	621
S:	Stuff opportunity bit	1
O:	Overhead communications bit	2
	Total bits	696

Note: S contains information when CCCCC = 00000.
S is a stuff bit when CCCCC = 11111.

Figure 8.24 Asynchronous 44.736-Mbps (DS3) payload mapping.

87(3) = 261 Bytes

X	12I	Y	12I	Y	12I	Y	12I

1	POH	13I					X	12I	Y	12I	Z	12I
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	POH	13I					X	12I	Y	12I	Z	12I

POH:	Path overhead byte	1
I:	Information (payload) byte	241
X:	CRRRRROO	5
Y:	RRRRRRRR	13
Z:	iiiiiiSR	1
	Total bytes:	261

C: Stuff control bit
R: Fixed stuff bit
O: Overhead communications bit
S: Stuff opportunity bit
i: Information bit

Note: S contains information when CCCCC = 00000.
S is a stuff bit when CCCCC = 11111.

Figure 8.25 Asynchronous 139.264-Mbps (E4) payload mapping.

(SPE-3c), not the 9 bytes of section and line overhead in each row. Notice that there is only one column of POH within the SPE-3c envelope. The POH bytes carry the 9 bytes of overhead as defined in Section 8.5.3.

The payload mapping in Figure 8.25 treats the 139.264-Mbps signal as a transparent data stream with no implied internal structure. Each row of a nine-row SPE-3c envelope contains $87 \times 3 = 261$ bytes, which can carry 1934 or 1935 data bits depending on the value of the C bits. Notice that this format also has five C bits, which allows for single and double bit error correction.

8.5.8 SONET Optical Standards

The optical interface standard [36] defined for "mid-span-meet" of SONET equipment allows for either NRZ or RZ line codes on single-mode fibers. Generation of the OC-N signal from the STS-N signal requires a scrambler as shown in Figure 8.26. The scrambler is synchronized to each STS-N frame by presetting the shift register to all 1's immediately after transmitting the last C1 byte of the STS-N section overhead. Thus, the frame codes (A1,A2) and STS-1 ID (C1) code are not scrambled. A minimum level of timing content is assured by the A1, A2, and C1 bytes along with the static overhead bits of the STS-N frame that are anticoincident with the scrambler sequence. Because the scrambler is preset at the same point of every frame, every bit position in successive frames experience the same scrambler value. Thus, when static overhead is "exclusive ored" with the scrambler, the same data values arise. (The scrambler merely converts fixed overhead data into a different pattern of fixed data.)

The BER objective is 1×10^{-10} for optical sections of 40 km or less. Equipment from separate manufacturers can be freely interchanged for applications with distances up to 25 km. Longer distances may require joint engineering. (Transmitters and receivers from separate suppliers must be jointly specified to support longer distances.)

SONET systems are specified to operate with central wavelengths at 1310 nm with SMF fibers or at 1550 nm with DS-SMF fibers. Operation at 1310 nm with DS-SMF fibers or at 1550 nm with SMF fibers is not disallowed but must be jointly engineered. A range of laser wavelength tolerances and maximum allowable spectral widths is

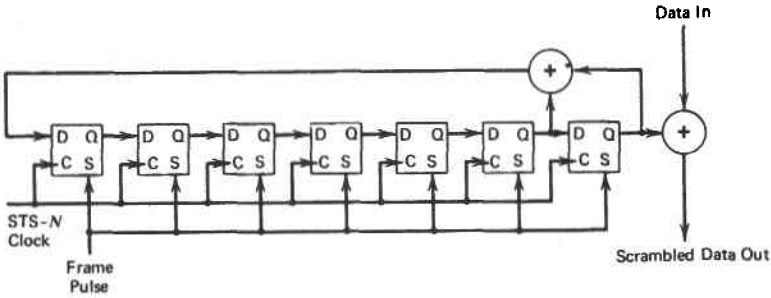


Figure 8.26 SONET scrambler.

TABLE 8.10 Representative Maximum Spectral Widths of SONET Sources[a,b]

SONET Source	Rate (Mbps)	< 25 km	< 40 km
OC-1	51.84	30nm	25.0 nm
OC-3	155.52	15nm	10.0 nm
OC-9	466.56	10nm	7.5 nm
OC-12	622.08	8nm	6.0 nm

[a]The spectral width of a source is determined as the wavelength difference between the peak mode and the farthest mode that is 10 dB below the peak.

[b]These specifications are for lasers operating within 10 nm of the central wavelengths (1310 and 1550 nm). Lasers with greater deviation from the central wavelength are allowed but have narrower spectral width specifications to compensate for greater fiber dispersion.

specified for both 1310 and 1550 nm. Table 8.10 provides representative values of the specifications.

8.5.9 SONET Networks

A basic block diagram of a SONET network is shown in Figure 8.27. Gateway network elements (GNEs) provide interfaces to external (asynchronous) digital signals. These signals are mapped (synchronized) and unmapped (desynchronized) by the gateway using the appropriate mapping format. At this point only bit stuffing is used to synchronize the asynchronous tributaries to SONET. No pointer adjustments occur in the GNE. As the STS-N signals propagate through the network, pointer adjustments in pointer processing (PP) interfaces may be applied at internal network elements (NEs), but the lower level interface mappings that occur at the GNEs are untouched. If a particular NE accesses VT payloads, VT payloads in the same VT group that pass through the node may experience VT pointer adjustments. Otherwise, VT pointer adjustments do not occur (only the STS-1 level signals are rate adjusted). The following paragraphs summarize pointer processing aspects of a SONET network:

1. Pointer justification events (PJEs) never occur in an originating GNE.
2. A desynchronizer experiences continuous PJEs only as a result of a synchronization difference between the originating GNE and the terminating

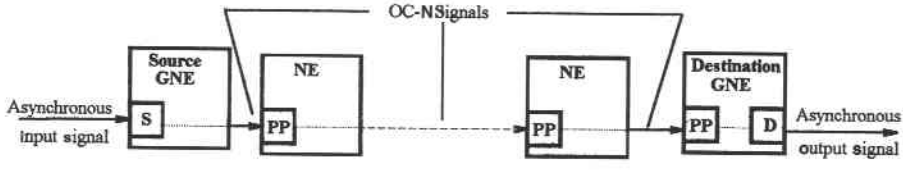


Figure 8.27 SONET network elements: S, synchronizer; PP, pointer processor; D, desynchronizer.

GNE.* Synchronization differences/failures at internal nodes of a SONET network produce continuous pointer adjustments, but these get removed when the SPE passes through a node that is synchronized to the source GNE.

3. PJE bursts occur for two possible reasons. The first is a result of a reference switch and a subsequent phase adjustment of a node's local clock to align it with the phase of the new reference. Bursts can also occur as a result of clock noise in multiple nodes producing near-simultaneous pointer adjustments. In order for all of these adjustments to propagate to a desynchronizing gateway, all of the elastic stores in the path must be at the appropriate threshold. This can only happen if the source GNE has previously produced some abnormal behavior such as a loss of a reference or sustained a rather large amount of wander.
4. A pointer adjustment at the SPE level does not affect a VT signal unless it is passed to a node that accesses the VT and that particular adjustment happens to cause a pointer movement at the VT level. Even when this occurs, the VT pointer adjustment must pass through the network (without absorption) to the desynchronizing gateway to affect the outgoing tributary signal. On average, one of every 30 PJEs at the STS-1 level produces a PJE at the VT1.5 level.

Figure 8.28 shows a block diagram of an SPE synchronization circuit (PP) depicting two halves of pointer processing: one half extracts (desynchronizes) the SPE payload from a received signal and the other half synchronizes the SPE to the local STS-1 frame rate. The RX pointer processing block extracts the payload data from the received signal and passes it to the elastic store. The TX pointer processing block monitors the fill level of the elastic store and makes pointer adjustments to maintain a nominal level of storage. The size of the elastic store only needs to be on the order of 8 bytes in length, not a full frame. The ability to use a relatively small elastic store (as compared to frame-length elastic stores in the asynchronous network) is one of the features of a pointer-based synchronization architecture: The payloads are allowed to float with respect to the STS-1 frame boundaries.

Frequency of Pointer Justification Events

If all NEs of a SONET island use a timing reference that is traceable to a common primary reference source (PRS), PJEs occur only as a result of distribution-induced clock wander that produces no sustained frequency offset. Thus, when all NEs are synchronized to the same reference, PJEs occur at random times and have equal numbers of positive and negative values over the long run.

Continuous PJEs occur only when there is a reference failure at some NEs within a SONET island or the island is intentionally designed to operate in a plesiochronous mode. If the reference failure occurs at some internal node of the SONET island, the resulting PJEs are removed at the next node in the path that is still locked to the same reference as the gateway NE. Thus, a tributary desynchronizer at a GNE must deal

*This statement assumes that the terminating GNE synchronizes incoming SPEs to a local clock before they arrive at the desynchronizer.

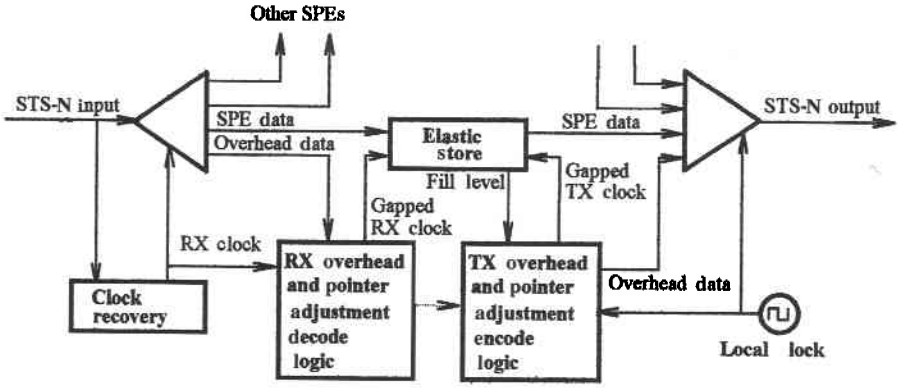


Figure 8.28 Block diagram of SPE synchronizing equipment: RX, receiver; TX, transmitter.

with continuous PJEs only when there is a reference failure in either the originating GNE or in the terminating GNE.

SONET Desynchronizers

SONET desynchronizers are necessarily designed with very low clock recovery bandwidths to smooth the effects of (1) isolated pointer adjustments, (2) continuous pointer adjustments, (3) pointer adjustment bursts, or (4) combinations of the latter two. A pointer burst is defined as the occurrence of multiple pointer adjustments of one polarity occurring within the decay time of the desynchronizer circuit (i.e., the reciprocal of the desynchronizer closed-loop PLL bandwidth). Thus, it is ironic that as the clock recovery bandwidth is narrowed to smooth the effect of a burst, the probability of a burst occurrence is increased (by definition only). Extremely narrow PLL bandwidths are easiest to implement using digital filtering techniques commonly referred to as bit leaking. Bit leaking is essentially a mechanism for converting byte-sized pointer adjustments into bit- (or fractional-bit-) sized timing adjustments.

Figure 8.29 shows a block diagram of a microprocessor-controlled DS1 desynchronizer. The microprocessor is used to perform long-term averaging of phase ad-

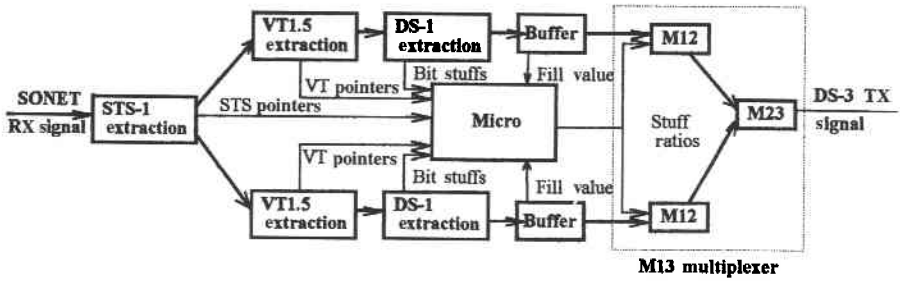


Figure 8.29 VT1.5 desynchronizer hardware functional components.

justments in lieu of dedicated logic that requires large counters and wide word sizes for low-bandwidth DSP filtering. The first function of the microprocessor is to determine the average DS1 payload frequency offset represented by all frequency adjustment events (bit stuffs, VT pointer adjustments, and STS pointer adjustments). After the average frequency adjustment is determined, a stuff ratio value is calculated that allows insertion into a DS3 signal as shown. (The M12 stage is embedded in the M13 multiplexer.) The elastic store fill level is used for very long term adjustments in the output frequency that arise from finite precision limits of the DSP calculations and for accommodating variations in the DS3 clock, which is typically not synchronized to the SONET line clock.

8.6 SONET RINGS

As has been mentioned earlier in this book, the development of large switching machines and transmission systems with extremely large cross sections has impacted telecommunications network architectures with a trend toward fewer hierarchical levels. An undesirable consequence of this trend is increased dependence on the operational status of individual switching machines and transmission paths. A SONET self-healing ring, or more simply a SONET ring, is a network architecture that specifically addresses network survivability. Two basic types of self-healing rings are shown in Figure 8.30: a unidirectional ring and a bidirectional ring. The main difference between the two types of rings is how the two directions of a duplex connection are established.

In a unidirectional ring a single time slot of the entire ring is assigned to both halves of a connection. As indicated in Figure 8.30*a*, traffic is normally carried only on the (unidirectional) working path with the counterrotating path used for protection. In the example, an STS-1 (out of an OC-48) might be carried directly from A to B, but the returning STS-1 would be carried from B through C and D to A. A bidirectional ring, on the other hand, establishes both halves of the duplex connection over the shortest path in the ring. Thus, no fiber is identified as a pure working fiber and another as a pure protection fiber. Because bidirectional rings provide shorter round trip delays for most connections and allow reuse of time slots on the ring, it is the preferred mode of operation for interoffice networks. Rings for subscriber access applications do not carry much traffic between ADM nodes and therefore are more suited to a unidirectional mode of operation.

8.6.1 Unidirectional Path-Switched Ring

As shown in Figure 8.31, a unidirectional path-switched ring (UPSR) [38] transmits the same information from A to B in both directions around the ring. Normally, only the working path is accessed by the receiving node: If a failure occurs, a node can select the data on the protection channel. Notice that in the example shown selection of the protection path actually leads to a shorter path for the connection from A to B.

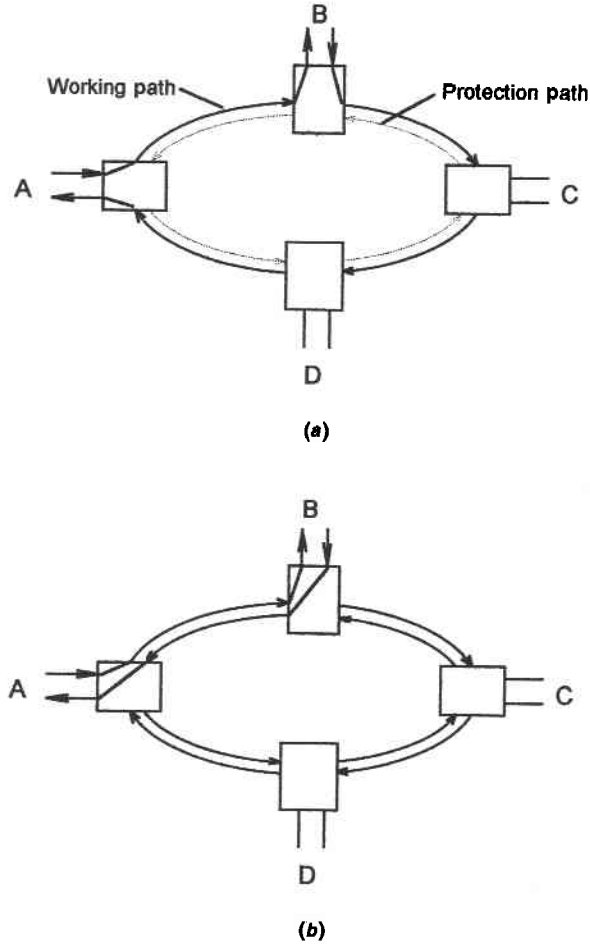


Figure 8.30 (*a*) Unidirectional and (*b*) bidirectional rings.

8.6.2 Bidirectional Line-Switched Ring

Bellcore defines two versions of bidirectional line-switched rings (BLSRs) [39]: a two-fiber BLSR and a four-fiber BLSR. On a two-fiber BLSR protection is provided by reserving bandwidth in each of two counterrotating fiber paths (Figure 8.32). If all traffic is to be protected, only 50% of the total system capacity can be used. Under normal conditions connections between two nodes utilize the shortest path between the nodes. If a fault in either direction of transmission occurs, the nodes adjacent to the fault perform ring switches as indicated. A ring switch involves switching traffic from working channels of the failed facility to spare channels of the other facility on the side of the node on which the fault occurs. The protection-switched traffic propagates all the way around the ring, being ignored by intervening nodes, until it is switched back

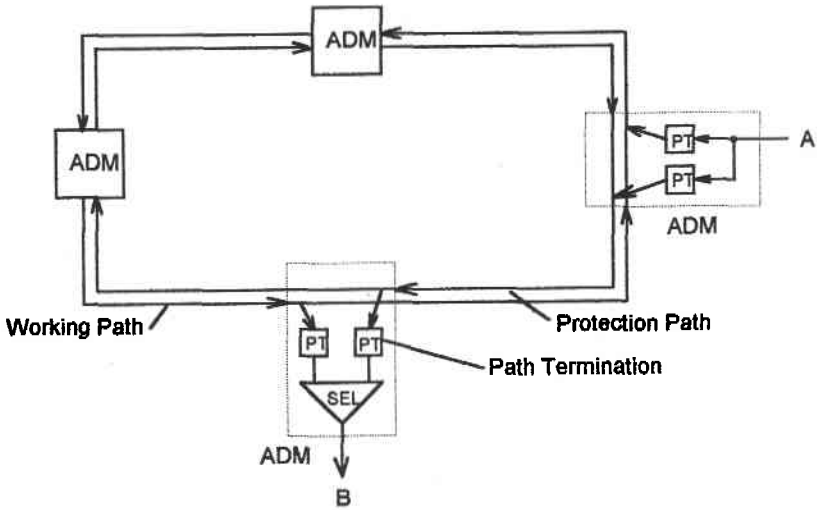


Figure 8.31 UPSR protection switching.

to the working channels by the other node next to the fault. Notice that all nodes (including the nodes adjacent to the fault) communicate on working channels in the same manner as they did before the protection switching. That is, the path terminations are not part of the protection path. The main impact of the protection switch is an increase in delay for affected traffic (and a momentary insertion of extraneous data when the switch occurs).

On a four-fiber BLSR (Figure 8.33) two pairs of fibers are provided for each direction of transmission—one bidirectional working pair and another pair for protection of the first pair. Thus, working and protection channels are carried on different physical facilities. Again, connections are normally set up to use the shortest distance of

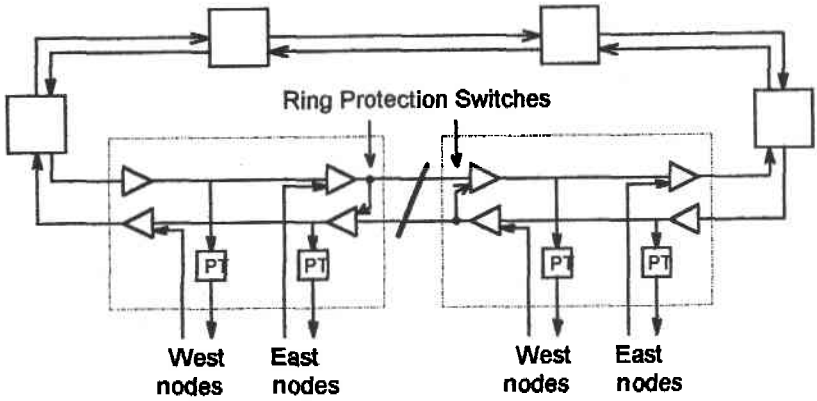


Figure 8.32 Two-Fiber BLSR protection switches.

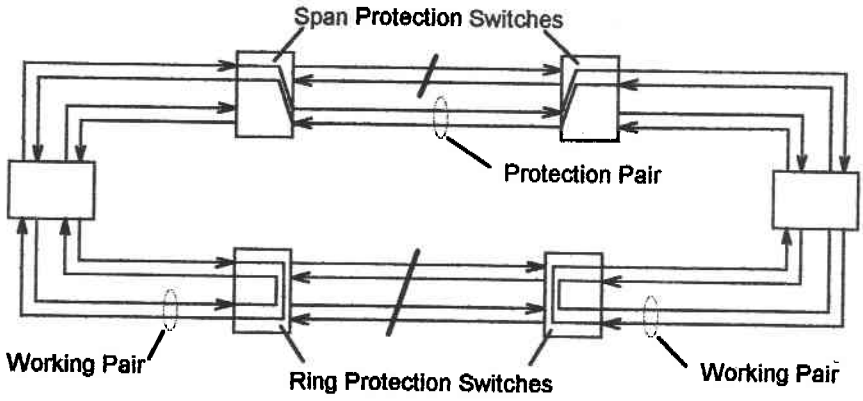


Figure 8.33 Four-Fiber BLSR protection switches. Note that a ring switch and a span switch cannot coexist without channel switching and a reduction in capacity.

travel for each side of a connection. If a failure occurs on only a working facility, protection switching occurs similar to "span switching" of a point-to-point system: The traffic is merely switched to and from the protection facility by nodes adjacent to the fault. However, if a fault affects both the working and the protection facilities, a ring switch is needed as shown. Again, protection-switched traffic propagates all the way around the ring without being accessed by intervening nodes. All traffic accesses still occur on the working channels even though the same information is passing through the nodes in the protection path.

A four-fiber BLSR obviously requires more facilities than a two-fiber BLSR but has numerous advantages. First, the protected capacity of the system is twice as large. Second, fiber failures on only the working pair can be accommodated by a span switch with minimal disruption to traffic. Third, multiple separate failures can occur on working pairs and be accommodated by multiple span switches. Fourth, the presence of a spare pair simplifies maintenance testing and possible upgrading of facilities. For these reasons, a four-fiber BLSR is generally favored.

REFERENCES

1 F. P. Kapron, D. B. Keck, and R. D. Maurer, "Radiation Losses in Glass Optical Waveguides," *Applied Physics Letters*, Nov. 1970, pp. 423–425.

2 T. Miya, Y. Terunuma, T. Hosaka, and T. Miyashita, "Ultimate Low-Loss Single Mode Fibre at 1.55 μm," *Electronic Letters*, Feb. 1979, pp. 106–108.

3 D. Large, "The Star-Bus Network: Fiber Optics to the Home," *CED Magazine*, Jan. 1989.

4 R. Olshansky, V. A. Lanzisera, and P. M. Hill, "Subcarrier Multiplexed Lightwave Systems for Broadband Distribution," *Journal of Lightwave Technology*, Sept. 1989, pp. 1329–1342.

5 T. Li, "Advances in Optical Fiber Communications: An Historical Perspective," *IEEE Journal on Selected Areas in Communications*, Apr. 1983, pp. 356–372.

6 P. S. Henry, "Lightwave Primer," *IEEE Journal of Quantum Electronics*, Dec. 1985, pp. 1862–1877.

7 J. R. Stauffer, "FT3C-A Lightwave System for Metropolitan and Intercity Applications," *IEEE Journal on Selected Areas in Communications*, Apr. 1983, pp. 413–419.

8 P. S. Henry, R. A. Linke, and A. H. Gnauck, "Introduction to Lightwave Systems," in *Optical Fiber Transmission* 11, S. Miller and I. P. Kaminow, Eds., Academic Press, San Diego, 1988, pp. 781–832.

9 D. C. Gloge and I. Jacobs, "Terrestrial Intercity Transmission Systems," in *Optical Fiber Transmission II*, S. Miller and I. P. Kaminow, Eds., Academic Press, San Diego, 1988.

10 D. L. Howells, "High Capacity Lightwave Technology Comes of Age," *AT&T Technology*, Vol. 3, No. 4, 1988, pp. 10–15.

11 W. C. Marra, J. S. Linnell, N. M. Denkin, and K. Ogawa, "FT Series G Lightwave Digital Transmission System Architecture and Upgrade Capability," *IEEE International Conference on Communications*, 1987, pp. 19B.3.1–19B.3.5.

12 K. Ogawa, "Analysis of Mode Partition Noise in Laser Transmission Systems," *IEEE Journal of Quantum Electronics*, May 1982, pp. 849–855.

13 J. E. Goell. "An Optical Repeater with High-Impedance Input Amplifier," *Bell System Technical Journal*, Apr. 1974, pp. 629–643.

14 N. Yoshikai, K. I. Katagiri, and T. Ito, "mB1C Code and Its Performance in an Optical Communications System," *IEEE Transactions on Communications*, Feb. 1984, pp. 163–168.

15 K. Hagishima, Y. Kobayashi, and K. Aida, "High-Speed CMI Optical Intraoffice Transmission System: Design and Performance," *IEEE Journal on Selected Areas in Communications*, Dec. 1986, pp. 1451–1457.

16 F. Ross, "Overview of FDDI: The Fiber Distributed Data Interface," *IEEE Journal on Selected Areas in Communications*, Sept. 1989, pp. 1043–1051.

17 P. Cochrane, R. Brooks, and R. Dawes, "A High-Reliability 565 Mbit/s Trunk Transmission System," *IEEE Journal on Selected Areas in Communications*, Dec. 1986, pp. 1396–1403.

18 R. L. Williamson and M. Chown, "NL1 Submarine System," *IEEE Journal on Selected Areas in Communications*, Apr. 1983, pp. 454–458.

19 P. A. Franaszek and A. X. Widmer, "Byte-Oriented DC Balanced (0.4) 8B/10B Partitioned Block Transmission Code," U.S. Patent 4,488,739, Dec. 4, 1984.

20 A. F. Benner, *Fibre Channel Gigabit Communications and I/O for Computer Networks*, McGraw-Hill, New York, 1996.

21 Y. Niro, Y. Ejiri, and H. Yamamoto, "The First Transpacific Optical Fiber Submarine Cable System," *IEEE International Conference on Communications*, June 1989, pp. 50.1.1–50.1.5.

22 N. Yoshikai, S. Nishi, and J. I. Yamada, "Line Code and Terminal Configuration for Very Large-Capacity Optical Transmission System," *IEEE Journal on Selected Areas in Communications*, Dec. 1986, pp. 1432–1437.

23 H. Kimura and K. Nakagawa, "F-1.6G System Overview," *Review of the Electrical Communications Laboratories NTT Japan*, Vol. 35, No. 3, 1987, pp. 219–225.

24 S. S. Wagner and H. Kobrinski, "WDM Applications in Broadband Telecommunications Networks," *IEEE Communications Magazine*, Mar. 1989, pp. 22–30.

25 R. J. Sanferrare, "Terrestrial Lightwave Systems," *AT&T Technical Journal*, Jan./Feb. 1987.

26 S. S. Wagner, H. L. Lemberg, H. Kobrinski, L. S. Smoot, and T. J. Robe, "A Passive Photonic Loop Architecture Employing Wavelength-Division Multiplexing," *IEEE Globecom Proceedings*, 1988, pp. 48.1.1–48.1.5.

27 N. K. Dutta, "III-V Device Technologies for Lightwave Applications," *AT&T Technical Journal*, Jan./Feb. 1989, pp. 5–18.

28 P. R. Trischitta and D. T. S. Chen, "Repeaterless Undersea Lightwave Systems," *IEEE Communications Magazine*, Mar. 1989, pp. 16–21.

29 S. C. Mettler and C. M. Miller, "Optical Fiber Splicing," in *Optical Fiber Transmission* 11, S. Miller and I. P. Kaminow, Eds., Academic Press, San Diego, 1988, pp. 263–300.

30 J. B. Haber and J. W. Rogers, "Fiber in the Loop," *AT&T Technology*, Vol. 3, No. 4, 1988, pp. 2–9.

31 N. Yamaguchi, K. Yarnane, and T. Kihara, "565-Mbit/s Optical Fiber Transmission System," *Fujitsu Science and Technology Journal*, Mar. 1989, pp. 60–63.

32 "Synchronous Optical Network (SONET)—Basic Description Including Multiplex Structure, Rates and Formats," T1.105-1995, ANSI, New York, 1995.

33 "Synchronous Optical Network (SONET)—Payload Mappings," T1.105.02-1995, ANSI, New York, 1995.

34 Recommendation G.707, "Network Node Interface for the Synchronous Digital Hierarchy (SDH)," ITU-T, Geneva, Switzerland, Mar. 1996.

35 N. B. Sandesara, T. H. Jones, and A. G. Edwards, "SONET Intra-Office Interconnect Signal," *IEEE Globecom Proceedings*, 1988, pp. 30.5.1–30.5.7.

36 "Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria," GR-253-Core, Bellcore, Morristown, NJ, Dec. 1995.

37 N. Zhang, K.-H. Liu, and E. C. Posner, "Reliable Payload Pointer Protocol for Synchronous Optical Network," *IEEE International Communications Conference Proceedings*, 1989, pp. 14.2.1–14.2.5.

38 "SONET Dual-Fed Undirectional Path Switched Ring (UPSR) Equipment Generic Criteria," GR-1400-Core, Bellcore, Jan. 1999.

39 "SONET Bidirectional Line-Switched Ring Equipment Generic Criteria," GR-1230-Core, Bellcore, Dec. 1993.

PROBLEMS

8.1 Determine the attenuation α in decibels per kilometer such that the loss limit of a 500-Mbps fiber system is exactly equal to the dispersion limit. Assume the transmitter output is 30 dB above the receiver threshold for the desired error rate and that the system has a BDP of 80 Gbps-km.

8.2 An 850-nm, graded-index, multimode, 50-Mbps fiber system with 2 dB/km loss in the fiber is to be used for a token-passing bus local area network. Assume the system uses the first entries in Tables 8.1 and 8.2 as a source–detector pair. If the BDP of the fiber is 500 Mbps-km, how many passive taps with 0.4 dB of loss can be inserted per kilometer without affecting the distance between transmitters and receivers.

8.3 What is the distance limit (without taps) of the system in Problem 8.2 if the data rate is reduced to 10 Mbps?

8.4 A graded, multimode fiber exhibits 100 psec/km nm of chromatic dispersion at 850 nm. Determine the BDP of a NRZ system using an LED having 40 nm of spectral width.

8.5 A WDM SMF system using 1300 and 1550 nm is dispersion limited at both wavelengths. How must the spectral bandwidths of the two sources be related? (Assume $D = 3.5$ psec/km nm at 1300 nm and $D = 17.5$ psec/km nm at 1550 nm.)

8.6 What is the minimum and maximum pulse density of mB1p and mB1C line codes?

8.7 What is the longest string of no pulses that can occur in mB1P and mB1C line codes?

8.8 A 4B6B line code can encode all 4 bits of source data with 6-bit line codes having exactly three pulses in every code. How many dc-balanced codes are left over for link control?

8.9 Determine the minimum and maximum data rate of an asynchronously mapped DS1 signal that can be carried within a VT1.5 envelope.

8.10 Determine the minimum and maximum data rates of an asynchronously mapped E1 signal that can be carried within a VT2 envelope.

8.11 Determine the minimum and maximum data rates of an asynchronous DS3 payload that can be carried within an STS-1 envelope.

8.12 Determine the minimum and maximum E4 data rates that can be accommodated within an STS-3c envelope.

8.13 Determine the average frame acquisition time of an STS-1 signal assuming all nonframing bits have equally likely random values of 0 or 1.

8.14 Repeat Problem 8.13 for an STS-3c signal.

8.15 Determine the payload capacity of a STS-3c (STM-1) signal.

9

DIGITAL MOBILE TELEPHONY

A basic concept of a cellular system is to provide ever-increasing capacity by dividing cells into smaller and smaller sizes to increase frequency reuse. Unfortunately, the cell division concept has proven to be impractical in terms of finding suitable locations for base station antennas and for getting repeated construction authorizations from governing organizations. The explosive demand for mobile telephones in the early 1990s within the United States and elsewhere in the world helped stimulate the development of new systems to accommodate the demand. Two basic approaches have been pursued: expanding the channel capacity of existing systems and allocating new frequency bands to cellular mobile phone service. All of the new systems utilize digital transmission in lieu of the analog FDM transmission used by analog cellular systems such as AMPS in North America, TACS in Great Britain, and NMT in Scandinavia.

Commercially viable digital cellular systems are enabled by the availability of low-cost digital signal processing technology to provide solutions to problems in several key areas. First, it is necessary to compress a digital speech signal into a low enough bit rate that digital voice transmission does not impose a spectrum penalty with respect to analog systems. As discussed in Chapter 3, speech compression algorithms have advanced to the point that digitization can, in some cases, provide greater spectrum efficiency than analog systems. Second, the application of digital transmission to a mobile environment requires a sophisticated equalizer to overcome the effects of dynamic multipath transmission impairments. Third, the susceptibility of the speech compression algorithms to channel errors requires sophisticated error correction and control measures. Lastly, the low-bit-rate voice coders and the digital system architectures introduce significant artificial delay into the voice channel, which imposes the need for echo cancelers for acceptable voice quality.

9.1 NORTH AMERICAN DIGITAL CELLULAR

North American Digital Cellular (NADC), also known as US Digital Cellular (USDC) or Digital-AMPS (D-AMPS), represents a digital upgrade from the previously deployed analog cellular system referred to as Advanced Mobile Phone Service

(AMPS). The D-AMPS system is designed to be compatible with AMPS. In fact, a D-AMPS system installation can coexist with an AMPS installation, thus allowing a graceful migration from an all-analog service to an all-digital service. An analog-to-digital migration is supported by a dual-mode phone that can operate as an AMPS phone in one call and as a D-AMPS phone on the next call. D-AMPS is standardized by EIA/TIA as Interim Standards IS-54 and IS-136.

9.1.1 D-AMPS Transmission Format

The most significant aspect of maintaining compatibility with AMPS is the need to adhere to the AMPS FDM channel structure. This channel structure uses 30-kHz-wide channels in the range 824–894 MHz. Within each 30-kHz FDM channel IS-54 defines six digital channels operating in a time division multiple-access (TDMA) mode of operation, as shown in Figure 9.1. Transmission from a base station to the mobiles is accomplished with a continuous TDM stream with six time slots. Transmission from each of the mobiles occurs in data bursts that are timed to arrive at the base station in separate, nonoverlapping time slots synchronized to the outgoing time slots. Associated with each burst from a mobile is a guard time to prevent overlap and provide a transmitter ramp-up preceding the data. The guard time between time slots is minimized by adjusting the transmit time of the mobiles with control messages from the base station. These adjustments are dynamic to accommodate mobility.

The TDMA digital transmission frame format within each 30-kHz channel contains six time slots with a total of 1944 bits. The repetition rate of each frame is 25 frames per second, which leads to an aggregate bit rate of 48.6 kbps in the 30 kHz of bandwidth. The modulation format is $\pi/4$ shifted, differentially encoded, quadrature phase shift keying. This format is essentially 4-PSK modulation with two four-point constellations offset from each other by $\pi/4$ radians. By alternating between constellations, a symbol transition of at least $\pi/4$ radians is always assured—a property that helps in symbol clock recovery.

Full-rate voice coding utilizes two time slots in each frame for the voice information. Thus, the system capacity with full-rate voice coding is three times that of an AMPS system since there are three TDM voice channels within each FDM channel.

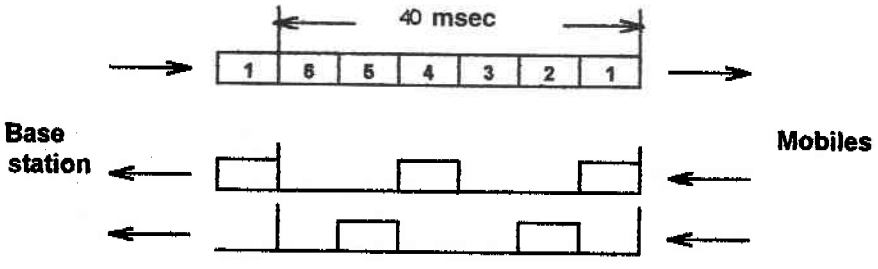


Figure 9.1 TDMA transmission format of North American digital cellular.

If half-rate voice coding gets implemented in the future, the capacity expansion will be sixfold.

9.1.2 D-AMPS Speech Coding

The speech-coding algorithm is vector sum excited linear predictive (VSELP) coding [1], which is described in Chapter 3. The D-AMPS VSELP algorithm processes the speech waveform in segments of 20 msec duration. Each speech segment is represented by 159 bits. Since there are two VSELP frames in each TDMA frame, the raw data rate of the voice is $2 \times 159 \times 25 = 7950$ bps. To the raw bit date is added 5050 bps of redundancy encoding for error correction and detection to produce a composite, aggregate data rate of 13 kbps for a voice channel. As shown in Figure 9.2, there are 260 data channel bits and 64 overhead bits in each time slot. Table 9.1 describes the basic use of each data field within the time slots.

9.1.3 D-AMPS Control Channel

In addition to providing a threefold capacity expansion, the digital nature of D-AMPS provides other advantages that are not possible or at least difficult to achieve in an analog system such as AMPS. The first of these is use of the CDVCC channel to ensure that a base station maintains connections with intended mobiles. AMPS utilizes a similar feature with supervisory audio tones (SATs). A SAT is a tone at 5970, 6000, or 6030 Hz that is inserted and removed from the audio signal specifically for detecting fades and ascertaining connection integrity. The availability of only three tones and the complexity of inserting, detecting, and repeating these tones are significant limitations of AMPS.

A more significant advance provided by the digital nature of D-AMPS involves the use of the SACCH channel embedded in each time slot. Because this channel is always present, it is quite useful for communicating control and supervision information while speech is actively in progress. Specific, advantageous uses of this control chan-

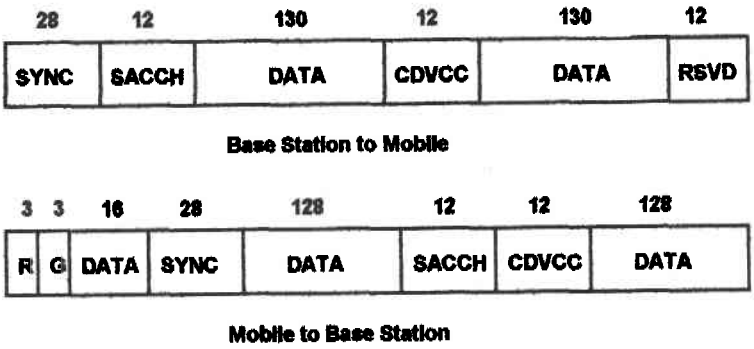


Figure 9.2 Time slot format of North American digital cellular.

TABLE 9.1 Data Field Functions of D-AMPS Time Slots

CDVCC	Coded digital verification color code. A unique code sent by a base station and returned by each mobile for base station confirmation of connection integrity.
DATA	Application bearer channel bits (voice or data). Can also be used for a fast associated control channel (FACC) when there is no active application or a situation arises when application transmission needs to be usurped.
G	Guard time. Mobile transmitter is off.
R	Ramp time. Mobile transmitter ramps up to assigned power level.
RSVD	Reserved (unused).
SACCH	Slow associated control channel. A continuous channel used to send control and supervisory information.
SYNC	Synchronization channel. Used for synchronization, equalizer training, and time slot identification.

nel involve authentication, additional connection integrity, transmit power control, channel quality measurement reports, handoffs to a new cell, keypad depressions, and calling number identification. The SACCH control channel is also used for timing adjustments specific to the TDMA operation.

A particular example of the usefulness of the SACCH channel is its support of mobile assisted hand off (MAHO). If a mobile with an established connection moves from one cell to another, the process of handing the mobile off is performed with much more control and reliability than is possible in AMPS. The MAHO process begins by the base station telling the mobile to make channel quality measurements on the current channel and on candidate channels for a potential handoff. Channel quality measurements involve received signal power levels and bit error rates (BERs). The TDMA nature of D-AMPS facilitates measurements of candidate channels by tuning to the candidate frequency during an inactive time slot. After each set of measurements the mobile sends the results to the base station (via SACCH) whereupon the base station can determine if a handoff is justified.

9.1.4 D-AMPS Error Control

There are three mechanisms incorporated into D-AMPS for mitigating the effects of channels errors: error correction, error detection, and interleaving. Error correction is implemented with a half-rate convolution coder for the perceptually most significant bits of the voice. There are 77 such bits out of a frame size of 159 bits. The half-rate convolution coding process therefore adds 77 more bits to the channel. Of the 77 bits, 12 are particularly important. A 7-bit CRC check sum is added for these bits to determine if any of these 12 bits are received in error. When a CRC error occurs, certain critical parameters from the previous error-free frame are used to reconstruct speech to avoid use of aberrant values. If several CRC errors are received in successive frames, the reconstructed speech is muted. The general term for these operations is *bad frame masking*.

The third error control mechanism involves separating the data in a single speech frame, interleaving it with data from adjacent speech frames, and transmitting it in two

time slots. This process reduces the possibility that a burst of errors will circumvent the error correction capabilities of the convolutional coding. A drawback to interleaving is the delay it adds to the channel, which must be accounted for in echo cancelers.

9.2 GLOBAL SYSTEM FOR MOBILE COMMUNICATIONS

Global System for Mobile Communications (GSM) [2] is a cellular mobile communications system developed in Europe and standardized by the European Telecommunication Standards Institute (ETSI). GSM has subsequently been adopted worldwide as the international digital mobile standard. Initial work on GSM standardization began in 1982. The first field trial of a GSM system occurred in 1991.

The International Telecommunication Union (ITU) allocated frequency spectrum at 935–960 MHz for the downlink (base station to mobile) and 890–915 MHz for the uplink (mobile to base station). Even though some of this spectrum was being used by established analog systems, there was no attempt to be compatible with the existing frequency plan. (The incumbent analog systems in the various European countries were incompatible with each other and a main goal of GSM was to establish a common standard so backward compatibility was not a consideration.)

9.2.1 GSM Channel Structure

Starting with an open-spectrum plan provided GSM system designers with more freedom in system design than was available to D-AMPS designers. The most significant difference in GSM with respect to D-AMPS is the use of 200-kHz-wide digital RF channels as opposed to the 30-kHz-wide D-AMPS RF channels. Each GSM RF channel operates at 270.833 Mbps using GMSK modulation. As mentioned in Chapter 6, GMSK modulation is fairly closely related to 4-PSK modulation as used in D-AMPS. GMSK modulation does, however, require more bandwidth than tightly filtered 4-PSK, as evidenced by an information density of 270.883/200 = 1.35 bps/Hz for GSM and 48.6/30 = 1.62 bps/Hz for D-AMPS. The GMSK modulation format of GSM provides a constant-envelope RF signal that is more efficient for RF power generation than is tightly filtered 4-PSK modulation used by D-AMPS. This efficiency is most important for hand-held battery life.*

As shown in Figure 9.3, a GSM RF channel utilizes digital TDMA with eight full-rate voice channels, in contrast to three full-rate voice channels supported by a single D-AMPS RF channel. The ability to terminate more TDMA channels on a single transceiver provides a cost advantage for GSM base stations. (The cost of a 270.833-kbps TDMA transceiver is no different than the cost of a 48.6-kbps TDMA transceiver.)

The GSM system carries eight full-rate voice channels in 200 kHz of bandwidth, which amounts to 25 kHz per voice channel, a spectrum efficiency identical to Euro-

*As discussed below, GSM uses less speech compression than does D-AMPS, which leads to a higher rate digital channel for voice and in turn requires more transmit power.

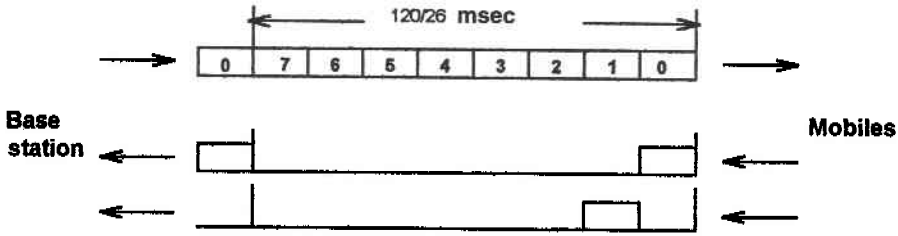


Figure 9.3 TDMA transmission format of GSM.

pean analog FDM systems of the time. Thus, the introduction of the GSM system did not provide spectrum efficiency improvements as D-AMPS did. A GSM system does, however, provide cellular system efficiencies in that digital transmission, in general, and strong error correction, in particular, allow operation at lower signal-to-noise ratios. Greater noise or interference tolerance leads to longer transmission distances and/or greater amounts of frequency reuse.

The burst period of a GSM system is 120/26/8 = 15/26 ms. This burst period is derived from a 120-ms superframe consisting of 26 TDMA frames and 8 bursts per TDMA frame. Twenty-four frames of a 26-frame super frame are allocated to traffic (e.g., voice) transmission while one of the frames is allocated to a SACCH control channel for each traffic channel. The last TDMA frame of a superframe is reserved. A unique aspect of GSM, with respect to D-AMPS, is that a TDMA burst format is used in both directions of transmission, as opposed to only on the uplink from the mobile to the base station. The format of these bursts is shown in Figure 9.4, where it can be seen that there are 148 bits of data and an idle guard time corresponding to the period of 8.25 bits. The burst transmission rate of a traffic channel can now be determined as 156.25/15/26 = 270.833 kbps. The fields within the burst are identified in Table 9.2.

The use of multiple bursts in the downlink direction, as opposed to continuous transmission, is advantageous in that it inherently allows turning off the base station transmitter during idle channels, which in turn reduces the total amount of interference between cells in a widespread and congested installation. An advantage of continuous transmission, as used in D-AMPS, is the relative ease of implementation and greater performance of the digital receiver in the mobile.

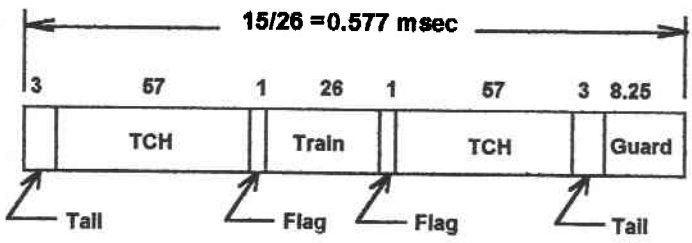


Figure 9.4 Time slot format of GSM burst.

TABLE 9.2 Data Field Functions of GSM Time Slot

Flag	A single bit used to signify voice or FACCH content in an associated TCH field
Guard	Idle period of 8.25 bits intervals timing margin between bursts
Tail	3 "0" bits for equalizer training
TCH	Field for transporting bearer data or FACCH data
Train	Field of fixed data pattern used to train equalizers and acquire a data clock for the entire burst

9.2.2 GSM Speech Coding

GSM uses regular pulse excited–linear predictive coding with a long-term predictor loop (RPE-LTP) [3]. The RPE-LTP algorithm is described in Chapter 3. Speech is divided into 20-msec samples, each of which is encoded as 260 bits, giving a total bit rate of 13 kbps. This is the original, full-rate speech-coding algorithm. An enhanced full-rate (EFR) speech-coding algorithm has been implemented by some North American GSM 1900 operators. EFR is said to provide improved speech quality using the existing 13-kbps bit rate.

9.2.3 GSM Channel Coding and Modulation

GSM utilizes error control mechanisms similar to D-AMPS. First of all, the 260 bits of a speech frame are divided into three classes:

Class Ia, 50 bits—most sensitive to bit errors

Class Ib, 132 bits—moderately sensitive to bit errors

Class II, 78 bits—least sensitive to bit errors

Class Ia bits have a 3-bit CRC added for error detection. If an error is detected, the frame is replaced by a slightly attenuated version of the previous correctly received frame. The 50 Class Ia, 3 CRC, 132 Class Ib, and a 4-bit tail sequence (a total of 189 bits) are processed by a half-rate convolutional encoder for error correction. The output of the convolutional encoder is added to the 78 Class II bits to produce an aggregate speech frame of 456 bits. Thus, the redundantly encoded speech rate is 456/0.020 = 22.8 kbps.

To further protect against the burst errors, each sample is interleaved. The 456 bits of output by the convolutional encoder are divided into eight blocks of 57 bits, and these blocks are spread across eight consecutive time slot bursts. Since a time slot burst carries two 57-bit blocks, each burst contains traffic from two different speech samples.

9.2.4 GSM Mobile Station

The GSM mobile station (MS) consists of the mobile equipment (the terminal) and a smart card called the Subscriber Identity Module (SIM). The SIM provides personal mobility, so that the user can have access to subscribed services irrespective of a specific terminal. By inserting the SIM card into another GSM terminal, the user is able

to receive calls at that terminal, make calls from that terminal, and receive other subscribed services.

The mobile equipment is uniquely identified by the International Mobile Equipment Identity (IMEI). The SIM card contains the International Mobile Subscriber Identity (IMSI) used to identify the subscriber to the system, a secret key for authentication, and other information. The IMEI and the IMSI are independent, thereby allowing personal mobility. The SIM card may be protected against unauthorized use by a password or personal identity number.

9.2.5 GSM Frequency Hopping

The mobile station is inherently frequency agile, meaning it can move between a transmit, receive, and monitor time slot within one TDMA frame, all of which are normally on different frequencies. GSM makes use of this inherent frequency agility to implement slow frequency hopping—the mobile and BTS transmit each TDMA frame on a different carrier frequency. The frequency-hopping algorithm is broadcast on the broadcast control channel. Since multipath fading is dependent on carrier frequency, slow frequency hopping helps alleviate the problem. In addition, cochannel interference is more random than when fixed-frequency allocations exist.

9.2.6 GSM Short Message Service

Short Message Service (SMS) is an integrated bidirectional messaging service that allows GSM cellular subscribers, and various PCS offerings, to send and receive data. Individual messages (with GSM) can be up to 160 bits in length. Because the SACCH is used for SMS data transmission, messages can be received or transmitted during a voice call. Initial applications of SMS focused on alphanumeric paging services with fundamental differences: SMS is bidirectional and message delivery is guaranteed. Subsequent applications served by SMS are voice mail notification, e-mail delivery, stock quotes, and downloading/updating of SIM cards.

9.3 CODE DIVISION MULTIPLE-ACCESS CELLULAR

The two digital cellular systems discussed in the previous sections, D-AMPS and GSM, utilize a combination of frequency division multiplexing and time division multiplexing* as a method of partitioning a block of allocated frequency spectrum into individual communication channels. This section discusses systems with a fundamentally different approach to channel definition—code division multiple access (CDMA)—that belong to a class referred to as spread spectrum communications systems [4]. The term spread spectrum refers to the fact that transmission bandwidth

*Within a particular GSM cell it is conceivable that a single FDM channel supporting eight TDMA channels is sufficient for sufficiently low traffic situations. In this case, the particular cell utilizes only TDMA. Nevertheless, the mobiles still support FDM operations so they can move to different cells and switch frequencies.

used by an individual channel is much wider than the inherent bandwidth required by the message being transmitted. Spread spectrum systems have traditionally been used in military applications where the increased complexity of implementation is justified by two particular features. First, it is relatively difficult to detect the presence of a spread spectrum signal because the signal energy is spread across a wide band and is often masked by background noise. Second, it is more difficult to jam a spread spectrum signal because the jamming signal energy must be spread across a wide band as opposed to being focused into a relatively narrow band. The performance of a spread spectrum receiver is comparable to the performance of receivers for traditional narrowband signals as long as the spread spectrum receiver knows and can synchronize to the method being used to spread the spectrum.

Two general categories of spread spectrum communications are frequency-hopping systems and direct-sequence modulation systems. A frequency-hopping system is one in which transmission at any particular instant is confined to a relatively narrow band of frequencies commensurate with the inherent bandwidth of the message. Instead of staying within one particular band as a conventional communications system does, a frequency-hopping system jumps between narrow bands of frequencies within a large block of spectrum in some prescribed manner. As mentioned in the previous section, a GSM cellular system has the ability to operate with frequency hopping (specifically, slow frequency hopping).

9.3.1 CDMA Channel Establishment

A contemporary cellular CDMA system, as developed by Qualcomm and standardized by EIA/TIA in Interim Standard IS-95 [5], uses direct-sequence spectrum spreading. One particular method of implementing a direct-sequence spectrum spreading system is shown in Figure 9.5. In this system the source data is "exclusive ored" with a relatively long digital codeword. In essence, the "exclusive or" process replaces a 1 of the source data with the given codeword while a 0 of the source data is replaced with the bitwise complement of the codeword. If a codeword contains n bits, the occupied spectrum of the transmitted signal is n times as wide as if the source data

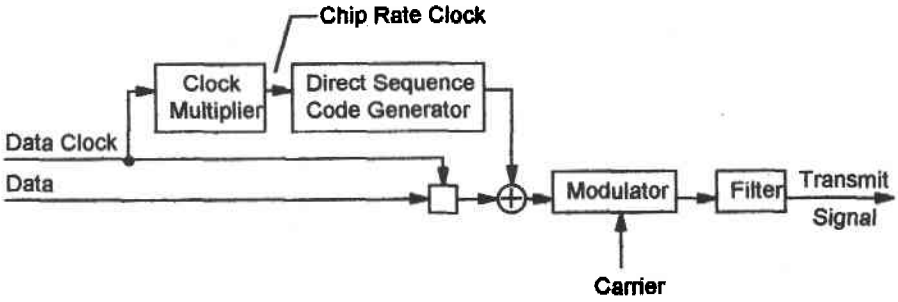


Figure 9.5 Direct-sequence generation of spread spectrum signal.

were directly transmitted. Signals of other CDMA channels occupy the same band of frequencies but do so with different spectrum-spreading codes, which allows separation of the signals in the signal processing circuitry of a receiver.

The basic process of separating CDMA channels in a receiver involves correlating a received signal with each of the various codewords (i.e., channels) assigned to the cell. The correlation process produces a correlation measurement by subtracting the number of mismatches in a codeword from the number of matches. Table 9.3 lists an example set of codewords with particularly useful correlation properties. As indicated in the table, a codeword has seven matches with itself and no mismatches for a net correlation of +7. A source data value of 0 produces no matches and seven mismatches for a net correlation of −7. The measurement weights for all other codewords are either +1 or −1, depending on the values of the source data. If all seven channels defined in Table 9.3 are active in a single cell, the worst-case interference between the codes produces an interference value of either +6 or −6. Thus, the desired data can be recovered with a discrimination threshold of 0 for each channel. (The worst-case net measurement values are +1 for a 1 and −1 for a 0.) Notice that same correlation properties exist for all channel codes with respect to the other codes in the table.

Example baseband waveforms for the seven-channel CDMA system of Table 9.3 are shown in Figure 9.6. Notice the channel 0 receiver measurement is positive, which implies a data value of 1. In this example interference from the adjacent channels actually enhanced the channel 0 measurement from an expected value of 7 to 9.

Discriminating between nominal measurement values of +7 for a desired signal and +6 for worst-case interference is obviously very tenuous, particularly because the individual channels will be received with different power levels. It is important to note, however, that an interference value of +6 can arise only if the data values of all interfering channels destructively coincide. On average, the composite interference has an average value of 0. The length of a spectrum-spreading code in IS-95 is actually 64 bits long, which means that the example worst-case interference would be 63 in relation to a desired channel value of 64. The chances of destructive interference from all 63 channels, or even a large number of channels, is astronomically small.

TABLE 9.3 Example Direct-Sequence Spectrum-Spreading Codes[a]

Channel Number	Channel Code	Number of Matches	Number of Mismatches	Net Correlation
0	1110010	7	0	+7
1	0111001	3	4	−1
2	1011100	3	4	−1
3	0101110	3	4	−1
4	0010111	3	4	−1
5	1001011	3	4	−1
6	1100101	3	4	−1

[a]Spectrum-spreading code of desired channel: 1110010.

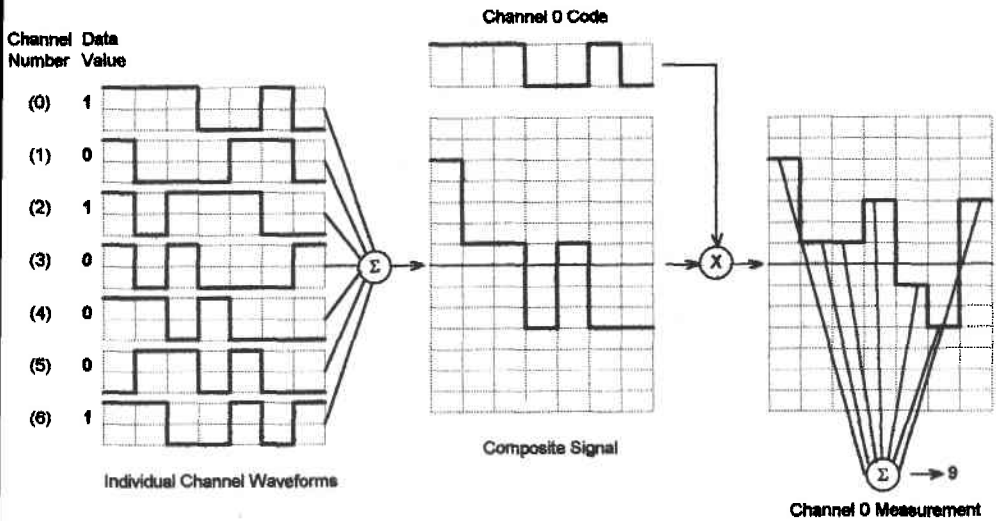


Figure 9.6 Example seven-channel CDMA encoding and decoding.

Example 9.1. Determine the probability of maximum interference of a 64-channel CDMA system with 64-bit spreading codes. Also determine the effective signal-to-interference power ratio of the same CDMA system. Assume all channels operate at the same effective power level at the receiver and that all channel codes have a cross correlation of ±1 bit.

Solution. The probability of 63 destructive interferers is merely the probability of occurrence of 63 equally likely binary events: prob(max interference) = $(0.5)^{63} = 1 \times 10^{-19}$. The value of a desired receive signal is the autocorrelation of a codeword with itself and can therefore be represented as a value of 64. The interference level is the sum of 63 binary random numbers with equally likely values of ±1. Even though a single interferer does not produce a Gaussian probability distribution, the sum of a large number of independent random variables approaches a Gaussian distribution (central limit theorem). The mean and variance of an individual interferer are easily determined to be 0 and 1, respectively. The mean and variance of a sum of 63 such variables are 0 and 63, respectively. The signal-to-interference ratio is now determined as

$$\text{SIR} = 10 \log_{10} \frac{64^2}{63} = 18\text{dB}$$

As presented in Chapter 6, the 18-dB SIR result of Example 9.1 is quite sufficient to support an acceptable error rate if the effective power level of all channels can be

maintained to be equal. In actual practice, a CDMA deployment does not use all possible codes in a single cell, just as an FDM system does not use all frequencies in one cell. Thus, the amount of interference in a cell is limited by the number of codes assigned to the cell and a lesser amount of interference from adjacent cells (assuming the adjacent cells do not use identical, synchronous spectrum-spreading codes). The effect of varying power levels in the interfering channels is covered in some problems at the end of the chapter.

9.3.2 CDMA Multipath Tolerance

A primary advantage of a CDMA transmission system is its robustness in the presence of multipath conditions. The basic reason for multipath tolerance can be appreciated by examining the example codes given in Table 9.3. Notice that each code is a cyclic shift of all other codes in the table. Because the selected codes have low correlation with each other, a delayed version of any particular code has the same low correlation with an undelayed version of itself. Thus, the effect of a multipath delay of more than one spread spectrum bit (referred to as a chip) is no more than the effect of the interference from another CDMA channel, even if the delayed version is at the same power level as the primary signal.

The effect of a multipath condition on a D-AMPS channel may be much more deleterious. Because a D-AMPS system operates at a relatively narrow band of frequencies, it is possible that a complete fade occurs for a particular channel at a particular physical location. If the user moves through the physical location, the effect of the fade is a momentary dropout. If the user dwells at the location, the connection can be lost unless a handoff occurs (to a new frequency and/or an adjacent cell). The slow frequency-hopping feature of GSM ameliorates the effect of a complete multipath fade to a greater degree, but not as effectively as a CDMA system. In GSM there may be momentary dropouts, but they are not long enough to cause a dropped connection and do not require a handoff, even if a user dwells in a physical location where one particular frequency is totally lost. In essence, the frequency-hopping process of GSM is equivalent to repeated and automatic handoffs to different frequencies.

From a somewhat philosophical point of view, the basic intent of a CDMA system is to equalize the performance of all channels in the system. In a FDM/TDMA system it is likely that some channels operate with very high performance while others operate at very low performance or cannot be used at all. The existence of high-performance channels does not compensate for the existence of low-performance channels. Thus, a system that equalizes the performance of all channels has a greater total capacity. The primary reason for variable channel performance in FDM/TDMA is multipath fading. An FDM/TDMA transmitter typically operates with a certain amount of excess power referred to as fade margin. The excess power is not much of a problem with adjacent-channel interference because it is relatively easy to isolate FDM channels with frequency guard bands and TDM channels with time guard bands. Cochannel interference from one cell to another is the crux of the problem. If a particular channel

is operating at an excess power level, that channel cannot be reused except at a relatively larger distance.

9.3.3 CDMA Power Control

Effective transmit power control of the mobile units in a CDMA system is both a requirement and a benefit. It is a requirement because a mobile transmitter that is close to a base station receiver will obliterate other mobiles in the same cell that are farther away. This is referred to as the *near–far problem* of CDMA. If the transmit power levels of all mobiles in a CDMA system are controlled to be no higher than absolutely necessary, the channels can be reused more often. Although power control is used in FDM/TDMA systems, it is not possible to operate with bare minimum power levels because the system cannot respond fast enough to adjust the power levels for fast multipath fades. A side benefit of using minimum power levels in a CDMA mobile is increased battery life.

The power levels of a mobile are controlled in two ways: open loop and closed loop. In the open-loop mode, a mobile can determine its transmit power level by measuring its received power level under the assumption that transmission losses are equal in both directions. This assumption is reasonable for a CDMA system but not for a FDM/TDMA system because the latter are much more susceptible to independent frequency-selective fading (multipath). Closed-loop power control involves base station measures of mobile received power and adjustments to the mobile power levels with the control channel. Power control in IS-95 CDMA is described in references [6–8].

Table 9.4 lists the basic parameters of the IS-95 CDMA digital cellular system for the 800-MHz cellular band. The speech compression algorithm uses QCELP (Qualcomm code excited linear prediction). The first commercial installation of CDMA occurred in Hong Kong in 1995.

9.3.4 CDMA Soft Handoff

A unique feature of a CDMA system is the ability of a mobile to simultaneously receive from more than one source. Because each cell in a CDMA cellular network

TABLE 9.4 IS-95 CDMA Mobile Telephone System Parameters

Channel bandwidth	1.25 MHz
Voice-Coding Rate	9.6 kbps (maximum)
Error control overhead	9.6 kbps (downlink)
	19.2 kbps (uplink)
Aggregate channel rate	19.2 kbps (downlink)
	28.8 kbps (uplink)
Code length	64 chips
Spread spectrum channel rate	1.2288 Mbps
Modulation format	QPSK (base to mobile), offset QPSK (mobile to base)

transmits in a common frequency, a common RF receiver inherently receives the spread spectrum signal from all adjacent base stations. Signals from multiple base stations can then be acquired with multiple baseband code correlators or by time sharing a single correlator with the separate codes. In a soft handoff operation the same voice signal is distributed to selected cells adjacent to a currently active cell. An active mobile can then compare the quality of the signals and switch to the best one before disconnecting from the established base station.

A "make-before-break" operation is not feasible in an FDM/TDMA system wherein adjacent cells utilize separate frequencies that require separate RF receivers. FDM/TDMA systems inherently use hard handoffs, which require disconnecting from one base station before connecting to a new one. Notice, however, that a soft handoff operation in a CDMA system increases the background interference because of the multiple active signals for a single connection.

9.4 PERSONAL COMMUNICATION SYSTEM

A personal communication system (PCS) is a cellular system operating in a band of frequencies at 1.9 GHz. The original concept for PCS included multiple, new features and services beyond those offered by a basic cellular system. Some of the envisioned features were single telephone number for multiple services (voice, data, fax) and user mobility for home or office use and location determination. Although some commercial PCS offerings provide some new user features, initial North American PCS systems are basically cellular systems utilizing a new band of frequencies.

When the FCC allocated the PCS frequencies for the United States, they did so without stipulating which type of system should be deployed. Thus, any organization that bids for and receives a franchise for PCS spectrum is free to choose whatever type of system it wants for providing service to the public. As a result, North American PCS systems have been developed with three different transmission formats: D-AMPS, GSM, and CDMA. The D-AMPS implementation follows EIA/TIA standard IS-136, which is basically a revision of IS-54 that incorporates digital control channels. (IS-54 defines the use of an analog control channel for compatibility with AMPS.)

9.5 VOICE PRIVACY AND AUTHENTICATION

Ensuring privacy of conversations and preventing fraud are two critical aspects of cellular telephone systems that are addressed more completely in the digital systems than in the original analog systems. An FDM analog system is particularly vulnerable to casual eavesdropping because a relatively simple scanner can be used to identify an active channel and tune into the conversation. If the scanner has two receivers, the eavesdropper can listen to both sides of the conversation, assuming the eavesdropper is in a high enough location or close enough to the active mobile that it can receive the uplink signal.

Digital signals are inherently more complicated to intercept because an eavesdropper may need to monitor the connection establishment process and not just tune in after a conversation has started. Explicit encryption parameters for optional voice privacy is established during call establishment, or possibly during a conversation, by exchanging privacy control messages in the control channel. An eavesdropper must receive the relevant information sent in both directions before eavesdropping is possible. Even without explicit encryption, an eavesdropper must be located somewhere near the base station to properly receive both sides of a TDMA conversation. At other locations within a cell the eavesdropper will encounter overlapping time slots that will inflict bit errors into the eavesdropper's received data. A CDMA system produces a similar effect when the phase of two spectrum-spreading codes from two separate mobiles coincide.

Authentication of a valid mobile station is significantly strengthened in digital systems. A major problem in analog cellular systems is the process of cloning, in which disreputable persons monitor call establishment handshake procedures to acquire valid mobile equipment electronic serial numbers (ESNs) and program them into counterfeit mobile units that are then used to place calls. Part of the improved fraud prevention involves maintaining a more up-to-date database of valid mobile ESNs. The strongest part of the fraud prevention involves determining authorization codes based on past call history in both the base station and the mobile. The authorization code determined by the mobile and sent to the base station must coincide with the authorization code calculated by the base station before service is allowed. Additional steps may require entry of access codes by the user or, in the case of GSM, the magnetic data card called the Subscriber Identification Module (SIM).

9.6 IRIDIUM

Iridium is a satellite-based system for telephone and two-way paging services. The satellite system is a Low Earth Orbit Satellite (LEOS) system, which means that signal powers and antenna sizes can be reduced with respect to conventional geostationary satellites. (The Iridium system orbits are less than 500 miles, versus the 23,000 miles for geostationary satellites.) In addition to enabling lower power transmission, a LEOS avoids the propagation delay of a geostationary satellite, which is a minimum of 500 msec. A basic disadvantage of a LEOS system is the continuous movement of the satellites with respect to ground locations. Because of this, Iridium provides a large number of satellites (66) so continuous coverage from at least one satellite is always available.

Iridium phones are dual-mode phones. A phone first makes an attempt to place a call over a terrestrial cellular system* but will default to the satellite network if local cellular coverage is not available. When communicating through the Iridium system, a user first gets connected to the nearest available satellite. From there the communication might return to the ground or be relayed through multiple satellites before re-

*Multiple versions of handhelds are available to operate on TDMA, GSM, or CDMA cellular networks.

TABLE 9.5 Iridium System Parameters

Number of satellites	66
Satellite orbit	485 miles (780 km)
Orbit period	100 min
Transmission protocol	FDMA/TDMA
RF modulation	QPSK
Speech data rate	2.4 kbps
Radio frequency	1.6 MHz handheld to satellite
	23 GHz satellite to satellite
	19 GHz uplink to satellite
	29 GHz downlink from satellite

turning to the ground. The second ground link might be direct to another Iridium user or it may involve a base station with interconnection to a public telephone network. System aspects of the Iridium system are provided in Table 9.5.

9.7 TRUNKED RADIO

The term *trunked radio* generally refers to Private Mobile Radio (PMR) communications networks. Originally, users of PMR equipment were allocated specific operating frequencies dedicated to each user (or organization). Such allocations were obviously inefficient in terms of bandwidth utilization when the users did not have continuous need for radio communications. Significant improvements in efficiency are achieved when the separate channels are placed in a group and shared by a larger group of users on an as-needed basis. When a user of a trunked radio system needs service, the radio equipment accesses an idle channel that becomes temporarily assigned to that user. As soon as the users of a particular channel cease to transmit, the channel becomes available for other users. Access to a communications channel utilizes a control channel and a centralized controller for resource allocation.

A trunked radio system is not a cellular system but does utilize a central node referred to as a repeater. The repeater receives a signal on one frequency, shifts it to another frequency, and transmits the signal on the new frequency. Thus, end users do not communicate directly with each other.* Transmission through the repeater is more effective because the tower is located at a high point in the coverage area utilizing a relatively high transmit power that facilitates communications between end users who are likely to not have an adequate path between them.

In a general sense, trunked radio systems are PMR systems that provide radio communications directly between users without use of the public telephone network. However, the distinction between a trunked radio system and a cellular system has be-

*Some trunked radio equipment does support a (special) two-way mode of operation in which the users communicate directly with each other. This mode of operation is generally reserved for communications in outlying areas where the repeater cannot provide service.

come blurred with U.S. offerings from companies like Nextel and Geotek. These companies utilize technology developed by Motorola that augments the use of spectrum previously used for private radio service referred to as Specialized Mobile Radio (SMR). SMR applications typically involve fleet operations such as taxi cabs and delivery vehicles in need of dispatch services wherein multiple mobiles simultaneously hear transmissions on a common channel. SMR radios utilize analog FM/FDM transmission with 25-kHz channels.

The new equipment, generically referred to as Enhanced Specialized Mobile Radio (ESMR), upgrades analog SMR transmission to digital TDM transmission in the same manner that D-AMPS upgrades AMPS. ESMR transmission protocols allow for six digital TDM channels in each 25-kHz channel. ESMR systems support conventional dispatch applications and other similar specialized services such as private group calling. Most importantly, for this chapter, an ESMR system can provide connections to the public telephone network—referred to as an interconnect feature. With this feature, an ESMR system can function as a cellular telephone system.

9.8 CELLULAR DIGITAL PACKET DATA

Cellular Digital Packet Data (CDPD) [9] is a system designed to provide data services up to 19.2 kbps as an overlay of an AMPS installation. The primary service concept is to offer fixed and mobile data services independent of the need to upgrade an AMPS system to a D-AMPS system. CDPD uses AMPS channels that are not being used for voice. Because it is a packet-oriented data service, multiple users share a single 30-kHz channel—a significant savings in spectrum usage as compared to the use of connection-oriented voice channels for data. Principal applications for CDPD are mobile Internet access and credit card verification.

Access to the shared channel is accomplished with Digital Sense Multiple Access with Collision Detection (DSMA/CD) which is similar to CSMA/CD of other radio systems and Ethernet LANs. CDPD uses GMSK modulation with RS(63,47) forward error correction.

REFERENCES

1 "Vector Sum Excited Linear Prediction (VSELP) 7950 Bit Per Second Voice Coding Algorithm," Technical Description, Motorola, Schaumburg, Illinois, Nov. 14, 1989.

2 M. Rahnema, "Overview of the GSM System and Protocol Architecture," *IEEE Communications Magazine*, Apr. 1993, pp. 92–100.

3 P. Vary, K. Hellwig, C. Galland, M. Russo, J. Petit, and D. Massaloux, "Speech Codec for the European Mobile Radio System," in *IEEE GLOBECOM 1989*, Nov. 1989, pp. 29.B.2.

4 R. Dixon, *Spread Spectrum Systems with Commercial Applications*, Wiley, New York, 1994.

5 "Mobile Station–Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System," EIA/TIA/IS-95, Washington, DC, July 1993.

6 A. Salamasi and K. S. Gilhousen, "On the System Design Aspects of Code Division Multiple Access (CDMA) Applied to Digital Cellular and Personal Communications Networks," *Proceeding of the Forty-First IEE Vehicular Technology Conference*, May 1991, pp. 57–62.

7 A. J. Viterbi and A. M. Viterbi, "Erlang Capacity of a Power Controlled CDMA System," *IEEE Journal on Selected Areas of Communications*, Nov. 1993, pp. 882–890.

8 K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley, "On the Capacity of a Cellular CDMA System," *IEEE Transactions on Vehicular Technology*, May 1991, pp. 303–312.

9 A. K. Salkintzis, "Packet Data over Cellular Networks: The CDPD Approach," *IEEE Communications Magazine*, June 1999, pp. 152–159.

PROBLEMS

9.1 What is the available bearer rate of a half-rate digital channel in a D-AMPS system?

9.2 What is the data rate of the slow associated control channel in a D-AMPS system?

9.3 What is the available bearer rate a full-rate digital channel in a GSM system?

9.4 What is the data rate of the slow associated control channel in a GSM system?

9.5 Determine the receiver output measurements for channels 1 and 2 for the CDMA example of Figure 9.6.

9.6 What is the effective signal-to-interference ratio of a single CDMA uplink channel operating at a distance that is twice as far from the base station as 62 other channels? Assume a code length of 64, cross correlations of ± 1, and all transmitters operate at identical power levels. (a) Assume all interferers are active. (b) Assume half the interferers operate with a 25% data rate because of no voice activity.

9.7 In a CDMA system with a code length of 64 determine the signal-to-interference ratio of a single uplink CDMA channel if there are 16 active interfering channels operating at an effective receive power level that is 12 dB higher that the desired channel.

10

DATA AND ASYNCHRONOUS TRANSFER MODE NETWORKS

The primary focus of this book is to describe how and why digital electronic technology is used to communicate voice, a fundamentally analog signal. Naturally, the digital telephone network can accommodate inherently digital information (data), but only if data can accept or be adapted to service characteristics established for voice. This chapter discusses digital technology and networks specifically directed to supporting data communications services. Ironically, the later part of this chapter describes how the data-oriented networks are adapted to carry voice.

Communication of discrete messages (data) is actually an ancient practice. In early times communication over distances greater than the range of the human voice was provided by sight or sound of discrete signals (e.g., heliographs, smoke signals, flags, and horns). Furthermore, the first practical communication system to use electricity, the telegraph, is inherently digital. As telegraphy evolved from the original, manually based systems to fully automated systems, they developed into what are commonly referred to as message-switching networks. Modern message switching is discussed in the first section of this chapter.

As the need for modern electronic data communications arose, it was only natural that the public telephone network would be used for data transmission services. The availability overshadowed numerous technical shortcomings of a network designed primarily for voice communications services. The main deficiencies of a conventional telephone network for data transmission are:

1. Need for signal transducers (modems) on analog access lines
2. Limited data rates
3. High error rates (in the older analog network)
4. Inefficient circuit utilizations

As data communications requirements increased, so did the justification for more cost-effective data communications solutions. One solution to reducing data transmis-

sion costs was to improve circuit utilizations through the use of packet-switching networks. The technology of packet switching was pioneered by the Advanced Research Projects Agency of the U.S. government. This agency developed a network referred to as the ARPANET [1]. In addition to ARPANET, which was used only by government, educational, and industrial research institutions, a number of public packet-switching networks were also developed in the United States and around the world. ARPANET developments eventually evolved into what is now known as the Internet.

A second approach to improving data communications involves developing separate networks specifically designed for digital transmission (no analog circuits with modems). The first major enterprise of this type in the United States was a nationwide digital microwave network developed by Digital Transmission Corporation (DATRAN). Because the data communications marketplace of the time could not support a separate, dedicated network for data, DATRAN ran into financial difficulty and declared bankruptcy in 1977.

Another, more successful, data-networking approach was the Dataphone Digital Service (DDS) offering of AT&T. This service utilizes digital technology of the telephone network for strictly data applications. DDS circuits are dedicated to data services, but the facilities and routes are shared with telephone network facilities. A major hurdle for DDS is achieving digital access to DDS circuits. If a subscriber is outside the range of digital transmission facilities of a DDS serving office, a voiceband modem over an analog line is required.

A fourth approach to satisfying data communications services involved developing means for directly accessing the digital transmission and switching equipment of the telephone network. The first widespread approach of the telephone companies for providing universal digital access is the Integrated Services Digital Network (ISDN). ISDN provides digital access to the digital facilities of the telephone network for voice or data services on a call-by-call basis. ISDN digital subscriber lines and other methods of digital access to digital networks such as the Internet are described in Chapter 11.

10.1 MESSAGE SWITCHING

As one telegraph system after another was installed in the countries around the world, nationwide communications networks evolved. A message could be sent from one point to another even if the two points were not serviced by a common telegraph line. In this case, telegraph operators at intermediate points would receive a message on one line and retransmit it on another. When a telegraph office had several lines emanating from it, the process of transferring a message from one line to another was, in essence, a switching function.

The process of relaying, or switching, a message from one telegraph line to another became semiautomated when teletypes with paper tape punches and readers were developed. An incoming message could be punched automatically onto a paper tape by one teletype and subsequently read by another teletype for transmission on the appro-

priate outgoing line. The process of transferring a message from one line to another in this manner led to these systems being referred to as torn-tape message switches.

One of the world's largest message switches was completely automated in 1963 when Collins Radio Company of Cedar Rapids, Iowa, installed a computer-based message switch for the airline companies of North America. This system and the more recent successors eliminate paper tape transfers by storing incoming messages directly into a computer memory (disk file) and forwarding them automatically to the appropriate output line when available. Hence this mode of operation is often referred to as *store-and-forward* message switching.

Included with each message is a header containing an address and possibly routing information so the message processor at each node can determine to which output line to switch the message. As indicated in Figure 10.1, the processor in each node maintains message queues for each outgoing link. These queues are normally serviced on a first-come, first-served basis. However, priority information can sometimes be included in each header to establish different classes, or grades of service, thereby allowing time-critical messages to be placed at the head of a queue.

A message-switching network is fundamentally different from a circuit-switching network in that the source and destination do not interact in real time. In fact, most message-switching networks could deliver a message on a delayed basis if a destination node is busy or otherwise unable to accept traffic. In a message-switching network there is no need to determine the status of the destination node before sending a message, as there is in circuit switching.

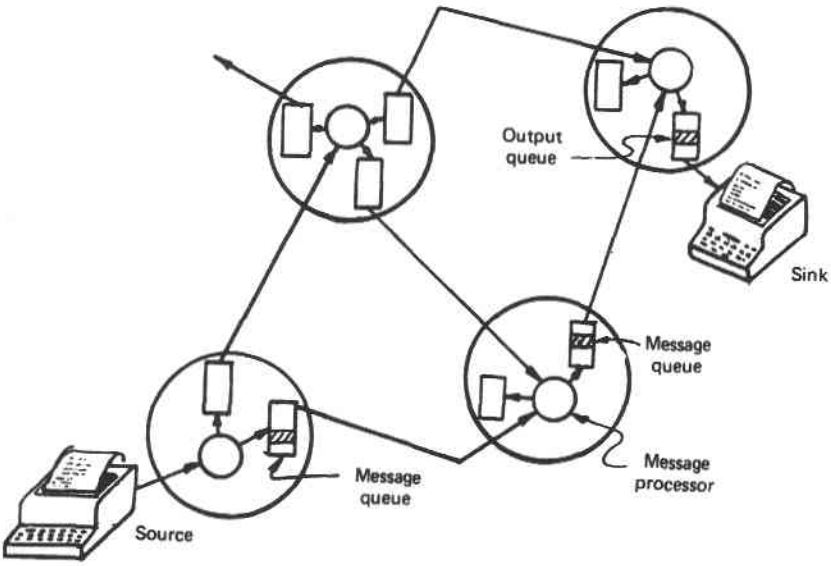


Figure 10.1 Message-switching network.

Message-switching networks are also fundamentally different from circuit-switching networks in their response to traffic overloads. A circuit-switching network blocks or rejects excess traffic while a message-switching network normally accepts all traffic but provides longer delivery times as a result of increased queue lengths.

Another important distinction of a message-switching network is that the transmission links are never idle while traffic is waiting to use them. In a circuit-switching network, a circuit may be assigned to a particular connection but not actually carrying traffic. Thus, some of the transmission capacity may be idle while some users are denied service. In contrast, utilization of the transmission links of a message-switching network is directly related to the actual flow of information. Arbitrarily high utilization efficiencies are possible if increased store-and-forward queuing delays are acceptable. Chapter 12 provides basic results of queuing theory that relate utilization efficiency to queuing delay.

10.2 PACKET SWITCHING

The circuit-switched telephone network is ill-suited to interactive data traffic because it is fundamentally designed for less frequent service requests with comparatively long holding times (3–4 min on average). Neither the control elements in the switches nor the capacity of the signaling channels are capable of accommodating frequent requests for very short messages. The result is that connection setup time may be many times greater than the holding time of a data message. Obviously, more efficient utilization of the network requires greater control channel bandwidth and increased call processing capacities in the switches. Beyond this, however, interactive data traffic with low-activity factors requires a network operation that is fundamentally different from a conventional circuit-switched network. The most appropriate mode of operation for traffic that comes in bursts is more closely related to a message-switched network than to a circuit-switched network.

Figure 10.2 depicts both the conceptual structure and the conceptual operation of a packet-switched network. A single message at the source is broken up into "packets" for transmission through the network. Included in each packet is a header containing address and other control information. Each packet is relayed through the network in a store-and-forward fashion similar to a message-switching network. At the destination node, the packets are reassembled into the original contiguous message and delivered.

The main feature of a packet-switching operation is the manner in which the transmission links are shared on an as-needed basis. Each packet is transmitted as soon as the appropriate link is available, but no transmission facilities are held by a source when it has nothing to send. In this manner, a large number of relatively inactive sources can share the transmission links. In essence, link utilization is improved at the expense of storage and control complexity in the nodes. A circuit-switched network has control overhead associated with connection setup but very little control thereafter. In contrast, packet-switching nodes must process the header information in each packet as it arrives. Thus a long message in a packet-switched network requires more

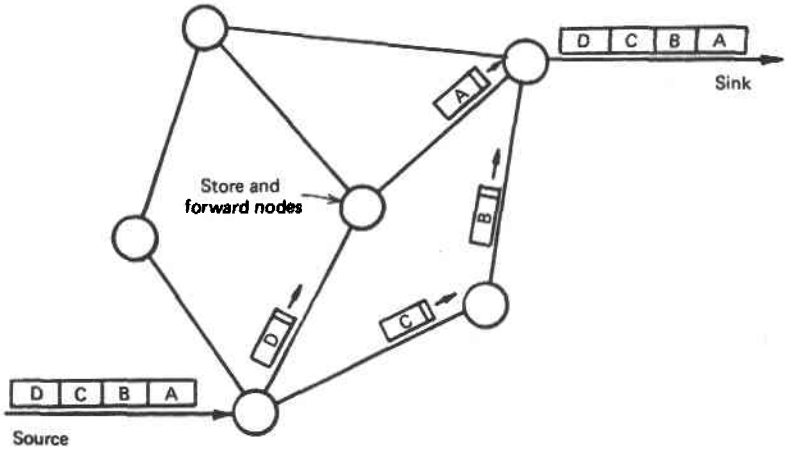


Figure 10.2 Packet-switching network.

control overhead than if it were serviced in a circuit-switched network. Considering the declining cost of digital memory and processing, the increased control complexity becomes less and less significant as digital technology advances. As discussed later in this chapter, one particular variation of packet switching, Asynchronous Transfer Mode (ATM), is designed to specifically support hardware implementations of control intensive functions, thereby supporting very high traffic volumes with low delay.

As the traffic load in a packet-switched network increases, the average transmission delay increases correspondingly. In contrast, a circuit-switched network either grants service or rejects it. There is no graceful degradation in service. Conversely, when only a few circuits are in use in a circuit-switched network, much network transmission capacity is idle. When there is a light load on a packet-switched network, the active users benefit by shorter than usual delay times. Hence from a user's grade-of-service point of view, the two network types are fundamentally different.

Using automatic repeat request (ARQ) error control, packet-switching networks (e.g., ARPANET) traditionally provided essentially error-free transmission for each node-to-node transfer. This process requires the receiving nodes to monitor redundant check bits appended to each packet to determine if the packet was received correctly. When errors are detected, a retransmission is requested (message NAK). Hence transmitting nodes must hold all transmitted packets in memory until a positive response (message ACK) is returned by the receiving terminal. Furthermore, an entire packet is usually received and checked for errors before forwarding it to another node.

Customers typically access packet networks by way of leased lines or dial-up connections. Dial-up connections are used by infrequent users, while leased lines are preferred by heavy users to achieve constant availability, higher data rates, and possibly lower error rates.

Despite the similarity to a message-switching operation, a packet-switching network is different in two important respects:

1. The store-and-forward delay through a packet-switched network is relatively short. Thus interactive communications can occur in much the same manner as if a dedicated, end-to-end circuit is established.
2. A packet-switched network does not provide storage of messages, except in an incidental manner while relaying packets from one node to another. The network is designed to provide switched communication between two nodes, both of which are actively involved in the communications process. A packet-switching network does not normally store a message for later delivery to an inactive or busy terminal.

One reason for breaking messages into packets is to allow transmission of the first segment of a long message while other segments are in transit. If the entire message had to be received at each node before forwarding it to the next node (as in message switching), the delays through the nodes might be too large. Another reason for breaking the messages into packets arises from operational simplifications derived from storing, processing, and transmitting smaller and possibly fixed-length blocks of data. In addition, if long messages are transmitted intact, short messages experience excessive delays when queued behind long messages. Packetization allows short messages to get through a transmission link without waiting behind long messages. This same principle occurs in multiprogrammed computers, which use time slicing to allow short jobs the opportunity of finishing before previously started long jobs. One more motivation for packetization is that when a transmission block is too long, it is unlikely that the entire message will be received correctly. Packetization provides a means of retransmitting only those portions of a message that need to be retransmitted.

10.2.1 Packet Formats

The format of a packet in a packet-switching network can vary significantly from one network to another. Some formats include numerous fields for control information while other systems rely more heavily on special control packets to transmit control information. Generally speaking, the control information associated with a particular message or link is included in the header of a message packet. Less frequent, network-related control information is communicated through special control packets.

As indicated in Figure 10.3, a packet contains three major fields: the header, the message, and the redundancy check bits. Some packets may not contain a message field if they are being used strictly for control purposes. Although a variety of techniques for generating redundancy checks are possible, the most popular technique uses cyclic redundancy checks (CRCs). Basically, a CRC is nothing more than a set of parity bits that cover overlapping fields of message bits. The fields overlap in such a way that small numbers of errors are always detected and the probability of not detecting the occurrence of 2 large number of errors is 1 in 2^M, where M is the number of bits in the check code.

A header typically contains numerous subfields in addition to the necessary address field. Additional fields sometimes included in a header are:

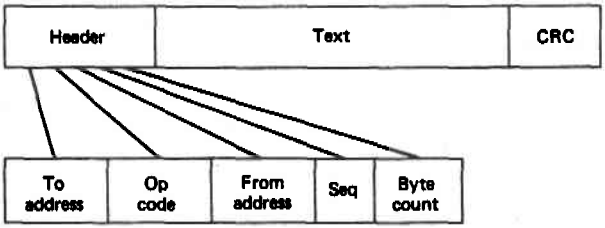


Figure 10.3 Typical packet format.

1. An operation code to designate whether the packet is a message (text) packet or a control packet. In a sense this field is a part of the destination address, with the address specifying the control element of a switching node.
2. A source address for recovery purposes or identification of packets at a destination node that is capable of simultaneously accepting more than one message.
3. A sequence number to reassemble messages at the destination node, detect faults, and facilitate recovery procedures.
4. A length code to indicate the length of a packet when less than a standard size packet is transmitted. Some protocols insert special delimiters (flags) at the end of a packet and therefore do not use a length count.

10.2.2 Statistical Multiplexing

The digital time division multiplexing techniques described in the previous chapters provide multiple channels by periodically assigning a time slot to each channel. The time slots are assigned whether or not the respective sources have anything to send. Channel assignments may be altered on a connection-by-connection basis, but for the duration of a "call" a particular time slot is dedicated to a respective connection. Because of the continuous manner in which time slots occur for each channel, this form of multiplexing is sometimes referred to as synchronous time division multiplexing (STDM). In this section we describe another form of multiplexing, variously referred to as asynchronous time division multiplexing (ATDM)* or statistical time division multiplexing (statmux). This form of multiplexing is mentioned here because of its close relationship to packet-switching techniques.

Statmuxes operate with framing formats that are basically identical to STDM framing formats. The major difference is that a statmux periodically redefines the length of its frames to change the number of time slots and, hence, the number of channels. Whereas an STDM system permanently assigns a time slot to each of its sources, a

*Here is another context-sensitive use of the terms synchronous and asynchronous. In the context of digital telephony, "synchronous multiplexing" refers to combining tributary signals that have been synchronized to each other as in upper layer SONET multiplexing. Correspondingly, "asynchronous multiplexing" refers to the use of pulse stuffing to accommodate unsynchronized tributaries.

statmux assigns a time slot only when a source becomes active. A time slot is eliminated (the frame shortened) when the respective source becomes inactive.

Statmux systems have been primarily used to provide line sharing for a multiple number of interactive terminals communicating with a host computer. When only a few sources (terminals) are active, each source has a relatively high data rate channel available to it. As the number of active sources increases, the frame length increases so individual channel rates decrease. Some systems limited the number of active sources to ensure certain minimum data rates.

The purpose and performance of statmux systems are very similar to the purpose and performance of a packet-switching link. The main difference is that a packet-switching link transmits larger blocks of data with a header included in each block. Each time slot of a statmux system is shorter and contains only source data. Figure 10.4 contrasts the basic operation of message switching, packet switching, and statistical time division multiplexing. The message switch transmits each message in its entirety in a first-come, first-served manner. Packet switching breaks messages up to allow interleaving of packets from other sources. Thus short messages never get queued behind long messages such as file transfers. The statmux system breaks the messages up into even finer blocks (words) of data and adds periodic frame definition messages so that receiving terminals can properly identify the individual time slots and switch the incoming data accordingly.

As indicated in Figure 10.4, a packet-switching operation becomes very similar to a statmux operation if the size of the packets is small. In fact, commercially available

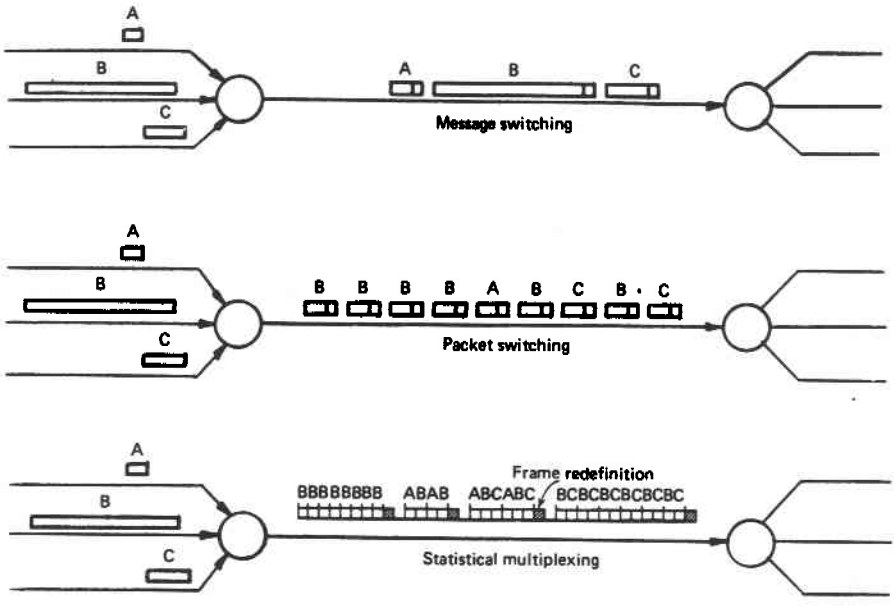


Figure 10.4 Comparison of message-switching, packet-switching, and statistical multiplexing.

statistical time division multiplexers can be used to build up a packet-switching network. In addition to the multiplexers, however, a packet-switching node needs sufficient storage to accumulate an entire packet, check it for errors, process it, and retransmit it. Strictly speaking, a statistical multiplexer does not accumulate an entire message or packet. It switches the incoming data as they are received. Error control functions (redundancy checking and requests for retransmission) are implemented between endpoints of a "connection" instead of between nodes. In summary, a statmux is strictly a multiplexer/demultiplexer. A packet-switching node provides multiplexing functions as well as message-level and network-level control functions.

10.2.3 Routing Control

Much discussion and experimentation have been undertaken regarding various procedures for routing packets from sources to sinks through a packet-switching network [2]. All routing techniques allow for a certain amount of adaptation or alternate routing to circumvent line or node failures. The various techniques differ, however, in how fast they respond to failures and whether or not they circumvent network congestion as well as equipment failures. The basic techniques are the following:

1. Dynamic routing
2. Virtual circuit routing
3. Fixed-path routing

Each of these techniques can be implemented in a variety of ways and can assume some of the characteristics of the other routing control procedures.

Dynamic Routing

Dynamic routing is implemented on a distributed basis with network nodes examining the destination address of each received packet to determine the appropriate outgoing link. The outgoing link is selected by processing locally stored information to determine which path provides minimum delay to the destination. The routing criteria are routinely updated to include both the operational status (health) and the size of the queues in the neighboring nodes. The routing decisions are adjusted rapidly enough that individual packets of a single message may follow *different* paths through the network. Dynamic routing, with its ability to respond quickly to changes in network topology or traffic conditions, is one of the original features espoused for packet switching. In fact, dynamic routing was sometimes considered to be inherent in the definition of a packet-switching network.

In spite of the obvious attraction of being able to adjust to rapid fluctuations in traffic patterns, dynamic routing has a number of significant drawbacks. One implication of allowing successive packets in a message to follow different routes through the network is that packets may arrive at a destination out of sequence. Although sequence numbers are used to rearrange the packets properly, the reassembly process is compli-

cated, particularly since the destination does not know if a missing packet is merely delayed or lost entirely.

Another drawback to dynamic routing is the possibility of oscillation occurring in the routing decisions. If the bandwidth assigned to updating the routing control algorithms is too small, a lightly loaded node will attract more traffic than it can handle before neighboring nodes are informed of the change in the traffic flow. In this instance, a packet might even wind up at a node from which it has previously been sent.

Purely distributed control, in general, and dynamic routing, in particular, also present difficulties with respect to flow control in a packet network. As mentioned in Chapter 7, alternate routing in the switched telephone network is sometimes suspended when the network as a whole becomes overly congested (alternate routes require more resources). Obviously, the same principle applies to a packet-switching network. Flow control in packet networks is discussed in later sections.

Dynamic routing is most appropriate for small networks or in a military environment where survivability of the network in the presence of multiple-node failures is a requirement. A military network is typically more concerned with reliable and timely completion of a few important messages than with achieving the highest possible throughput from a given amount of resources.

Virtual Circuits

A virtual circuit network embodies some of the basic aspects of both circuit switching and packet switching. The transmission capacity is dynamically assigned on an "as-needed" basis, but all packets of a multipacket message follow the *same* route through the network. Before interactive communication begins, a route is established through the network, and all participating nodes are informed of the "connection" and how to route the individual packets that follow. From then on, all packets flowing between the two end points follow the same route through the network. In essence, a virtual circuit is a logical concept involving addresses and pointers in the nodes of the network but no dedicated transmission facilities. At the end of a connection (or "session" in data communications terminology), a virtual circuit is released by a "disconnect" message propagating through the network.

Except in the case of permanent virtual circuits, separate connections, or sessions, involving the same two endpoints do not necessarily use identical paths through the network. Each virtual circuit is established during the call establishment phase depending on the traffic patterns at the time. Thus a virtual circuit network can respond to network failures or changing traffic patterns, but in a longer time frame than a dynamically routed network. When virtual circuits are changed from one connection to the next, the mode of operation is sometimes referred to as a *switched virtual circuit network* by direct analogy to conventional circuit switching.

Virtual circuits can be established using either distributed or centralized control. When distributed control is used, the call establishment message propagates through the network with each node making a local decision as to which outgoing link should be selected.

As discussed in Chapter 7 concerning conventional circuit-switched networks, centralized call establishment has the basic advantage of being able to set up circuits with a networkwide view of network status and traffic patterns. The TYMNET network of the United States [3] and the DATAPAC network of Canada [4] are examples of virtual circuit, packet-switched networks with centralized control.

Since the call establishment phase of a virtual circuit represents excessive overhead for single-packet messages, a virtual circuit mode of operation is obviously most useful when a network services a relatively large number of multiple-packet messages or sessions. Thus a dual-mode network is suggested requiring virtual circuits for longer messages and immediate transmission with dynamic routing for single-packet messages. In this instance, the single-packet messages are usually referred to as *datagrams.*

One of the main advantages of a virtual circuit operation is its ability to provide more orderly control of packet delivery. If a node in a virtual circuit never forwards a packet pertaining to a particular connection until the previous packet has been acknowledged, packets cannot arrive out of order. A second advantage of a virtual circuit is the reduced addressing requirements of individual packets. Once a virtual circuit has been established, complete destination addresses are no longer needed in the packets. In place of destination addresses, virtual circuit identifiers can be used that are local to each link. In essence, virtual circuit identifiers are pointers to memory addresses in the call processors of the packet-switching nodes or to look-up tables in ATM nodes. The designated memory addresses contain the pertinent information concerning the actual source, destination, and routing of the packets. Another important feature of a virtual circuit operation is its inherent ability to aid in flow control, as discussed in the next section.

The main disadvantage of a virtual circuit operation is the possibility of greater transmission delays. When a path for a virtual circuit is established, it is chosen to minimize the delay through the network under the traffic patterns at that time. If the traffic patterns change, packets pertaining to a particular virtual circuit may experience long queuing delays on some links while alternate links are more lightly loaded. Yum and Schwartz [5] report that analyses of routing techniques for several small network configurations indicate a packet delay improvement of 10–27% is possible when adaptive routing is used in lieu of a fixed routing rule.

When first considered, call establishment might seem to be a disadvantage of a virtual circuit network. Actually, however, flow control considerations require some type of query/response message to determine the status of the destination nodes before a source is allowed to begin sending a packet stream. Thus the control overhead and the delay associated with setting up a "connection" is usually a fundamental requirement, even in a dynamically routed network.

Fixed-Path Routing

Fixed-path routing embodies the same basic concepts of a virtual circuit network except successive connections, or sessions, between any two endpoints always use the

same path. In essence a fixed-path network is one that assigns permanent virtual circuits to each pair of endpoints.

One attractive feature of a fixed-path network is the absence of the call establishment phase of a virtual circuit network. However, unless the necessary resources are permanently allocated, a "clear-to-transmit" message is needed before sending packets. Inactive virtual circuits do not tie up network resources in the same manner as conventional circuits, but each virtual circuit does require transmission capacity and store-and-forward buffers in a probabilistic sense. If minimum grades of service (delay times) are to be guaranteed, a network must limit the number of virtual circuits existing at any particular time. Hence clear-to-transmit signals are needed before a fixed-path circuit becomes active. Of course, a network can provide two modes of operation: permanently active (hot line) virtual circuits and virtual circuits activated and deactivated as needed.

An obvious disadvantage of a purely-fixed-path network, as described so far, is its vulnerability to node or link failures. To circumvent this problem, a network control center usually assigns semipermanent paths through the network that are changed only when necessary for failure survivability or maintenance. Some older message-switching networks were implemented with fixed-path routing in the strict sense. This mode of operation was more appropriate to message-switching networks because message delivery was less time critical and could be deferred while repairs were undertaken for inoperative equipment.

10.2.4 Flow Control

As discussed in Chapter 7 concerning conventional circuit-switched networks, routing and flow control are two closely related operational requirements of any communications network. The same basic principle for controlling the flow in circuit-switched networks also applies to packet-switched networks. Namely, whenever the network is experiencing excessive stress, either from a loss of capacity due to failures or from an abnormally heavy demand for services, new service requests must be blocked at the periphery of the network before they tie up common resources and compound the congestion. In a packet-switched network, the common resources are store-and-forward buffers and transmission links.

Flow control in a packet network is primarily concerned with buffer management. For example, if all store-and-forward buffers in adjacent nodes become filled with packets destined to each other, the nodes are unable to receive additional packets, and a deadlock exists. Recall that, in pure packet switching, a node does not release a buffer as soon as it transmits a packet. The buffer is released when an acknowledgment is returned from the adjacent node. If a receiving node has no available buffers, it cannot accept a new packet and therefore cannot acknowledge it. (Frame relay and ATM networks described later do not provide node-by-node error control so they do not save copies of messages in internal nodes of the network.)

Flow control requirements imply that interface nodes in a packet-switched network are aware of overload conditions and refuse new requests for service until the conges-

tion is relieved. A particularly attractive feature of a virtual circuit network is that the call establishment phase provides an automatic means of determining whether or not a particular request should be serviced. If no path through the network can be established because the interior nodes are too congested, the request is rejected. On the other hand, if a virtual circuit is established, there is reasonable expectation that the entire request will be serviced in a timely manner.

Unless network nodes are very conservative in accepting requests for new virtual circuits, the ability to set up a circuit does not guarantee the avoidance of excessive congestion or deadlocks. A node accepts a new virtual circuit based on an expected load and its capacity to service that load. If the traffic volume and patterns happen to exceed the expected load, excessive delay or congestion is possible.

Virtual circuits are an effective means of controlling the flow of multiple-packet messages, but they represent too much overhead for single-packet or datagram flow control. If a virtual circuit network must support a significant number of single-packet messages, it can allow immediate transmission of these messages and forego the call establishment phase. In terms of transmission overhead, a datagram is not much greater than a call establishment packet. Thus, from this point of view, the packet might as well be sent immediately and be considered its own circuit setup message. In terms of store-and-forward buffers, however, a single-packet message is much different from a call establishment packet. A call establishment packet requires a certain amount of storage and processing by the call processor of each network node it reaches, but it does not compete for store-and-forward buffers as does a message packet. If necessary, a call establishment message can be ignored by an overloaded node, and the originating node times out waiting for the network response and reissues the request. The time out should be long enough that the network has had a chance to relax.

In contrast to call establishment packets, if a message packet is ignored by an overloaded node, congestion migrates to the node that last transmitted the packet, since this node is holding a copy of the rejected packet in a buffer. The buffer cannot be released until an acknowledgment is received. Hence datagrams cannot be allowed to enter the network unless a reasonable chance for complete passage exists. When the network is heavily congested, attempts to set up virtual circuits might also be suspended.

A conventional circuit-switched network is unconcerned with flow control between the endpoints of a connection since, once the circuit is established, the activity or inactivity of the endpoints has no effect on other connections or on the network as a whole. End users necessarily administer flow control between themselves so the source does not overrun the receive buffers of a sink. These procedures concern only the endpoints.

In contrast, the very nature of a packet-switching network implies direct involvement with endpoint activity. If a source pumps excessive traffic into a network, other users experience degraded performance. Hence interfaces to a packet-switching network necessarily include flow control for respective sources. Source flow control establishes a maximum data rate for a network. If a sink accepts data at a lower rate for a sustained period of time, this fact must be communicated to the network node serving the source in order to slow the source down.

Store-and-forward buffers of a packet-switching network are needed for communications purposes and are not used as a storage medium for messages. If a packet-switching network also provides message-switching services, message storage functions should be implemented separately from the communications buffers. Then, the packet-switching network is used to transfer data to the message storage facility.

Another implication of maximizing store-and-forward buffer utilization is the need to wait until a sink is ready to accept data before a source begins sending. If a sink is not ready, packets get stuck in store-and-forward buffers at the far end of a network and cause congestion. Thus some form of request-to-transmit/clear-to-transmit sequence is needed before message transmission begins. This requirement is independent of the routing algorithms employed in the network. Hence, in actual practice, the setting up of a virtual circuit may not represent a time penalty.

Single-packet messages (datagrams) can be an exception. If they are transmitted without a clear-to-transmit signal, they may be discarded at the destination node if the sink is inactive or has no receive buffers available. The destination node then needs to return a rejection message to the source indicating the status of the sink. In this manner, the message itself is a request-to-transmit signal. The ARPANET used a flow control strategy wherein single-packet messages serve as their own requests for buffer storage but multiple-packet messages require preallocated buffers in groups of eight at the destination [6, 7].

Flow control in TYMNET is implemented in a different manner because of its exclusive use of virtual circuits for all messages. Before any node in the network can send a packet to a neighboring node, it must receive a clear-to-transmit signal from the neighboring node for the particular virtual circuit. The clear-to-transmit signal is an indication that a specified number of store-and-forward buffers are being held in reserve for that particular virtual circuit. After a node sends the specified number of packets, it sends no more until another clear-to-transmit signal is received (indicating the previous packets have been forwarded or more buffers have been allocated to the virtual circuit). By using the same node-to-node method of flow control at the interface between the network and the users, a very solid end-to-end flow control strategy is established.

The TYMNET flow control strategy is somewhat conservative in that it may allocate store-and-forward buffers to one virtual circuit while another virtual circuit has more use for them (possibly causing a decrease in line utilization). This conservatism, however, provides a number of useful features:

1. Networkwide flow control is established automatically by the flow control within each virtual circuit.
2. Under light traffic conditions, packet flow within each virtual circuit adjusts automatically to the maximum rate that the source and sink can support. If only a few virtual circuits exist, a relatively large number of buffers can be assigned to each circuit, allowing the return of more frequent clear-to-transmit signals.
3. If a sink stops accepting packets for some reason, this condition propagates back to the source by way of cessation of clear-to-transmit signals. Thus the source

stops sending when all allocated buffers are full. This principle of operation is sometimes referred to as *back pressure*.

4. As long as a node never overcommits its buffers, store-and-forward lockups cannot occur. If several sinks stop accepting data, the store-and-forward buffers assigned to the particular virtual circuits get filled and become unavailable. Other virtual circuits, however, can maintain transmission through their own assigned buffers.
5. The mechanism is fail-safe in the sense that positive indications (clear-to-transmit signals) are needed before packets are forwarded to neighboring nodes. When network links are overloaded, flow control bandwidth requirements are minimal. If flow control signals stop altogether, packet transmission stops.

10.2.5 X.25

X.25 is an ITU standard protocol established in 1976 for accessing a public packet-switching network. The data link layer of X.25 is link access procedure balanced (LAPB) using high level data link control (HDLC). HDLC is a bit-oriented protocol based on the previously defined Synchronous Data Link Protocol (SDLC) established by IBM for Synchronous Network Architecture (SNA) networks. HDLC has become the data link protocol of choice for numerous data communications applications. The balanced aspect of LAPB refers to a peer-oriented operation between the two ends of the link. In an unbalanced mode of operation a primary node controls the link for one or more secondary nodes.

The basic format of an HDLC packet is shown in Figure 10.5. As indicated, packets are delimited by a starting and an ending flag (01111110). The data stream between the flags is precluded from including an inadvertent flag by a *zero bit insertion algorithm.* Whenever a string of five 1s occur in the data stream a 0 is inserted by the source node. The receiving node removes a 0 occurring after five 1s. If a 1 is present after five 1s, it must be the sixth 1 of a flag and is therefore not removed.

The address field is typically 8 bits long but can be extended in increments of 8 bits. The information field can be any number of bits long, but specific applications typically define a maximum length. The information field is nonexistent in a control packet. The frame check field contains a 16-bit CRC but can be 32 bits long in some

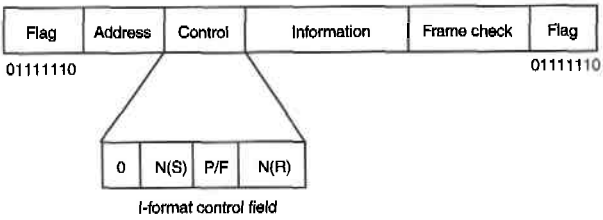


Figure 10.5 HDLC frame format.

applications. The control field is 8 or 16 bits long. Information in the control field is formatted in various ways depending on whether the packet is an information packet or a control packet. Control formats are designated by a leading 1. Figure 10.5 shows the format for a information packet (I-format) as designated by the leading 0 in the control field. The P/F field is a single bit used for polling.

The N(S) and N(R) fields contain sequence numbers used for both error control and flow control. The length of N(S) and N(R) is 3 or 7 bits depending on the size of control field in use.* For illustrative purposes, the following discussion assumes the use of 3-bit fields. Every time a source sends a packet containing information, it increments N(S) modulo 8. When a destination receives a packet with a valid check sum and the next expected sequence number, it returns that sequence number to the source in its outgoing N(R) field. Thus, a source node knows that the destination has correctly received every transmitted packet up to and including the packet with the sequence number N(R) in its incoming packets. If one or more packets have been received in error, the value of N(R) returned to a destination does not change until the source times out and retransmits the packets beginning with the last received N(R) value plus 1. [All packets after a lost or corrupted packet are retransmitted because the destination ignores packets with an out-of-sequence value in the incoming N(S)].

With a 3-bit length for N(S) a source can have a maximum of seven packets outstanding at any particular time. (A particular sequence number cannot be reused until an acknowledgment for the previous one has been received.) Beyond the maximum number of outstanding packets determined by the length of the sequence fields, particular applications may be configured for a smaller number of outstanding packets for flow control and reducing the amount of memory needed by a source to hold copies of unacknowledged packets. Whatever the value of the maximum number of outstanding packets is, a destination is expected to be able to receive that number as a burst. The destination does not have to immediately acknowledge a valid received packet if it is congested and wants to free up resources before receiving more packets. If a system is configured with a maximum number of outstanding packets that often exceeds the capacity of a receiver, the system works but the source spends a lot of time retransmitting packets that were discarded by the destination because it had no resources (memory) to store them.

X.25 permits a user on an X.25 network to communicate with a number of remote locations simultaneously. Connections occur on logical channels of two types: switched virtual circuits (SVCs) or permanent virtual circuits (PVCs). SVCs require a connection establishment process before data can be transferred. A PVC is similar to a leased line in that the connection is permanently established by network administration. Therefore, data can be sent without connection setup.

To establish a connection on an SVC, the originator sends an X.25 call request packet, which includes the address of the remote destination. The destination decides whether or not to accept the call (the call request packet includes the originator's address and possibly other information that the destination can use to decide to accept

*Use of HDLC on high-speed fiber links has led to the use of even larger sequence numbers because a large number of frames can be in transit on the link.

the call). A call is accepted by returning a call accepted packet or rejected by returning a clear request packet.

10.2.6 Frame Relay

As local area networks and file servers evolved to provide data processing functions within a corporation, some means of interconnecting LANs at separate sites became necessary. Figure 10.6 illustrates three ways of accomplishing LAN interconnection to implement a corporate wide area network (WAN). The first method utilizes a public (X.25) packet-switching network to relay messages between the LANs. This approach is a useful one for limited amounts of inter-LAN traffic. It also facilitates interconnecting two or more dissimilar LANs (such as Ethernets, token rings, or token buses)—

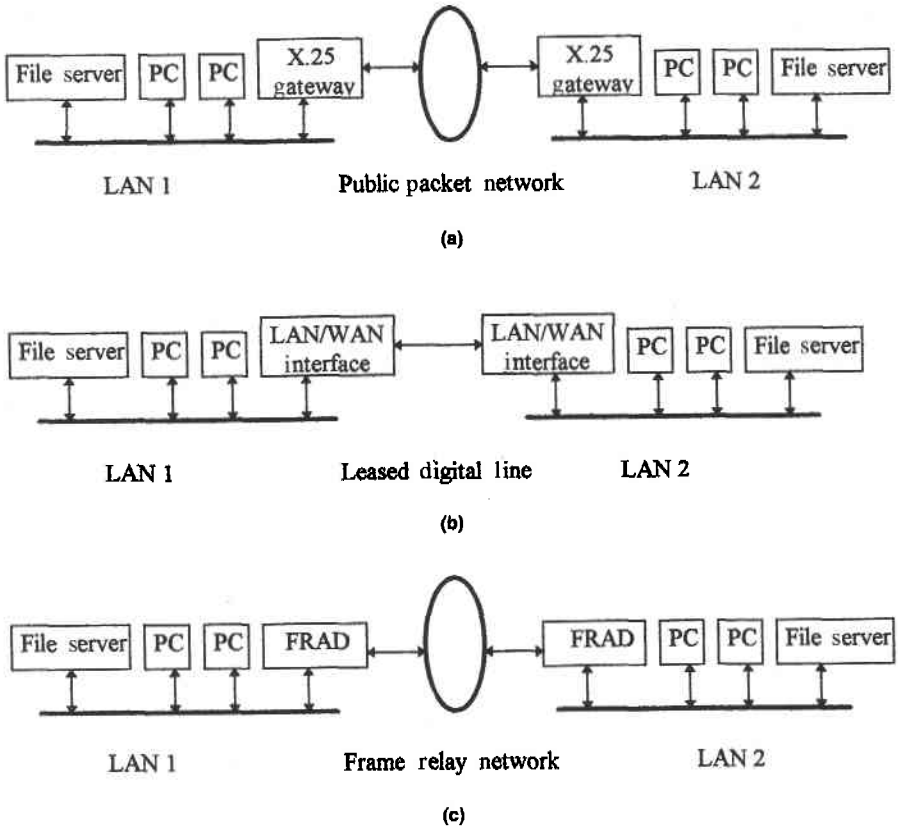


Figure 10.6 LAN-to-LAN interconnection methods: (*a*) through X.25 packet-switching netowrk; (*b*) through leased line; (*c*) through frame relay network.

because the X.25 interface of each LAN acts as a data link layer protocol converter. The main drawback is the cost and slow response for high-bandwidth transfers.

The second approach of interconnecting the two LANs is to use a leased line. Obviously this approach is more cost effective if large amounts of data are transferred between the two locations. The LAN/WAN interface equipment used to connect to the leased line will vary depending on the nature of the individual LANs and how the LANs and WAN are administered. Functional possibilities for the LAN–WAN interface are bridges, routers, or switches. The speed chosen for the leased line inevitably requires compromising performance objectives with cost.

The third method of interconnecting the LANs is to use a frame relay service of a common carrier. Frame relay services were developed specifically to address LAN interconnection and are available from all major common carriers. Frame relay is a packet-switching protocol but is faster than traditional X.25 networks because it does not provide error control. Error control at the data link layer (e.g., HDLC of X.25) requires receipt of an entire packet before it can be forwarded from one link to another. Frame relay also supports shortened address processing with a data link connection identifier (DLCI) field in the header which identifies a PVC. The frame relay protocol is defined by the ANSI T1.618 standard and ITU Recommendation Q.922.

The PVC feature of frame relay permits the functional equivalent of a leased-line connection but much more cost effectively for high-rate, bursty traffic. Furthermore, a single frame relay access device (FRAD) can achieve connectivity to multiple remote LANs by using a different DLCI for each remote LAN.

Frame relay nodes can process the simpler address format and switch an incoming message directly to an outgoing line as it is being received. The elimination of error control* at the data link layer of a network protocol reflects the fact that the underlying transmission facility, fiber, is essentially error free. In this environment it is more efficient to move error control to a higher level (e.g., an application layer). When infrequent errors occur, the application can invoke error control as appropriate to the application.

Example 10.1. Determine the amount of transmission capacity required to retransmit frames of 1000 bits of data in a network with 10 tandem links each of which has bit error probability 10^{-8}. (a) Assume link-by-link error control. (b) Assume end-to-end error control. (c) Repeat the calculations for a bit error probability of 10^{-5}.

Solution.

(a) With link-by-link error control the probability of a bit error in a frame is $1000 \times 10^{-8} = 10^{-5}$. The expected number of bits of transmission capacity required to retransmit is $10^{-5} \times 1000 = 0.01$ bit/link.

(b) With end-to-end error control the probability of a corrupted frame is $10 \times 10^{-5} = 10^{-4}$. The expected retransmission capacity required is $10^{-4} \times 1000 = 0.1$ bit/link.

*CRCs are included in frames of frame relay data, but a frame relay node does not request a retransmission if an error is detected.

(c) With a bit error rate of 10^{-5}, the respective calculations are 10 bits/link for link-by-link control and 100 bits/link for end-to-end control.

Example 10.1 illustrates that with very low bit error probabilities both forms of error control require insignificant amounts of transmission capacity for frame retransmissions. Thus, from a transmission capacity point of view there is no reason to use link-by-link error control. On transmission links with higher rates, as in T1 lines with marginal error performance, link-by-link error control is obviously desirable, particularly in a packet network where packets might be larger than 1000 bits.

10.2.7 TCP/IP

The original networking protocol of the ARPANET was designated Network Control Protocol (NCP). While this protocol addressed the needs of a single network, it could not be used as a means of communicating from one network to another. The desire to develop a protocol that would allow communication from a node in one network to a node in another network led to the eventual development of Transmission Control Protocol/Internet Protocol (TCP/IP) [8]. TCP/IP replaced NCP in ARPANET hosts on January 1, 1983.

TCP/IP is a protocol that can be used within a network, but more importantly, it can be used as a protocol to communicate between two networks.* Thus, the Internet is composed of many networks that communicate with each other using TCP/IP. Foremost among these many networks are LANs. The Internet is yet another way for interconnecting LANs. In contrast to the creation of a corporate WAN using frame relay, LAN interconnection via the Internet is primarily intended to cross corporate boundaries.

As defined by the Internet Engineering Task Force (IETF), TCP/IP uses variable-length packets that are processed with basically one class of service: best-effort routing. A user datagram protocol (UDP) replaces TCP when real time services such as voice or video are carried. UDP has no retransmissions. Before Voice on the Internet (VoIP) can be realized for something approaching universal service, the routers must be extended to support a new class of service: one with controlled delay, delay variation, and delivery guarantees. A first attempt to address the quality-of-service issue was a proposed modification of TCP/IP with a ReSerVation Protocol (RSVP). This approach was determined unsatisfactory because it required changing all nodes in a network simultaneously [9]. A more practical approach is a "differentiated service" (DiffServ) enhancement to TCP/IP.

When considering the viability and expected quality of service of packetized speech, it is worthwhile to consider the difference between IP telephony and Internet telephony. IP telephony refers to carrying packetized speech in an IP network. If this network is wholly contained within one organization, commonly referred to as an intranet, it is more likely that a relatively high level quality of service can be provided. Packetized speech carried over the Internet that traverses multiple, autonomous net-

*The most apparent aspect of TCP/IP to a user is the Universal Resource Locator (URL), with the following format: filetype://www.address.tld/resource. The address.tld (.top level.domain) identifies a location on the web while everything after the back slash identifies a resource (e.g., a file) inside the location.

works is much more difficult to manage, particularly when economic considerations for allocation of costs and revenues are considered. For a description of many of the technical considerations for packetized voice see reference [10].

10.3 ASYNCHRONOUS TRANSFER MODE NETWORKS

In contrast to TCP/IP, the Asynchronous Transfer Mode (ATM) network architecture incorporated features for supporting real-time traffic such as voice and video in the initial implementation. The principal aspects of ATM directed to real-time support are short, fixed-sized packets (cells), short headers, and no link-by-link error control.*

ATM is a standardized architecture of packet-oriented transmission and switching originally proposed for a Broadband Integrated Services Digital Network (BISDN). ATM has since been expanded in scope to support a wide variety of service types: wideband, narrowband, bursty, non–real time, and real time. The synchronous TDM (circuit-switched) network that evolved primarily for voice services supports the same services with external adaptations, but the adaptations come from a variety of suppliers necessitating the need for multiple, nonintegrated, nonstandardized equipment and support. ATM standardizes the wide range of services by defining quality-of-service requirements for various traffic types. The quality-of-service parameters specifically intended for voice services are maximum delay, delay variation, and cell loss probability.

10.3.1 ATM Cells

The ATM transmission format consists of fixed-size cells of 53 bytes each. As shown in Figure 10.7, there are five overhead bytes within each cell, which leads to 48 bytes of payload per cell. The use of fixed-sized cells for all applications facilitates hardware-only implementations of switching functions (queuing and transfer).

Cell Loss Priority

The cell loss priority (CLP) bit in the cell header is used to identify two basic classes of service with respect to discarding of cells within a network. When a cell with a CLP value of 1 arrives at a heavily congested network element, that cell can be discarded to relieve congestion for higher priority traffic (e.g., cells with CLP = 0). Cells with a CLP value of 0 are never discarded unless all cells with lower priority are already discarded.

10.3.2 ATM Service Categories

ATM services are categorized (by the ATM forum [11, 12]) into the following categories related to the statistical nature of the data rate requirements of the respective sources and the quality of service (QoS) that the network can provide for those services:

*TCP/IP also defers error control to the application but has variable-sized packets and a relatively complicated header format.

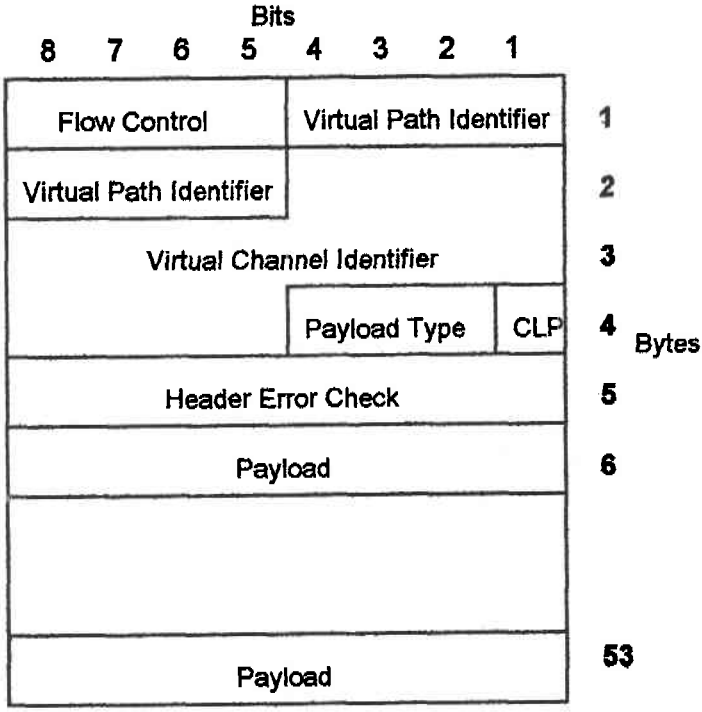


Figure 10.7 ATM cell format: CLP, cell loss priority.

Constant-bit-rate (CBR) services
Variable-bit-rate (VBR) services
Available bit rate (ABR) services
Unspecified bit rate (UBR) services

Constant-Bit-Rate Services

An application requesting a CBR service is guaranteed a specified data rate assuming the service request is accepted (i.e., the network can preallocate sufficient resources for the request). Thus, a CBR service is similar to a synchronous TDM service. CBR services are specifically designed to emulate circuit switching. A CBR application cannot exceed the specified rate (i.e., the peak cell rate) but need not necessarily send data at the specified rate. In this sense, CBR services are different from synchronous TDM services—if a source has periods of time in which there are no data available, the channel capacity may be released for other applications.

Variable-Bit-Rate Services

VBR services are separated into two distinct categories: real time (rt-VBR) and non–real time (nrt-VBR). The rt-VBR services are for bursty applications with tightly constrained delay and delay variation requirements such as voice or video. The main QoS

parameter of an rt-VBR service is the maximum cell transfer delay (maxCTD). The nrt-VBR services are for bursty data transmission applications with no particular delay requirements.

Whenever an rt-VBR or nrt-VBR service request is accepted, the following three parameters are agreed upon for source transmission statistics:

Peak cell rate (PCR)
Sustainable cell rate (SCR)
Maximum burst size (MBS)

As long as the source adheres to the above statistical parameters, the network is committed to carrying the traffic.

Available Bit Rate Services

ABR services are intended for non-real-time applications that operate between a minimum and maximum data rate specified by a minimum cell rate (MCR) parameter and a PCR parameter. An essential aspect of ABR services is the need for feedback signals within the network to control the flow of data from the ABR sources. The network dynamically allocates additional resources to ABR services (beyond those for MCR support) when the additional resources are available and so informs the ABR sources with the feedback mechanism.

Unspecified Bit Rate Services

UBR services are inherently intended for non-real-time applications wherein no specific quality of service is desired or implied. Neither delay parameters nor cell loss ratios (CLRs) are specified. If a network becomes overly congested, cells from UBR applications are discarded by the network (as opposed to controlling the flow of the UBR source). When a UBR application negotiates a connection, a PCR parameter value may be provided by the network, which provides the application with an indication of the throughput that can be expected. The following indicates which QoS parameters are relevant to the ATM classes of service.

QoS Parameter \ Class of Service	CBR	rt-VBR	nrt-VBR	ABR	UBR
Cell loss ratio (CLR)	X	X	X	X	
Cell transit delay (CTD)	X	X			
Cell delay variation (CDV)	X	X	X		
Peak cell rate (PCR)	X	X	X		X
Sustained cell rate (SCR)		X	X		
Burst tolerance (BT)		X	X		
Flow control				X	

10.3.3 ATM Connections

ATM networks are connection oriented, which means that a source must make a request for service before transmitting cells. Network control determines a route through the network and allocates resources (in a statistical or probabilistic sense) in response to each service request. In contrast, LANs typically operate in a connectionless mode wherein a source sends unconstrained data to a destination whenever the need arises.

ATM supports two types of connections: virtual channel connections (VCCs) and virtual path connections (VPCs). A VCC is a virtual circuit between two users representing a single stream of cells. When there are multiple VCCs between two common endpoints, a VPC can be created as a bundle of VCCs. The use of VPCs simplifies network management and the operations of intermediate nodes of an ATM network because a VPC is processed as a single entity as opposed to individual VCCs. Thus, a VPC is analogous to a trunk group in a circuit-switched network.

10.3.4 ATM Switching

The basic requirement of an ATM switching fabric is to transfer cells arriving on one physical link to another link. Because incoming headers must be processed before an incoming cell can be transferred through a switch and because an outgoing link is likely to be busy, ATM structures always provide some amount of delay in the switching process. Thus, ATM switches provide both space and time switching functions in a similar manner as TDM circuit switches described in Chapter 5.

A significant difference, and complication, of an ATM switch servicing stream traffic such as voice or video is that, in contrast to TDM channels, ATM cells can arrive at irregular intervals. This phenomenon greatly increases the likelihood that a particular cell encounters blocking on a shared interstage link of a switching network. The basic solution to this problem is to add cell buffers as a queue for blocked cells until the necessary path is free. Simple first-in, first-out (FIFO) queuing operations are readily implemented in hardware but better network efficiency occurs if the FIFO discipline is modified to allow a cell in the queue with an open path to bypass a blocked entry [13, 14]. Thus, more complicated hardware implementations are desirable.

The principal aspects of an ATM switch that are different from a TDM circuit switch are:

1. There is no preassigned incoming time slot so header information needs to be extracted and processed to know where to transfer incoming cells.
2. There is no preassigned outgoing time slot for outgoing cells so the delay involved in transferring cells is a variable queuing delay associated with waiting until the outgoing link is idle.
3. The header information in an outgoing cell is generally different than what is received so the switch must insert appropriate header information into the traffic stream.

ATM Memory Switch

As is the case of TDM circuit switches, ATM switches use memory switching to the maximum extent possible. As long as the memory speeds support the traffic, the most economical design is one in which all traffic is written into and read from a common memory. In addition to the switching fabric, a high-speed ATM switch must implement other cell processing functions in hardware. These functions are header extraction, header error checking, table lookup for route information using received virtual path indicator (VPI) or VCC, recording of traffic usage statistics, cell queuing/rejection, insertion of outgoing VPI/VCC, and outgoing checksum generation.

Example 10.2. Determine the memory speed required for an ATM switch fabric using a shared-memory architecture in support of 12 STS-3 (OC-3) bidirectional ports. Also determine the number of voice connections that can be carried by this switch. Assume that an active voice signal requires 12 kbps and that the activity factor is 40%.

Solution. Except for the 9 bytes of path overhead, all bytes of an STS-3 payload can be used for ATM cells. Thus, each STS-3 input link provides $260 \times 9/53 = 44$ cells/frame. (A partial frame is not carried so cell boundaries can be aligned with the start of the STS-3 payload.) If a 53-byte-wide memory is used for storing the cells, the total access rate for 12 bidirectional ports is $12 \times 2 \times 44 \times 8000 = 8.448$ MHz.

Since each ATM cell provides 48 bytes of payload, the number of voice connections that can be supported by one STS-3 is $44 \times 48 \times 8 \times 8000/12{,}000/0.4 = 28{,}160$ connections. The total capacity of the switch is 337,920 voice calls.

It is interesting to compare the results of Example 10.2 with a circuit switch design for 64-kbps voice channels. Using a standard SONET multiplexing format, a single STS-3 carries three DS3 signals (2016 voice channels). Similarly, a STS-3 (e.g., an STM-1) carries a single E4 signal with 1920 voice channels. In either case, the capacity of each ATM link is over 10 times the capacity of the conventional circuit-switched architecture using the same data rate. The advantage comes from a combination of compressing the speech and taking advantage of voice inactivity.

A memory switch for $12 \times 2016 = 24{,}192$ channels requires an access rate of 387 MHz, which is much greater than the 8.448-MHz access rate of the ATM switch, even though the total capacity is much less. The reason for the high access rate of the circuit switch is that only 8 bits are being switched at a time. If multiple bytes were accumulated and switched through a wider memory, as in the ATM case, the access rate of the circuit switch would be reduced accordingly. Circuit switches have not been implemented with wider transfers to prevent the insertion of extra circuit delay. Voice interfaces to ATM networks necessarily require an echo canceler to accommodate the delay.

Example 10.2 is a nonblocking switch as far as internal operations of the switch are concerned. However, if a large burst of traffic arrives at one particular output port, it may be that some cells are discarded because the queue for that particular port is full.

Notice that if traffic is being queued for some ports, other outgoing ports are operating below their capacity.

The amount of variation in the traffic flow on outgoing ATM links is a direct function of the connection admission control (CAC) process involved in setting up connections. If all connections are VPCs and these VPCs have relatively constant arrival rates (because the aggregate data rate of multiple, independent variable-rate channels is fairly constant), the transmission resources of the network will be efficiently utilized with modest queue sizes. If the traffic arrivals are highly variable, larger queues (and consequently larger delays) are required to achieve high transmission link utilization. When there is a mixture of traffic types, the time-critical types are given priority so cell loss and long delays are confined to non-real-time-critical cells.

Space–Memory Switch

The implementation of large ATM switches requires multiple stages of space and times switching (e.g., buffering) in the same manner that large time division switches described in Chapter 5 are implemented. One particularly useful architecture involves a front-end TDM bus as a space switch followed by separate output queues, as shown in Figure 10.8. The input bus is synchronized to series-to-parallel (S/P) circuits of the input ports so that a time slot is assigned for every received cell. Thus, the bus must operate at the total speed of all incoming signals. Each header is decoded via a table look-up to select the appropriate output so the output memories store only cells for respective output ports.

The main attraction of the multiple-output-memory architecture shown in Figure 10.8 is its ability to scale to a large range of switch sizes. Furthermore, broadcasting and multicasting are easily supported by having the header decode logic enable simultaneous writes into multiple buffers. Unfortunately, this architecture does not do much to solve memory speed problems. Output accesses only need to support the output port transmission rate, but input accesses must support successive time slots on the bus. Memory speed restrictions can be somewhat alleviated with high-speed caches acting as speed buffers between the bus and larger, slower output buffers, as shown in Figure 10.9. The caches do not add much cell delay but do represent a source of cell loss when one fills up because of a burst to the respective output line.

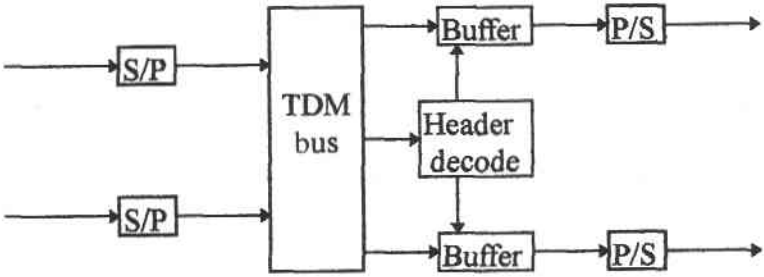


Figure 10.8 Input bus/output buffer ATM switch.

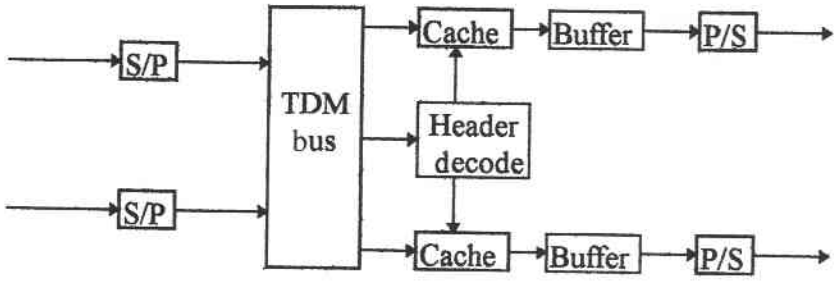


Figure 10.9 Input bus/output buffer ATM switch with caches.

Memory–Space Switch

Another version of a two-stage ATM switch is shown in Figure 10.10 wherein all incoming cells are available to all output ports via separate paths for each input. Each input path in each output port module is terminated in a separate cell buffer. Associated with each buffer is a table look-up function that selects only those cells destined to the respective output ports. Output data are selected from the separate buffers as a common queue, thereby performing a space switch function.

The main advantage of the configuration shown if Figure 10.10 is that all data paths and memories operate at the external link speed. The basic disadvantage is the need for N^2 buffers (and N^2 header processors), where N is the number of ports. Cell loss occurs when an incoming cell encounters a full buffer. The probability of this occurring is controlled by how much traffic for a particular output is accepted from a particular input. The use of multiple, individual queues leads to smaller group sizes, which leads to more total memory for a given cell loss probability. However, because the memories operate at relatively low speeds, the cost impact is minimized. Notice that delay probabilities are determined by the total traffic accepted for an output port in the same manner as previous configurations.*

Memory–Space–Memory Switch

ATM switch architectures with single stages of space and memory switching are impractical in terms of speed and/or hardware requirements when very large systems are desired. As is the case with circuit-switching architectures, more practical/economical structures require multiple memory stages or multiple space stages—or both. One such architecture is a counterpart to a TST switch, the memory–space–memory (MSM) implementation shown in Figure 10.11. The input memory stage of an MSM buffers complete cells of information that are queued for transfer through the space stage and then buffered again for transfer from the output memory. Successive cells of a particular connection are not necessarily transferred through the space stage at any particular predetermined time. As each cell is received into an input buffer, a queue entry is made for the appropriate transfer through the space stage. Support of multiple

*This assumes that the separate buffers associated with each output link are managed as a single FIFO queue for output purposes.

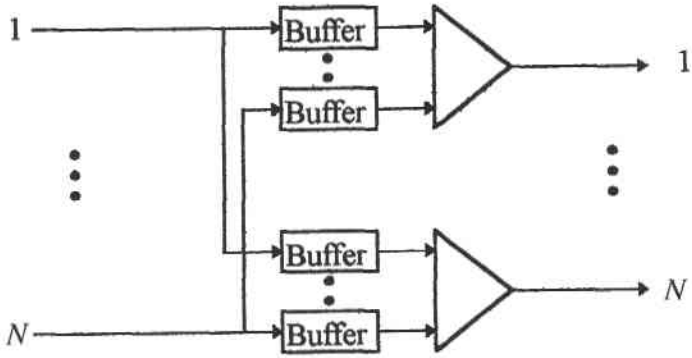


Figure 10.10 Multiple–memory ATM switch.

classes of service require multiple input queues (usually sharing a common memory element).

Before a cell in an input buffer can be transferred through the space stage, the desired output stage must be available for input. If the desired output memory is busy, the cell cannot be transferred and head-of-line blocking occurs. If another cell in the input queue can be selected when one or more of the head-of-line cells are blocked, the impact of the blocking is minimized. Such operations are easy to implement in computer-controlled store-and-forward switches but must be implemented in hardware in a large, high-speed ATM switch. An alternative approach to determining if blocking exists before a transfer is made is to blindly transfer cells through the switch and have blocked cells reenter an input port through a recirculation path, as shown in Figure 10.12. Determining the delay/cell loss performance of such a system is very complicated and dependent on the probability of blocking and the number of cells that can enter the recirculation buffer(s).

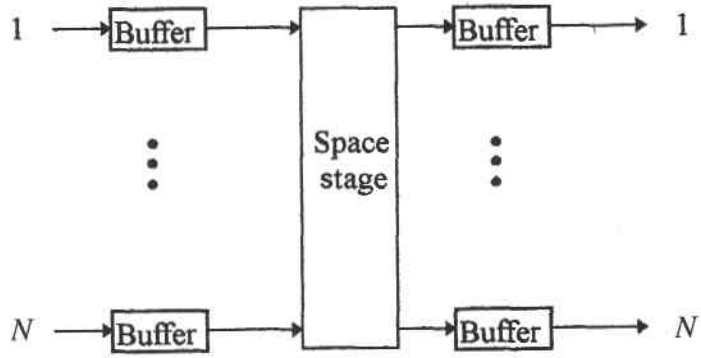


Figure 10.11 Memory–space–memory ATM switch.

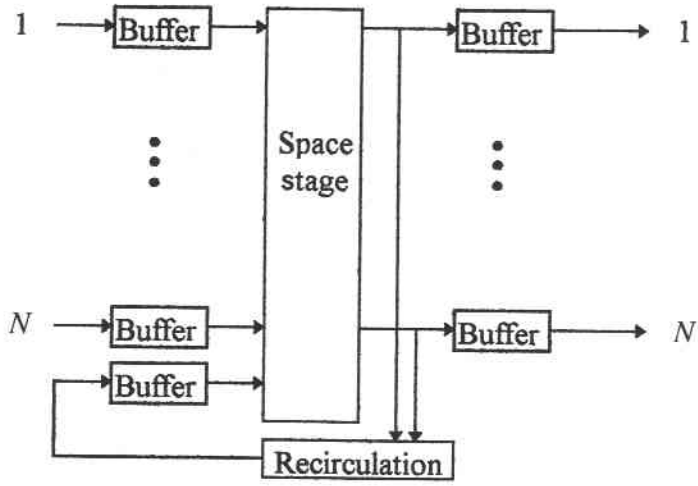


Figure 10.12 Memory–space–memory switch with recirculation.

For best performance, the input cells from all input lines of an MSM switch should be distributed across all input buffers using a front-end shuffle distribution switch [15]. This process decorrelates the entries in an input queue from the cell arrivals of any one particular input line. Recall that a similar operation is advantageous in TST switches to reduce blocking probabilities for individual channels of a busy trunk group.

The design of a space stage for an MSM switch matrix can follow any one of a variety of approaches, including a single-stage square matrix or a multiple-stage Clos network, as described in Chapter 5. With either of these designs the probability and treatment of output blocking can be minimized by running the space stage faster than the cell rate of the external ports (i.e., with the MSM equivalent of time expansion). The existence of expanded opportunities for space stage transfers reduces the probability of contention for an output buffer and minimizes the delay when contention occurs. A very significant consideration in the space stage design is the complexity of hardware control logic, particularly when multiple service priorities and multiple paths through the switch are accommodated.

Banyan Network Switch

One approach to implementing modular space stage structures is a banyan network [16]. As indicated in Figure 10.13, a banyan network is implemented by interconnecting 2×2 switching elements in multiple, recursive stages. Notice that the 8×8 structure is implemented as two 4×4 structures and an additional front-end stage to select either of the two 4×4 structures from each input. Repeated replications and first-stage additions allow implementation of arbitrary sized networks.

A banyan network provides full accessibility in that any input can be connected to any output. Notice that there is a unique path between any input and any output. How-

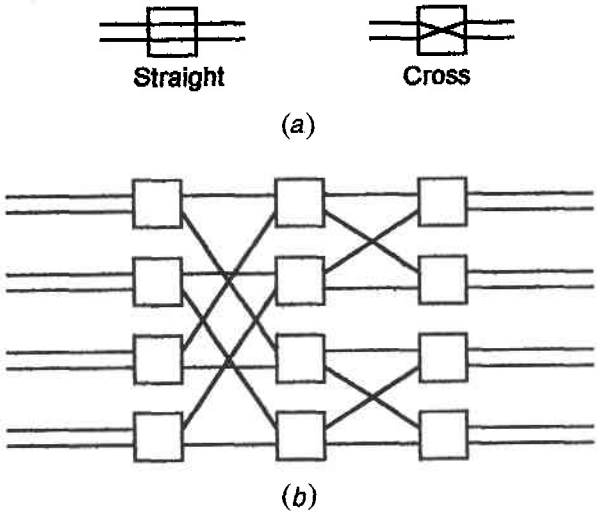


Figure 10.13 Banyan switching network: (*a*) 2×2 switching element connection states; (*b*) 8×8 network.

ever, the networks are not nonblocking. If the first input is connected to the first output, for example, the second input can only be connected to one of the bottom four outputs. One method of overcoming the blocking aspects of the banyan network involves cascading two such sections to produce what are generally referred to as Benes networks [17]. A Benes network is strictly nonblocking if rearrangement of connections is utilized. Another augmentation of a banyan network is to concatenate a Batcher sorting network [18] with a banyan network. The front-end sorting process orders the cells so they can all be transferred through the switch without blocking. Another modification to a banyan network to address blocking problems is to add parallel networks referred to as switch planes to provide more paths to an output. This latter approach is identical to adding more center stages to a multiple-stage switch described in Chapter 5. Unfortunately, this approach compromises the simplicity of unique path routing.

The basic attraction of a banyan network is the modular simplicity of its control. When a cell it to be transferred through the network, the table look-up process produces a sequence of bits for internal routing through the network. The first bit defines the selection of the first-stage element. The second bit defines the selection of the second stage and so on. The routing information is appended to the front of the cell and the composite message is transferred through the switch one stage at a time, with leading bits deleted as they are used. Thus, a banyan network utilizes direct progressive control reminiscent of an electromechanical step-by-step switch, described in Chapter 5. Unfortunately, modifications to the basic banyan architecture to provide multiple paths compromises the simplicity and distributed aspect of the hardware control.

A wide variety of other ATM switch architectures have been proposed and developed [19–23].

10.3.5 ATM Applications

Although the basic ATM architecture is primarily directed to servicing data communications, the ATM architecture includes provisions for other services such as voice, video, and multimedia. Support for the various services is included in various versions of ATM adaptation layers (AALs). Service adaptation occurs at the periphery of a network in edges switches. Internal nodes of an ATM network (core switches) are unconcerned with the nature of the traffic other than supporting the QoS objectives of the general classes of service. The role of AALs are shown in the functional layers of Figure 10.14 of the (ATM) Multi-service Switching Forum (MSF) [24]. The general classes of service and the adaptation layers to support these services are identified as follows:

AAL1: Transport of CBR traffic (program audio, video, and emulation of TDM-based circuits (i.e., DS1, E1).

AAL2: Protocol standard for low-bit-rate and time-dependent variable-bit-rate (VBR-rt) connection-oriented traffic (compressed voice and packetized video).

AAL3/4: Protocol standard for supporting both connectionless and connection-oriented VBR traffic [primary application, Switched Multimegabit Data Service (SMDS)].

AAL5: Transport of VBR traffic and signaling messages (packet data, IP, frame relay, LAN emulation).

Circuit Emulation Services

A circuit emulation service (CES) [25] is used to transport TDM-based channels such as DS0, DS1, or E1 signals. CESs are processed according to an AAL1 protocol. A

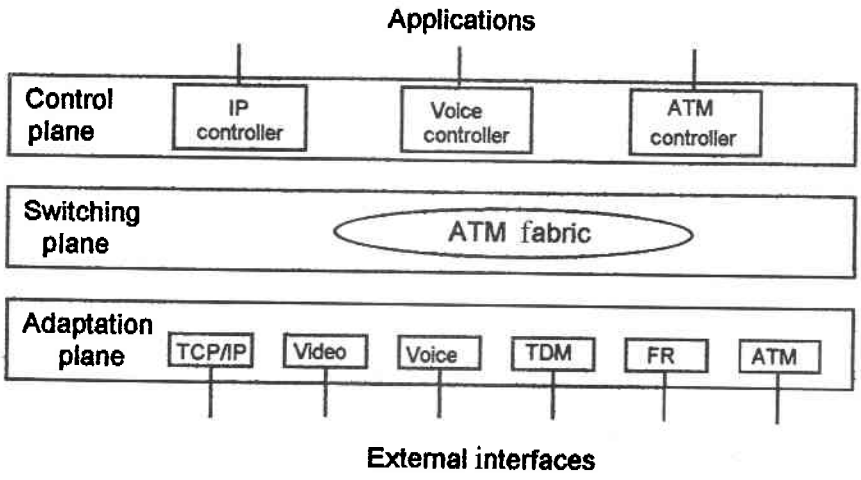


Figure 10.14 ATM functional implementation layers.

critical aspect of a CES service involves recovery of the source clock at the egress point of the ATM network. Figure 10.15 depicts three different clock recovery scenarios dependent on the nature of the network synchronization. The most desirable situation is shown in Figure 10.15*a*, in which the entire network, including the data source, is synchronized to a common primary reference clock (PRC). In this case the output data clock is derived from the PRC and therefore has the same average value as the source. As long as the ATM network meets cell delay variation (CDV) criteria, the cell buffer in the egress AAL1 equipment absorbs cell arrival variations and no data are lost. If cell arrival variations exceed CDV requirements, data slips may occur. As is the case in TDM networks the slips should be controlled to consist of integral numbers of frames of the underlying data (e.g., 193 bits for DS1 or 256 bits for E1) to preclude loss of framing in the underlying circuit.

Figure 10.15*b* shows a network configuration wherein the ATM network is synchronized to a common PRC but the data source is timed from some other reference. For this situation, ATM networks utilize synchronous residual time stamps (SRTSs). SRTSs are periodically inserted into the overhead of a cell data stream at an ingress AAL1. The egress AAL1 can determine the average data rate of the source by dividing the amount of source data received between successive SRTSs by the time interval defined by the SRTSs. Notice that the use of SRTSs depends on the ingress and egress equipment being synchronized to a common PRC.

Figure 10.15*c* depicts the most undesirable case in which the ingress and egress nodes are controlled by separate references (PRC1 and PRC2). In this case, adaptive clocking is utilized to recover the source data clock. Adaptive clocking involves a phase-locked loop (PLL) with the fill level of the cell buffer providing a phase measurement. A very low bandwidth filter is used to determine the long-term average fill level, which in turn establishes the recovered output clock of the VCO. Jitter in the derived clock is easily controlled by using a large-cell-buffer and narrow-bandwidth PLL. However, low-frequency clock variations (wander) caused by irregular cell arrival times are unavoidably transferred to the derived clock. Depending on the application, the wander may or may not cause downstream problems. If a DS1 signal, for example, is immediately demultiplexed into analog signals, no impairments to voice channels will arise but high-speed voiceband modems may be adversely affected (because of impairments to the carrier reference phase). If the DS1 signal is instead passed to a switching machine, the wander may contribute to slips in the elastic store interface of the switch. Again, slips do not impair voice signals but significantly corrupt voiceband data.

LAN Emulation

To provide compatibility with LAN protocols and established applications, the ATM Forum created a protocol, called LAN Emulation (LANE), which defines how Ethernet (IEEE 802.3) and token ring LAN (IEEE 802.5) data can be carried over an ATM network. Figure 10.16 show how an ATM network can provide LAN-to-LAN connectivity as well as connectivity directly to workstations (ATM to the desktop). The connectionless nature of contemporary LANs is simulated in a manner that first

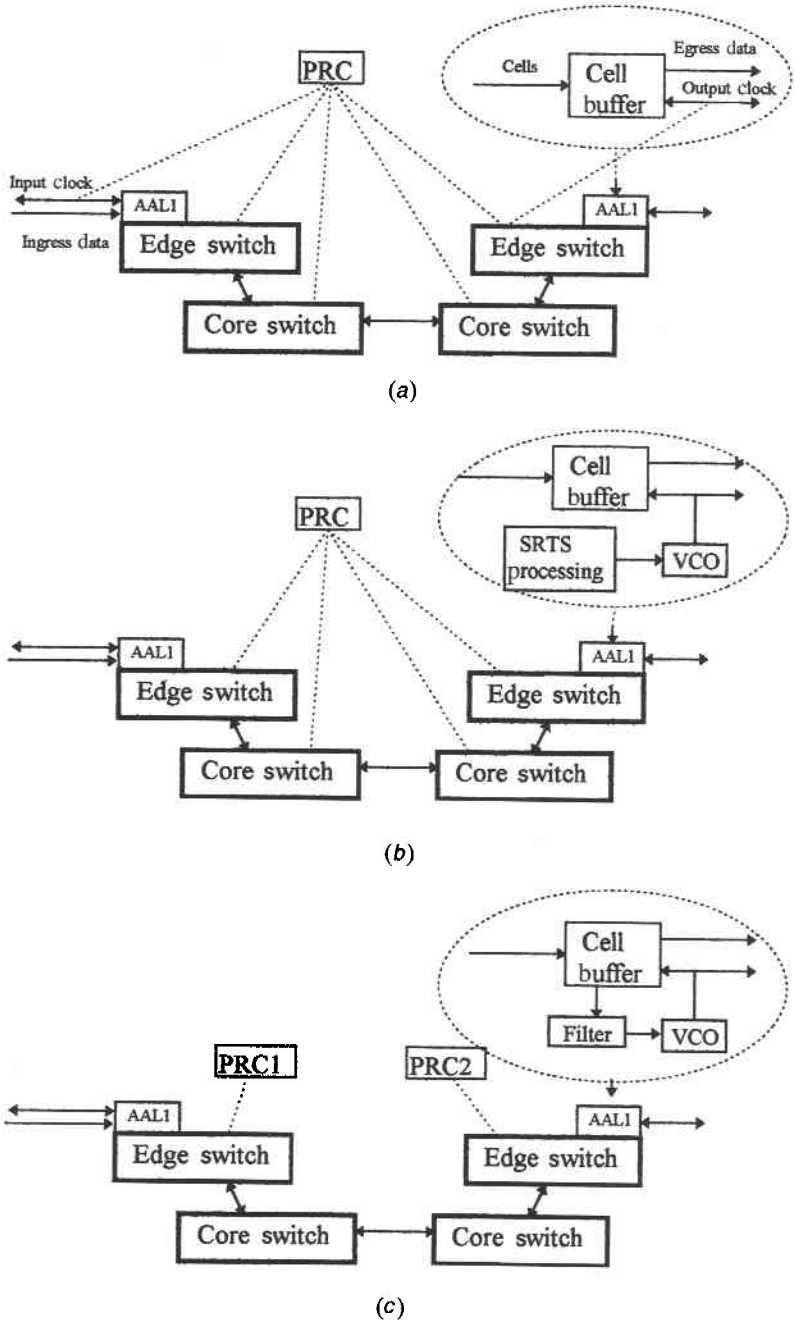


Figure 10.15 Synchronization of circuit emulation services: (*a*) fully synchronized network; (*b*) unsynchronized source (SRTs derived clock); (*c*) unsynchronized ATM gateways (adaptive clocking).

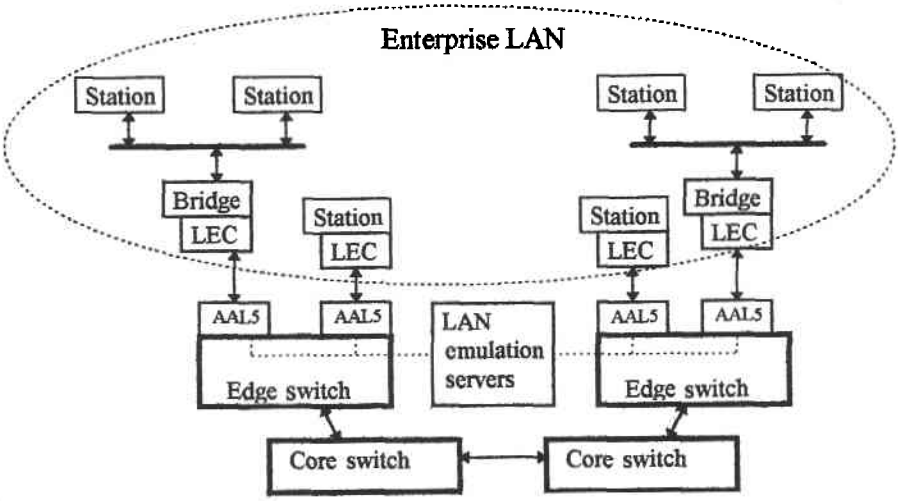


Figure 10.16 LAN interconnection using ATM LAN emulation.

determines if a connection (SVC or PVC) to an addressed station already exists. If so, an ATM message is sent to the station. Otherwise, an SVC is first established and then the message is sent. A LAN Emulation client (LEC) is established in every LAN attachment to the ATM network. Several types of LAN Emulation servers are used to configure the LECs and provide contemporary LAN functions such as multicast and broadcast services.

Low-Bit-Rate Voice

Although voice can be carried as 64-kbps PCM channels in a CES mode, great savings in bandwidth are achieved by transmitting compressed voice. Transport of compressed voice is particularly appropriate for transmission to and from digital cellular base stations where voice is already compressed or about to be compressed. Similarly, transport of a group of compressed voice channels between PBXs of a private network is advantageous. As shown in Figure 10.17, individual voice channels can be serviced as VBR-rt (AAL2) switched virtual circuits by packing multiple frames of low-bit-rate speech into ATM cells. The particular example shown in Figure 10.17 assumes G.729 Conjugate Structured Algebraic Code Excited LPC (CS-ACELP) voice compression with 80 bits per 10-msec processing frame. Notice that cell packing puts 4.8 ACELP frames into an ATM cell, which increases the voice latency by as much as 40 msec in each direction of transmission.

The voice latency can be reduced by forming a trunk group and packing multiple compressed voice packets into individual 48-byte cell payloads. A single cell can then carry four 10-byte CS-ACELP packets in a frame and not add a lot of delay to the in-

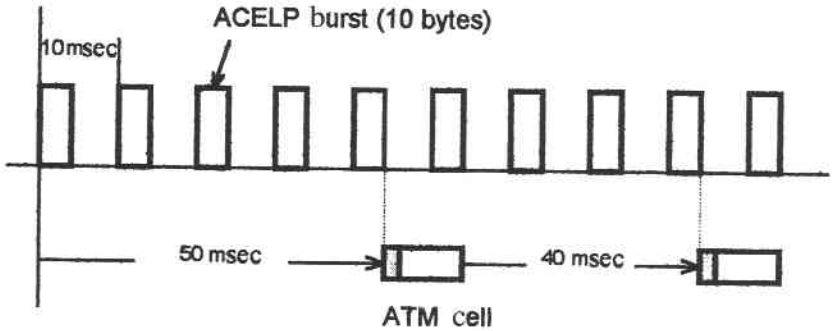


Figure 10.17 Packing low-bit-rate voice into ATM.

dividual channels. This approach is particularly useful for fixed network transport of digital cellular traffic, which inherently involves multiple connections to and from base stations.

Example 10.3. Determine the probability that the delay of an ATM voice cell exceeds 10 msec if the cell is carried on a DS1 access link to an ATM network. Assume the access link is 90% utilized on average.

Solution. The solution to this problem utilizes queuing theory as presented in Chapter 12 where more problems of this type are provided. It involves determining the probability that the DS1 access queue contains enough cells to represent 10 msec of transmission time. Assuming all but the framing bit of the DS1 signal is used to carry ATM cells, the service time (t_m) of a single cell is determined as

$$t_m = \frac{53 \times 8}{192 \times 8000} = 276\ \mu\text{sec}$$

Therefore, 10 msec of delay represents 10/0.276 = 36.2 cell times. Equation 12.25 of Chapter 12 is used to get a solution:

$$p(>t) = \rho e^{-(1-\rho)t/t_m}$$

$$p(>10\ \text{msec}) = (0.9)e^{-(1-0.9)36.2} = 0.025$$

Example 10.3 indicates that 2.5% of the cells will be delayed by more than 10 msec. Ten milliseconds of delay on an access link is quite acceptable considering there are presumably only two access links in a connection. Delays of this magnitude are not acceptable on internal links of an ATM network, but the internal links have very high bandwidths and as such very short service times so queuing delay of voice cells

is not much of a consideration—as long as voice cells get adequate priority and the CAC limits the number of voice calls on a link.

The solution to Example 10.3 involves a couple of simplifying shortcuts. The first of these is the use of Equation 12.25, which is intended for exponential service times as opposed to constant services times implied by fixed-length cells. As discussed in Chapter 12, constant-service-time analyses are generally not appropriate for ATM cell traffic because the cells are not statistically independent. A large file or extended duration talkspurt creates a burst of cells that have an effect comparable to a variable-length message—hence the use of an exponential service time model.

A second simplification of the solution of Example 10.3 is an assumption of homogeneous traffic—all voice cells. Multiple voice sources are reasonably well behaved from a traffic statistics point of view because no one source can produce an instantaneous burst of cells as can a file transfer. When a mixture of traffic types with a mixture of arrival statistics and service priorities exist, the analysis is quite involved and typically requires simulation (as opposed to time-honored analytical solutions). Other factors that influence the analyses are the queue size, the queue discipline (priorities and order of service), cell discard algorithms, flow control algorithms, and the specific CAC algorithm (what percentage of a link is allocated to real time versus discretionary traffic). An extensive amount of research has been undertaken to analyze the performance of various ATM switch structures and processing algorithms. References [26–28] are some good examples.

The condition of 90% loading of voice channels in Example 10.3 is required to achieve a controlled amount of delay. If 100% loading of voice channels is attempted, as implied by Example 10.2, the average queuing delay is unbounded. The channel can still be 100% loaded if the additional 10% is discretionary, non-real-time traffic. The solution requires voice traffic to be given priority in a manner that never sends discretionary traffic if there are any voice cells in the queue. Otherwise, the discretionary traffic impacts the queuing analysis. Cell loss for the discretionary (e.g., best effort) traffic will be significant if a link is allocated at close to 100% of capacity.

Traffic Shaping

Traffic shaping is a term used to denote the control of sources of an access node so that they conform to a particular set of traffic statistics (e.g., SCR and PCR). Traffic shaping may involve flow control for some types of traffic or cell discarding for others. Because the statistics of a group of sources is less variable than the statistics of individual sources, traffic shaping can be implemented with less impact on individual traffic streams if a group of sources (VCCs) are packed into a group and shaped as a VPC. If a VPC creation process has access to the sources of the individual VCCs (e.g., 64-kbps PCM voice channels), additional possibilities for traffic shaping occur. In this case the compression processes for the individual VCCs can be controlled to achieve a constant composite data rate with speech quality degradation distributed across each of the channels as opposed to degradation occurring to an individual connection as a result of cell loss or excessive delay. For a description of other techniques of traffic shaping and an associated analysis of delay and cell loss, see references [29 and 30].

TABLE 10.1 Uncompressed Bit Rates of Digitized Standards

Standard	Pixels/Line	Lines/Frame	Frames/sec	Pixels/sec	Mbps
NTSC	640	480	30	9.216	221
PAL	580	575	50	16.675	400
HDTV	1920	1080	30	62.208	1493

This principle can also be applied to MPEG-2 video encoding of multiple, real-time video signals [31, 32]. Maintaining a constant composite data rate of multiple independent channels complicates the multiplexing process because a figure of quality most be prepared and compared across the multiple sources to determine which source is least affected by a lower bandwidth allocation. Undoubtedly, this process also adds delay to the channels since double-pass encoding may be required.

Video

Table 10.1 identifies three major television standards and the bit rates required to digitally encode them with no compression and using 24 bits per pixel. The three standards are NTSC for North American broadcast TV, PAL for European broadcast TV, and high-definition TV (HDTV).

Two versions of digital compression encoding have been defined by the Motion Pictures Expert Group: MPEG1 and MPEG2. MPEG1 is intended for VHS-quality video and audio. MPEG2 addresses higher quality requirements of broadcast-quality video and audio as well as HDTV. Because of the encoding complexity, MPEG2 is primarily used in broadcast applications where the encoding cost is shared across multiple destinations. (MPEG2 decoders are relatively inexpensive.) Data rates are variable depending on the nature of the source material (amount of movement). The mean bit rate of MPEG-1 is 1.544 Mbps while MPEG-2 has a mean bit rate of 5 Mbps.

The ATM Forum has defined how MPEG-2 TS packets are to be packed into a AAL5-CBR frame. The use of AAL1 layer is advocated by the ITU (ITU-T J.82). However, ITU-T has agreed to incorporate ATM Forum defined AAL5-CBR packing into its document. According to the ATM Forum, AAL5 services are adequate for MPEG-2 streams because MPEG-2 includes a program clock reference (PCR) and therefore does not need the SRTSs provided by AAL1.

10.4 INTERNET PROTOCOL TRANSPORT

The Internet evolved from a packet-switching network consisting of host computers connected with leased lines to Internet Protocol (IP) routers that were, in turn, interconnected with leased lines. As this evolution occurred, high-speed digital networks also evolved. Figure 10.18 depicts the situation wherein these two developments have come together in various ways. What is shown is five different means of providing IP transport using various combinations of networks: ATM, SONET, frame relay, and direct physical layer fiber.

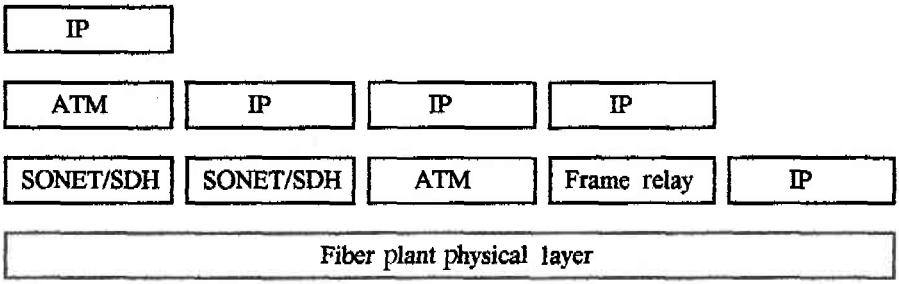


Figure 10.18 Internet transmission layer alternatives.

Other than IP transport directly over fiber, IP over SONET [33] is the most efficient. The most inefficient situation is IP on top of ATM. A straightforward approach to reconciling the two protocol layers is to use a relatively static configuration for the ATM network with permanent virtual circuits. In this manner the IP network thinks it is using leased lines. An AAL5 interface converts the IP packets into fixed-length cells and ships them through the ATM network to an ATM destination node. At the destination node the packet is reconstructed and passed to the IP layer. If this IP node is the final destination, all is well and good. More often than not this is not a final destination because a direct virtual connection between all nodes in a large network is very difficult to manage. (The number of paths grows as the square of the number of nodes and nodes are constantly being added and removed.) When the first ATM node is an intermediate node, the IP packet is passed back down through an AAL5 interface to get back into the ATM network, and so on.

An alternative procedure involves having the ATM network set up SVCs for each packet so the ATM destination coincides with the IP destination. This approach avoids the multiple trips up and down the adaptation layers but introduces an extreme amount of control overhead within the ATM network.

To minimize IP over ATM problems, Cisco systems has developed a networking procedure referred to as "tag switching" that merges IP and ATM protocols. A tag is essentially an ATM destination address that gets attached to each packet (address and all) at a tag edge router. The composite message is passed through the ATM network, which provides tag switching, until it gets reconstructed and delivered to the destination where the tag is removed. The simplification occurs because a tag is easier to process than IP addressing and routing. The IETF has standardized the basic concept of tag switching as multiprotocol label switching (MPLS).

As indicated on the far right of Figure 10.18, the overhead of multiple network layers is avoided by directly connecting one IP node to another IP node with a dedicated fiber link. There are, however, limitations with direct connections. IP networks have been traditionally implemented with leased circuits that are managed and maintained by a common carrier. These leased-line services include provisioning, performance monitoring, and protection switching. Common-carrier provisioning allows rearrang-

ing network interconnections using SONET cross-connect equipment. Performance monitoring allows automated identification and isolation of faulty or marginal equipment. Physical layer protection switching assures an application with a reliable connection despite equipment failures or cable cuts. When direct physical layer interconnect is used, the IP application must implement application layer protection switching to circumvent transmission link failures.

Another, possibly more significant limitation with IP "over glass" is the potential for a lack of carrier-supported signal regeneration. Although transparent optical transmission (on fibers or wavelengths) is a seemingly attractive service offering, a transparent channel, by its definition, cannot be supported with regeneration. The carriers could offer transparent optical channels for limited distance applications, but long-distance applications through optical multiplexing and switching systems cannot be realized without regeneration. Regeneration is what allows a telephone user to "hear a pin drop."

REFERENCES

1 R. E. Kahn, "Resource-Sharing Computer Communications Networks," *Proceedings of IEEE*, Nov. 1972, pp. 1397–1407.

2 H. Rudin and H. Müller, "More on Routing and Flow Control," *National Telecommunications Conference*, 1979, pp. 34.5.1–34.5.9.

3 J. Rinde, "Routing and Control in a Centrally Directed Network," *National Computer Conference*, 1977, pp. 603–608.

4 S. C. K. Young and C. I. McGibbon, "The Control System of the Datapac Network," *International Conference on Communications*, 1976, pp. 137–141.

5 T. S. Yum and M. Schwartz, "Comparison of Adaptive Routing Algorithms for Computer Communications Networks," *National Telecommunications Conference*, 1978, pp. 4.1.1–4.1.5.

6 F. E. Heart, R. E. Kahn, S. M. Ornstein, W. R. Crowther, and D. C. Walden, "The Interface Message Processor for the ARPA Computer Network," *Spring Joint Computer Conference*, 1970, pp. 551–556.

7 L. Kleinrock, "Principles and Lessons in Packet Communications," *Proceedings of IEEE*, Nov. 1978, pp. 1320–1329.

8 Internet Standard 3, "Requirements for Internet Hosts: IETF RFC 1122," Internet Engineering Task Force, 1998.

9 J. McQuillan, "Beyond 'Best-Effort' Routing," *Business Communications Review*, May 1998.

10 M. Hamidi, O. Verscheure, J. P. Hubaux, I. Dalgic, and P. Wang, "Voice Service Interworking for PSTN and IP Networks," *IEEE Communications Magazine*, May 1999, pp. 104–111.

11 The ATM Forum TC, "Traffic Management Specification Version 4.0," April 1996.

12 N. Ghani, S. Nanukul, and S. Dixit, "ATM Traffic Management Considerations for Facilitating Broadband Access," *IEEE Communications Magazine*, Nov. 1998, pp. 98–105.

13 P. Newman, "A Fast Packet Switch for the Integrated Services Backbone Network," *IEEE Journal on Selected Areas of Communications,* Dec. 1988.

14 K. Sarkies, "The Bypass Queue in Fast Packet Switching," *IEEE Transactions on Communications,* May 1991, pp. 766–774.

15 F. M. Chiussi, J. G. Kneuer, and V. P. Kumar, "Low-Cost Scalable Switching Solutions for Broadband Networking: The ATLANTA Architecture and Chipset," *IEEE Communications Magazine,* Dec. 1997, pp. 44–53.

16 S. C. Knauer and A. Huang, "STARLITE: A Wideband Digital Switch," *IEEE Globecom,* Nov. 1984.

17 V. E. Benes, *Mathematical Theory of Connecting Networks,* Academic Press, New York, 1965.

18 K. E. Batcher, "Sorting Networks and Their Applications," *Proceedings of Spring Joint Computer Conference,* 1968, pp. 307–314.

19 Y-S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," *IEEE Journal on Selected Areas of Communications,* Oct. 1987, pp. 1274–1283.

20 E. W. Zegura, "Architectures for ATM Switching Systems," *IEEE Communications Magazine,* Feb. 1993, pp. 28–37.

21 A. Pattavina, "Nonblocking Architectures for ATM Switching," *IEEE Communications Magazine,* Feb. 1993, pp. 38–48.

22 M. de Prycker, *Asynchronous Transfer Mode Solution for Broadband ISDN,* Prentice Hall, Englewood CLiffs, NJ, 1995.

23 R. Y. Awdeh and H. T. Mouftah, "Survey of ATM Switch Architectures," *Computer Networks and ISDN Systems,* Vol. 27, 1995, pp. 1567–1613.

24 "Multiservice Switching Forum Implementation Agreement," msf-architecture-01.00, ATM Forum, Nov. 1998.

25 *Circuit Emulation Service Interoperability Specification Version 2.0,* af-vtoa-0078.000, ATM Forum, Jan. 1997.

26 B. Steyaert, Y. Xiong, and H. Bruneel, "An Efficient Solution Technique for Discrete-Time Queues Fed by Heterogeneous Traffic," *International Journal of Communications Systems,* Mar./Apr. 1997, pp. 73–86.

27 A. la Corte, A. Lombardo, and G. Schembra, "Analysis of Packet Loss in a Continuous-Time Finite-Buffer Queue with Multimedia Traffic Stream," *International Journal of Communications,* Mar./Apr. 1997, pp. 103–115.

28 J. Choe and N. B. Shroff, "A Central-Limit-Theorem-Based Approach for Analyzing Queue Behavior in High-Speed Network," *IEEE/ACM Transactions on Networking,* Oct. 1998, pp. 659–671.

29 K. Sriram, T. G. Lyons, and Y. T. Wang, "Anomalies Due to Delay and Loss in AAL2 Packet Voice Systems: Performance Models and Methods of Mitigation," *IEEE Journal on Selected Areas in Communications,* Jan. 1999, pp. 4–17.

30 K. Sriram and Y. T. Wang, "Voice over ATM Using AAL2 and Bit Dropping: Performance and Call Admission Control," *IEEE Journal on Selected Areas in Communications,* Jan. 1999, pp. 18–28.

31 S. Gringeri, K. Shuaib, R. Egorov, A. Lewis, B. Khasnabish, and B. Basch, "Traffic Shaping, Bandwidth Allocation, and Qualilty Assessment for MPEG Video Distribution over Broadband Networks," *IEEE Network,* Nov./Dec. 1998.

32 M. Krunz, "Bandwidth Allocation Strategies for Transporting Variable-Bit-Rate Video Traffic," *IEEE Communications Magazine,* Jan. 1998, pp. 40–46.

33 J. Manchester, J. Anderson, B. Doshi, and S. Dravida, "IP over SONET," *IEEE Communications Magazine,* May 1998, pp. 136–142.

PROBLEMS

10.1 Determine the average transmission capacity required to retransmit 50-kbyte messages if they are transmitted intact across a single transmission link with a bit error rate of 10^{-6}. What is the average retransmission capacity required if the messages are packetized into 2-kbyte packets? (Ignore the size of the packet headers.)

10.2 Repeat Problem 10.1 with 2 bit error probability of 10^{-4}.

10.3 Determine the total number of bits in an HDLC frame if an information field contains nothing more than 4 bytes of all 1's. Assume the minimum size for all fields of overhead. Include the opening and closing flag in the calculation.

10.4 Determine the number of 2×2 switching elements in a 32×32 banyan switch matrix.

10.5 Determine the transmission efficiency that results when a single channel of G.723.1 compressed voice is packed into ATM cells and transmitted with minimal delay. G723.1 utilizes 30 ms processing frames. Assume the higher of the two standard rates (6.3 kbps). How much delay is added to the voice channel if the ATM cells are packed with close to 100% of capacity? (Ignore speech activity considerations.)

11

DIGITAL SUBSCRIBER ACCESS

By the early 1990s the public telephone networks of the world had converted to digital technology in virtually all of the internal transmission and switching equipment. Toward the latter half of the 1990s new digital applications for traditional analog technology focused on providing customer access to the digital network. The Integrated Services Digital Network (ISDN) digital subscriber loops, defined in the mid 1980s, represent one approach to digital access. Although ISDN has been deployed in many markets, it has not achieved widespread penetration for various reasons. This chapter contrasts ISDN access with alternate approaches for accessing the digital network.

Broadly speaking, there are four basic transmission technologies used in providing local digital access: copper wire pairs, fiber, coaxial cables, and wireless. Copper access utilizes telephony wire pairs by replacing analog transmission with digital transmission for both voice and data (e.g., ISDN), adding digital transmission onto conventional analog voice transmission (e.g., ADSL), or carrying digital signals in the existing analog loop using modems. The rapid advancement of modems for realizing transmission rates of 33.6 kbps in a V.34 voiceband modem or roughly 56 kbps for a V.90 modem is a main reason for less than expected ISDN usage.

Wide-bandwidth transmission on copper pairs installed for analog POTS service is complicated by several factors. First, loading coils on longer loops must be removed. Second, it may be necessary to remove bridged taps for adequate performance. ISDN basic rate access typically requires bridged tap removal, but newer approaches such as ADSL utilize more sophisticated channel characterization and equalization techniques to accommodate bridged taps. Crosstalk and interference at high frequencies are other considerations. A few wire pairs in a cable can carry the high-frequency signals, but if a large number of pairs try to be active at the same time, crosstalk becomes a significant limitation. Interference from external sources such as AM radio stations is a consideration at higher frequencies (e.g., 1 MHz). Some higher rate access systems are designed to detect narrowband interference and mask it out.

Coaxial cable access involves adding data transmission onto unused frequency bands of a cable TV system. Fiber access has the potential for providing very wide bandwidths but requires installation of new transmission facilities dedicated to the new services. Wireless access involves microwave distribution systems like mul-

tichannel multipoint distribution service (MMDS) and local microwave distribution system (LMDS) or satellite-based services such as DSS.

11.1 INTEGRATED SERVICES DIGITAL NETWORK

In addition to the digitization of the internal portions of public telephone networks, a lesser known but also significant change involved the development of common-channel signaling (CCS) for network control. Both the digitization and the use of CCS started at internal portions of the network and migrated toward the periphery. Except for some special data service offerings and a few network-based features derived from the signaling network, these facilities provided no direct benefit to the end users. As shown in Figure 11.1, ISDN is a service offering that extends access to both of these facilities to the end user. Access to the digital transport facilities occurs on 64-kbps bearer (B) channels while access to the signaling network occurs on 16- or 64-kbps signaling (D) channels. Major features or benefits made available by these channels are listed in Tables 11.1 and 11.2, respectively.

Two levels of digital access to the ISDN network have been standardized: basic rate access and primary rate access. As shown in Figure 11.2, the (worldwide) basic rate interface (BRI) standard is also referred to as a 2B + D interface. In North America, the primary rate interface (PRI) standard is sometimes referred to as 23B + D while the ITU-T counterpart is 30B + D. The North American PRI is fundamentally a 1.544-Mbps DS1 signal with the D channel replacing one of the 24 message channels (usually the last one). To achieve a 64-kbps clear channel capability, a B8ZS line code is used to eliminate one's density requirements and common-channel signaling frees up the signaling bits so the full 64-kbps bandwidth is available for user data. The ITU-T

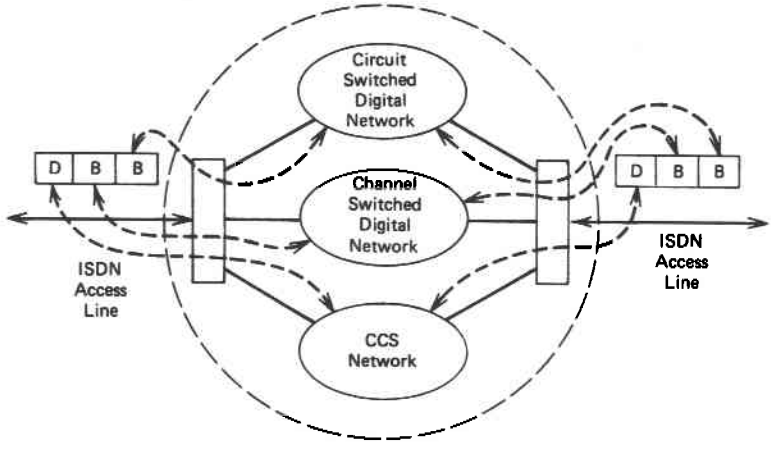


Figure 11.1 Integrated Services Digital Network access to circuits, channels, leased lines, and common-channel signaling.

TABLE 11.1 Features of ISDN B Channels

1. End-to-end four-wire digital circuits: no loss or echoes for voice circuits using digital instruments
2. Shared network access for voice, data, and leased lines
3. Relatively high bandwidth data channels (64 kbps)
4. Lower error rates than typical voiceband modems
5. In-service performance monitoring
6. Possible expansion of speech bandwidth because elimination of tandem encodings allows greater one-time quantization errors

PRI is a 2.048-Mbps E1 digital signal with the D channel occupying the signaling channel (time slot 16). Because a single D channel can support more than one PRI, 24B and 31B interfaces are allowed for additional PRIs in a group of PRIs.

11.1.1 ISDN Basic Rate Access Architecture

An ISDN basic rate access line is a standard copper pair that has been specially conditioned to support a bidirectional 160-kbps aggregate data rate. Transmission technology required for basic rate access is generally referred to as the digital subscriber loop (DSL). Complications arise when using existing analog pairs. The principal considerations are bridged taps and wire gauge changes, both of which cause reflections that impact higher speed digital signals. To allow flexibility in the selection and deployment of the DSL, the ITU-T basic rate specification [1] does not define a two-wire transmission standard. Instead, it establishes an interface standard that assumes the presence of a network termination module that converts any chosen transmission system to the standard interface. In the interest of supporting deregulated customer premises equipment, the Exchange Carriers Standard Association in the United States established a basic rate transmission standard [2] so CPE equipment could connect directly to the transmission link or select network termination modules from alternate vendors. Figure 11.3 depicts the architecture and associated terminology of a North American BRI.

TABLE 11.2 Features of ISDN D Channels

1. Signaling simultaneous with active connections
2. Calling number identification
3. Far-end supervision
4. User-to-user message transfer
5. Telemetry for fire alarms, security, meter reading, etc.
6. Access to packet-switching network
7. Support for network services such as multiple directory numbers sharing one or more B channels, trunk group blocking statistics, and identification of calling number for abandoned or blocked calls

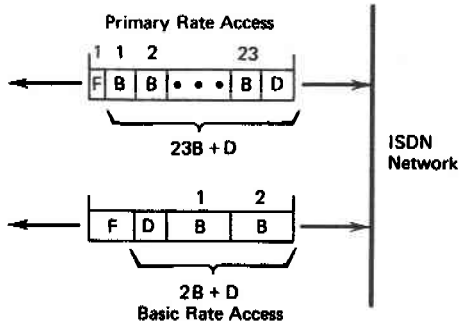


Figure 11.2 Basic rate and primary rate access to ISDN.

Module Definitions

NT1: A network termination module for layer 1 functions that provides physical and electrical termination of the transmission link only. In essence, the NT1 isolates the user from the transmission technology but does not demultiplex or process D channel messages.

NT2: A second level of network termination that implements functions associated with layers 2 and 3 of the OSI protocol stack. Thus, NT2 equipment extracts and processes D channel messages. Representative NT2 equipment includes PBXs, multiplexers, or LAN gateways.

TE1: Type 1 terminal equipment such as a digital telephone that complies with the ISDN S interface recommendation.

TA: Terminal adapter used to convert from an arbitrary (R) interface to the ISDN S interface.

TE2: A non-ISDN terminal that requires a terminal adapter to interface to the ISDN S interface. Prevalent examples of a TE2 equipment are analog telephones or asynchronous (RS-232) data terminals.

Reference Points

U: Interface to the two-wire transmission line.

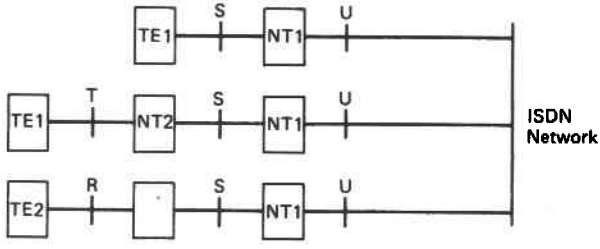


Figure 11.3 Basic rate interface architecture.

T: CCITT ISDN interface defined in Recommendation I.430.

S: Interface to NT2 equipment identical to a T interface.

R: A non-ISDN interface such as an analog tip and ring.

11.1.2 S/T Interface

The S/T interface is defined in ITU-T recommendation 1.430 to be supplied by network termination equipment (NT2/NT1). It is intended for customer premises installations only. (No overvoltage protection is prescribed.) The most significant aspects of the S/T interface are:

1. Four-wire facility (one pair for each direction)
2. One kilometer maximum required distance
3. Alternate space inversion line code (which is the inverse of an AMI line code: see Figure 11.4)
4. Point-to-point or point-to-multipoint configurations
5. Data rate of 192 kbps with 48 kbps of framing, control, and synchronization

The frame structure at reference points S and T is shown in Figure 11.5. As indicated, each 250-msec frame contains 48 bits. Thirty-eight of these bits are common to both directions of transmission and are defined as follows:

16 bits in first B channel (B1)

16 bits in second B channel (B2)

4 bits in the D channel

1 bit in the framing channel F

1 bit in the auxiliary framing channel F_A

The remaining 10 bits are assigned different functions depending on the direction of transmission. From the TE to the NT all remaining 10 bits are defined as L bits, which are used to maintain dc balance. The definition of the 10 bits from the NT to the TE are:

2 L bits for maintaining dc balance

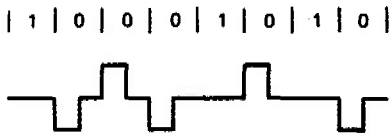


Figure 11.4 Alternate space inversion line code.

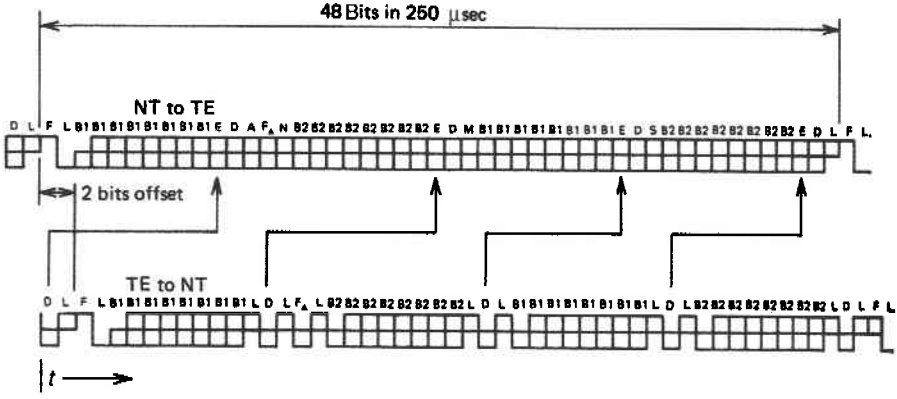


Figure 11.5 S/T frame structure.

4 E bits that echo D bits received from the TE(s)

1 A bit for activation

1 N bit, which is the complement of the F_A bit

1 M bit for multiframe identification

1 S bit for S channel

Figure 11.5 indicates that the framing bit F is always a binary 0. Even though a positive voltage level is indicated, either a positive or negative voltage is allowed so the receivers are not sensitive to wiring polarity. A transmitter always produces the same level, however, so the receiver always receives the same polarity in every framing bit. As an aid in rapid acquisition of the framing pattern, the framing bit always represents a line code violation (it is the same polarity as the previous 0). To maintain dc balance, an L bit with the opposite polarity of the F bit always follows the F bit. The first 0 in a data block following the L bit is encoded with the same polarity as the L bit, which implies another line code violation. Direct-current balancing of this violation is the purpose of the L bit at the end of each data block, which also assures that the next (fixed-polarity) F bit produces a line code violation.

The reason for the additional L bits in the frame from the TE to the NT arises because more than one TE can be connected to the S interface as a passive bus (Figure 11.6). Because the TEs transmit independently of each other, each individual transmission (D channel bits and B channel bytes) is individually dc balanced.

Passive bus operations are also the reason for the existence of the NT-to-TE E bits. Multiple station access to the D channel is controlled by having a terminal wait for an idle code on the NT-to-TE D channel before transmitting on the TE-to-NT D channel. When a terminal begins D channel transmission, it monitors the incoming E bits. If an incoming E bit does not match the previously transmitted D bit, that terminal stops transmitting and waits to seize the channel at a later time. Two levels of priority are defined for accessing the D channel. Signaling information is the highest priority

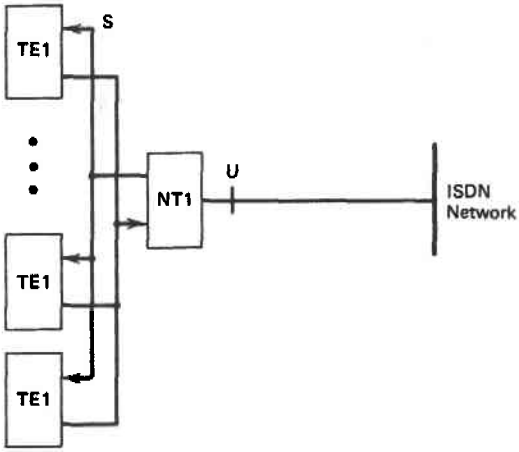


Figure 11.6 S-bus connections.

while user packet messages are the lower priority. All terminals on a passive bus have equal priorities within each level.

Example 11.1. Determine the distance limit imposed by the need to echo E bits in a BRI S/T interface. Assume the speed of transmission of a signal on a pair of wires is one-third the speed of light in a vacuum. Ignore transmitter and receiver filter delays and assume no appreciable delays in the NT logic circuitry.

Solution. From Figure 11.5 it can be seen that the minimum delay between a terminal transmitting a D bit and receiving it back in the following E bit is seven bit times (this is the TE to NT D bit following the first B2 byte). At a 192-kbps data rate the duration of a bit is 5.2 μsec. Thus, the total round-trip propagation time is 7 × 5.2 = 36.4 μ sec. Assuming no appreciable circuitry delays in the NT,

$$\text{Maximum wire length} = (36.4 \times 10^{-6}) \times \tfrac{1}{3}(3 \times 10^{8})\ \text{m}$$

Because round-trip propagation involves both directions of transmission

$$\text{Maximum distance} = \tfrac{1}{2} \times 3.64 = 1.82\ \text{km}$$

Example 11.1 shows that the BRI standard has a fundamental distance limitation that is not far above the minimum transmission distance specification of 1 km.

11.1.3 ISDN U Interface

Prior to the establishment of a standard U interface by the Exchange Carriers Standards Association in North America a number of basic rate transmission systems were developed by telephone equipment suppliers around the world and put into service.

One of these involves the use of time compression multiplex (TCM). TCM, developed by NEC in Japan [3], provides full-duplex transmission on a single pair of wires by alternately transmitting bursts of data in each direction. For this reason it is sometimes referred to as "ping-pong" transmission. One big advantage of TCM transmission is that near-end crosstalk is avoided because a station is never receiving while transmitting. The big disadvantage is that the burst data rate must be more than twice the desired data throughput.

AT&T in the United States also developed a BRI transmission system for the No. 5 ESS end office switching system available with generic releases 5E4 and 5E5 [4]. This system has a 160-kbps data rate utilizing a 50% duty cycle AMI line code. Full-duplex transmission is achieved by simultaneous transmission in both directions using hybrids and echo cancelers (ECs) to separate the two signals, as indicated in Figure 11.7. Beginning with generic 5E6 the No. 5 ESS supports both the AT&T (Lucent) U interface (referred to as a 5E4/5E5 U interface) and the ANSI U interface.

Like the 5E4/5E5 U interface, the ANSI U interface uses simultaneous transmission in both directions with echo cancelers and a data rate of 160 kbps. The major difference is the use of a four-level line code referred to as 2B1Q (two binary digits in one quaternary digit). Thus, the symbol (baud) rate on the line is 80,000 symbols/sec. Because the line code itself does not prevent dc wander, dc restoration is necessary. Reference [5] describes an implementation with adaptive quantized feedback as being the best approach. A 2B1Q line code was chosen primarily because the lower symbol rate minimizes the two dominant transmission limitations in this application: intersymbol interference and near-end crosstalk [6].

The frame format and superframe structure of the ANSI U interface are shown in Figure 11.8. Each frame consists of 240 bits containing 18 framing bits, 216 payload bits (12 fields of 18 2B + D data bits), and 6 overhead bits. Because the frame rate is 667 frames/sec the data rate is 160 kbps. The 6 overhead bits are organized as a block of 48 bits in an eight-frame superframe. Functions included in the overhead bits are 24 bits of embedded operations channel, 1 activation bit, 1 deactivation bit, 1 far-end block error bit, 12 CRC bits, and 9 fixed 1 bits. All bits except the framing bits are scrambled for transmission.

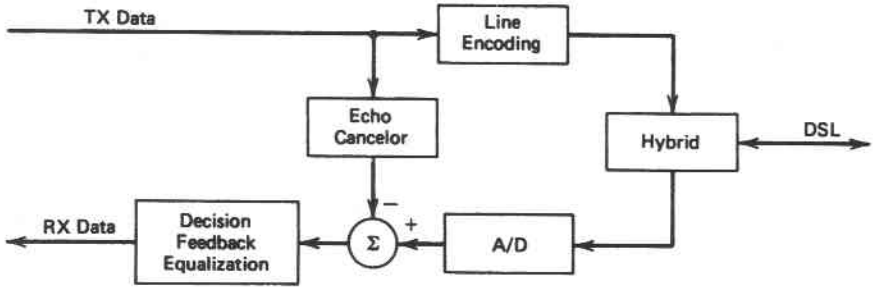


Figure 11.7 ISDN DSL TX/RX block diagram.

	18	18	18	240 ...	18	6
1	ISW	2B+D	2B+D		2B+D	M_1–M_6
2	SW	2B+D	2B+D		2B+D	M_1–M_6
3	SW	2B+D	2B+D		2B+D	M_1–M_6
4	SW	2B+D	2B+D	...	2B+D	M_1–M_6
5	SW	2B+D	2B+D		2B+D	M_1–M_6
6	SW	2B+D	2B+D		2B+D	M_1–M_6
7	SW	2B+D	2B+D		2B+D	M_1–M_6
8	SW	2B+D	2B+D		2B+D	M_1–M_6

SW = Sync Word = +3+3−3−3−3+3−3+3+3
ISW = Inverted SW = −3−3+3+3+3−3+3−3−3
2B+D = | B_1 | B_2 | D | (|8|8|2)
M_1–M_6 = Overhead bits
Data are encoded as 00 = −3, 01 = −1, 11 = +1, 10 = +3

Figure 11.8 ANSI U interface frame and superframe **structure.**

11.1.4 ISDN D Channel Protocol

The D channel protocol is defined in two separate series of ITU-T recommendations: the I series and the Q series. The data link layer (LAPD) is defined in I.441 or Q.921. This protocol is similar to LAPB of the X.25 standard except Q.921 allows more than one logical link. (Therefore separate "connections" can exist for signaling, packet network, or far-end terminals.) The main functions of the data link layer are message sequencing, error checking and retransmission, and data layer link recovery. The network layer of the D channel protocol is defined in I.451 or Q.931. This layer provides connection setup, alerting, routing, and release of ISDN calls. When a B channel accesses a packet network, the X.25 protocol is used while in the connected state.

11.2 HIGH-DATA-RATE DIGITAL SUBSCRIBER LOOPS

A basic rate ISDN digital subscriber loop provides an aggregate, bidirectional data rate of 160 kbps on a single pair of wires. This section describes several transmission techniques that allow much larger bandwidths on copper wire pairs. These new techniques are enabled by the availability of low-cost, high-performance digital signal processing. Several versions of high-speed digital subscriber lines have been developed. The various versions are collectively referred to as xDSL (see Table 11.3).

11.2.1 Asymmetric Digital Subscriber Line

ADSL allows for high data rates to the subscriber and moderate to low data rates from the subscriber to the network. ADSL technology was originally conceived as a means of delivering switched digital video services over a copper loop [7], which obviously do not require high data rates from the subscriber. Although video applications did not

TABLE 11.3 Versions of Digital Subscriber Lines

DSL	Digital Subscriber Line (ISDN basic rate)
ADSL	Asymmetric DSL (9 Mbps downstream, 640 kbps upstream)[a]
HDSL	High-bit-rate DSL (T1/E1 service on two pairs)
SDSL	Single-line DSL (T1/E1 service on one pair)
VDSL	Very high bit rate DSL (52 Mbps downstream, 2.3 Mbps upstream)[a]

[a]Listed data rates are maximum possible values. Downstream is toward the subscriber. Upstream is toward the network.

materialize, the asymmetric data rate is also suited for Internet access. The philosophy behind the asymmetric data rates is that subscribers typically need to receive high-bandwidth data (for Internet file downloads) but normally need to transmit (query) at a relatively low data rate. The subscriber's receive data rate on ADSL varies between 1.5 and 9 Mbps while the subscriber transmit rate varies between 16 and 640 kbps.* The specific data rate utilized depends on the transmission quality of the particular wire pair.

In contrast to ISDN channels, which represent extensions of the digital facilities of the public switched telephone network (PSTN), ADSL channels are separated from the public network at the line interface of the PSTN. As shown in Figure 11.9, ADSL lines terminate at a telephone company central office (or remote terminal) where the data streams are forwarded to and received from a facility that is separate from the telephone network. Transmission between the line interface and an Internet service provider, for example, is typically provided with an ATM connection.

ADSL has two major advantages over ISDN access. First, ADSL data rates provided to the subscriber are significantly higher than the 128-kbps ISDN basic rate. Second, ADSL piggy backs digital transmission on a standard analog telephone wire pair. Thus, existing analog telephones are retained on ADSL but are either replaced by digital phones or are connected through conversion devices when ISDN is utilized.

Two versions of ADSL transmission links have been developed: carrierless amplitude and phase (CAP) modulation and discrete multitone (DMT) modulation. CAP is the first version deployed but DMT has been selected as the standard [8]. Because DMT makes more intensive use of DSP, a DMT implementation typically requires more power—a significant consideration for remote terminal deployment. CAP, on the other hand, its not generally considered to be as flexible as DMT in achieving maximum data rates on some wire pairs or in some interference environments. For a thorough comparison of the two alternatives, see reference [9].

DMT Implementation

Basic parameters of the standard ADSL DMT implementation are provided in Table 11.4. A block diagram of an end-to-end DMT implementation is shown in Figure 11.10. As mentioned in Chapter 6, a DMT implementation utilizes an inverse FFT as

*A less ambitious version of ADSL referred to as G.Lite only attempts to achieve 1.544 Mbps downstream and 384 kbps upstream.

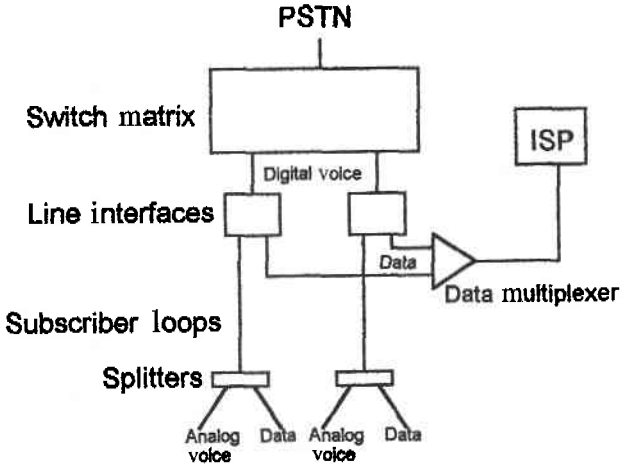


Figure 11.9 ADSL network configuration.

TABLE 11.4 ADSL DMT Implementation Parameters

Subchannel separation	4.3125 kHz
Maximum bits/subchannel	15
	Downlink
Number of subchannels[a]	255
FFT sample size	512 samples
Cyclic prefix	32 samples
Total number of samples	544 (512 + 32)
Sample rate	2.208 MHz (512 × 4312.5)
FFT frame duration	246.377×10^{-6} sec ($544/2.208 \times 10^6$)
Pilot frequency	276 kHz (subchannel 64)
	Uplink
Number of subchannels	31
FFT sample size	64 samples
Cyclic prefix	4 samples
Total number of samples	68 (64 + 4)
Sample rate	276 kHz (64 × 4312.5)
FFT frame duration	246.377×10^{-6} sec ($68/0.276 \times 10^6$)
Pilot frequency	69 kHz (subchannel 16)

[a]The number of subchannels depends on the amount of margin for POTS filtering and on choice of echo canceling or FDM to isolate the two directions of transmission.

a modulator and an FFT as the corresponding demodulator. The following paragraphs identify the basic function of each block in Figure 11.10:

PRS Source: A pseudorandom sequence generator provides a prescribed sequence for characterizing the channel during a training period. Characteristics of the channel that are determined during the training period are attenuation and phase distortion across the band, noise/interference levels across the band, and the information capacity of individual subchannels.

Channel Allocation: Provides assignment of data bits to individual subchannels according to the subchannel capacity determinations obtained during training.

FEC: A combination of Reed–Solomon and convolutional coding.

IFFT QAM Modulation: Conversion of data values to quadrature channel signal amplitudes and conversion to a time-domain waveform using an inverse FFT.

D/A: Digital-to-analog conversion.

TX Filter: Bandpass filter to prevent interference into the voiceband and to smooth the discrete DSP samples.

POTS Splitter: Used at both ends of the line to separate the analog voice from the data. (Simpler versions of ADSL incorporate the splitter function in the modems to simplify installation.)

AGC: Automatic gain control to adjust overall receive level.

Slope/Delay Equalization: A front-end equalizer to partially flatten the frequency response and equalize extreme delay variations in the channel [10].

A/D: Analog-to-digital conversion.

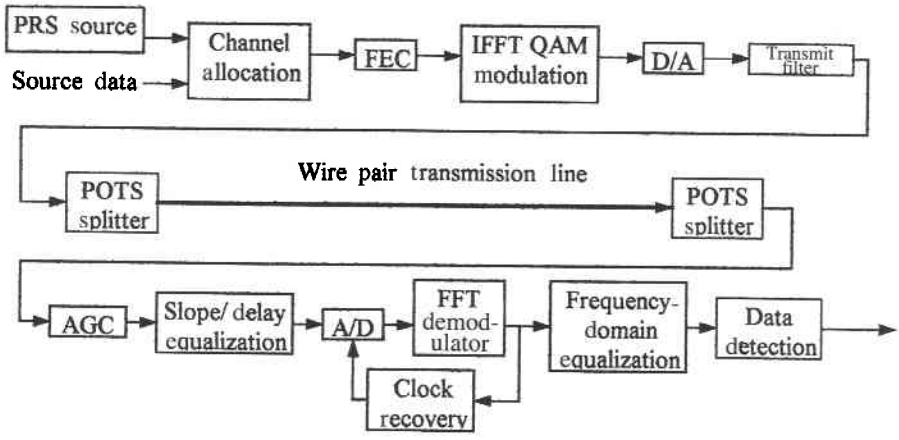


Figure 11.10 ADSL DMT block diagram.

FFT QAM Demodulation: FFT conversion of time-sampled waveform to frequency domain where data values are related to amplitudes of quadrature carriers.

Clock Recovery: A/D sample timing obtained by locking to pilot frequency. The desired sampling rate is eight times the pilot frequency so 1-of-8 phase ambiguity has to be determined by monitoring framing/synchronization bit integrity.

Frequency-Domain Equalizer: Multiplication of complex (quadrature) frequency spectrum by amplitude and phase equalization parameters obtained during training period.

Data Detection and Interleaving: Slicing of quadrature amplitude values to decode data and subsequent generation of composite stream identical to original source data.

11.2.2 VDSL

VDSL is an expanded version of ADSL to achieve even higher bandwidths on particularly short lines as would be available from remote terminals of fiber to the curb systems. A primary motivation for VDSL is potential distribution of HDTV signals. Although several different modulation techniques have been proposed for VDSL, a DMT version seems to be favored [11].

11.3 DIGITAL LOOP CARRIER SYSTEMS

As discussed in Chapter 1, the primary purpose of a digital loop carrier (DLC) system is to reduce or eliminate copper pairs from a central office to the vicinity of a group of subscribers. Even though the transmission link from the central office to the DLC remote terminal is digital, the transmission links from the remote terminal to the subscribers are typically conventional analog loops. Thus, the main purpose of a DLC is not to provide digital subscriber access. Nevertheless, some DLC systems (particularly fiber-based DLC systems) provide options for T1, ISDN, or xDSL digital interfaces. Moreover, the fact that the DLC remote terminal is relatively close to the subscriber locations facilitates the use of these interfaces. (A short copper drop from the remote terminal has low attenuation and crosstalk and is less likely to have loading coils, bridged taps, or multiple sections of wire with varying gauges.) Although a DLC is a natural mechanism to provide enhanced services through ISDN or xDSL digital interfaces, mechanical packaging and power considerations of these interfaces represent unique requirements with respect to strictly POTS applications.

11.3.1 Universal Digital Loop Carrier Systems

As shown in Figure 11.11, a universal digital loop carrier (UDLC) system [12] can be interfaced to any switching system: analog or digital. The interface between the local

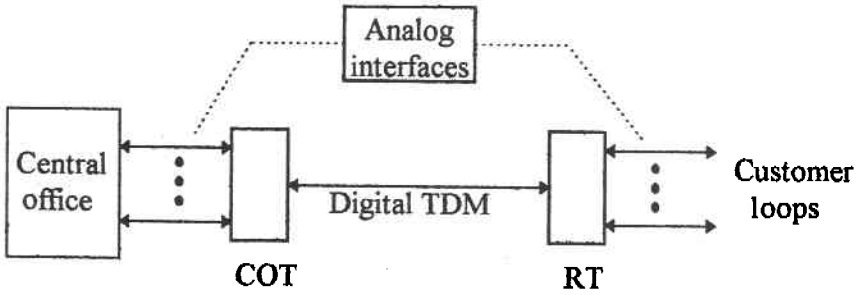


Figure 11.11 Universal digital loop carrier system.

switching system (end office) of the public network and the central office terminal (COT) involves individual circuits (e.g., individual analog tip and ring connections). The multiplexed digital transmission links between the COT and the remote terminal (RT) can be wire pairs or fiber. Each interface of a COT is paired with a corresponding subscriber interface at the RT so the use of a UDLC is transparent to both the switch and the subscriber. In its simplest mode of operation, the UDLC uses pure multiplexing between the COT and the RT so that there is a one-to-one correspondence between a particular TDM channel and the COT/RT interface pair. Some systems can also be configured with concentration wherein the COT/RT pairs are dynamically assigned transmission channels. If the number of requested conversations exceeds the number of channels, blocking occurs. The possibility of blocking introduces nontransparency and implies that some means of returning a reorder tone (fast busy) is needed in the RT.

UDLC installations are configured to match each particular interface of the central office switch with a complimentary interface in the RT. A fully capable system must provide a wide variety of interfaces such as loop-start line, ground-start PBX trunk, foreign exchange lines, and coin telephone interfaces. In some early systems the configuration process involved nothing more than physically installing matching interfaces in respective equipment slots of the COT and RT. More recent systems typically utilize line units with multiple-service capabilities. These systems can be configured electronically (i.e., no straps) with either a local or a remote management interface.

11.3.2 Integrated Digital Loop Carrier Systems

Whenever a universal digital loop carrier system is interfaced with a digital switch, obvious inefficiencies occur in terms of back-to-back demultiplexing–multiplexing and D/A–A/D conversion. As shown in Figure 11.12, an integrated digital loop carrier (IDLC) system eliminates the inefficiency by directly connecting the DLC TDM link to the digital matrix. Typically the direct digital connections are either DS1 or E1 cross-connect signals. Thus, a fiber-based DLC system will typically interface with the switch through multiplexing/demultiplexing equipment as some number of digital cross-connect signals.

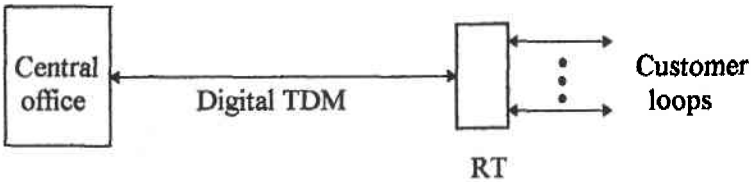


Figure 11.12 Integrated digital loop carrier system.

From a functional point of view an early IDLC systems was nothing more than a distributed switching system wherein some line interfaces of the switch are moved to remote locations. Most central office switch manufacturers provide such a capability—sometimes as simply as using channel banks for analog line interfaces that can be colocated with the matrix or remoted. Central office switch vendors also offer remote switching modules wherein some portion of the matrix itself is remotely located. These configurations provide remote concentration for efficient use of the connecting transmission link and, in some cases, provide local switching in the remote module so connections between two ports of a remote module do not use the transmission link. (If remote local switching is not provided, a connection between two ports of a remote module requires two channels of the transmission link.)

Remote modules of a particular switching system vendor are often implemented with proprietary signaling and management processes that preclude use of DLC equipment from other vendors (except when a DLC vendor develops the ability to emulate a particular switch vendor's remote modules). In the interest of opening IDLC applications to competing vendors, Bellcore established an IDLC standard referred to as GR-303 [13, 14] that compliant switching system vendors must support (possibly in addition to a proprietary IDLC capability). Included in the GR-303 specification are definitions for signaling, provisioning, testing, alarm surveillance, and performance monitoring.

Due, in part, to a move to unbundle LEC local loop services GR-303 has assumed a much broader scope than just an IDLC application. Because a GR-303 capability includes being able to define and administer a myriad of switch interface types, the GR-303 standard can be used for interfacing other types of equipment such as xDSL equipment. ETSI has established a similar IDLC standard for international (ITU) switching equipment referred to as a V5 interface [15, 16].

11.3.3 Next-Generation Digital Loop Carrier Systems

The term next-generation digital loop carrier (NGDLC) has been adopted within the industry to refer to DLC systems that adhere to GR-303 and provide additional configuration options and interfaces. There is no precise definition of what constitutes an NGDLC system. Four basic attributes are adherence to GR-303, optical fiber transmission capabilities (e.g., SONET/SDH), generally larger line sizes, and an ability to interface with a central office operational support system for diagnostics, alarms, and

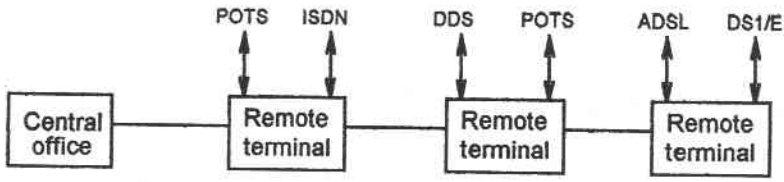


Figure 11.13 Next-generation digital loop carrier.

remote provisioning [14]. Other aspects commonly available in NGDLC systems are depicted in Figure 11.13.

The most important aspect of an NGDLC system from a services point of view is the availability of new services such as copper or fiber digital subscriber interfaces in addition to the conventional POTS interfaces. The most desired digital interfaces are T1/E1, primary rate and basic rate ISDN, ADSL, and VDSL. Other digital interfaces may also be provided for services like broadband data and digital video. Although the diversity of interfaces belies use of a single multiple service line interface for total electronic provisioning, the system must report inconsistencies between installed hardware and the electronic database. A particularly desirable feature of the system shown in Figure 11.13 is drop-and-insert capability, which allows distributed access to a single backbone digital route. GR-303 identifies star configurations, linear ADM distribution, and ADM rings as desirable topologies. NGDLC systems that support advanced interfaces or topologies necessarily use a COT to separate out the special services from the POTS. In essence, the COT performs cross-connect operations to groom and distribute various types of traffic.

11.4 FIBER IN THE LOOP

Fiber in the loop (FITL) is a generic term that refers to one of three more specific descriptions of the use of fiber for local distribution. The first category is fiber to the cabinet or fiber to the neighborhood. These systems are often installed by local telephone companies as part of the feeder portion of their local distribution facilities. Traditional copper pairs for voice service extend from the cabinet to subscriber residences. Because a relatively long copper pair is used for "the last mile," they do not provide much opportunity for enhanced service offerings. Their use and justification are strictly based on reducing the cost of POTS distribution.

The second category of FITL is a fiber-to-the-curb (FTTC) system. As the name implies, these systems are designed to reach within 1000 feet of a subscriber residence. An FTTC system is generally intended to provide enhanced services such as video or high-speed data using ADSL or VDSL. Distribution of the enhanced services from the "curb" location is carried over wire pairs or coaxial cable. These systems are essentially identical to advanced DLC systems with optical transport.

The third category of FITL is fiber to the home (FTTH). These systems obviously offer opportunities for extremely large bandwidths to the home but have significant

deployment obstacles. First, installation in established neighborhoods is expensive because underground installations (under streets and driveways) are normally required. Second, providing network power to subscriber equipment is a major problem. The lack of power for enhanced service applications is not much of a consideration but power to telephones is. Local exchange carriers go to great efforts to ensure independence from commercial power systems for both their office switching systems and the connected subscribers. (A central office typically maintains enough batteries to keep a system up for 24 hours if commercial power is lost. If power is out for longer than this, diesel generators are available.) A further complication of providing telephone service over FTTH is the need to convert a digital voice channel to analog for interfacing to conventional telephones—an additional expense and power problem. Batteries are often offered as an alternative to line powered telephones but logistical problems remain: who maintains the battery and what happens if the battery is dead when commercial power is lost and someone wants to dial 911?

11.5 HYBRID FIBER COAX SYSTEMS

The cable TV systems installed around the country can be augmented with downstream data transmission to subscribers by merely adding "cable modems" utilizing unused or displaced TV channel bandwidth. Upstream transmission from the subscriber to a cable TV head end is much more difficult. Although many cable TV systems were designed and installed with upstream transmission as an option, the bandwidth available to the upstream channels is generally limited and often subject to very high noise and interference levels. Upstream transmission limitations can be somewhat alleviated by utilizing a conventional telephone connection with voiceband modems with data rates up to 28.8 kbps. The telephone modem connection is used in the same basic manner as the upstream channel of an ADSL while relative high-bandwidth downstream data are carried on the cable. The disadvantages of this solution include the need for a subscriber telephone line, the cost of large numbers of telephone channel connections into a service provider, and the need to coordinate dial-up telephone connections with particular cable channel users.

The basic configuration of a hybrid fiber coax (HFC) system is depicted in Figure 11.14. The coaxial cable portions of CATV systems are configured as tree-and-branch topologies with all customers receiving the same, multichannel broadcast signal. Amplifiers are inserted wherever the signal level gets unacceptably low from attenuation and branching loses. Bidirectional amplifiers are shown in Figure 11.14 under the assumption that this system is an application with a return path from the residences. In TV-only applications the return path (if there is one) is used for premium channel selection. In expanded service applications the return path carries voice or data with frequency division multiplexed cable modems in a band from 5 to 42 MHz.

The optical fiber transmission portion of an HFC system represents a replacement of relatively long haul coaxial cable sections with numerous amplifiers. For this reason, the optical transmitters and receivers are designed to carry a wideband analog signal. Notice that the optical links are shared by a large number of customers (anywhere

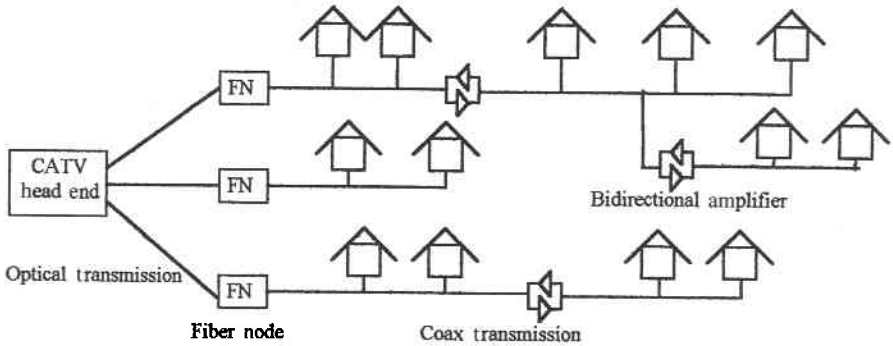


Figure 11.14 Hybrid fiber coax system configuration.

from 100 to 1500). Upgrading an HFC system for new services typically requires greater penetration of the optical fiber portions so that fewer households are connected to a common coaxial cable segment. In the limiting situation, wherein each household is connected through a dedicated coaxial cable, an HFC system becomes an FTTC system.

Downstream digital services can utilize cable modems that typically pack 30–40 Mbps into a 6-MHz analog TV channel. 64-QAM modulation is commonly used. In newer HFC systems, new downstream digital services can be carried at frequencies above 450 MHz while the band from 54 to 450 MHz is reserved for traditional analog TV.

A major impediment to upgrading an HFC system for return channel services is the shared use of a coaxial cable segment common to some number of households. The network termination within each home is passive and bidirectional, which means that all noise and interference within a home is passed onto the common cable to all other homes. Thus, a single source of interference can disrupt the signal to all other homes served by the common coaxial cable. Furthermore, the noise and interference of all households are additive, indicating the need to limit the number of households served by a single coax segment. An additional drawback of the shared cable is the need for some form of encryption for content security. TDMA return channels help minimize the interference problem by blocking all output energy from a residence except when an allotted time slot occurs.

Using an HFC system for POTS has the same basic drawbacks as an FTTH system in that there is no inherent facility for line powering the telephones. Thus, HFC might not be used for primary ("lifeline") POTS distribution but could be quite effective in providing secondary telephone applications. The main attribute of HFC for enhanced services is the ability to provide dynamic assignment of high-rate digital downstream channels and relatively low rate full-period upstream channels.

11.6 VOICEBAND MODEMS

Voiceband modem technology improved dramatically in the early 1990s with the culmination of 33.6 kbps becoming standard with ITU recommendation V.34 [17]. The

rapid advance of voiceband modem performance was due to two primary factors: the availability of economical DSP technology for equalization/echo canceling and the improved quality of the network in terms of lower noise and distortion resulting from the near-all-digital implementation. In an all-digital network the only significant source of noise is the quantization noise of the analog-to-digital conversion.

Example 11.2. Determine the theoretical maximum data rate of a perfectly equalized voiceband modem undergoing a single A/D and D/A conversion.

Solution. The signal-to-quantizing-noise ratio (SQR) is derived in Chapter 3 to be on the order of 36 dB, which corresponds to a power ratio of 3981. Using this value in Shannon's theorem for the theoretical capacity of a channel yields

$$\begin{aligned} C &= W\log_2(1 + \text{SNR}) \quad \text{bps} \\ &= 3100\log_2(1 + 3981) \\ &= 37 \text{ kbps} \end{aligned}$$

where the channel is limited by the transformer coupling, 60 Hz elimination filters, and D/A smoothing filters to extend from 300 to 3400 Hz.

11.6.1 PCM Modems

As the previous problem indicates, a V.34 voiceband modem provides data rates that are near the theoretical limit imposed by quantization noise alone. Recognition that the principal source of noise in the end-to-end connection is the quantization noise of the A/D converters leads to alternative modem implementations that directly utilize the digital 64-kbps channel and eliminate the quantization noise [18]. These modems are commonly referred to as PCM modems.

As shown in Figure 11.15, a V.90 PCM modem relies on the digital network to carry an unaltered digital signal from a digital source to a digital-to-analog conversion device (codec) at an analog subscriber interface. The codec converts the PCM codewords to PAM samples that are detected by the receiving customer premises modem and converted back to the original PCM data. Successful data detection requires the receiving circuitry to adequately equalize the combined distortion of the D/A smoothing filter and the transmission link, to know the quantization levels of the codec, and to become synchronized to the D/A conversion clock. The equalization and quantization requirements are determined during an initialization process while clock synchronization requires processing of data transitions in the multilevel received waveform.

It might seem that a PCM modem could provide a data rate of 64 kbps. However, several factors restrict the data rate to something less than 64 kbps. The first of these is the bandpass filter in the D/A codec (for 60 Hz elimination and sample smoothing). A second constraint is the possibility that the digital path through the network might

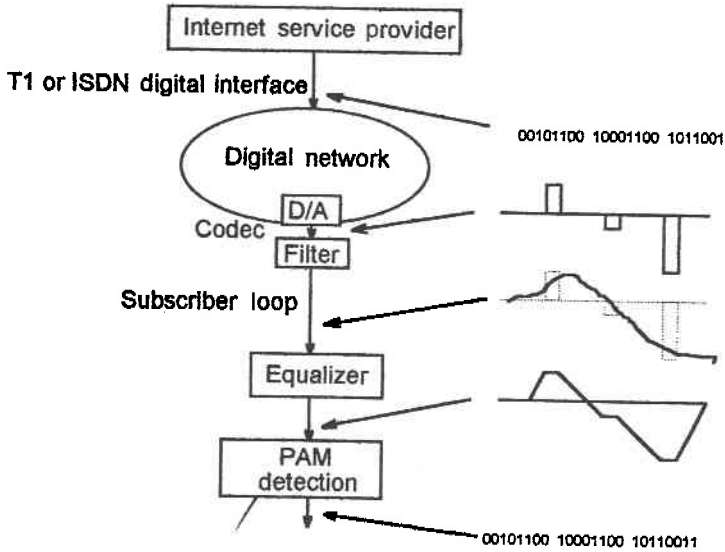


Figure 11.15 Downstream V.90 modem concept.

include a digital pad for changing the signal level of the assumed analog signal.* A third constraint is the possibility (in North America) that robbed bit signaling might be in use on one or more of the digital links.

The fact that the overall bandwidth of the channel is slightly over 3kHz implies that the maximum, intersymbol, interference-free pulse rate is just over 6 kHz. Thus, the actual sample rate of 8 kHz implies that some amount of intersymbol interference is inevitable (assuming the PCM samples are independent).

The lack of a low-frequency transmission response is accommodated by V.90 modems by utilizing every eighth PCM sample solely for dc restoration. Thus, the maximum data rate is 56 kbps. If robbed bit signaling is present, its effects are minimized by determining, during initialization, which frames in the six frame sequences are signaling frames and then using only 7 bits per sample in those frames and 8 bits per sample in the nonsignaling frames. Digital pads can be accommodated by detecting their presence and modifying the digital codewords according to the particular amount of attenuation being inserted by the pad.

The V.90 uplink direction is implemented as a conventional (V.34) modem, thereby implying asymmetry in the data rates. It is conceivable that PAM signaling could be used in both directions, but the uplink is more complicated to implement and is often unnecessary because most applications (e.g., Internet access) are inherently asymmetric in the data rates required. It is also possible to utilize analog PAM on both

*Digital pads are often included in the codecs for digital setting of gain levels to desired voiceband signal levels.

ends of the connection (as opposed to just one end), but these implementations are considerably more complicated [19, 20].

The main attraction of PCM modems is that they provide almost the same data rate as an ISDN B channel but do not require changes in the line interfaces or special treatment of the customer loop (e.g., the elimination of bridged taps). When the customer loop is a short drop from a remote terminal of a digital loop carrier, the maximum data rate of 53–56 kbps is assured.

11.7 LOCAL MICROWAVE DISTRIBUTION SERVICE

The FCC has allocated two separate microwave frequency bands for digital radio access communications services. The first of these is Multichannel Multipoint Distribution Service (MMDS), which operates at 2 GHz. MMDS is essentially a wireless cable system and, as such, provides only one-way transmission. Reverse channel communications requires dial-up modem connections through the telephone network.

Local Microwave Distribution Service (LMDS) operates at 28 GHz and provides cell-based, two-way communications. Bandwidths available with LMDS are dependent on a particular service provider's frequency plan. Typically, these plans provide bidirectional data rates on the order of 1.5–50 Mbps, although the total allocated bandwidth of almost 1 GHz can be partitioned to provide higher rates or asymmetric rates if desired. International LMDS allocations have been made in a range of 25–31 GHz. These systems are also referred to as Local Multipoint Communications Systems (LMCS) in Canada and Brazil.

The principal applications supported by LMDS are voice, video, and high-bandwidth data. The immense amount of bandwidth available is the main attraction. The major cost of an LMDS system is associated with the subscriber radio equipment. Thus, after the cell sites are established, additional costs are incurred only as subscribers begin paying for service. This situation is in contrast with new fiber-based facilities that require major, up-front investment before revenues are realized. The most immediate opportunity for LMDS is to offer an alternative to high-speed digital leased lines for businesses. High-speed Internet access and HDTV to individual residences are secondary opportunities.

The major disadvantages of LMDS are the need for FCC licensing, line-of-sight transmission, distance limits of 2–3 miles, extreme attenuation in heavy rainfalls, potential interference from other services such as satellites in the same bands, and the need for local power in all subscriber (telephone) sites. Fade margins or dynamic power control on the order of 40 dB overcome all but the most severe rainfall rates.

The requirement for line-of-sight transmission means that a single base station cannot communicate with all locations in a cell (except maybe in West Texas or the Australian outback). Overlapping cells provide greater coverage, but some locations are sure to be in the shadow of all base stations unless reflectors can be installed. Line-of-sight transmission can be further compromised by the appearance of buildings or trees after initial deployment.

11.8 DIGITAL SATELLITE SERVICES

Direct Satellite Service (DSS) has recently become a viable alternative to cable TV services. Because the DSS system utilizes digital transmission for its TV channels, it is straightforward for a DSS provider to offer data communications services to the TV subscribers. However, because DSS is a one-way communications service, dial-up modem connections through the telephone network are required for two-way data communications. DSS is also limited to the total bandwidth (400 MHz) of the satellite transponders, which must be shared by the TV transmissions and any allocated data channels.

In contrast to DSS, which uses geostationary satellite orbits, Low Earth Orbit Satellites of Iridium and Teledesic provide opportunities for lower cost, bidirectional user terminals. As mentioned in Chapter 9, Iridium is primarily oriented to relatively narrow band applications such as voice, messaging, and paging. Teledesic, on the other hand, is oriented to wide-bandwidth data applications.

REFERENCES

1 Recommendation I.430 ISDN User-Network Interface: Layer I Recommendations, Fascicle III.8, *CCITT Blue Book,* 1988, pp. 171–240.

2 American National Standards for Telecommunications, "Integrated Services Digital Network—Basic Access Interface for Use on Metallic Loops for Application on the Network Side of the NT-Layer I Specification," ANSI TI.601-1988.

3 R. Komiya, K. Yoshida, and N. Tamaki, "The Loop Coverage Comparison between TCM and Echo Canceller under Various Noise Considerations," *IEEE Transactions on Communications*, Nov. 1986, pp. 1058–1067.

4 5ESS®. Switch ISDN Basic Rate Interface Specification, 5E5 Generic Program Issue 1.00, AT&T, Dec. 1987, pp. II-B-1/15.

5 S. Sugimoto, K. Hayashi, and F. Mano, "Design of 2B1Q Transceiver for ISDN Subscriber Loops," *IEEE International Conference on Communication*, June 1989, pp. 8.1.1–8.1.5.

6 J. W. Lechleider, "Line Codes for Digital Subscriber Lines," *IEEE Communications Magazine*, Sept. 1989, pp. 25–32.

7 W. Chen and D. Waring, "Applicability of ADSL to Support Video Dial Tone in the Copper Loop," *IEEE Communications Magazine,* May 1994, pp. 102–109.

8 "Network and Customer Installation Interfaces—Asymmetric Digital Subscriber Line (ADSL) Metallic Interface," ANSI T1.413-1995, Aug. 1995.

9 B. Saltzberg, "Comparison of Single-Carrier and Multitone Digital Modulation for ADSL Applications," *IEEE Communications Magazine,* Nov. 1998, pp. 114–122.

10 W. Y. Chen, *DSL Simulation Techniques and Standards Development for Digital Subscriber Line Systems,* Macmillan Technical Publishing, Indianapolis, 1998.

11 J. M. Cioffi, V. Oksman, J. J. Werner, T. Pollett, P. M. P. Spruyt, J. S. Chow, and K. S. Jacobsen, "Very-High-Speed Digital Subscriber Lines," *IEEE Communications Magazine,* Apr. 1999, pp. 72–79.

12 "Functional Criteria for Digital Loop Carrier Systems," Technical Reference, TR-NWT-000057, Bellcore, Morristown, NJ, Jan. 1993.

13 "Integrated Digital Loop Carrier System Generic Requirements, Objectives and Interface," Generic Requirement, GR-NWT-000303, Bellcore, Morristown, NJ, Dec. 1998.

14 "Generic Operations Interfaces Using OSI Tools: Information Model for Integrated Digital Loop Carrier and Fiber in the Loop Systems," Generic Requirements, GR-NWT-002833, Bellcore, Morristown, NJ, Dec. 1996.

15 Recommendation G.964 (06/94), "V-Interfaces at the Digital Local Exchange (LE)—V5.1-Interface (Based on 2048 kbit/s) for the Support of Access Network (AN). ITU, Geneva, Switzerland.

16 Recommendation G.965 (03/95)—V-Interfaces at the Digital Local Exchange (LE)-V5.2 Interface (Based on 2048 kbit/s) for the Support of Access Network (AN). ITU, Geneva, Switzerland.

17 G. D. Forney, L. Brown, M. V. Eyuboglu, and J. L. Moran, "The V.34 High-Speed Modem Standard," *IEEE Communications Magazine,* Dec. 1996, pp. 28–33.

18 B. Townsend, "High Speed Communications System for Analog Subscriber Connections," U.S. Patent 5,801,695, Sept. 1, 1998.

19 I. Kalet, J. E. Mazo, and B. R. Saltzberg, "The Capacity of PCM Voiceband Channels," *International Communications Conference,* 1993, pp. 507–511.

20 E. Ayanoglu, G. Golden, R. Jones, J. Mazo, and D. Shaw, "High Speed Quantization-Level-Sampling Modem with Equalization Arrangement," U.S. Patent 5,528,625, June 18, 1986.

PROBLEMS

11.1 Determine the maximum loop length of a ping-pong transmission system utilizing 8-kHz data bursts of 50 μsec duration in each direction. Assume the velocity of propagation is one-third the speed of light.

11.2 Determine the maximum theoretical data rate that can be achieved by a voiceband modem passing through the public network and a single universal digital loop carrier system. Assume there are no signal imperfections in the digital portions of the facilities.

11.3 Determine the range of data rates achievable with a V.90 modem passing through six T1 facilities with "robbed bit" signaling.

11.4 Assume a multipair cable system is used to carry bidirectional high-bandwidth AMI signals and that near-end crosstalk coupling from one pair to another is 1% (–20 dB). If the system is crosstalk limited, determine the maximum number of active pairs for a bit error rate of 10^{-6}.

11.5 Repeat Problem 11.4 with a 3-dB interference margin for noise and other imperfections.

12

TRAFFIC ANALYSIS

Except for station sets and their associated loops, a telephone network is composed of a variety of common equipment such as digit receivers, call processors, interstage switching links, and interoffice trunks. The amount of common equipment designed into a network is determined under an assumption that not all users of the network need service at one time. The exact amount of common equipment required is unpredictable because of the random nature of the service requests. Networks conceivably could be designed with enough common equipment to instantly service all requests except for occurrences of very rare or unanticipated peaks. However, this solution is uneconomical because much of the common equipment is unused during normal network loads. The basic goal of traffic analysis is to provide a method for determining the cost-effectiveness of various sizes and configurations of networks.

Traffic in a communications network refers to the aggregate of all user requests being serviced by the network. As far as the network is concerned, the service requests arrive randomly and usually require unpredictable service times. The first step of traffic analysis is the characterization of traffic arrivals and service times in a probabilistic framework. Then the effectiveness of a network can be evaluated in terms of how much traffic it carries under normal or average loads and how often the traffic volume exceeds the capacity of the network.

The techniques of traffic analysis can be divided into two general categories: loss systems and delay systems. The appropriate analysis category for a particular system depends on the system's treatment of overload traffic. In a loss system overload traffic is rejected without being serviced. In a delay system overload traffic is held in a queue until the facilities become available to service it. Conventional circuit switching operates as a loss system since excess traffic is blocked and not serviced without a retry on the part of the user. In some instances "lost" calls actually represent a loss of revenue to the carriers by virtue of their not being completed.

Store-and-forward message or packet switching obviously possesses the basic characteristics of a delay system. Sometimes, however, a packet-switching operation can also contain certain aspects of a loss system. Limited queue sizes and virtual circuits both imply loss operations during traffic overloads. Circuit-switching networks also incorporate certain operations of a delay nature in addition to the loss operation

of the circuits themselves. For example, access to a digit receiver, an operator, or a call processor is normally controlled by a queuing process.

The basic measure of performance for a loss system is the probability of rejection (blocking probability). A delay system, on the other hand, is measured in terms of service delays. Sometimes the average delay is desired, while at other times the probability of the delay exceeding some specified value is of more interest.

Some of the analyses presented in this chapter are similar to those presented in Chapter 5 for the blocking probabilities of a switch. Chapter 5 is concerned mostly with matching loss—the probability of not being able to set up a connection through a switch under normal or average traffic volumes. This chapter, however, is mostly concerned with the probability that the number of active sources exceeds some specified value. Typically, the specified value is the number of trunk circuits in a route.

12.1 TRAFFIC CHARACTERIZATION

Because of the random nature of network traffic, the following analyses involve certain fundamentals of probability theory and stochastic processes. In this treatment only the most basic assumptions and results of traffic analysis are presented. The intent is to provide an indication of how to apply results of traffic analysis, not to delve deeply into analytical formulations. However, a few basic derivations are presented to acquaint the user with assumptions in the models so they can be appropriately applied.

In the realm of applied mathematics, where these subjects are treated more formally, blocking probability analyses are referred to as congestion theory and delay analyses are referred to as queuing theory. These topics are also commonly referred to as traffic flow analysis. In a circuit-switched network, the "flow" of messages is not so much of a concern as are the holding times of common equipment. A circuit-switched network establishes an end-to-end circuit involving various network facilities (transmission links and switching stages) that are held for the duration of a call. From a network point of view, it is the holding of these resources that is important, not the flow of information within individual circuits.

On the other hand, message-switching and packet-switching networks are directly concerned with the actual flow of information, since in these systems traffic on the transmission links is directly related to the activity of the sources.

As mentioned in Chapter 7, circuit switching does involve certain aspects of traffic flow in the process of setting up a connection. Connect requests flow from the sources to the destinations acquiring, holding, and releasing certain resources in the process. As was discussed, controlling the flow of connect requests during network overloads is a vital function of network management.

The unpredictable nature of communications traffic arises as a result of two underlying random processes: call arrivals and holding times. An arrival from any particular user is generally assumed to occur purely by chance and be totally independent of arrivals from other users. Thus the number of arrivals during any particular time interval is indeterminate. In most cases holding times are also distributed randomly. In some applications this element of randomness can be removed by assuming constant hold-

ing times (e.g., fixed-length packets). In either case the traffic load presented to a network is fundamentally dependent on both the frequency of arrivals and the average holding time for each arrival. Figure 12.1 depicts a representative situation in which both the arrivals and the holding times of 20 different sources are unpredictable. The bottom of the figure depicts activity of each individual source while the top displays the instantaneous total of all activity. If we assume that the 20 sources are to be connected to a trunk group, the activity curve displays the number of circuits in use at any particular time. Notice that the maximum number of circuits in use at any one time is 16 and the average utilization is a little under 11 circuits. In general terms, the trunks are referred to as servers, and a trunk group is a server group.

Traffic Measurements

One measure of network capacity is the volume of traffic carried over a period of time. Traffic volume is essentially the sum of all holding times carried during the interval. The traffic volume represented in Figure 12.1 is the area under the activity curve (approximately 84 call minutes).

A more useful measure of traffic is the traffic intensity (also called traffic flow). Traffic intensity is obtained by dividing the traffic volume by the length of time during which it is measured. Thus traffic intensity represents the average activity during a period of time (10.5 in Figure 12.1). Although traffic intensity is fundamentally dimensionless (time divided by time), it is usually expressed in units of erlangs, after the Danish pioneer traffic theorist A. K. Erlang, or in terms of hundred (century) call seconds per hour (CCS). The relationship between erlangs and CCS units can be derived by observing that there are 3600 sec in an hour:

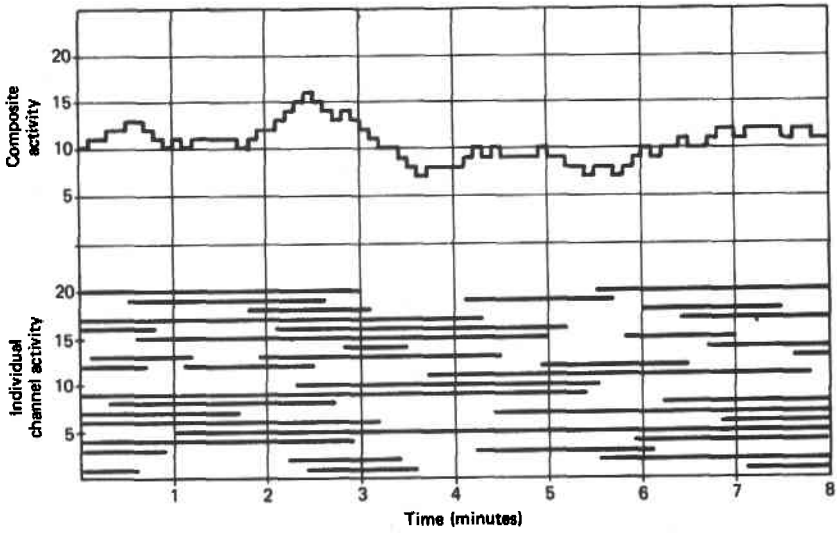


Figure 12.1 Activity profile of network traffic (all calls carried).

$$1 \text{ erlang} = 36 \text{ CCS}$$

The maximum capacity of a single server (channel) is 1 erlang, which is to say that the server is always busy. Thus the maximum capacity in erlangs of a group of servers is merely equal to the number of servers. Because traffic in a loss system experiences infinite blocking probabilities when the traffic intensity is equal to the number of servers, the average activity is necessarily less than the number of servers. Similarly, delay systems operate at less than full capacity, on average, because infinite delays occur when the average load approaches the number of servers.

Two important parameters used to characterize traffic are the average arrival rate λ and the average holding time t_m. If the traffic intensity λ is expressed in erlangs, then

$$A = \lambda t_m \tag{12.1}$$

where λ and t_m are expressed in like units of time (e.g., calls per second and seconds per call, respectively).

Notice that traffic intensity is only a measure of average utilization during a time period and does not reflect the relationship between arrivals and holding times. That is, many short calls can produce the same traffic intensity as a few long ones. In many of the analyses that follow the results are dependent only on the traffic intensity. In some cases, however, the results are also dependent on the individual arrival patterns and holding time distributions.

Public telephone networks are typically analyzed in terms of the average activity during the busiest hour of a day. The use of busy-hour traffic measurements to design and analyze telephone networks represents a compromise between designing for the overall average utilization (which includes virtually unused nighttime hours) and designing for short-duration peaks that may occur by chance or as a result of TV commercial breaks, radio call-in contests, and so on.

Busy-hour traffic measurements indicate that an individual residential telephone is typically in use between 5 and 10% of the busy hour. Thus each telephone represents a traffic load of between 0.05 and 0.10 erlangs. The average holding time is between 3 and 4 min, indicating that a typical telephone is involved in one or two phone calls during the busy hour.

Business telephones usually produce loading patterns different from residential phones. First, a business phone is generally utilized more heavily. Second, the busy hour of business traffic is often different from the busy hour of residential traffic. Figure 12.2 shows a typical hourly variation for both sources of traffic. The trunks of a telephone network are sometimes designed to take advantage of variations in calling patterns from different offices. Toll connecting trunks from residential areas are often busiest during evening hours, and trunks from business areas are obviously busiest during midmorning or midafternoon. Traffic engineering depends not only on overall traffic volume but also on time–volume traffic patterns within the network.

A certain amount of care must be exercised when determining the total traffic load of a system from the loading of individual lines or trunks. For example, since two tele-

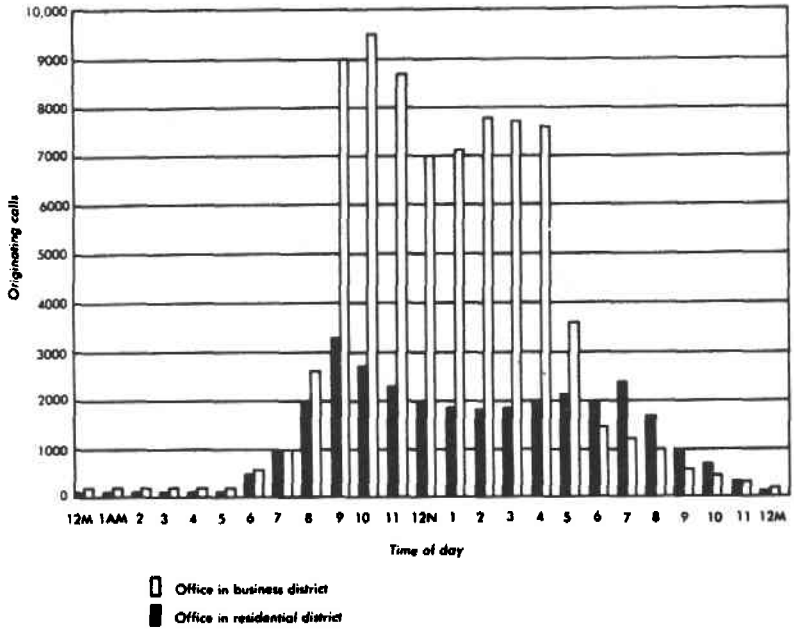


Figure 12.2 Traffic volume dependence on time of day.

phones are involved in each connection, the total load on a switching system is exactly one-half the total of all traffic on the lines connected to the switch. In addition, it may be important to include certain setup and release times into the average holding times of some common equipment. A 10-sec setup time is not particularly significant for a 4-min voice call but can actually dominate the holding time of equipment used for short data messages. Common equipment setup times also become more significant in the presence of voice traffic overloads. A greater percentage of the overall load is represented by call attempts since they increase at a faster rate than completions.

An important distinction to be made when discussing traffic in a communications network is the difference between the offered traffic and the carried traffic. The offered traffic is the total traffic that would be carried by a network capable of servicing all requests as they arise. Since economics generally precludes designing a network to immediately carry the maximum offered traffic, a small percentage of offered traffic typically experiences network blocking or delay. When the blocked calls are rejected by the network, the mode of operation is referred to as blocked calls cleared or lost calls cleared. In essence, blocked calls are assumed to disappear and never return. This assumption is most appropriate for trunk groups with alternate routes. In this case a blocked call is normally serviced by another trunk group and does not, in fact, return.

The carried traffic of a loss system is always less than the offered traffic. A delay system, on the other hand, does not reject blocked calls but holds them until the necessary facilities are available. With the assumption that the long-term average of offered traffic is less than the capacity of the network, a delay system carries all offered

traffic. If the number of requests that can be waiting for service is limited, however, a delay system also takes on properties of a loss system. For example, if the queue for holding blocked arrivals is finite, requests arriving when the queue is full are cleared.

12.1.1 Arrival Distributions

The most fundamental assumption of classical traffic analysis is that call arrivals are independent. That is, an arrival from one source is unrelated to an arrival from any other source. Even though this assumption may be invalid in some instances, it has general usefulness for most applications. In those cases where call arrivals tend to be correlated, useful results can still be obtained by modifying a random arrival analysis. In this manner the random arrival assumption provides a mathematical formulation that can be adjusted to produce approximate solutions to problems that are otherwise mathematically intractable.

Negative Exponential Interarrival Times

Designate the average call arrival rate from a large group of independent sources (subscriber lines) as λ. Use the following assumptions:

1. Only one arrival can occur in any sufficiently small interval.
2. The probability of an arrival in any sufficiently small interval is directly proportional to the length of the interval. (The probability of an arrival is $\lambda\,\Delta t$, where Δt is the interval length.)
3. The probability of an arrival in any particular interval is independent of what has occurred in other intervals.

It is straightforward [1] to show that the probability distribution of interarrival times is

$$P_0(\lambda t) = e^{-\lambda t} \tag{12.2}$$

Equation 12.2 defines the probability that no arrivals occur in a randomly selected interval t. This is identical to the probability that t seconds elapse from one arrival to the next.

Example 12.1. Assuming each of 10,000 subscriber lines originate one call per hour, how often do two calls arrive with less than 0.01 sec between them?

Solution. The average arrival rate is

$$\lambda = 3600/10{,}000 = 2.78 \text{ arrivals/sec}$$

From Equation 12.2, the probability of no arrival in a 0.01-sec interval is

$$P_0(0.0278) = e^{-0.0278} = 0.973$$

Thus 2.7% of the arrivals occur within 0.01 sec of the previous arrival. Since the arrival rate is 2.78 arrivals per second, the rate of occurrence of interarrival times less than 0.01 sec is

$$2.78 \times 0.027 = 0.075 \text{ times/sec}$$

The first two assumptions made in deriving the negative exponential arrival distribution can be intuitively justified for most applications. The third assumption, however, implies certain aspects of the sources that cannot always be supported. First, certain events, such as television commercial breaks, might stimulate the sources to place their calls at nearly the same time. In this case the negative exponential distribution may still hold but for a much higher calling rate during the commercial.

A more subtle implication of the independent arrival assumption involves the number of sources, not just their calling patterns. When the probability of an arrival in any small time interval is independent of other arrivals, it implies that the number of sources available to generate requests is constant. If a number of arrivals occur immediately before any subinterval in question, some of the sources become busy and cannot generate requests. The effect of busy sources is to reduce the average arrival rate. Thus the interarrival times are always somewhat larger than what Equation 12.2 predicts them to be. The only time the arrival rate is truly independent of source activity is when an infinite number of sources exist.

If the number of sources is large and their average activity is relatively low, busy sources do not appreciably reduce the arrival rate. For example, consider an end office that services 10,000 subscribers with 0.1 erlang of activity each. Normally, there are 1000 active links and 9000 subscribers available to generate new arrivals. If the number of active subscribers increases by an unlikely 50% to 1500 active lines, the number of idle subscribers reduces to 8500, a change of only 5.6%. Thus the arrival rate is relatively constant over a wide range of source activity. Whenever the arrival rate is fairly constant for the entire range of normal source activity, an infinite source assumption is justified.

Actually, some effects of finite sources have already been discussed in Chapter 5 when analyzing blocking probabilities of a switch. It is pointed out that Lee graph analyses overestimate the blocking probability because, if some number of interstage links in a group are known to be busy, the remaining links in the group are less likely to be busy. A Jacobaeus analysis produces a more rigorous and accurate solution to the blocking probability, particularly when space expansion is used. Accurate analyses of interarrival times for finite sources are also possible. These are included in the blocking analyses to follow.

Poisson Arrival Distribution

Equation 12.2 merely provides a means of determining the distribution of interarrival times. It does not, by itself, provide the generally more desirable information of how

many arrivals can be expected to occur in some arbitrary time interval. Using the same assumptions presented, however, the probability of j arrivals in an interval t can be determined [1] as

$$P_j(\lambda t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \tag{12.3}$$

Equation 12.3 is the well-known Poisson probability law. Notice that when $j = 0$, the probability of no arrivals in an interval t is $P_0(t)$, as obtained in Equation 12.2.

Again, Equation 12.3 assumes arrivals are independent and occur at a given average rate λ, irrespective of the number of arrivals occurring just prior to an interval in question. Thus the Poisson probability distribution should only be used for arrivals from a large number of independent sources.

Equation 12.3 defines the probability of experiencing exactly j arrivals in t seconds. Usually there is more interest in determining the probability of j or more arrivals in t seconds:

$$\begin{aligned} P_{\geq j}(\lambda t) &= \sum_{i=j}^{\infty} P_i(\lambda t) \\ &= 1 - \sum_{i=0}^{j-1} P_i(\lambda t) \\ &= 1 - P_{<j}(\lambda t) \end{aligned} \tag{12.4}$$

where $P_i(\lambda t)$ is defined in Equation 12.3.

Example 12.2. Given a message-switching node that normally experiences four arrivals per minute, what is the probability that eight or more arrivals occur in an arbitrarily chosen 30-sec interval?

Solution. The average number of arrivals in a 30-sec interval is

$$\lambda t = 4 \times \frac{30}{60} = 2$$

The probability of eight or more arrivals (when the average is 2) is

$$P_{\geq 8}(2) = \sum_{i=8}^{\infty} P_i(2)$$

$$= 1 - \sum_{i=0}^{7} P_i(2)$$

$$= 1 - e^{-2}\left(1 + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \ldots + \frac{2^7}{7!}\right)$$

$$= 0.0011$$

Example 12.3. What is the probability that a 1000-bit data block experiences exactly four errors while being transmitted over a transmission link with a bit error rate (BER) of 10^{-5}?

Solution. Assuming independent errors (a questionable assumption on many transmission links), we can obtain the probability of exactly four errors directly from the Poisson distribution. The average number of errors (arrivals) $\lambda t = 10^3 \times 10^{-5} = 0.01$. Thus

$$\text{prob(4errors)} = P_4(0.01) = \frac{(0.01)^4}{4!} e^{-0.01} = 4.125 \times 10^{-10}$$

An alternative solution can be obtained from the binomial probability law:

$$\text{prob(4errors)} = \binom{1000}{4} p^4 (1-p)^{996}$$

$$= 4.101 \times 10^{-10} \qquad \text{where } p = 10^{-5}$$

As can be seen, the two solutions of Example 12.3 are nearly identical. The closeness of the two answers reflects the fact that the Poisson probability distribution is often derived as a limiting case of a binomial probability distribution. Because it is easier to calculate, a Poisson distribution is often used as an approximation to a binomial distribution.

12.1.2 Holding Time Distributions

The second factor of traffic intensity as specified in Equation 12.1 is the average holding time t_m. In some cases the average of the holding times is all that needs to be known about holding times to determine blocking probabilities in a loss system or delays in a delay system. In other cases it is necessary to know the probability distribution of the holding times to obtain the desired results. This section describes the two most commonly assumed holding time distributions: constant holding times and exponential holding times.

Constant Holding Times

Although constant holding times cannot be assumed for conventional voice conversations, it is a reasonable assumption for such activities as per-call call processing requirements, interoffice address signaling, operator assistance, and recorded message playback. Furthermore, constant holding times are obviously valid for transmission times in fixed-length packet networks.

When constant holding time messages are in effect, it is straightforward to use Equation 12.3 to determine the probability distribution of active channels. Assume, for the time being, that all requests are serviced. Then the probability of j channels being busy at any particular time is merely the probability that j arrivals occurred in the time interval of length t_m immediately preceding the instant in question. Since the average number of active circuits over all time is the traffic intensity $A = \lambda t_m$, the probability of j circuits being busy is dependent only on the traffic intensity:

$$P_j(\lambda t_m) = P_j(A)$$

$$= \frac{A^j}{j!} e^{-A} \tag{12.5}$$

where λ = arrival rate
t_m = constant holding time
A = traffic intensity (erlangs)

Exponential Holding Times

The most commonly assumed holding time distribution for conventional telephone conversations is the exponential holding time distribution:

$$P(>t) = e^{-t/t_m} \tag{12.6}$$

where t_m is the average holding time. Equation 12.6 specifies the probability that a holding time exceeds the value t. This relationship can be derived from a few simple assumptions concerning the nature of the call termination process. Its basic justification, however, lies in the fact that observations of actual voice conversations exhibit a remarkably close correspondence to an exponential distribution.

The exponential distribution possesses the curious property that the probability of a termination is independent of how long a call has been in progress. That is, no matter how long a call has been in existence, the probability of it lasting another t seconds is defined by Equation 12.6. In this sense exponential holding times represent the most random process possible. Not even knowledge of how long a call has been in progress provides any information as to when the call will terminate.

Combining a Poisson arrival process with an exponential holding time process to obtain the probability distribution of active circuits is more complicated than it was for constant holding times because calls can last indefinitely. The final result, however, proves to be dependent on only the average holding time. Thus Equation 12.5 is

valid for exponential holding times as well as for constant holding times (or any holding time distribution). Equation 12.5 is therefore repeated for emphasis: The probability of j circuits being busy at any particular instant, assuming a Poisson arrival process and that all requests are serviced immediately, is

$$P_j(A) = \frac{A^j}{j!} e^{-A} \tag{12.7}$$

where A is the traffic intensity in erlangs. This result is true for any distribution of holding times.

Example 12.4. Assume that a trunk group has enough channels to immediately carry all of the traffic offered to it by a Poisson process with an arrival rate of one call per minute. Assume that the average holding time is 2 min. What percentage of the total traffic is carried by the first five circuits, and how much traffic is carried by all remaining circuits? (Assume that the traffic is always packed into the lowest numbered circuits.)

Solution. The traffic intensity (offered load) of the system is $A = 1 \times 2 = 2$ erlangs. The traffic intensity carried by i active circuits is exactly i erlangs.

Hence the traffic carried by the first five circuits can be determined as follows:

$$A_5 = 1P_1(2) + 2P_2(2) + 3P_3(2) + 4P_4(2) + 5P_5(2)$$

$$= e^{-2}\left(2 + \frac{2 \times 2^2}{2!} + \frac{3 \times 2^3}{3!} + \frac{4 \times 2^4}{4!} + \frac{5 \times 2^5}{5!}\right)$$

$$= 1.89 \text{ erlangs}$$

All of the remaining circuits carry

$$2 - 1.89 = 0.11 \text{ erlang}$$

The result of Example 12.4 demonstrates the principle of diminishing returns as the capacity of a system is increased to carry greater and greater percentages of the offered traffic. The first five circuits in Example 12.4 carry 94.5% of the traffic while all remaining circuits carry only 5.5% of the traffic. If there are 100 sources, 95 extra circuits are needed to carry the 5.5%.

12.2 LOSS SYSTEMS

Example 12.4 provides an indication of the blocking probabilities that arise when the number of servers (circuits) is less than the maximum possible traffic load (number of sources). The example demonstrates that 94.5% of the traffic is carried by only five circuits. The implication is that the blocking probability, if only five circuits are available to carry the traffic, is 5.5%. Actually, Example 12.4 is carefully worded to indicate that all of the offered traffic is carried but that only the traffic carried by the first five circuits is of interest. There is a subtle but important distinction between the probability that six or more circuits are busy (as can be obtained from Equation 12.7) and the blocking probability that arises when only five circuits exist.

The basic reason for the discrepancy is indicated in Figure 12.3, which depicts the same traffic pattern arising from 20 sources as is shown previously in Figure 12.1. Figure 12.3, however, assumes that only 13 circuits are available to carry the traffic. Thus the three arrivals at $t = 2.2$, 2.3, and 2.4 min are blocked and assumed to have left the system. The total amount of traffic volume lost is indicated by the shaded area, which is the difference between all traffic being serviced as it arrives and traffic being carried by a blocked calls cleared system with 13 circuits. The most important feature to notice in Figure 12.3 is that the call arriving at $t = 2.8$ is not blocked, even though the original profile indicates that it arrives when all 13 circuits are busy. The reason it is not blocked is that the previously blocked calls left the system and therefore reduced the congestion for subsequent arrivals. Hence the percentage of time that the original traffic profile is at or above 13 is not the same as the blocking probability when only 13 circuits are available.

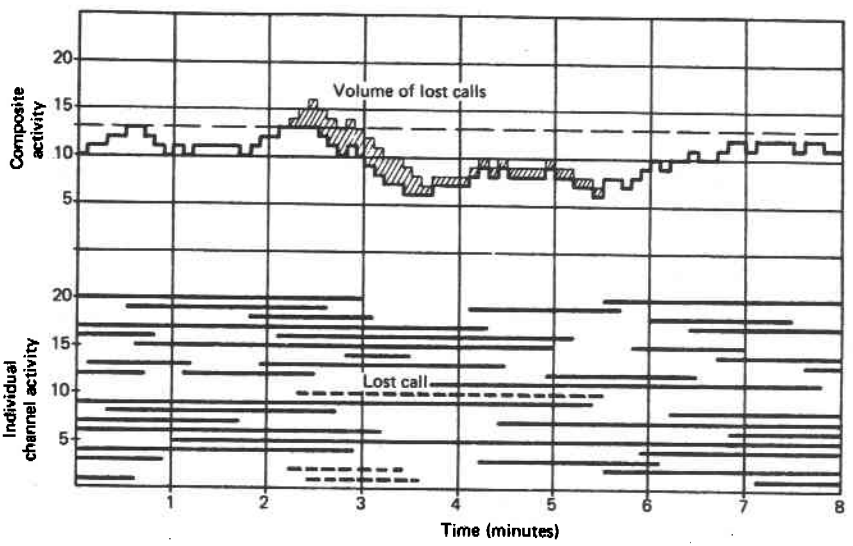


Figure 12.3 Activity profile of blocked calls cleared (13 channels).

12.2.1 Lost Calls Cleared

The first person to account fully and accurately for the effect of cleared calls in the calculation of blocking probabilities was A. K. Erlang in 1917. In this section we discuss Erlang's most often used result: his formulation of the blocking probability for a lost calls cleared system with Poisson arrivals. Recall that the Poisson arrival assumption implies infinite sources. This result is variously referred to as Erlang's formula of the first kind, $E_{1,N}(A)$; the Erlang-B formula; or Erlang's loss formula.

A fundamental aspect of Erlang's formulation, and a key contribution to modern stochastic process theory, is the concept of statistical equilibrium. Basically, statistical equilibrium implies that the probability of a system's being in a particular state (number of busy circuits in a trunk group) is independent of the time at which the system is examined. For a system to be in statistical equilibrium, a long time must pass (several average holding times) from when the system is in a known state until it is again examined. For example, when a trunk group first begins to accept traffic, it has no busy circuits. For a short time thereafter, the system is most likely to have only a few busy circuits. As time passes, however, the system reaches equilibrium. At this point the most likely state of the system is to have $A = \lambda t_m$ busy circuits.

When in equilibrium, a system is as likely to have an arrival as it is to have a termination. If the number of active circuits happens to increase above the average A, departures become more likely than arrivals. Similarly, if the number of active circuits happens to drop below A, an arrival is more likely than a departure. Thus if a system is perturbed by chance from its average state, it tends to return.

Although Erlang's elegant formulation is not particularly complicated, it is not presented here because we are mostly interested in application of the results. The interested reader is invited to see reference [2] or [3] for a derivation of the result:

$$B = E_{1,N}(A) = \frac{A^N}{N!\,\sum_{i=0}^{N}(A^i/i!)} \tag{12.8}$$

where N = number of servers (channels)
A = offered traffic intensity, λt_m (erlangs)

Equation 12.8 specifies the probability of blocking for a system with random arrivals from an infinite source and arbitrary holding time distributions. The blocking probability of Equation 12.8 is plotted in Figure 12.4 as a function of offered traffic intensity for various numbers of channels. An often more useful presentation of Erlang's results is provided in Figure 12.5, which presents the output channel utilization for various blocking probabilities and numbers of servers. The output utilization ρ represents the traffic carried by each circuit:

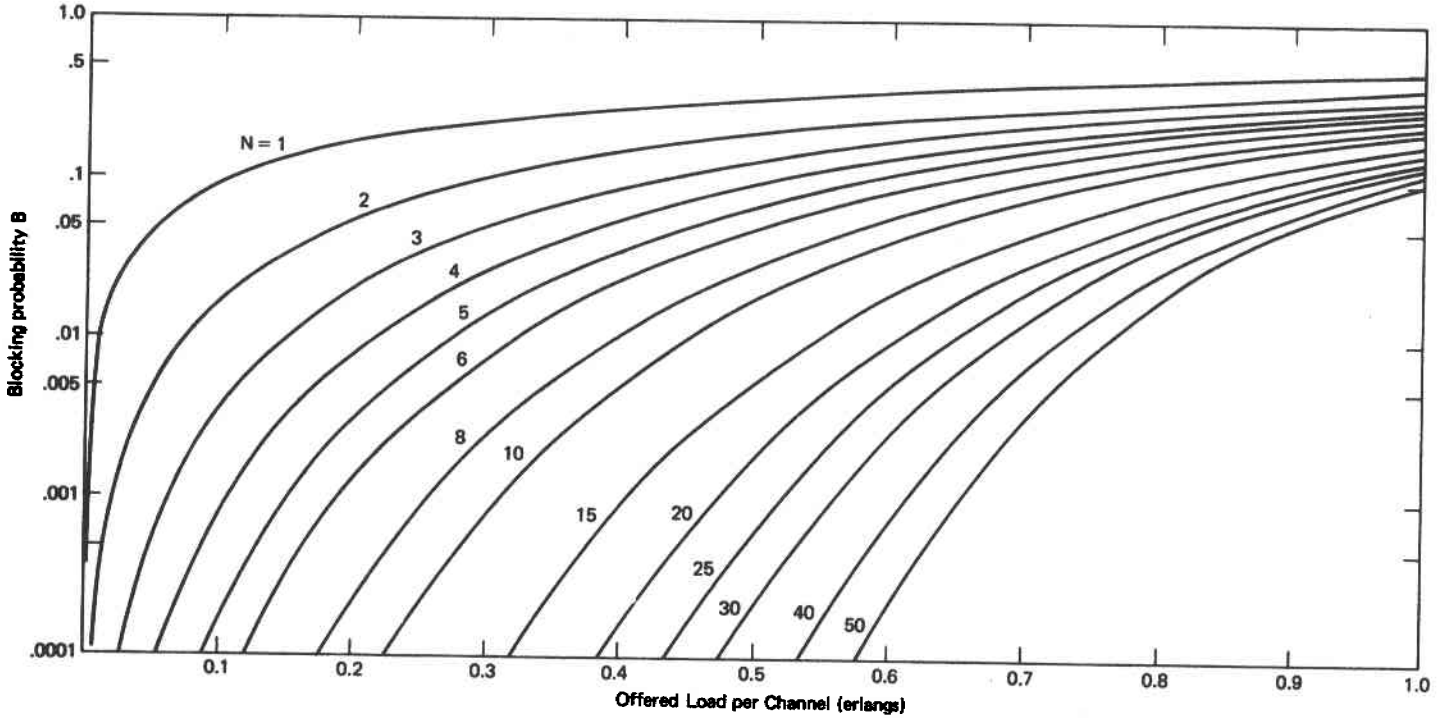


Figure 12.4 Blocking probability of lost calls cleared system.

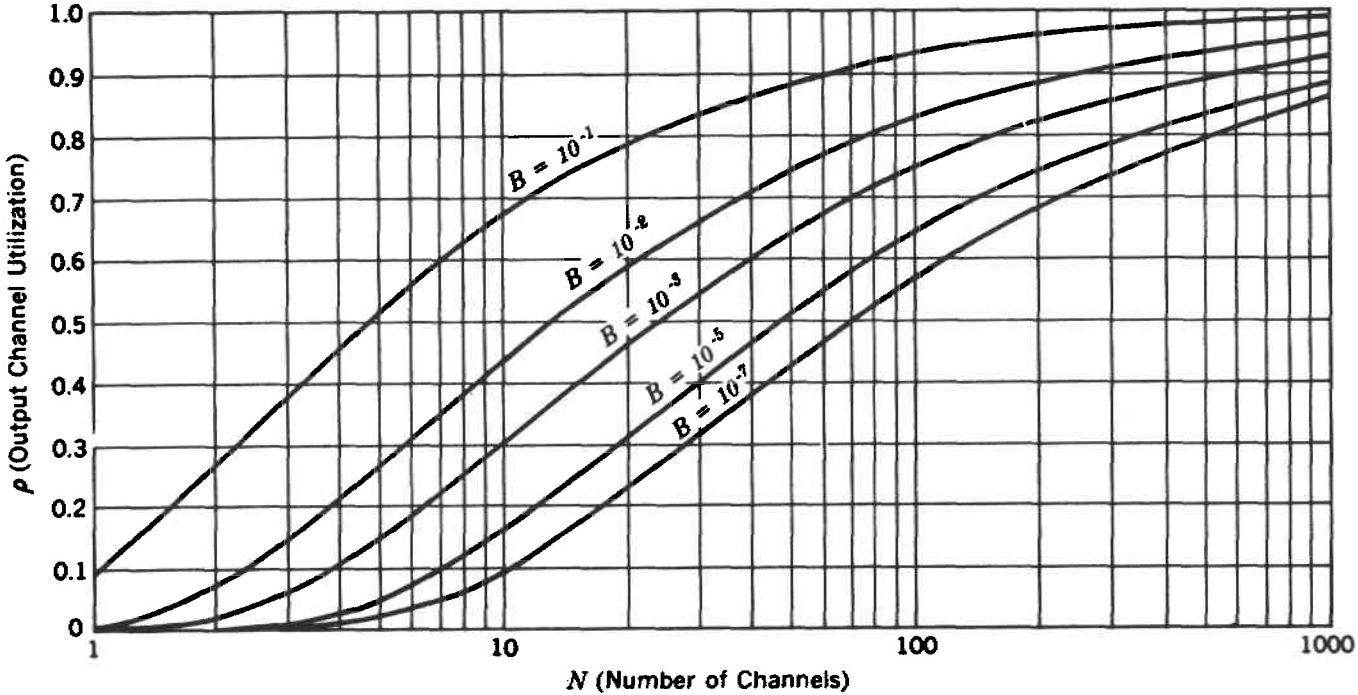


Figure 12.5 Output channel utilization of lost calls cleared system.

$$\rho = \frac{(1-B)A}{N} \tag{12.9}$$

where A = offered traffic
N = number of channels
B = blocking probability
$(1-B)A$ = carried traffic

Blocking probabilities are also provided in tabular form in Appendix D.

Example 12.5. A T1 line is to be used as a tie-line trunk group between two PBXs. How much traffic can the trunk group carry if the blocking probability is to be 0.1? What is the offered traffic intensity?

Solution. From Figure 12.5 it can be seen that the output circuit utilization for $B = 0.1$ and $N = 24$ is 0.8. Thus the carried traffic intensity is $0.8 \times 24 = 19.2$ erlangs. Since the blocking probability is 0.1, the maximum level of offered traffic is

$$A = \frac{19.2}{1-0.1} = 21.3 \text{ erlangs}$$

Example 12.6. Four clusters of data terminals are to be connected to a computer by way of leased circuits, as shown in Figure 12.6. In Figure 12.6*a* the traffic from the clusters uses separate groups of shared circuits. In Figure 12.6*b* the traffic from all clusters is concentrated onto one common group of circuits. Determine the total number of circuits required in both cases when the maximum desired blocking probability is 5%. Assume that 22 terminals are in each cluster and each terminal is active 10% of the time. (Use a blocked calls cleared analysis.)

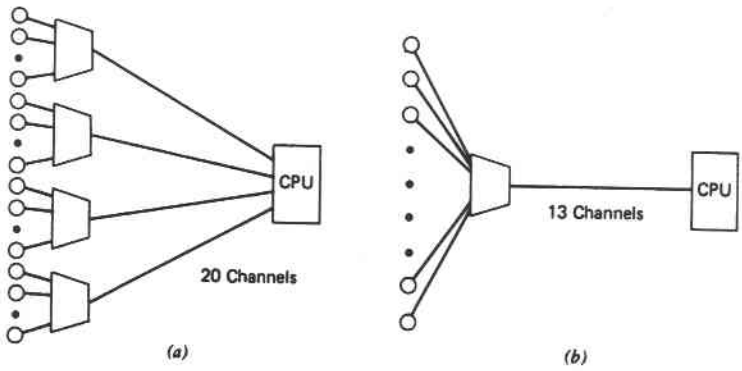


Figure 12.6 Data terminal network of Example 10.6: (*a*) four separate groups; (*b*) all traffic concentrated into one group.

Solution. The offered traffic from each cluster is $22 \times 0.1 = 2.2$ erlangs. Since the average number of active circuits is much smaller than the number of sources, an infinite source analysis can be used. Using Table D.1, the number of circuits required for B = 5% at a loading of 2.2 erlangs is 5. Thus the configuration of Figure 12.6*a* requires a total of 20 circuits.

The total offered traffic to the concentrator of the configuration of Figure 12.6*b* is $4 \times 2.2 = 8.8$ erlangs. From Table D.1, 13 circuits are required to support the given traffic load.

Example 12.6 demonstrates that consolidation of small traffic groups into one large traffic group can provide significant savings in total circuit requirements. Large groups are more efficient than multiple small groups because it is unlikely that the small groups will become overloaded at the same time (assuming independent arrivals). In effect, excess traffic in one group can use idle circuits in another group. Thus those circuits that are needed to accommodate traffic peaks but are normally idle are utilized more efficiently when the traffic is combined into one group. This feature is one of the motivations mentioned in Chapter 10 for integrating voice and data traffic into a common network. The total savings in transmission costs is most significant when the individual traffic intensities are low. Hence it is the peripheral area of a network that benefits the most by concentrating the traffic.

The greater circuit efficiency obtained by combining traffic into large groups is often referred to as the advantage of large group sizes. This efficiency of circuit utilization is the basic motivation for hierarchical switching structures. Instead of interconnecting a large number of nodes with rather small trunk groups between each pair, it is more economical to combine all traffic from individual nodes into one large trunk group and route the traffic through a tandem switching node. Figure 12.7 contrasts a mesh versus a star network with a centralized switching node at the center. Obviously, the cost of the tandem switch becomes justified when the savings in total circuit miles is large enough.

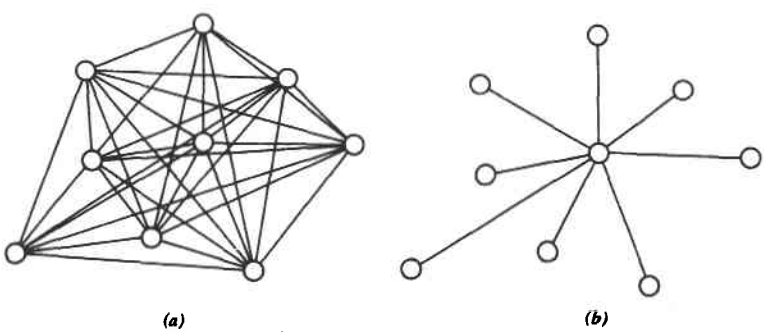


Figure 12.7 Use of tandem switching to concentrate traffic: (*a*) mesh; (*b*) star.

Example 12.7. What happens to the blocking probabilities in Figure 12.6*a* and *b* discussed in Example 12.6 when the traffic intensity increases by 50%?

Solution. If the traffic intensity of each group increases from 2.2 to 3.3 erlangs, the blocking probability of the configuration of Figure 12.6*a* increases from 5% to almost 14%.

In the configuration of Figure 12.6*b* a 50% increase in the traffic intensity causes a 400% increase in the blocking probability (from 5 to 20%).

Example 12.7 demonstrates some important considerations in network design. As indicated, blocking probabilities are very sensitive to increases in traffic intensities, particularly when the channels are heavily utilized. Because large trunk groups utilize their channels more efficiently, they are more vulnerable to traffic increases than are a number of smaller groups designed to provide the same grade of service. Furthermore, failures of equal percentages of transmission capacity affect the performance of a large group more than the performance of several small groups. In both cases the vulnerability of the large groups arises because large groups operate with less spare capacity than do multiple small groups.

A second aspect of blocking analyses demonstrated in Example 12.7 is that the calculated results are highly dependent on the accuracy of the traffic intensities. Accurate values of traffic intensities are not always available. Furthermore, even when accurate traffic measurements are obtainable, they do not provide an absolute indication of how much growth to expect. Thus only limited confidence can be attached to calculations of blocking probabilities in an absolute sense. The main value of these analyses is that they provide an objective means of comparing various network sizes and configurations. The most cost-effective design for a given grade of service is the one that should be chosen, even if the traffic statistics are hypothetical. If a network is liable to experience wildly varying traffic patterns or rapid growth, these factors must be considered when comparing design alternatives. A network with a somewhat larger initial cost may be more desirable if it can absorb or grow to accommodate unanticipated traffic volumes more easily.

12.2.2 Lost Calls Returning

In the lost calls cleared analyses just presented, it is assumed that unserviceable requests leave the system and never return. As mentioned, this assumption is most appropriate for trunk groups whose blocked requests overflow to another route and are usually serviced elsewhere. However, lost calls cleared analyses are also used in instances where blocked calls do not get serviced elsewhere. In many of these cases, blocked calls tend to return to the system in the form of retries. Some examples are subscriber concentrator systems, corporate tie lines and PBX trunks, calls to busy telephone numbers, and access to WATS lines (if DDD alternatives are not used). This

section derives blocking probability relationships for lost calls cleared systems with random retries.

The following analysis involves three fundamental assumptions regarding the nature of the returning calls:

1. All blocked calls return to the system and eventually get serviced, even if multiple retries are required.
2. The elapsed times between call blocking occurrences and the generation of retries are random and statistically independent of each other. (This assumption allows the analysis to avoid complications arising when retries are correlated to each other and tend to cause recurring traffic peaks at a particular waiting time interval.)
3. The typical waiting time before retries occur is somewhat longer than the average holding time of a connection. This assumption essentially states that the system is allowed to reach statistical equilibrium before a retry occurs. Obviously, if retries occur too soon, they are very likely to encounter congestion since the system has not had a chance to "relax." In the limit, if all retries are immediate and continuous, the network operation becomes similar to a delay system discussed in later sections of this chapter. In this case, however, the system does not queue requests—the sources do so by continually "redialing."

When considered in their entirety, these assumptions characterize retries as being statistically indistinguishable from first-attempt traffic.* Hence blocked calls merely add to the first-attempt call arrival rate.

Consider a system with a first-attempt call arrival rate of λ. If a percentage B of the calls is blocked, B times λ retries will occur in the future. Of these retries, however, a percentage B will be blocked again. Continuing in this manner, the total arrival rate λ' after the system has reached statistical equilibrium can be determined as the infinite series

$$\begin{aligned}\lambda' &= \lambda + B\lambda + B^2\lambda + B^3\lambda + \ldots \\ &= \frac{\lambda}{1-B}\end{aligned} \tag{12.10}$$

where B is the blocking probability from a lost calls cleared analysis with traffic intensity $A' = \lambda' t_m$.

Equation 12.10 relates the average arrival rate λ', including the retries, to the first-attempt arrival rate and the blocking probability in terms of λ'. Thus this relationship does not provide a direct means of determining λ' or B since each is expressed in terms of the other. However, the desired result can be obtained by iterating the lost calls

*First-attempt traffic is also referred to as demand traffic: the service demands assuming all arrivals are serviced immediately. The offered traffic is the demand traffic plus the retries.

cleared analysis of Equation 12.8. First, determine an estimate of B using λ and then calculate λ'. Next, use λ' to obtain a new value of B and an updated value of λ'. Continue in this manner until values of λ' and B are obtained.

Example 12.8. What is the blocking probability of a PBX to a central office trunk group with 10 circuits servicing a first-attempt offered traffic load of 7 erlangs? What is the blocking probability if the number of circuits is increased to 13? Assume random retries for all blocked calls.

Solution. It can be assumed that the 7 erlangs of traffic arise from a large number of PBX stations. Thus an infinite source analysis is justified. The blocking probability for $A = 7$ erlangs and $N = 10$ servers is about 8%. Thus the total offered load, including retries, is approximately 7.6 erlangs. With $N = 10$ and $A' = 7.6$, the blocking probability is 10%. Two more iterations effectively produce convergence at $A' = 8$ erlangs and $B = 12\%$. If the number of circuits in the trunk group is increased to 13, the blocking probability of a lost calls cleared system is 1.5%. Thus a first approximation to the returning traffic intensity is 7/0.985 = 7.1 erlangs. Hence the blocking probability including all returning traffic increases only slightly above the 1.5%.

Example 12.8 demonstrates that the effect of returning traffic is insignificant when operating at low blocking probabilities. At high blocking probabilities, however, it is necessary to incorporate the effects of the returning traffic into the analysis. This relationship between lost calls cleared and lost calls returning is shown in Figure 12.8.

When measurements are made to determine the blocking probability of an outgoing trunk group, the measurements cannot distinguish between first-attempt calls (demand traffic) and retries. Thus if a significant number of retries are contained in the measurements, this fact should be incorporated into an analysis of how many circuits must be added to reduce the blocking of an overloaded trunk group. The apparent offered load will decrease as the number of servers increases because the number of re-

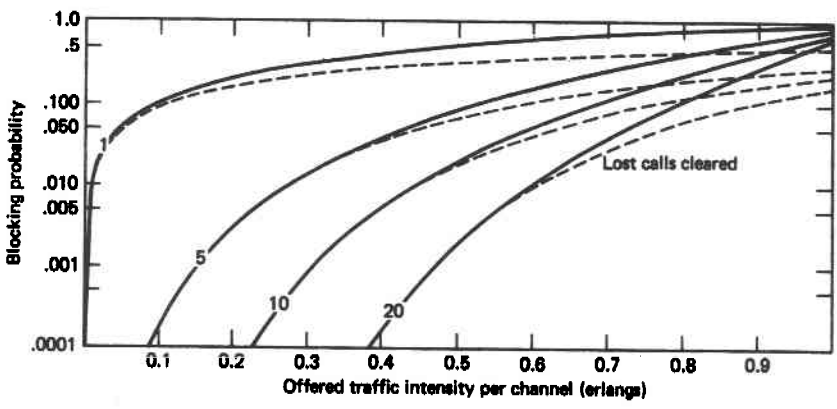


Figure 12.8 Blocking probability of lost calls returning.

tries decreases. Thus fewer additional circuits are needed than if no retries are contained in the measurements.

12.2.3 Lost Calls Held

In a lost calls held system, blocked calls are held by the system and serviced when the necessary facilities become available. Lost calls held systems are distinctly different from the delay systems discussed later in one important respect: The total elapsed time of a call in the system, including waiting time and service time, is independent of the waiting time. In essence, each arrival requires service for a continuous period of time and terminates its request independently of its being serviced or not. Figure 12.9 demonstrates the basic operation of a lost calls held system. Notice that most blocked calls eventually get some service, but only for a portion of the time that the respective sources are busy.

Although a switched telephone network does not operate in a lost calls held manner, some systems do. Lost calls held systems generally arise in real-time applications in which the sources are continuously in need of service, whether or not the facilities are available. When operating under conditions of heavy traffic, a lost calls held system typically provides service for only a portion of the time a particular source is active.

Even though conventional circuit switching does not operate according to the theoretical model of lost calls held, Bell System traffic engineers have used it to calculate blocking probabilities for trunk groups [4]. A lost calls held analysis always produces a larger value for blocking than does Erlang's loss formula. Thus the lost calls held analysis produces a conservative design that helps account for retries and day-to-day

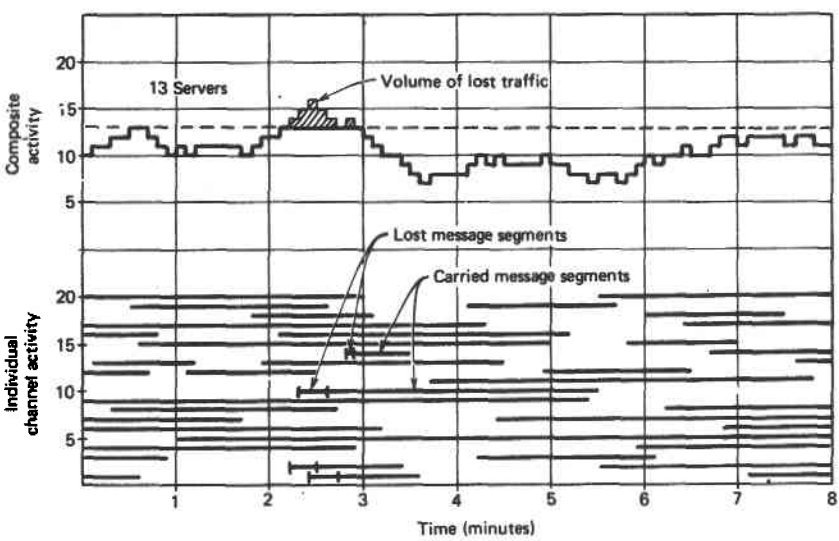


Figure 12.9 Activity profile of lost calls held.

variations in the busy-hour calling intensities. In contrast, CCITT recommendations [5] stipulate Erlang-B formulas should be used in determining blocking probabilities.

One example of a system that closely fits the lost calls held model is time assignment speech interpolation (TASI). A TASI system concentrates some number of voice sources onto a smaller number of transmission channels. A source receives service (is connected to a channel) only when it is active. If a source becomes active when all channels are busy, it is blocked and speech clipping occurs. Each speech segment starts and stops independently of whether it is serviced or not. TASI systems were originally used on analog long-distance transmission links such as undersea cables. More modern counterparts of TASI are referred to as digital circuit multiplication (DCM) systems. In contrast to the original TASI systems, DCM systems can delay speech for a small amount of time, when necessary, to minimize the clipping. In this case, a lost calls held analysis is not rigorously justified because the total time a speech segment is "in the system" increases as the delay for service increases. However, if the average delay is a small percentage of the holding time, or if the coding rate of delayed speech is reduced to allow the transmission channel time to "catch up," a lost calls held analysis is still justified. Recall that controlling the coding rate is one technique of traffic shaping used for transporting voice in an ATM network.

Lost calls held systems are easily analyzed to determine the probability of the total number of calls in the system at any one time. Since the duration of a source's activity is independent of whether it is being serviced, the number in the system at any time is identical to the number of active sources in a system capable of carrying all traffic as it arises. Thus the distribution of the number in the system is the Poisson distribution provided earlier in Equation 12.3. The probability that i sources requesting service are being blocked is simply the probability that $i + N$ sources are active when N is the number of servers. Recall that the Poisson distribution essentially determines the desired probability as the probability that $i + N$ arrivals occurred in the preceding t_m seconds. The distribution is dependent only on the product of the average arrival rate λ and the average holding time t_m.

Example 12.9. What is the probability that a talkspurt experiences clipping in a TASI system with 10 sources and 5 channels? With 100 sources and 50 channels? Assume that the activity factor of each talker is 0.4. (Ignore finite sources.)

Solution. For the first case, the clipping probability can be determined as the probability that five or more sources are busy in a Poisson process with an average of $A = 0.4 \times 10 = 4$ busy servers. Using Equation 12.7,

$$\text{prob(clipping)} = \sum_{j=5}^{9} P_j(4) = e^{-4}\left(\frac{4^5}{5!} + \frac{4^6}{6!} + \frac{4^7}{7!} + \frac{4^8}{8!} + \frac{4^9}{9!}\right) = 0.36$$

With 100 sources, the average number of busy circuits is $A = 0.4 \times 100 = 40$. A speech segment is clipped if 50 or more talkers are active at once. Thus the clipping probability can be determined as

$$\text{prob(clipping)} = \sum_{j=50}^{99} P_j(40) = 0.04$$

Example 12.9 demonstrates that TASI systems are much more effective for large group sizes than for small ones. The 36% clipping factor occurring with 5 channels produces unacceptable voice quality. On the other hand, the 4% clipping probability for 50 channels can be tolerated when the line costs are high enough.

In reality, the values for blocking probabilities obtained in Example 12.9 are overly pessimistic because an infinite source assumption was used. The summations did not include the case of all sources being active because there needs to be at least one idle source to create an arrival during the time congestion. A more accurate solution to this problem is obtained in a later section using a finite source analysis.

12.2.4 Lost Calls Cleared—Finite Sources

As mentioned previously, a fundamental assumption in the derivation of the Poisson arrival distribution, and consequently Erlang's loss formula, is that call arrivals occur independently of the number of active callers. Obviously, this assumption can be justified only when the number of sources is much larger than the number of servers. This section presents some fundamental relationships for determining blocking probabilities of lost calls cleared systems when the number of sources is not much larger than the number of servers. The blocking probabilities in these cases are always less than those for infinite source systems since the arrival rate decreases as the number of busy sources increases.

When considering finite source systems, traffic theorists introduce another parameter of interest called time congestion. Time congestion is the percentage of time that all servers in a group are busy. It is identical to the probability that all servers are busy at randomly selected times. However, time congestion is not necessarily identical to blocking probability (which is sometimes referred to as call congestion). Time congestion merely specifies the probability that all servers are busy. Before blocking can occur, there must be an arrival.

In an infinite source system, time congestion and call congestion are identical because the percentage of arrivals encountering all servers busy is exactly equal to the time congestion. (The fact that all servers are busy has no bearing on whether or not an arrival occurs.) In a finite source system, however, the percentage of arrivals encountering congestion is smaller because fewer arrivals occur during periods when all servers are busy. Thus in a finite source system, call congestion (blocking probability) is always less than the time congestion. As an extreme example, consider equal numbers of sources and servers. The time congestion is the probability that all servers are busy. The blocking probability is obviously zero.

The same basic techniques introduced by Erlang when he determined the loss formula for infinite sources can be used to derive loss formulas for finite sources [3]. Us-

ing these techniques, we find the probability of n servers being busy in a system with M sources and N servers is

$$P_n = \frac{\binom{M}{n}(\lambda' t_m)^n}{\sum_{i=0}^{N}\binom{M}{i}(\lambda' t_m)^i} \tag{12.11}$$

where λ' is the calling rate per *idle* source and t_m is the average holding time. Equation 12.11 is known as the truncated Bernoullian distribution and also as the Engset distribution.

Setting $n = N$ in Equation 12.11 produces an expression for the time congestion:

$$P_N = \frac{\binom{M}{N}(\lambda' t_m)^N}{\sum_{i=0}^{N}\binom{M}{i}(\lambda' t_m)^i} \tag{12.12}$$

Using the fact that the arrival rate when N servers are busy is $(M - N)/M$ times the arrival rate when no servers are busy, we can determine the blocking probability for lost calls cleared with a finite source as follows:

$$B = \frac{\binom{M-1}{N}(\lambda' t_m)^N}{\sum_{i=0}^{N}\binom{M-1}{i}(\lambda' t_m)^i} \tag{12.13}$$

which is identical to P_N (the time congestion) for $M - 1$ sources.

Equations 12.11, 12.12, and 12.13 are easily evaluated in terms of the parameters λ' and t_m. However, λ' and t_m do not, by themselves, specify the average activity of a source. In a lost calls cleared system with finite sources the effective offered load decreases as the blocking probability increases because blocked calls leave and do not return. When a call is blocked, the average activity of the offering source decreases, which increases the average amount of idle time for that source. The net result is that λ' decreases because the amount of idle time increases. If the average activity of a source assuming no traffic is cleared is designated as $\rho = \lambda t_m$, the value of $\lambda' t_m$ can be determined as

$$\lambda' t_m = \frac{\rho}{1 - \rho(1 - B)} \tag{12.14}$$

where B is the blocking probability defined by Equation 12.13.

The difficulty with using the unblocked source activity factor ρ to characterize a source's offered load is now apparent. The value of $\lambda' t_m$ depends on B, which in turn depends on $\lambda' t_m$. Thus some form of iteration is needed to determine B when the sources are characterized by ρ (an easily measured parameter) instead of λ'. If the total offered load is considered to be $M\rho$, the carried traffic is

$$A_{\text{carried}} = M\rho(1 - B) \tag{12.15}$$

A table of traffic capacities for finite sources is provided in Appendix D.2, where the offered load $A = M\rho$ is listed for various combinations of M, N, and B. Some of the results are plotted in Figure 12.10, where they can be compared to blocking probabilities of infinite source systems. As expected, infinite source analyses (Erlang-B) are acceptable when the number of sources M is large.

Example 12.10. A group of callers generate requests at a rate of five calls per hour per telephone (including incoming and outgoing calls). Assuming the average holding time is 4 min, what is the average calling rate of each idle source? How many callers can be supported by a 12-channel concentrator/multiplexer if the maximum acceptable blocking probability is 1%?

Solution. Since each caller is typically active for 20 min of every hour and places an average of five calls during the 40 min of inactivity, the calling rate for idle sources $\lambda' = 5/40 = 0.125$ calls per minute. The offered load for M sources, assuming all traffic is carried, is 0.33 M. Table D.2 must be searched to find the largest M such that 0.33 M is less than or equal to the maximum offered load for $B = 1\%$ and $N = 12$. Using interpolation for $M = 21$ reveals that 12 servers can support 7.11 erlangs at $B = 1\%$. Since $21 \times 0.33 = 6.93$ is the offered load, 21 sources is an acceptable solution. If 22

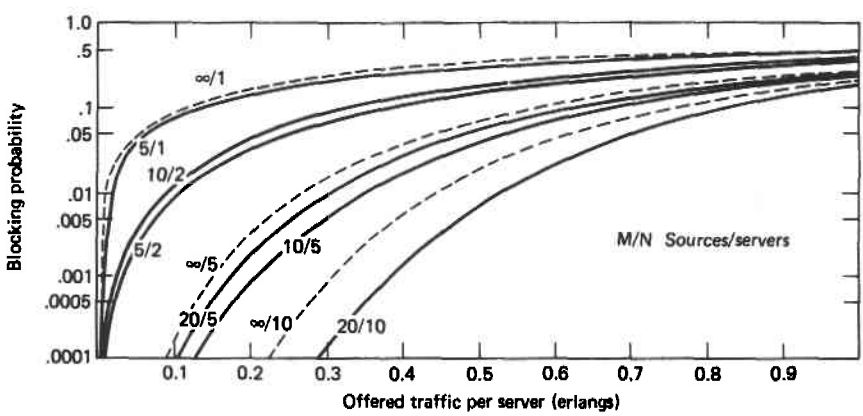


Figure 12.10 Blocking probability of lost calls cleared with finite sources.

sources are used, the offered load of 7.26 erlangs is higher than the 7.04 erlangs obtainable from interpolation in Table D.2 as the maximum offered load for $B = 1\%$.

It is worthwhile comparing the result of Example 12.10 to a result obtained from an infinite source analysis (Erlang-B). For a blocking probability of 1%, Table D.1 reveals that the maximum offered load for 12 servers is 5.88 erlangs. Thus the maximum number of sources can be determined as $5.88/0.333 \approx 17.64$. Hence in this case an infinite source analysis produces a result that is conservative by 15%.

12.2.5 Lost Calls Held—Finite Sources

A lost calls held system with finite sources is analyzed in the same basic manner as a lost calls held systems with infinite sources. At all times the number of calls "in the system" is defined to be identical to the number of calls that would be serviced by a strictly nonblocking server group. Thus Equation 12.11 is used to determine the probability that exactly n calls are in the system:

$$P_n = \frac{\binom{M}{n}(\lambda' t_m)^n}{\sum_{i=0}^{M}\binom{M}{i}(\lambda' t_m)^i} = \frac{\binom{M}{n}(\lambda' t_m)^n}{(1+\lambda' t_m)^M} \tag{12.16}$$

Because no calls are cleared, the offered load per idle source is not dependent on B:

$$\lambda' t_m = \frac{\lambda t_m}{1-\lambda t_m} = \frac{\rho}{1-\rho} \tag{12.17}$$

Combining Equations 12.16 and 12.17 produces a more useful expression for the probability that n calls are in the system:

$$P_n = \binom{M}{n}\rho^n(1-\rho)^{M-n} \tag{12.18}$$

If there are N servers, the time congestion is merely the probability that N or more servers are busy:

$$P_N = \sum_{n=N}^{M} P_n \tag{12.19}$$

The blocking probability, in a lost calls held sense, is the probability of an arrival encountering N or more calls in the system:

$$B_h = \frac{\sum_{n=N}^{M} P_n \text{ prob(arrival } \mid n \text{ sources are busy)}}{\text{average arrival rate}}$$

$$= \sum_{n=N}^{M-1} \binom{M-1}{n} \rho^n (1-\rho)^{M-1-n} \tag{12.20}$$

where ρ = offered load per source
M = number of sources
N = number of servers

Example 12.11. Determine the probability of clipping for the TASI systems described in Example 12.9. In this case, however, use a lost calls held analysis for finite sources.

Solution. In this example we are concerned only with the probability that a speech utterance is clipped for some period of time until a channel becomes available. Thus Equation 12.20 provides the desired answer using $\rho = 0.4$ for the offered load per source. In the first case, with 10 sources and 5 channels,

$$B_h = \sum_{n=5}^{9} \binom{9}{n} (0.4)^n (0.6)^{9-n} = 0.27$$

In the second case with 100 sources and 50 servers,

$$B_h = \sum_{n=50}^{99} \binom{99}{n} (0.4)^n (0.6)^{99-n} = 0.023$$

The results of Example 12.11 show again that a TASI system requires large group sizes to provide low clipping probabilities. When compared to the results of Example 12.9, these results indicate that an infinite source analysis overestimates the clipping probability in both cases (0.36 versus 0.27 and 0.04 versus 0.023). Notice that the percentage error in the infinite source analysis is almost identical for the 10-source system and the 100-source system. Hence the validity of choosing an infinite source model is more dependent on the ratio of sources to servers than it is on the number of sources.

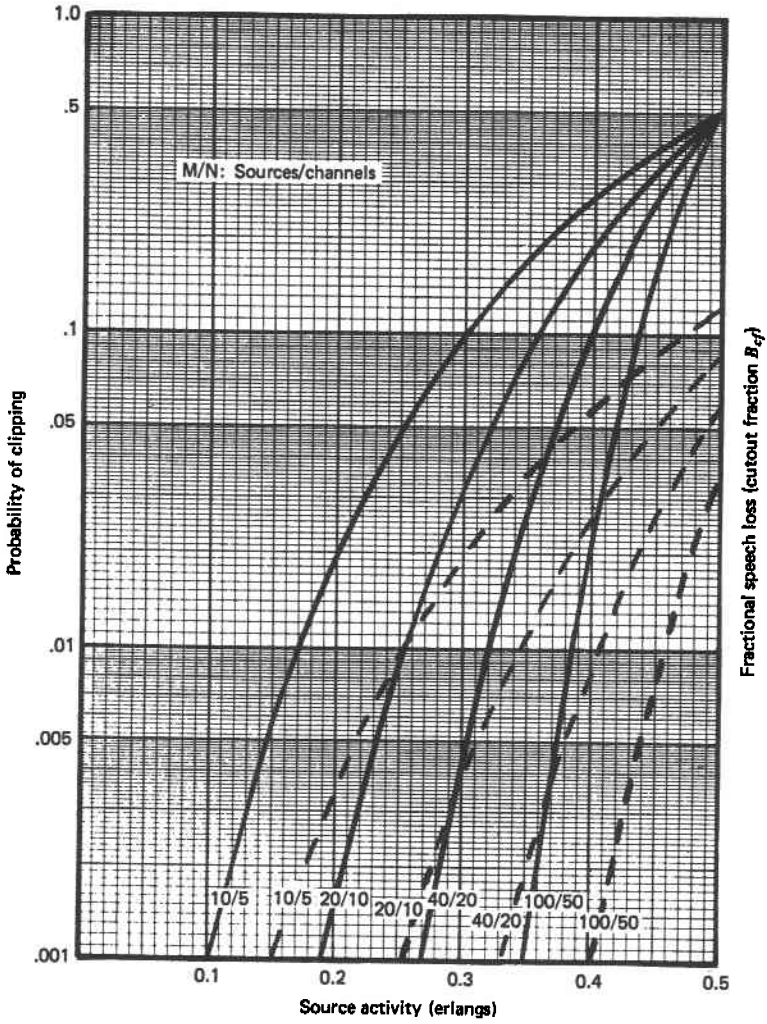


Figure 12.11 Clipping probability and clipping duration of TASI.

Example 12.11 merely determines the probability that a speech segment encounters congestion and is subsequently clipped. A complete analysis of a TASI (or DCM) system must consider the time duration of clips in addition to their frequency of occurrence. In essence, the desired information is represented by the amount of traffic volume in the clipped segments. Weinstein [6] refers to the clipped segments as "fractional speech loss" or simply "cutout fraction." This is not the same as the lost traffic, since a conventional lost calls held analysis considers any arrival that encounters congestion as being completely "lost"—even if it eventually receives some service. The cutout fraction is determined as the ratio of untransmitted traffic intensity to offered traffic intensity:

$$B_{cf} = \frac{1}{A} \sum_{n=N+1}^{M} (n - N) P_n \tag{12.21}$$

where M = number of sources
A = offered load, = $M\rho$
N = number of servers
P_n = probability of n calls in the system (Equation 12.18)

Example 12.12. Determine the average duration of a clip in the two TASI systems of Example 12.11. Assume the average duration of a speech segment is 300 msec. (The average length of a speech segment is dependent on the activity threshold, which also influences the activity factor ρ.)

Solution. In the first case,

$$B_{cf} = \frac{1}{4} \sum_{n=6}^{10} (n - 5) \binom{10}{n} (0.4)^n (0.6)^{10-n} = 0.059$$

Thus, on average, 5.9%, or 17.7 msec, of every 300-msec speech segment is clipped. Since 27% of the segments experience clipping (Example 12.11), the average duration of a clip for clipped segments is 0.059/0.27 = 22%, or 66 msec—an obviously intolerable amount. In the second case for 100 sources and 50 channels,

$$B_{cf} = \frac{1}{40} \sum_{n=51}^{100} (n - 50) \binom{100}{n} (0.4)^n (0.6)^{100-n} = 0.001$$

Thus in this case only 0.1% of all speech is clipped, which implies that when clipping occurs, $300 \times 0.001/0.023 = 13$ msec of the segment is lost.

Example 12.12 shows that large group sizes not only greatly reduce the clipping probability of TASI systems but also reduce the duration of the clips. The relationship of clipping probabilities and clipping duration (fractional speech loss) to group size and source activity is provided in Figure 12.11. As shown, the clipping probability is extremely sensitive to source activity (offered load). For a discussion of the effects of clipping on speech quality, see references [7] and [8].

12.3 NETWORK BLOCKING PROBABILITIES

In the preceding sections basic techniques of congestion theory are presented to determine blocking probabilities of individual trunk groups. In this section techniques of calculating end-to-end blocking probabilities of a network with more than one route

between endpoints is considered. In conjunction with calculating the end-to-end blocking probabilities, it is necessary to consider the interaction of traffic on various routes of a network. Foremost among these considerations is the effect of overflow traffic from one route onto another. The following sections discuss simplified analyses only. More sophisticated techniques for more complex networks can be obtained in references [9], [10], and [11].

12.3.1 End-to-End Blocking Probabilities

Generally, a connection through a large network involves a series of transmission links, each one of which is selected from a set of alternatives. Thus an end-to-end blocking probability analysis usually involves a composite of series and parallel probabilities. The simplest procedure is identical to the blocking probability (matching loss) analyses presented in Chapter 5 for switching networks. For example, Figure 12.12 depicts a representative set of alternative connections through a network and the resulting composite blocking probability.

The blocking probability equation in Figure 12.12 contains several simplifying assumptions. First, the blocking probability (matching loss) of the switches is not included. In a digital time division switch, matching loss can be low enough that it is easily eliminated from the analysis. In other switches, however, the matching loss may not be insignificant. When necessary, switch blocking is included in the analysis by considering it a source of blocking in series with the associated trunk groups.

When more than one route passes through the same switch, as in node C of Figure 12.12, proper treatment of correlation between matching losses is an additional complication. A conservative approach considers the matching loss to be completely correlated. In this case the matching loss is in series with the common link. On the other hand, an optimistic analysis assumes that the matching losses are independent, which implies that they are in series with the individual links. Figure 12.13 depicts these two approaches for including the matching loss of switch C into the end-to-end blocking probability equation of Figure 12.12. In this case, the link from C to D is the common link.

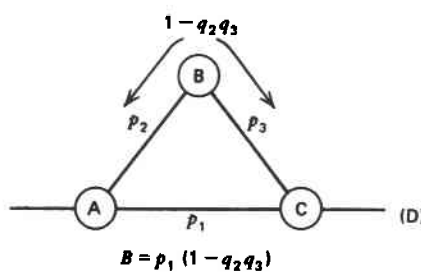


Figure 12.12 Probability graph for end-to-end blocking analysis.

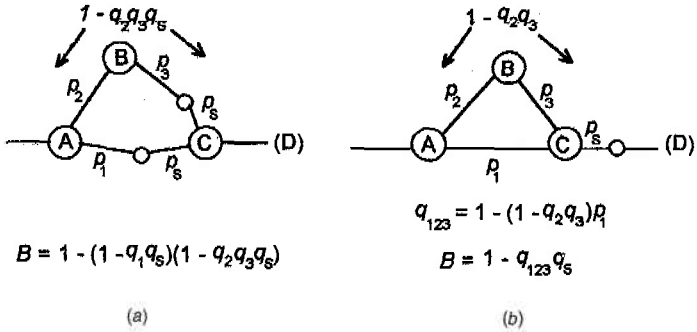


Figure 12.13 Incorporating switch-matching loss into end-to-end blocking analysis: (*a*) independent switch blocking; (*b*) correlated switch blocking.

A second simplifying assumption used in deriving the blocking probability equation in Figure 12.12 involves assuming independence for the blocking probabilities of the trunk groups. Thus the composite blocking of two parallel routes is merely the product of the respective probabilities (Equation 5.4). Similarly, independence implies that the blocking probability of two paths—in series—is 1 minus the product of the respective availabilities (Equation 5.5). In actual practice individual blocking probabilities are never completely independent. This is particularly true when a large amount of traffic on one route results as overflow from another route. Whenever the first route is busy, it is likely that more than the average amount of overflow is being diverted to the second route. Thus an alternate route is more likely to be busy when a primary route is busy.

In a large public network, trunks to tandem or toll switches normally carry traffic to many destinations. Thus no one direct route contributes an overwhelming amount of overflow traffic to a particular trunk group. In this case independent blocking probabilities on alternate routes are justified. In some instances of the public network, and often in private networks, overflow traffic from one route dominates the traffic on tandem routes. In these cases failure to account for the correlation in blocking probabilities can lead to overly optimistic results.

Example 12.13. Two trunk groups are to be used as direct routes between two switching systems. The first group has 12 channels and the second group has 6 channels. Assume 10.8 erlangs of traffic is offered to the 12-channel group and overflows are offered to the 6-channel group when the first group is busy. What is the blocking probability of the first group, and how much traffic overflows to the second group? Using the overflow traffic volume as an offered load, determine the blocking probability of the second trunk group. What is the probability that both trunk groups are busy? Compare this answer to the blocking probability of one 18-channel trunk group.

Solution. Using a lost calls cleared analysis with an infinite source, we determine that blocking of the first group is 15% ($A = 10.8$, $N = 12$). Therefore the overflow traffic is $10.8 \times 0.15 = 1.62$ erlangs. The blocking probability (assuming random arrivals?) of the second group is 0.5% ($A = 1.62$, $N = 6$). The probability that both trunk groups are busy simultaneously can be determined (assuming independence?) as

$$B = 0.15 \times 0.005 = 0.00075$$

In contrast, the correct blocking probability of an 18-channel trunk group with an offered load of 10.8 erlangs is

$$B = 0.013$$

The question marks in the solution of Example 12.13 point to two sources of error in the determination of the first blocking probability value. One error is the assumption of independence of blocking in the two trunk groups. A second error results from the use of an analysis predicated on purely random (Poisson) arrivals for overflow traffic into the second trunk group. Resolution of this error is discussed in the next section.

Separating the 18 channels of Example 12.13 into two groups is an obvious artifice. This example is useful in that it demonstrates an extreme case of correlation between blocking probabilities of two trunk groups. When correlation exists, the composite blocking probability of a direct route and an alternate route should be determined as follows:

$$B = (B_1)\,(B_2|1) \tag{12.22}$$

where B_1 = blocking probability of group 1
$B_2|1$ = blocking probability of group 2 given that group 1 is busy

In the artificial case of dividing a trunk group into two subgroups, the conditional blocking probability can be determined as

$$B_2|1 = B(N|N_1)$$

$$= \text{prob } (N \text{ servers are busy when } N_1 \text{ are known to be busy})$$

$$= \frac{P_N}{\sum_{n=N_1}^{N} P_n} = \frac{A^N/N!}{\sum_{n=N_1}^{N}(A^n/n!)} \tag{12.23}$$

where P_n is the probability that exactly n of N servers are busy (Equation 12.3).

Evaluating Equation 12.23 for $A = 10.8$, $N_1 = 12$, and $N = 18$ reveals that the appropriate conditional probability $B_2|1$ for Example 12.13 is 0.033. Thus the composite blocking probability per Equation 12.22 is $B = 0.15 \times 0.033 = 0.005$. The remaining inaccuracy (0.005 versus 0.013 actually) is due to nonrandom characteristics of overflow traffic.

Equation 12.23 is valid only for the contrived case of an alternate route carrying overflow traffic from only one primary route. It can be used, however, as a worst-case solution to situations where overflow from one route tends to dominate the traffic on an alternate route.

The correlations between the blocking probabilities of individual routes arise because congestion on one route produces overflows that tend to cause congestion on other routes. External events stimulating networkwide overloads also cause the blocking probabilities to be correlated. Thus a third assumption in the end-to-end blocking probability equation of Figure 12.12 is that traffic throughout the network is independent. If fluctuations in the traffic volume on individual links tend to be correlated (presumably because of external events such as television commercials, etc.), significant degradation in overall performance results.

12.3.2 Overflow Traffic

The second source of error in Example 12.13 occurred because an Erlang-*B* analysis used the average volume of overflow traffic from the first group to determine the blocking probability of the second trunk group. An Erlang-*B* analysis assumes traffic arrivals are purely random, that is, they are modeled by a Poisson distribution. However, a Poisson arrival distribution is an erroneous assumption for the traffic offered to the second trunk group. Even though arrivals to the first group may be random, the overflow process tends to select groups of these arrivals and pass them on to the second trunk group. Thus instead of being random the arrivals to the second group occur in bursts. This overflow effect is illustrated in Figure 12.14, which portrays a typical random arrival pattern to one trunk group and the overflow pattern to a second group. If a significant amount of the traffic flowing onto a trunk group results as overflow from other trunk groups, overly optimistic values of blocking probability arise when all of the traffic is assumed to be purely random.

The most common technique of dealing with overflow traffic is to relate the overflow traffic volume to an equivalent amount of random traffic in a blocking probability sense. For example, if the 1.62 erlangs of overflow traffic in Example 12.12 is equated to 2.04 erlangs of random traffic, a blocking probability of 1.3% is obtained for the second trunk group. (This is the correct probability of blocking for the second group since both groups are busy if and only if the second group is busy.)

This method of treating overflow traffic is referred to as the equivalent random theory [12]. Tables of traffic capacity are available [13] that incorporate the overflow effects directly into the maximum offered loads. The Neal–Wilkinson tables used by Bell System traffic engineers comprise one such set of tables. The Neal–Wilkinson tables, however, also incorporate the effects of day-to-day variations in the traffic load.

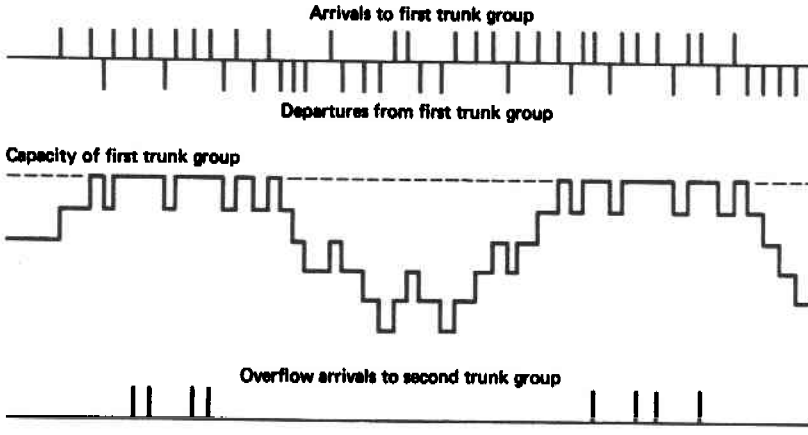


Figure 12.14 Bursty characteristic of overflow traffic.

(Forty erlangs on one day and 30 erlangs on another is not the same as 35 erlangs on both days.) These tables are also used for trunk groups that neither generate nor receive overflow traffic. The fact that cleared traffic does not get serviced by an alternate route implies that retries are likely. The effect of the retries, however, is effectively incorporated into the value of B by equivalent randomness.

12.4 DELAY SYSTEMS

The second category of teletraffic analysis concerns systems that delay nonserviceable requests until the necessary facilities become available. These systems are variously referred to as delay systems, waiting-call systems, and queuing systems. Call arrivals occurring when all servers are busy are placed in a queue and held until service commences. The queue might consist of storage facilities in a physical sense, such as blocks of memory in a message-switching node, or the queue might consist only of a list of sources waiting for service. In the latter case, storage of the messages is the responsibility of the sources themselves.

Using the more general term *queueing theory,* we can apply the following analyses to a wide variety of applications outside of telecommunications. Some of the more common applications are data processing, supermarket check-out counters, aircraft landings, inventory control, and various forms of service bureaus. These and many other applications are considered in the field of operations research. The foundations of queuing theory, however, rest on fundamental techniques developed by early telecommunications traffic researchers. In fact, Erlang is credited with the first solution to the most basic type of delay system. Examples of delay system analysis applications in telecommunications are message switching, packet switching, statistical time division multiplexing, multipoint data communications, automatic call distribution, digit receiver access, signaling equipment usage, and call processing. Furthermore, many

PBXs have features allowing queued access to corporate tie lines or WATS lines. Thus some systems formerly operating as loss systems now operate as delay systems.

In general, a delay operation allows for greater utilization of servers (transmission facilities) than does a loss system. Basically, the improved utilization is achieved because peaks in the arrival process are "smoothed" by the queue. Even though arrivals to the system are random, the servers see a somewhat regular arrival pattern. The effect of the queuing process on overload traffic is illustrated in Figure 12.15. This figure displays the same traffic patterns presented earlier in Figures 12.1, 12.3, and 12.9. In this case, however, overload traffic is delayed until call terminations produce available channels.

In most of the following analyses it is assumed that all traffic offered to the system eventually gets serviced. One implication of this assumption is that the offered traffic intensity A is less than the number of servers N. Even when A is less than N, there are two cases in which the carried traffic might be less than the offered traffic. First, some sources might tire of waiting in a long queue and abandon the request. Second, the capacity for storing requests may be finite. Hence requests may occasionally be rejected by the system.

A second assumption in the following analyses is that infinite sources exist. In a delay system, there may be a finite number of sources in a physical sense but an infinite number of sources in an operational sense because each source may have an arbitrary number of requests outstanding (e.g., a packet-switching node). There are instances in which a finite source analysis is necessary, but not in the applications considered here.

An additional implication of servicing all offered traffic arises when infinite sources exist. This implication is the need for infinite queuing capabilities. Even

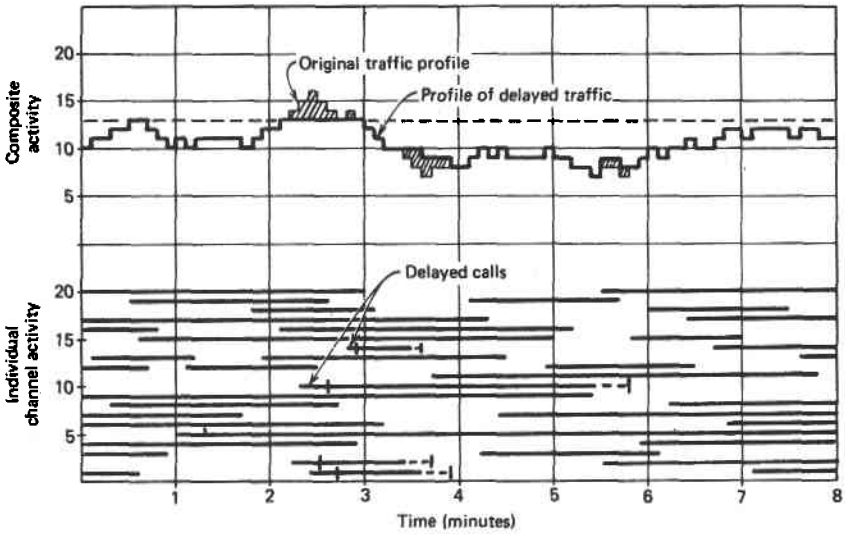


Figure 12.15 Activity profile of blocked calls delayed (13 servers).

though the offered traffic intensity is less than the number of servers, no statistical limit exists on the number of arrivals occurring in a short period of time. Thus the queue of a purely lossless system must be arbitrarily long. In a practical sense, only finite queues can be realized, so either a statistical chance of blocking is always present or all sources can be busy and not offer additional traffic.

When analyzing delay systems, it is convenient to separate the total time that a request is in the system into the waiting time and the holding time. In delay systems analysis the holding time is more commonly referred to as the service time. In contrast to loss systems, delay system performance is generally dependent on the distribution of service times and not just the mean value t_m. Two service time distributions are considered here: constant service times and exponential service times. Respectively, these distributions represent the most deterministic and the most random service times possible. Thus a system that operates with some other distribution of service times performs somewhere between the performance produced by these two distributions.

The basic purpose of the following analyses is to determine the probability distribution of waiting times. From the distribution, the average waiting time is easily determined. Sometimes only the average waiting time is of interest. More generally, however, the probability that the waiting time exceeds some specified value is of interest. In either case, the waiting times are dependent on the following factors:

1. Intensity and probabilistic nature of the offered traffic
2. Distribution of service times
3. Number of servers
4. Number of sources
5. Service discipline of the queue

The service discipline of the queue can involve a number of factors. The first of these concerns the manner in which waiting calls are selected. Commonly, waiting calls are selected on a first-come, first-served (FCFS) basis, which is also referred to as first-in, first-out (FIFO) service. Sometimes, however, the server system itself does not maintain a queue but merely polls its sources in a round-robin fashion to determine which ones are waiting for service. Thus the queue may be serviced in sequential order of the waiting sources. In some applications waiting requests may even be selected at random. Furthermore, additional service variations arise if any of these schemes are augmented with a priority discipline that allows some calls to move ahead of others in the queue.

A second aspect of the service discipline that must be considered is the length of the queue. If the maximum queue size is smaller than the effective number of sources, blocking can occur in a lost calls sense. The result is that two characteristics of the grade of service must be considered: the delay probability and the blocking probability. A common example of a system with both delay and loss characteristics is an automatic call distributor with more access circuits than attendants (operators or reservationists). Normally, incoming calls are queued for service. Under heavy loads,

however, blocking occurs before the ACD is even reached. Reference [14] contains an analysis of a delay system with finite queues and finite servers.

To simplify the characterization of particular systems, queuing theorists have adopted a concise notation for classifying various types of delay systems. This notation, which was introduced by D. G. Kendall, uses letter abbreviations to identify alternatives in each of the categories listed. Although the discussions in this book do not rely on this notation, it is introduced and used occasionally so the reader can relate the following discussions to classical queuing theory models. The interpretation of each letter is specified in Figure 12.16.

The specification format presented in Figure 12.16 actually represents an extension of the format commonly used by most queuing theorists. Thus this format is sometimes abbreviated by eliminating the last one or two entries. When these entries are eliminated, infinite case specifications are assumed. For example, a single-server system with random input and negative exponential service times is usually specified as M/M/1. Both the number of sources and the permissible queue length are assumed infinite.

12.4.1 Exponential Service Times

The simplest delay system to analyze is a system with random arrivals and negative exponential service times: M/M/N. Recall that a random arrival distribution is one with negative exponential interarrival times. Thus in the shorthand notation of queuing theorists, the letter M always refers to negative exponential distributions (an M is used because a purely random distribution is memoryless).

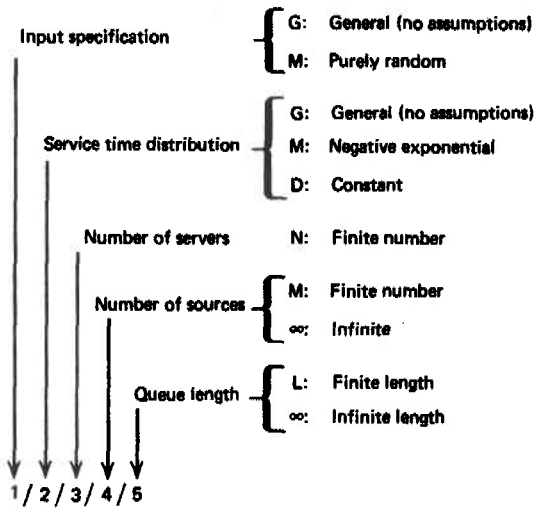


Figure 12.16 Queueing system notation.

In the M/M/1 system and all other systems considered here, it is assumed that calls are serviced in the order of their arrival. The following analyses also assume that the probability of an arrival is independent of the number of requests already in the queue (infinite sources). From these assumptions, the probability that a call experiences congestion and is therefore delayed was derived by Erlang:

$$\text{prob(delay)} = p(>0) = \frac{NB}{N - A(1 - B)} \tag{12.24}$$

where N = number of servers
A = offered load (erlangs)
B = blocking probability for a lost calls cleared system (Equation 12.8)

The probability of delay $p(>0)$ is variously referred to as Erlang's second formula, $E_{2,N}(A)$; Erlang's delay formula; or the Erlang-C formula. For single-server systems ($N = 1$) the probability of delay reduces to ρ, which is simply the output utilization or traffic carried by the server. Thus the probability of delay for a single-server system is also equal to the offered load λt_m (assuming $\lambda t_m < 1$).

The distribution of waiting times for random arrivals, random service times, and a FIFO service discipline is

$$p(>t) = p(>0)e^{-(N-A)t/t_m} \tag{12.25}$$

where $p(>0)$ = probability of delay given in Equation 12.24
t_m = average service time of negative exponential service time distribution

Equation 12.25 defines the probability that a call arriving at a randomly chosen instant is delayed for more than t/t_m service times. Figure 12.17 presents the relationship of Equation 12.25 by displaying the traffic capacities of various numbers of servers as a function of acceptable delay times. Given a delay time objective t/t_m, Figure 12.17*a* displays the maximum traffic intensity if the delay objective is to be exceeded for only 10% of the arrivals. Similarly, Figure 12.17*b* displays the maximum traffic intensity if the delay objective is exceeded for only 1% of the arrivals. Notice that at $p(>t)$ = 0.01, the server systems do not approach their maximum capacity (number of servers) unless the acceptable delay is several times larger than t_m.

By integrating Equation 12.25 over all time, the average waiting time for all arrivals can be determined as

$$\bar{t} = \frac{p(>0)t_m}{N - A} \tag{12.26}$$

Notice that $\bar{t}$ is the expected delay for all arrivals. The average delay of only those ar-

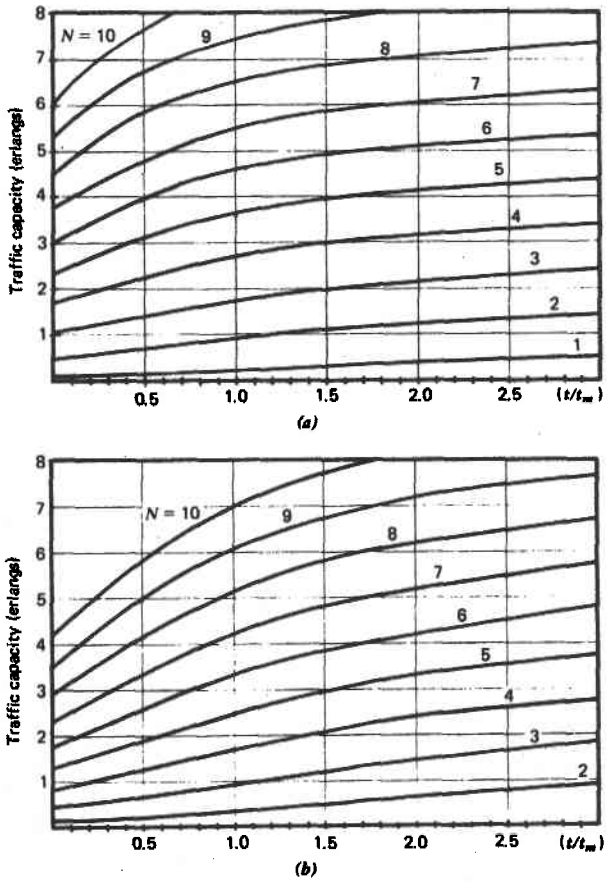


Figure 12.17 Traffic capacity of multiple-server delay systems with exponential service times: (*a*) probability of exceeding t, $p(>t) = 10\%$; (*b*) probability of exceeding t, $p(>t) = 1\%$.

rivals that get delayed is commonly denoted as

$$t_w = \frac{t_m}{N - A} \tag{12.27}$$

Example 12.14. A message-switching network is to be designed for 95% utilization of its transmission links. Assuming exponentially distributed message lengths and an arrival rate of 10 messages per minute, what is the average waiting time, and what is the probability that the waiting time exceeds 5 min?

Solution. Assume that the message-switching network uses a single channel between each pair of nodes. Thus there is a single server and a single queue for each transmission link. Since p is given to be 0.95 and $\lambda = 10$ arrivals per minute, the

average service time can be determined as $t_m = 0.95/10 = 0.095$ min. The average waiting time (not including the service time) is easily determined as

$$\bar{t} = \frac{0.95 \times 0.095}{1 - 0.95} = 1.805 \text{ min}$$

Using Equation 12.25, we can determine the probability of the waiting time exceeding 5 min as

$$p(>5) = (0.95)e^{-(1-0.95)5/0.095} = 0.068$$

Thus 6.8% of the messages experience queuing delays of more than 5 min.

Example 12.15. Determine the number of digit receivers required to support 1000 telephones with an average calling rate of two calls per hour. Assume the dialing time is exponentially distributed with an average service time of 6 sec. The grade of service objective is to return dial tone within 1 sec of the off-hook signal for 99% of the call attempts. Compare the answer obtained from a delay system analysis to an answer obtained from a loss system analysis at $B = 1\%$. If the blocking probability is less than 1%, fewer than 1% of the calls are delayed.

Solution. The calling rate λ and the offered traffic intensity A are easily determined as 0.555 calls per second and 3.33 erlangs, respectively. Since the number of servers N cannot be solved for directly from the equations, Figure 12.17*b* is used to obtain a value of eight servers for $t/t_m = \frac{1}{6}$.

Examination of Table D.1 reveals that 99.5% of the call attempts can be serviced immediately if there are nine digit receivers. Thus in this case the ability to delay service provides a savings of only one server.

Example 12.15 demonstrates that a blocking probability analysis produces approximately the same results as a delay system analysis when the maximum acceptable delay is a small percentage of the average service time. The two results are almost identical because, if a digit receiver is not immediately available, there is only a small probability that one will become available within a short time period. (With an average service time of 6 sec, the expected time for one of eight digit receivers to be released is 6/8 = 0.75 sec. Hence the delay operation in this case allows a savings of one digit receiver.)

Because a digit receiver must be available within a relatively short time period after a request is generated, digit receiver group sizing is often determined strictly from a blocking probability analysis. The fact that digit receiver access is actually operated as a delay system implies the grade of service is always better than that calculated.

12.4.2 Constant Service Times

This section considers delay systems with random arrivals, constant service times, and a single server (M/D/1). Again, FIFO service disciplines and infinite sources are as-

sumed. The case for multiple servers has been solved [3] but is too involved to include here. Graphs of multiple-server systems with constant service times are available in reference [15].

The average waiting time for a single server with constant service times is determined as

$$\bar{t} = \frac{\rho t_m}{2(1-\rho)} \tag{12.28}$$

where $\rho = A$ is the server utilization. Notice that Equation 12.28 produces an average waiting time that is exactly one-half of that for a single-server system with exponential service times. Exponential service times cause greater average delays because there are two random processes involved in creating the delay. In both types of systems, delays occur when a large burst of arrivals exceeds the capacity of the servers. With exponential service times, however, long delays also arise because of excessive service times of just a few arrivals. (Recall that this aspect of conventional message-switching systems is one of the motivations for breaking messages up into packets in a packet-switching network.)

If the activity profile of a constant service time system (M/D/1) is compared with the activity profile of an exponential service time system (M/M/1), the M/D/1 system is seen to be active for shorter and more frequent periods of time. That is, the M/M/1 system has a higher variance in the duration of its busy periods. The average activity of both systems is, of course, equal to the server utilization ρ. Hence the probability of delay for a single-server system with constant service times is identical to that for exponential service times: $p(>0) = \lambda t_m$.

The probability of congestion for larger N is relatively close to that for exponential service times. Thus Equation 12.25 can be used as a close approximation for $p(>0)$ for multiple-server systems with arbitrary service time distributions.

For single-server systems with constant holding times, the probability of delay greater than an arbitrary value t is

$$\begin{aligned} p(>t) &= p[>(k+r)t_m] \\ &= 1-(1-\rho)\sum_{i=0}^{k} \frac{\rho^i (i - t/t_m)^i e^{-\rho(i-t/t_m)}}{i!} \\ &= 1-(1-\rho)e^{\lambda t}\sum_{i=0}^{k} \frac{(i\rho - \lambda t)^i e^{-i\rho}}{i!} \end{aligned} \tag{12.29}$$

where k = largest integral quotient of t/t_m
r = remainder of t/t_m
ρ = server utilization, $= \lambda t_m$

Comparisons of the waiting time distributions for single-server systems with exponential and constant service times are shown in Figure 12.18. For each pair of curves, the upper one is for exponential service times and the lower one is for constant service times. Since all other service time distributions produce delay probabilities between these extremes, Figure 12.18 provides a direct indication of the range of possible delays.

Example 12.16. A packet-switching node operates with fixed-length packets of 300 bits on 9600-bps lines. If the link utilization is to be 90%, what is the average

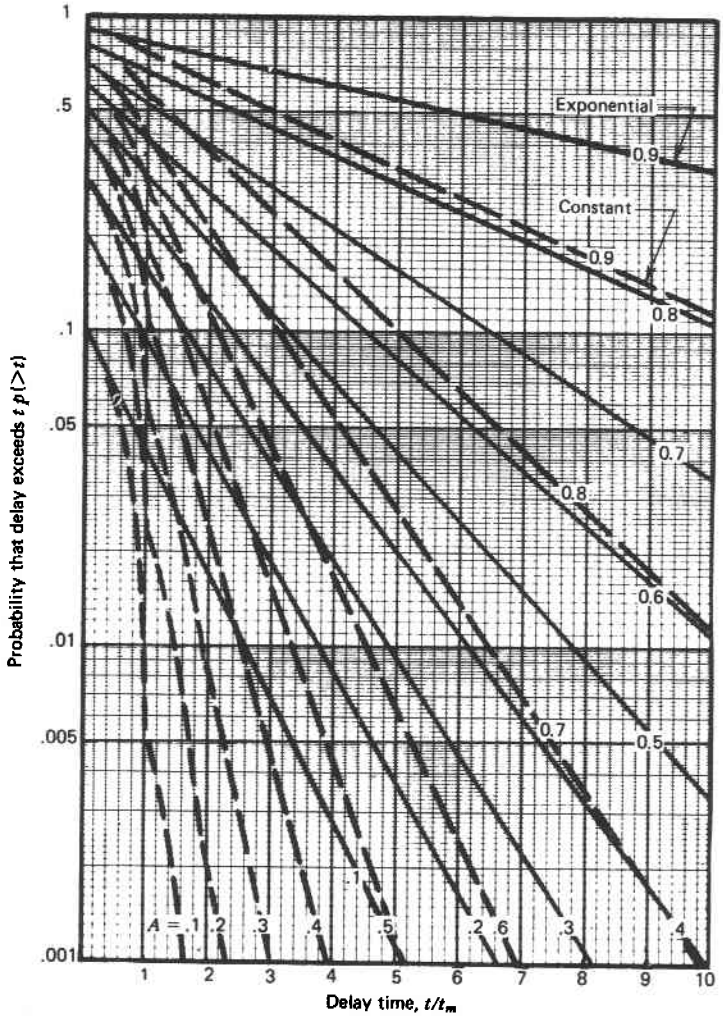


Figure 12.18 Delay probabilities of single-server systems (exponential and constant service times).

delay through a node? What percentage of packets encounter more than 0.35 sec of delay? What is the average delay if the offered load increases by 10%?

Solution. Message lengths of 300 bits and a data rate of 9600 bps imply that the fixed-length service time is 300/9600 = 0.031 sec. From Equation 12.28, the average waiting time is

$$\bar{t} = \frac{0.9 \times 0.031}{2(1 - 0.9)} = 0.14 \text{ sec}$$

The total average delay through the node, excluding processing, is obtained by adding the average waiting time to the service time:

$$\text{Average delay} = 0.140 + 0.031 = 0.171 \text{ sec}$$

Since the service time is 0.031 sec, 0.35 sec of delay occurs when the waiting time is 0.35 − 0.031 = 0.319. This corresponds to 0.319/0.031 = 10 service times. From Figure 12.18, the probability of delay for $t/t_m = 10$ is approximately 0.12. Thus 12% of the packets experience delays of greater than 0.35 sec. An increase of 10% in the traffic intensity implies that the new offered load is 0.99 erlang. From Equation 12.30, the average waiting time becomes

$$\bar{t} = \frac{0.99 \times 0.031}{2(1 - 0.99)} = 1.53 \text{ sec}$$

Thus when the offered load increases by only 10%, the average delay through the node increases ninefold to a value of 1.53 + 0.031 = 1.56 sec!

Example 12.16 demonstrates the same characteristic for heavily utilized delay systems that was demonstrated for loss systems: The performance is very sensitive to increases in traffic intensity. Thus, as discussed in Chapter 10, flow control is a critical aspect of a packet-switching operation, particularly when there are real-time delivery objectives.

12.4.3 Finite Queues

All of the delay system analyses presented so far have assumed that an arbitrarily large number of delayed requests could be placed in a queue. In many applications this assumption is invalid. Examples of systems that sometimes have significantly limited queue sizes are store-and-forward switching nodes (e.g., packet switches and ATM switches), automatic call distributors, and various types of computer input/output devices. These systems treat arrivals in three different ways, depending on the number "in the system" at the time of an arrival:

1. Immediate service if one or more of N servers are idle
2. Delayed service if all servers are busy and less than L requests are waiting
3. Blocked or no service if the queue of length L is full

In finite-queue systems the arrivals getting blocked are those that would otherwise experience long delays in a pure delay system. Thus an indication of the blocking probability of a combined delay and loss system can be determined from the probability that arrivals in pure delay systems experience delays in excess of some specified value. However, there are two basic inaccuracies in such an analysis. First, the effect of blocked or lost calls cleared is to reduce congestion for a period of time and thereby to reduce the delay probabilities for subsequent arrivals. Second, delay times do not necessarily indicate how many calls are "in the system." Normally, queue lengths and blocking probabilities are determined in terms of the number of waiting requests, not the amount of work or total service time represented by the requests. With constant service times, there is no ambiguity between the size of a queue and its implied delay. With exponential service times, however, a given size can represent a wide range of delay times.

A packet-switching node is an example of a system in which the queue length is most appropriately determined by implied service time and not by the number of pending requests. That is, the maximum queue length may be determined by the amount of store-and-forward memory available for variable-length messages and not by some fixed number of messages.

For a system with random input, exponential service times, N servers, an infinite source, and a maximum queue length of L (M/M/N/∞/L), the probability of j calls in the system is

$$P_j(A) = \begin{cases} P_0(A)\dfrac{A^j}{j!} & 0 \le j \le N \\ \dfrac{P_0(A)A^j}{N!N^{j-N}} & N \le j \le N+L \end{cases} \tag{12.30}$$

where A = offered load (erlangs), $= \lambda t_m$
N = number of servers
L = maximum number in the queue

Here, $P_0(A)$ is chosen to make the sum of all $P_j(A) = 1$:

$$P_0(A) = \left(\sum_{j=0}^{N} \frac{A^j}{j!} + \sum_{j=N+1}^{N+L} \frac{A^j}{N!N^{j-N}} \right)^{-1}$$

$$= \left(\sum_{j=0}^{N} \frac{A^j}{j!} + \frac{A_N}{N!} \sum_{j=1}^{L} \frac{A^j}{N^j} \right)^{-1} \tag{12.30a}$$

The time congestion, or probability of an arrival being delayed or blocked, can be determined from Equation 12.30 as

$$P(>0) = \sum_{j=N}^{N+L} P_j(A) = P_N(A)\frac{1-\rho^{L+1}}{1-\rho} \tag{12.31}$$

where $\rho = A/N$ is the offered load per server.

The loss, or blocking probability, is determined as

$$B = P_{N+L}(A) = \frac{P_0(A)A^{N+L}}{N!N^L} \tag{12.32}$$

It is worth noting that if there is no queue ($L = 0$), these equations reduce to those of the Erlang loss equation (12.8). If L is infinite, Equation 12.31 reduces to Erlang's delay formula, Equation 12.24. Thus these equations represent a general formulation that produces the pure loss and pure delay formulas as special cases.

The waiting time distribution [3] is

$$p(>t) = P_N(A)\sum_{j=0}^{L-1}\frac{\rho^j}{j!}\int_{Nt/t_m}^{\infty} x^j e^{-x}dx \tag{12.33}$$

from which the average delay can be determined as

$$\bar{t} = \frac{[p(>0) - P_{N+L}(A)]t_m}{N-A} \tag{12.34}$$

Again, Equation 12.34 is identical to Equation 12.26 for an infinite queue ($L = \infty$).

Single-Server Equations

Because most queuing applications involve single-server configurations, the previous equations are listed explicitly for N = 1:

Prob(j calls in system) (12.30):

$$P_j(\rho) = P_0(\rho)\rho^j \tag{12.35}$$

$$P_0(\rho) = \left(1 + \rho + \rho\sum_{j=1}^{L}\rho^j\right)^{-1} = \left(\sum_{j=0}^{L+1}\rho^j\right)^{-1} = \frac{1-\rho}{1-\rho^{L+2}} \tag{12.36}$$

Probability of delay (12.31):

$$p(>0) = \frac{P_0(\rho)\rho(1-\rho^{L+1})}{1-\rho} = \frac{\rho(1-\rho^{L+1})}{1-\rho^{L+2}} \tag{12.37}$$

Probability of loss (12.32):

$$B = \frac{(1-\rho)\rho^{L+1}}{1-\rho^{L+2}} \tag{12.38}$$

Average delay (12.34):

$$\bar{t} = \frac{[p(>0) - P_{L+1}(\rho)]t_m}{1-\rho} = \frac{\rho(1-\rho^L)t_m}{(1-\rho)(1-\rho^{L+2})} \tag{12.39}$$

The blocking probability of a single-server system ($N = 1$) is plotted in Figure 12.19. When using Figure 12.19, keep in mind that the blocking probability (Equation 12.38) is determined by the number of waiting calls and not by the associated service time. Furthermore, since the curves of Figure 12.19 are based on exponential service times, they overestimate the blocking probabilities of constant holding time systems (e.g., fixed-length packet networks). However, if fixed-length packets arise primarily from longer, exponentially distributed messages, the arrivals are no longer independent, and the use of Figure 12.19 (or Equation 12.38) as a conservative analysis is more appropriate.

ATM Cell Queues

Analysis of queuing delays and cell loss in an ATM switching node is complicated. The cells have a fixed length of 53 bytes so it would seem that a constant service time analysis would be appropriate. This assumption is valid for voice traffic inserted onto wide-bandwidth signals such as 155-Mbps STS-1s. In this case the service time is much shorter that the duration of a speech burst (e.g., 2.7 μsec versus several tens of milliseconds). Even though correlated arrivals occur from individual sources, the arrival times are separated by many thousands of service times so they appear independent.

When ATM voice is carried in CBR trunk groups, a different situation results. In this case the service times of the voice cells may be only slightly smaller than the interval between voice cell generation, and the average delay would indicate that two or more cells from the same source could be present in the queue at one time. Thus, a queuing analysis that assumes exponentially distributed service times is more appropriate even though the variable-length talk spurts are broken up into fixed-length cells.

Example 12.17. A 64-kbps CBR virtual path in an ATM network is to carry 18 voice signals that are compressed to 7.25 kbps during voice spurts. Assuming a speech activity factor of 40%, determine each of the following:

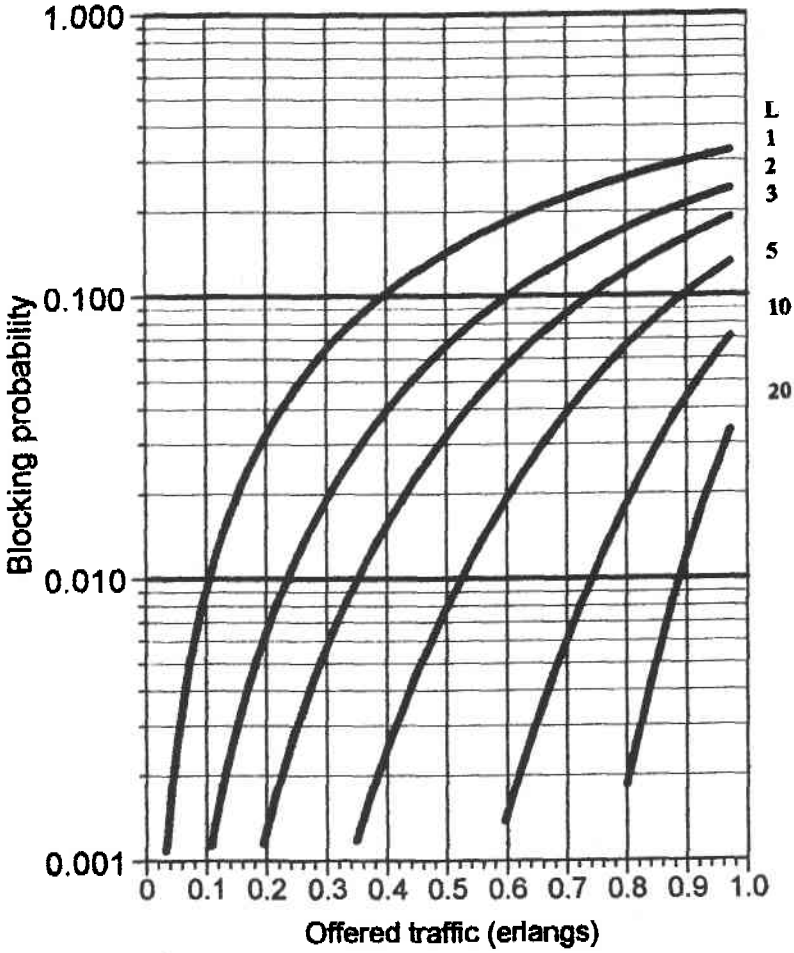


Figure 12.19 Blocking probability of single server loss/delay system (exponential service times).

(a) Output channel loading
(b) Service time
(c) Probability of delay with an infinite queue
(d) Probability of delay with a queue of length 20 ATM cells
(e) Average time in queue for exponential service times with an infinite queue
(f) Average time in queue for exponential service times with a finite queue of 20
(g) Average time in queue for constant service times (infinite queue)
(h) Probability of cell loss (assuming exponential service times)

Solution

(a) An ATM cell consists of 48 bytes of payload and 5 bytes of overhead. Thus, the offered load to the channel = (18 × 7.25 kbps × 0.4) (53/48)/64 kbps = 0.9 erlangs.

(b) The service time of a cell is 53 × 8/64 kbps = 6.625 msec.

(c) The probability of delay from Equation 12.24 is 0.9.

(d) The probability of delay from Equation 12.37 is 0.89.

(e) The average queuing delay from Equation 12.26 is 59.6 msec.

(f) The average queuing delay from Equation 12.34 is 58.2 msec.

(g) The average queuing delay from Equation 12.28 is 29.8 msec.

(h) The probability of cell loss from Equation 12.38 is 0.012.

The results of Example 12.17 illustrate several important points. First, (c) and (d) indicate that the probability of delay in a finite-queue system is smaller than that of an infinite-queue system—because some traffic is rejected. With a reasonably sized queue the effect on the delay probability is small. Similarly, a comparison of (e) and (f) illustrates that the average queuing delay in a reasonably sized queue is not much different than that of an infinite queue. As discussed earlier, there is a significant difference between the average delay of a system with exponentially distributed service times as opposed to constant service times. Before assuming that a constant service time analysis should be used, the average delay must be compared to the delay between arrivals of voice cells. The duration of a speech segment carried in the ATM cells of Example 12.17 can be determined as 48 × 8/7250 = 53 msec. The average delay of a constant service time analysis (29.8 msec) indicates that cells from a particular source are usually serviced before a subsequent arrival from the same source, but certainly not always. If Equation 12.34 is used to determine the average delay, the 58.2-msec result indicates that, on average, an arriving cell from a source encounters a previous cell from the same source. Thus the assumption of independent arrivals is marginal. To be safe, a cell loss calculation assuming exponential service times accounts for correlated arrivals. The cell loss probability of 1.2% is marginal from a voice quality point of view. Notice that this value of cell loss relates to the CBR gateway (AAL1 adaptation layer), which is presumably the only significant source of cell loss.

The solution to example problem 12.17 necessarily used a simple model for the traffic and the queue. A thorough analysis of an ATM or packet-switching network is much more complicated. Factors that must be considered are the switching node architectures and queue implementation (e.g., shared queues versus dedicated queues), server disciplines (e.g., priorities), cell discard algorithms, call admission control algorithms, and traffic statistics. Some of these issues are covered in references [16–18].

12.4.4 Tandem Queues

All of the equations provided in previous sections for delay system analysis have dealt with the performance of a single queue. In many applications a service request undergoes several stages of processing, each one of which involves queuing. Thus it is often

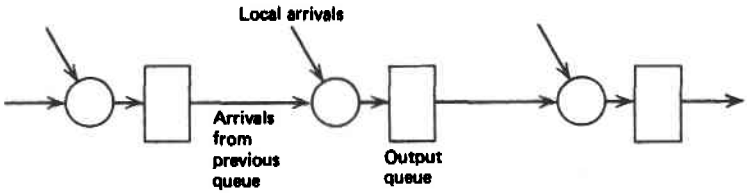


Figure 12.20 Tandem queues.

desirable to analyze the performance of a system with a number of queues in series. Figure 12.20 depicts a series of queues that receive, as inputs, locally generated requests and outputs from other queues. Two principal examples of applications with tandem queues are data processing systems and store-and-forward switching networks.

Researchers in queuing theory have not been generally successful in deriving formulas for the performance of tandem queues. Often, simulation is used to analyze a complex arrangement of interdependent queues arising in systems like store-and-forward networks. Simulation has the advantage that special aspects of a network's operation—like routing and flow control—can be included in the simulation model. The main disadvantages of simulation are expense and, often, less visibility into the dependence of system performance on various design parameters.

One tandem queuing problem that has been solved [19] is for random inputs and random (negative exponential) holding times for all queues. The solution of this system is based on the following theorem: In a delay system with purely random arrivals and negative exponential holding times, the instants at which calls terminate is also a negative exponential distribution.

The significance of this theorem is that outputs from an M/M/N system have statistical properties that are identical to its inputs. Thus a queuing process in one stage does not affect the arrival process in a subsequent stage, and all queues can be analyzed independently. Specifically, if a delay system with N servers has exponentially distributed interarrival times with average $1/\lambda$, and if the average service time is t_m, calls leave each of the servers according to exponentially distributed intercompletion times with average $1/\lambda N$.

Although independent analysis of tandem queues can be rigorously justified only for purely random arrivals and service times, independence is often assumed in other cases. Before using such assumptions, however, the systems in question should be examined closely to determine if the state of one queue can influence the operation of another queue in the system.

REFERENCES

1 W. Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1968.

2 A. A. Collins and R. D. Pederson, *Telecommunications—a Time for Innovation*, Merle Collins Foundation, Dallas, TX, 1973.

3 R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1965.

4 Technical Staff, Bell Telephone Laboratories, *Engineering and Operations in the Bell System*, Western Electric, Indianapolis, 1977.

5 "Determination of the Number of Circuits in Automatic and Semi-Automatic Operation," CCITT Recommendation E. 520, *Orange Book*, Vol. II.2, 1976, p. 211.

6 C. J. Weinstein, "Fractional Speech Loss and Talker Activity Model for Packet Switched Speech," *IEEE Transactions on Communications Technology*, Aug. 1978, pp. 1253–1256.

7 H. Miedema and M. Schachtman, "TASI Quality-Effect of Speech Detection and Interpolation," *Bell System Technical Journal*, July 1962, pp. 1455–1473.

8 G. Szarvas and H. Suyderhoud, "Voice-Activated-Switch Performance Criteria," *Comsat Technical Review*, Spring 1980, pp. 151–177.

9 S. S. Katz, "Improved Network Administration Process Utilizing End-to-End Service Considerations," *International Teletraffic Conference*, 1979.

10 R. Dayem, "Alternate Routing in High Blocking Communications Networks," *National Telecommunications Conference*, 1979, pp. 28.4.1–28.4.6.

11 P. R. Boorstyn and H. Frank, "Large-Scale Network Topological Optimization," *IEEE Transactions on Communications*, Jan. 1977, pp. 29–47.

12 R. I. Wilkinson, "Theories for Toll Traffic Engineering in U.S.A.," *Bell System Technical Journal*, Mar. 1956.

13 "Calculation of the Number of Circuits in a Group Carrying Overflow Traffic," CCITT Recommendation E.521, *Orange Book*, Vol. 2, No. 2, Geneva, Switzerland, p. 218.

14 J. A. Morrison, "Analysis of Some Overflow Problems with Queueing," *Bell System Technical Journal*, Oct. 1980, pp. 1427–1462.

15 *Telephone Tragic Theory, Tables, and Charts*, Siemens Aktiengesellschaft, Munich, 1970.

16 Y.-S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," *IEEE Journal on Selected Areas in Communications*, Oct. 1987, pp. 1274–1283.

17 K. Sriram, T. G. Lyons, and T.-T. Wang, "Anomalies Due to Delay and Loss in AAL2 Packet Voice Systems: Performance Models and Methods of Mitigation," *IEEE Journal on Selected Areas in Communications*, Jan. 1999, pp. 4–17.

18 K. Sriram and Y. T. Yang, "Voice over ATM Using AAL2 and Bit Dropping: Performance and Call Admission Control," *IEEE Journal on Selected Areas in Communications*, Jan. 1999, pp. 18–28.

19 L. Kleinrock and R. Gail, *Queueing Systems, Problems and Solutions*, Wiley, New York, 1996.

PROBLEMS

12.1 A central-office-to-PBX trunk group contains four circuits. If the average call duration is 3 min and the busy-hour offered traffic intensity is 2 erlangs, determine each of the following:

(a) Busy-hour calling rate
(b) Probability that two arrivals occur less than 1 sec apart
(c) Blocking probability assuming a lost calls cleared operation
(d) Amount of lost traffic
(e) Proportion of time the fourth circuit is in use (assuming fixed-order selection)

12.2 A T1 line is used to carry traffic from a remote concentrator to a central office. How many 10 CCS subscribers can the concentrator system support at 0.5% blocking. Compare an infinite source analysis to a finite source analysis. Assume blocked calls cleared.

12.3 Two switching offices experience 20 erlangs of average busy-hour traffic load between them. Assume a single T1 line provides 24 direct trunks between the offices. How much busy-hour traffic overflows to a tandem switch?

12.4 A PBX with 200 stations has five trunks to the public network. What is the blocking probability if each station is involved in three external calls per 8-hr working day with an average duration of 2 min per call? Assume the average calling rate is constant during the day (no busy hour) and blocked calls return with random retries. What is the offered load? What is the demand traffic?

12.5 How many dial-up input–output (I/O) ports are needed for a computer center to support 40 users with a blocking probability limit of 5%? Assume each user averages four calls per day with an average session duration of 30 min. If three users remain connected all day, what is the grade of service for the remaining 37 users?

12.6 A 24-channel trunk group is divided into two groups of 12 one-way trunks in each direction. (A one-way trunk is one that can only be seized at one end.) How many erlangs of traffic can this system support at 0.5% blocking? How many erlangs can be supported if all 24 trunks are two-way trunks? (That is, every trunk can be seized at either end.)

12.7 The following 10 A.M. to 11 A.M. busy-hour Erlang (E) statistics have been observed on a 32-channel interoffice trunk group. What is the overall blocking probability? What is the blocking probability for the same busy hour if day-to-day fluctuations are averaged together?

Monday	Tuesday	Wednesday	Thursday	Friday
20E	19E	22E	19E	30E

12.8 Traffic measurements on a PBX-to-central-office trunk group indicate that during the busiest hour of the day the trunks are 80% utilized. If there are eight trunks in the group, what is the blocking probability, assuming blocked calls do not return? How many trunks must be added to achieve a maximum blocking probability of 5%?

12.9 Repeat Problem 12.8 assuming blocked calls return with random retries.

12.10 A small community with 400 subscribers is to be serviced with a community dial office switch. Assume that the average subscriber originates 0.1 erlang of traffic. Also assume that 20% of the originations are local (intracommunity)

calls and that 80% are transit calls to the serving central office. How many erlangs of traffic are offered to the community-dial-office-to-central-office trunk group? How many trunks are needed for 0.5% blocking of the transit traffic?

12.11 For the community of Problem 12.10 determine the number of concentrator channels required if local calls are not switched locally but are merely concentrated into pair-gain systems and switched at the central office.

12.12 Repeat Problems 12.10 and 12.11 if 80% of the originations are intracommunity calls and 20% are transit calls.

12.13 A group of eight remote farm houses are serviced by four lines. If each of the eight families utilizes their telephones for 10% of the busy hour, compare the blocking probabilities of the following configurations:

(a) Four party lines with two stations per line
(b) An 8-to-4 concentration system

12.14 A PBX provides queuing and automatic call back for access to outgoing WATS lines. If there are 20 requests per hour for the WATS lines, and if the average call is 3 min in length, how many WATS lines are needed to provide delays of less than 1 hr for 90% of the requests?

12.15 A call processor has 50% of its time available for servicing requests. If each request requires 50 msec of processing time, what arrival rate can be supported if only 1% of the service requests are delayed by more than 1 sec? Assume that processor time is sliced into 500-msec time slots. (That is, 500 msec are allocated to call processing and then 500 msec to overhead functions.)

12.16 A group of 100 sources offers messages with exponentially distributed lengths to a 1200-bps line. The average message length is 200 bits, including overhead, and each source generates one message every 20 sec. Access to the line is controlled by message-switching concentration with an infinite queue. Determine the following:

(a) Probability of entering the queue
(b) Average queuing delay for all arrivals
(c) Probability of being in the queue for more than 1 sec
(d) Utilization of the transmission link

12.17 An airline company uses an automatic call distributor to service reservations and ticket purchases. Assume that the processing time of each inquiry is randomly distributed with a 40-sec average. Also assume that if customers are put on hold for more than 2 min, they hang up and call another airline. If each of 200 inquiries per hour produces $30 worth of sales, on average, what is the optimum number of reservationists? Assume each reservationist costs the company $20/hr (including overhead).

12.18 A radio station talk show solicits the listening public for comments on the ineptness of government (I assume this will be a topical subject for the life of this book). Assume that each caller talks for a random length of time with an average duration of 1 min. (Either the show is unpopular or the public has given up

on the government.) How many incoming lines must the radio station have to keep the idle time below 5% if the call arrival rate is 3 calls/min?

12.19 Repeat Example 12.17 for a queue length $L = 40$. (Although a rigorous solution requires calculation of a 20-term summation, only the first few terms are significant.)

APPENDIX A

DERIVATION OF EQUATIONS

A.1 QUANTIZING NOISE POWER: EQUATION 3.2

The probability density function of a noise v sampled is assumed to be uniform:

$$p(\eta) = \begin{cases} \dfrac{1}{q} & -\frac{1}{2}q \le \eta \le \frac{1}{2}q \\ 0 & \text{otherwise} \end{cases}$$

The average or expected value of noise power is determined as

$$\text{Quantization noise power} = \int_{-q/2}^{q/2} \left(\frac{1}{q}\right)\eta^2 \, dn$$

$$= \left(\frac{1}{12}\right)q^2$$

A.2 NRZ LINE CODE: EQUATION 4.1

$$f(t) = \begin{cases} 1 & |t| \le \frac{1}{2}T \\ 0 & \text{otherwise} \end{cases}$$

$$F(j\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} \, dt$$

$$= \int_{-T/2}^{T/2} e^{-j\omega t} \, dt$$

$$= \left(\frac{1}{j\omega}\right)(e^{j\omega T/2} - e^{-j\omega T/2})$$

$$= (T)\frac{\sin(\omega T/2)}{(\omega T/2)}$$

Note: $F(j\omega)$ is the spectrum of a single pulse; $|F(j\omega)^2\ (1/T)$ is the power spectral density of a random pulse train assuming positive and negative pulses are equally likely and occur independently.

A.3 DIGITAL BIPHASE: FIGURE 4.13

$$f(t) = \begin{cases} 1 & -\frac{1}{2}T < t < 0 \\ -1 & 0 < t < \frac{1}{2}T \\ 0 & \text{otherwise} \end{cases}$$

$$F(j\omega) = \int_{-(1/2)T}^{0} e^{-j\omega t}\,dt - \int_{0}^{(1/2)T} e^{-j\omega t}\,dt$$

$$= \left(\frac{-1}{j\omega}\right)(1 - e^{j(1/2)\omega T} - e^{-j(1/2)\omega T} + 1)$$

$$= \left(\frac{j4}{\omega}\right)\sin^2\left(\frac{\omega T}{4}\right)$$

A.4 FRAME ACQUISITION TIME OF SINGLE-BIT FRAME CODE: EQUATION 4.10

Framing is established by successively examining one bit position after another until a sufficiently long framing pattern is detected. In this derivation it is assumed that the framing pattern alternates 1's and 0's. Furthermore, it is assumed that when beginning to test a particular bit position for framing, the value of the first appearance is saved and compared to the second appearance. Thus the minimum time to reject an invalid framing position is one frame time. If we denote by p the probability of a 1 and by $q = 1 - p$ the probability that a 0 is received first, the average number of frames required to receive a mismatch is

$$
\begin{aligned}
A_0 &= (1)\ (\text{probability of mismatch at end of first frame}) \\
&\quad + (2)\ (\text{probability of mismatch at end of second frame}) \\
&\quad + (3)\ (\text{probability of mismatch at end of third frame}) \ldots \\
&= (1)q + (2)(1-q)p + (3)(1-q)(1-p)q \\
&\quad + (4)(1-q)^2(1-p)p + \ldots \\
&= (1)q + (2)p^2 + (3)pq^2 + (4)p^3q + (5)p^2q^3 + \ldots \\
&= (q + 2p^2 + pq^2)[1 + (2)pq + (3)p^2q^2 + (4)p^3q^3 + \ldots] \\
&= [1 - p + 2p^2 + p(1-p)^2](1 + pq + p^2q^2 + p^3q^3 + \ldots)^2 \\
&= \frac{1+p^3}{(1-pq)^2} \\
&= \frac{1+p}{1-pq}
\end{aligned}
$$

Similarly, if a 1 is received first, the expected number of frames before receiving a mismatch is

$$A_1 = \frac{1+q}{1-pq}$$

The overall average number of frames required to detect a mismatch is

$$
\begin{aligned}
A &= qA_0 + pA_1 \\
&= \frac{1+2pq}{1-pq}
\end{aligned}
$$

If we assume a random starting point in a frame with N bits, the average number of bits that must be tested before the true framing bit is encountered is

$$
\begin{aligned}
\text{Frame time} &= (1/2N)(A \cdot N) + 1/2N \\
&= 1/2N(A \cdot N + 1) \text{ bit times}
\end{aligned}
$$

If 1's and 0's are equally likely ($p = q = \frac{1}{2}$), $A = 2$ so frame time $= N^2 + 1/2N$ bit times (Equation 4.10).

A.5 FRAME ACQUISITION TIME OF SINGLE-BIT FRAMING CODE USING A PARALLEL SEARCH: EQUATION 4.11

This framing algorithm assumes that all bit positions in a frame are simultaneously scanned for the framing pattern. If we assume that an alternating-bit frame code is used and that 1's and 0's in the information bits are equally likely, the probability that a particular information bit does not produce a framing violation in n frames is

$$p_n = \left(\frac{1}{2}\right)^n$$

The probability that a framing violation has been received in n or less frames is $1 - p_n$. The probability that all $N - 1$ information bits in a frame produce a framing violation in n or less frames is

$$p_n = \left[1 - \left(\frac{1}{2}\right)^n\right]^{N-1} \tag{4.11}$$

A.6 FRAME ACQUISITION TIME OF MULTIBIT FRAME CODE: EQUATION 4.13

N = length of frame including framing code
L = length of framing code
$p = (\frac{1}{2})^L$ = probability of matching frame code with random data

The expected number of frames examined before a particular frame position mismatches the frame code follows:

$$\begin{aligned} A &= (0)(1-p) + (1)p(1-p) + (2)p^2(1-p) + \ldots \\ &= (1-p)p(1 + 2p + 3p^2 + 4p^3 + \ldots) \\ &= (1-p)p(1 + p + p^2 + p^3 + \ldots)^2 \\ &= (1-p)p\left(\frac{1}{1-p}\right)^2 \\ &= \frac{p}{1-p} \end{aligned}$$

The average number of bits that pass before the frame position is detected (assuming a random starting point and the test field is moved one bit position when a mismatch occurs) is given as

$$T = \left(\frac{1}{2}N\right)(A)(N) + \frac{1}{2}N$$

$$= \frac{N^2(1/2)^{L+1}}{1-(1/2)^L} + \frac{1}{2}N \tag{4.13}$$

Note: Equation 4.13 with $L = 1$ is not identical to Equation 4.10 because Equation 4.13 assumes a fixed frame code while Equation 4.10 assumes an alternating code.

A.7 PATHFINDING TIME: EQUATION 5.11

Assume that all paths through a switch are independently busy with probability p. Let the probability that a path is not busy be denoted by $q = 1 - p$. The probability p_i that exactly i paths are tested before an idle one is found is the probability that the first $i - 1$ are busy and the ith is not:

$$p_i = p^{(i-1)}q$$

The expected number of paths tested before an idle path is found is

$$N_p = (1)q + (2)pq + (3)p^2q + \ldots + (k)p^{k-1}q + (k)p^k$$

where the last term represents the expectation that all possible paths k are unavailable. A closed form for A is determined as

$$N_p = (1-p)(1 + 2p + 3p^2 + \ldots + kp^{k-1} + kp^k)$$

$$= 1 + p + p^2 + p^3 + \ldots + p^{k-1}$$

$$= \frac{1}{1-p} - p^k\left(\frac{1}{1-p}\right)$$

$$= \frac{1-p^k}{1-p} \tag{5.11}$$

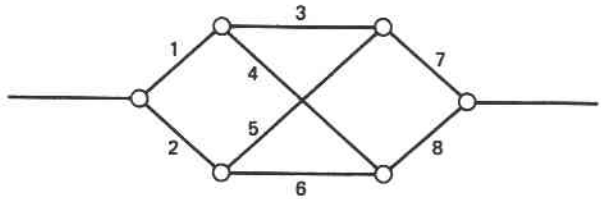


Figure A.1. Probability graph of No. 4 ESS four-stage space switch.

A.8 LESS SPACE STAGE BLOCKING PROBABILITY 5.21

The most straightforward way to calculate the blocking probability of the four-stage space structure shown in the probability graph of Figure A.1 is to list all elementary, mutually exclusive probability terms, determine their probabilities of occurrence, and generate the sum of those that represent blocking. Because there are eight links, which are assumed to be independently busy or idle, there are $2^8 = 256$ elementary terms. Rather than laboriously list them all, the analysis can be greatly simplified, with a risk of miscounting blocking terms, by grouping the terms according to the number of busy links. The following table lists the groups and the corresponding numbers of terms that block and do not block.

Number of Busy Links i	Combinations $\binom{8}{i}$	Number That Block	Number that Do Not Block
0	1	0	1
1	8	0	8
2	28	2	26
3	56	16	40
4	70	50	20
5	56	52	4
6	28	28	0
7	8	8	0
8	1	1	0
	256	157	99

Entries in the last two columns are determined by analyzing the topology of the network. For example, when two links are busy, only 2 of the 28 combinations produce blocking (1 and 2 and 7 and 8). When three links are busy, 16 of the combinations produce blocking. To determine the remaining entries in column 3, it is easier to determine the number of combinations that do not block and subtract from the total. For example, when only three links are idle, there are 4 of 56 combinations that do not block. Using the entries in column 3, the blocking probability is determined as

$$B = 2p^2q^6 + 16p^3q^5 + 50p^4q^4 + 52p^5q^3 + 28p^6q^2 + 8p^7q + p^8$$

where p is the probability that a link is busy and $q = 1 - p$ is the probability that it is idle. All of the interstage links are equally loaded because there is no concentration or expansion in the stages.

APPENDIX B

ENCODING/DECODING ALGORITHMS FOR SEGMENTED PCM

B.1 EIGHT-BIT μ255 CODE

The encoded representations of μ255 PCM codewords use a sign–magnitude format wherein 1 bit identifies the sample polarity and the remaining bits specify the magnitude of the sample. The 7 magnitude bits are conveniently partitioned into a 3-bit segment identifier S and a 4-bit quantizing step identifier Q. Thus the basic structure of an 8-bit μ255 PCM codeword is shown in Figure B.1.

In the following descriptions of encoding and decoding algorithms, it is assumed, for convenience in using integer representations, that analog input signals are scaled to a maximum amplitude of 8159. Furthermore, all amplitudes and segment identifiers are assumed to be encoded using conventional binary representations. The actual encoders used in T1 transmission systems, however, complement the codewords to increase the density of 1's in a transmitted bit stream.

B.1.1 Algorithm 1: Direct Encoding (Table B.1)

$$\text{Polarity bit } P = \begin{cases} 0 & \text{for positive sample values} \\ 1 & \text{for negative sample values} \end{cases}$$

Given a sample value with a magnitude x, the first step in the magnitude encoding process is to determine the segment identifier S. The major segments are identified by the segment endpoints: 31, 95, 223, 479, 991, 2015, 4063, and 8159. Thus S can be

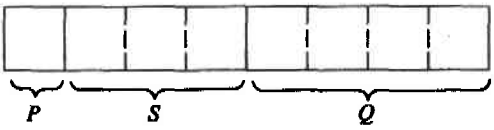


Figure B.1. Eight-bit μ225 PCM code format.

TABLE B.1 Piecewise Linear Approximation to μ255 Compounding[a]

Quantization Endpoints by Segment Code S								Quantization Code Q	
000	001	010	011	100	101	110	111		
0	31	95	223	479	991	2015	4063	0000	0
1	35	103	239	511	1055	2143	4319	0001	1
3	39	111	255	543	1119	2271	4575	0010	2
5	43	119	271	575	1183	2399	4831	0011	3
7	47	127	287	607	1247	2527	5087	0100	4
9	51	135	303	639	1311	2655	5343	0101	5
11	55	143	319	671	1375	2783	5599	0110	6
13	59	151	335	703	1439	2911	5855	0111	7
15	63	159	351	735	1503	3039	6111	1000	8
17	67	167	367	767	1567	3167	6367	1001	9
19	71	175	383	799	1631	3295	6623	1010	10
21	75	183	399	831	1695	3423	6879	1011	11
23	79	191	415	863	1759	3551	7135	1100	12
25	83	199	431	895	1823	3679	7391	1101	13
27	87	207	447	927	1887	3807	7647	1110	14
29	91	215	463	959	1951	3935	7903	1111	15
31	95	223	479	991	2015	4063	8159		

[a]Sample values are referenced to a full-scale value of 8159. Negative samples are encoded in sign–magnitude format with a polarity bit of 1. In actual transmission the codes are inverted to increase the density of 1's when low signal amplitudes are encoded. Analog output samples are decoded as the center of the encoded quantization interval. Quantization error is the difference between the reconstructed output value and the original input sample value.

determined as the smallest endpoint that is greater than the sample value x. Here, S is equal to the smallest a such that

$$x < 64 \cdot 2^a - 33 \qquad a = 0, 1, \ldots, 7$$

After the major segment containing the sample value has been determined, the particular quantization interval within the major segment must be identified. As a first step a residue r is determined as the difference between the input amplitude and the lower endpoint of the segment:

$$r = \begin{cases} x & S = 0 \\ x - (32 \cdot 2^S - 33) & S = 1, 2, \ldots, 7 \end{cases}$$

The value of Q can now be determined as the quantization interval containing the residue r. Here, Q is equal to the smallest b such that

$$r < \begin{cases} 2b + 1 & S = 0 \\ (2^{S+1})(b + 1) & S = 1, 2, \ldots, 7 \end{cases}$$

where $b = 0, 1, \ldots, 15$. Notice that this process identifies quantization intervals in segment $S = 0$ as having upper endpoints at $1, 3, 5, \ldots, 31$ while the other segments

have quantization endpoints that are multiples of 4, 8, 16, 32, 64, 128, 256 for $S = 1$, 2, 3, 4, 5, 6, 7, respectively.

After S and Q have been determined, they are binary encoded into 3 and 4 bits, respectively. The concatenation of S and Q produces a 7-bit word that can be conveniently represented as an integer between 0 and 127. In essence, this integer identifies one of the 128 quantization intervals of a compressed signal amplitude.

The decoding process involves assigning the designated polarity to an analog output sample at the midpoint of the nth quantization interval $n = 0, 1, \ldots, 127$. Using the values of S and Q directly, we can determine a discrete output sample value as

$$y_n = (2Q + 33)(2^S) - 33$$

where n is the integer obtained by concatenating the binary representations of S and Q.

Example B.1.

An input sample of +242 produces the following codeword:

$$0,\ 3,\ 1 = \boxed{0}\boxed{011}\boxed{0001}$$

The decoder output becomes

$$y_{49} = (2 \cdot 1 + 33)(2^3) - 33$$

$$= 247$$

which is the midpoint of the forty-ninth quantization interval from 239 to 255.

B.1.2 Algorithm 2: Linear Code Conversion

The fundamental reason for using a μ-law compounding characteristic with μ255 is the ease with which the segmented approximation can be digitally converted to and from a uniform code. This section describes the basic algorithms that implement the conversions. The first algorithm provides the means of implementing a μ255 PCM encoder using a 13-bit uniform encoder followed by digital logic to provide the compression function. The second algorithm indicates how to implement the decoder function of first expanding a compressed code into a 13-bit linear code to be used in generating the output samples.

Just as in algorithm 1, the polarity bit P is determined as

$$P = \begin{cases} 0 & \text{for positive sample values} \\ 1 & \text{for negative sample values} \end{cases}$$

The simplicity of converting from a linear code to a compressed code is most evident if the linear code is biased by adding the value 33 to the magnitude of all samples. Notice that this bias shifts the encoding range from 0–8159 to 33–8192. The addition process can be performed directly on the analog samples before encoding or with digital logic after encoding. In either case, the general form of all biased linear code patterns and the corresponding compressed codes are as follows:

μ255 Encoding Table

Biased Linear Input Code													Compressed Code						
0	0	0	0	0	0	0	1	w	x	y	z	a	0	0	0	w	x	y	z
0	0	0	0	0	0	1	w	x	y	z	a	b	0	0	1	w	x	y	z
0	0	0	0	0	1	w	x	y	z	a	b	c	0	1	0	w	x	y	z
0	0	0	0	1	w	x	y	z	a	b	c	d	0	1	1	w	x	y	z
0	0	0	1	w	x	y	z	a	b	c	d	e	1	0	0	w	x	y	z
0	0	1	w	x	y	z	a	b	c	d	e	f	1	0	1	w	x	y	z
0	1	w	x	y	z	a	b	c	d	e	f	g	1	1	0	w	x	y	z
1	w	x	y	z	a	b	c	d	e	f	g	h	1	1	1	w	x	y	z

From the foregoing table it can be seen that all biased linear codes have a leading 1 that indicates the value of the segment number S. Specifically, the value of S is equal to 7 minus the number of leading 0's before the 1. The value of Q is directly available as the 4 bits (w, x, y, z) immediately following the leading 1. All trailing bits (a–h) are merely ignored when generating a compressed code.

In reverse fashion the following table indicates how to generate a biased linear code from a compressed code. An unbiased output can be obtained by subtracting 33 from the biased code:

μ255 Decoding Table

Compressed Code							Biased Linear Output Code												
0	0	0	w	x	y	z	0	0	0	0	0	0	0	1	w	x	y	z	1
0	0	1	w	x	y	z	0	0	0	0	0	0	1	w	x	y	z	1	0
0	1	0	w	x	y	z	0	0	0	0	0	1	w	x	y	z	1	0	0
0	1	1	w	x	y	z	0	0	0	0	1	w	x	y	z	1	0	0	0
1	0	0	w	x	y	z	0	0	0	1	w	x	y	z	1	0	0	0	0
1	0	1	w	x	y	z	0	0	1	w	x	y	z	1	0	0	0	0	0
1	1	0	w	x	y	z	0	1	w	x	y	z	1	0	0	0	0	0	0
1	1	1	w	x	y	z	1	w	x	y	z	1	0	0	0	0	0	0	0

Both of these tables indicate that 13 bits of the linear code are used to represent the magnitude of the signal. In Chapter 3 it is mentioned that a μ255 PCM coder has an amplitude range equivalent to 12 bits. The discrepancy occurs because the first quantization interval has length 1 while all others in the first segment are of length 2. Thus the extra bit is needed only to specify the first quantization interval. Notice further that the least significant bit in the tables carries no information but is included only to facilitate the integer relationships. In particular, the least significant bit in the encoding table is completely ignored when determining a compressed code (assuming that the

bias is added to the analog sample). Furthermore, the least significant bit in the output codes is completely specified by the segment number S. It is a 0 for segment zero and a 1 for all other segments.

Example B.2.
An input code word of +242 is biased to produce a value of 275. The binary representation of 275 is

0 0 0 0 1 0 0 0 1 0 0 1 1 (biased linear code)

Hence from the encoding table $S = 011$ and wxyz = 0001, and the compressed code is

0	011	0001

(compressed code)

Using the decoding table, this compressed code produces the following biased linear output code:

0 0 0 0 1 0 0 0 1 1 0 0 0 (biased linear output)

The decimal representation of the foregoing code is 280, which corresponds to an unbiased output equal to +247.

B.2 EIGHT BIT *A*-LAW CODE

The following compounding algorithms for segmented A-law codes use the same basic procedures as those presented for the μ255 codes. One difference that does occur, however, involves the elimination of a bias in the linear code for conversion to and from a compressed code. Another difference occurs in the use of the integer 4096 as the maximum amplitude of a sample in an A-law representation. If desired, the scale factors for the two systems can be brought into close agreement by doubling the A-law scale to 8192.

As mentioned in Chapter 3, the segmented A-law code is usually referred to as a 13-segment code owing to the existence of seven positive and seven negative segments with the two segments near the origin being colinear. In the following descriptions, however, the first segment for each polarity is divided into two parts to produce eight positive and eight negative segments. This point of view permits a code format that is identical to the μ255 code format. Thus, a compressed codeword consists of a sign bit P followed by 3 bits of a segment identifier S and 4 bits of quantizer level Q.

B.2.1 Algorithm 1: Direct Encoding

The segment endpoints of an A-law code are 32, 64, 128, 256, 512, 1024, 2048, and 4096. Thus for a sample with magnitude x the major segment identifier S can be determined as the smallest a such that

TABLE B.2 Segmented *A*-Law Encoding Table

Quantization Endpoints by Segment Code								Quantization Code	
000	001	010	011	100	101	110	111		
0	32	64	128	256	512	1024	2048		
2	34	68	136	272	544	1088	2176	0000	0
4	36	72	144	288	576	1152	2304	0001	1
6	38	76	152	304	608	1216	2432	0010	2
8	40	80	160	320	640	1280	2560	0011	3
10	42	84	168	336	672	1344	2688	0100	4
12	44	88	176	352	704	1408	2816	6101	5
14	46	92	184	368	736	1472	2944	0110	6
16	48	96	192	384	768	1536	3072	0111	7
18	50	100	200	400	800	1600	3200	1000	8
20	52	104	208	416	832	1664	3328	1001	9
22	54	108	216	432	864	1728	3456	1010	10
24	56	112	224	448	896	1792	3584	1011	11
26	58	116	232	464	928	1856	3712	1100	12
28	60	120	240	480	960	1920	3840	1101	13
30	62	124	248	496	992	1984	3968	1110	14
32	64	128	256	512	1024	2048	4096	1111	15

Note: Every other bit is inverted for transmission.

$$x < 32 \cdot 2^a \qquad a = 0, 1, \ldots, 7$$

After S has been determined, the residue r can be obtained as

$$r = \begin{cases} x & S = 0 \\ x - 16 \cdot 2s & S = 1, 2, \ldots, 7 \end{cases}$$

The value of Q can then be determined as the smallest b such that

$$r < \begin{cases} 2(b+1) & S = 0 \\ 2^S(b+1) & S = 1, 2, \ldots, 7 \end{cases}$$

Just as in the case for μ255 coding, an A-law magnitude can be represented as an integer $n = 0, 1, \ldots, 127$ derived from the concatenation of 3 S-bits and 4 Q-bits. An output magnitude y^n can then be expressed as

$$y_n = \begin{cases} 2Q + 1 & S = 0 \\ 2^S(Q + 16\tfrac{1}{2}) & S = 1, 2, \ldots, 7 \end{cases}$$

Example B.3.

An input sample of +121 produces the following codeword:

0	010	1110

= +46 in decimal

The decoder output becomes

$$y_{46} = 2^S(14 + 16\tfrac{1}{2})$$

$$= 122$$

which is the midpoint of the quantization interval from 120 to 124.

B.2.2 Algorithm 2: Linear Code Conversion

The following tables indicate how to convert a 12-bit linear code directly into a compressed *A*-law code. The algorithm is basically the same as for the μ255 conversion except that biasing the linear code is unnecessary and a first segment code does not have a leading 1. Thus the segment number S can be determined as 7 minus the number of leading zeros as before. The Q field data are obtained as the 4 bits (wxyz) immediately following the leading 1, except when $S = 0$, in which case the Q field is contained in the 4 bits following the seven leading 0's.

A-Law Encoding Table

Linear Code												Compressed Code						
0	0	0	0	0	0	0	w	x	y	z	a	0	0	0	w	x	y	z
0	0	0	0	0	0	1	w	x	y	z	a	0	0	1	w	x	y	z
0	0	0	0	0	1	w	x	y	z	a	b	0	1	0	w	x	y	z
0	0	0	0	1	w	x	y	z	a	b	c	0	1	1	w	x	y	z
0	0	0	1	w	x	y	z	a	b	c	d	1	0	0	w	x	y	z
0	0	1	w	x	y	z	a	b	c	d	e	1	0	1	w	x	y	z
0	1	w	x	y	z	a	b	c	d	e	f	1	1	0	w	x	y	z
1	w	x	y	z	a	b	c	d	e	f	g	1	1	1	w	x	y	z

The following table provides the means of generating a linear codeword directly from a compressed codeword. The output value corresponds to the middle of the quantization interval designated by S and Q.

A-Law Decoding Table

Compressed Code							Linear Output Code											
0	0	0	w	x	y	z	0	0	0	0	0	0	0	w	x	y	z	1
0	0	1	w	x	y	z	0	0	0	0	0	0	1	w	x	y	z	1
0	1	0	w	x	y	z	0	0	0	0	0	1	w	x	y	z	1	0
0	1	1	w	x	y	z	0	0	0	0	1	w	x	y	z	1	0	0
1	0	0	w	x	y	z	0	0	0	1	w	x	y	z	1	0	0	0
1	0	1	w	x	y	z	0	0	1	w	x	y	z	1	0	0	0	0
1	1	0	w	x	y	z	0	1	w	x	y	z	1	0	0	0	0	0
1	1	1	w	x	y	z	1	w	x	y	z	1	0	0	0	0	0	0

Each of these tables relates 12 bits of magnitude in a linear code to a compressed code with 7 bits of magnitude. Notice, however, that the least significant bit of the encoder is always ignored. Thus the encoder needs only 11 bits of resolution if all of its outputs are immediately compressed. If any signal processing (such as adding two signals together) is to take place before compression, however, the extra bit is useful in reducing the composite quantization error.

Example

The previously used sample value 121 is represented in binary form as 000001111001. From the encoding table $S = 010$ and $Q = 1110$. Thus the compressed code word is

0	010	1110

Using the decoding table, the linear output can be determined as 000001111010, which is 122 decimal.

APPENDIX C

ANALYTIC FUNDAMENTALS OF DIGITAL TRANSMISSION

C.1 PULSE SPECTRA

This section presents the frequency spectra of common pulse waveforms used for digital transmission. These are square pulses as generated at a source. Since the spectra of square pulses have infinite frequency content, the spectra presented here do not correspond to pulse responses at the output of a channel where the pulse shapes have been altered by bandlimiting filters. In the next section channel output pulse responses are described. Then the necessary combinations of input pulse shapes and filter designs to produce particular output pulses are considered.

The various pulse shapes and corresponding frequency spectra are presented in Figure C.1. In deriving the spectra, the following conditions and assumptions were made:

1. All pulses have equal energy.
2. All systems signal at rate $1/T$.
3. The waveforms shown are used to encode a 1.
4. Opposite polarities are used for a 0.
5. It is equally likely for 1's and 0's to occur and to occur at random (complete independence).

C.2 CHANNEL OUTPUT PULSE RESPONSES

Although a digital transmission system can be designed to produce a variety of output pulse responses, the most common is defined as

$$y_{rc}(t) = \frac{\sin(\pi t/T)}{\pi t/T} \frac{\cos(\alpha \pi t/T)}{1-(2\alpha t/T)^2} \tag{C.1}$$

where $1/T$ is the signaling rate and α is an excess bandwidth factor between 0 and 1. Equation C.1 represents the response of a raised-cosine channel—so-called because the frequency spectrum corresponding to $y_{rc}(t)$ in Equation C.1 is

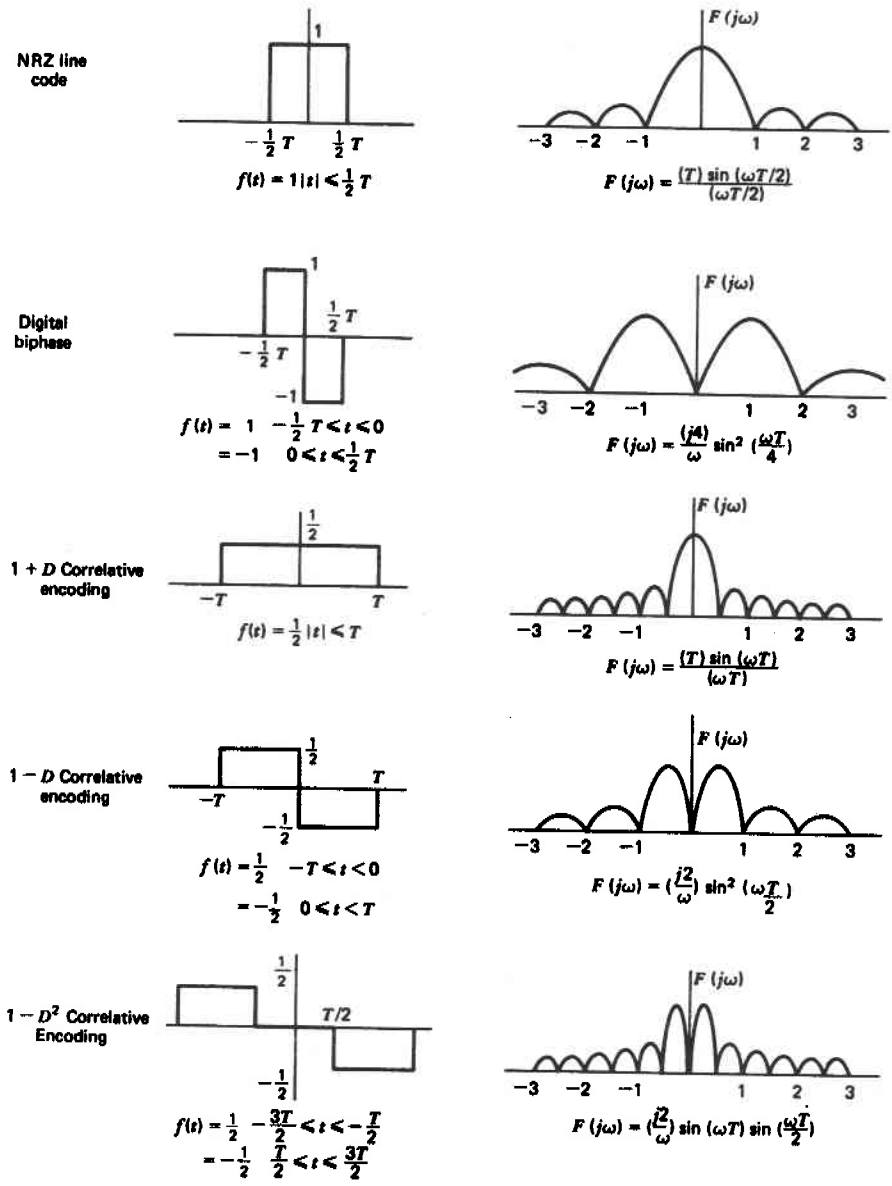


Figure C.1. Spectra of common pulse shapes.

$$
\begin{aligned}
Y_{rc}(f) &= 1 && |f| \le \frac{1-\alpha}{2T} \\
&= \frac{1}{2}\left[1+\cos\left(\frac{\pi|f|T}{\alpha}-\frac{\pi(1-\alpha)}{2\alpha}\right)\right] && \frac{1-\alpha}{2T} \le |f| \le \frac{1+\alpha}{2T} \\
&= \cos^2\left(\frac{\pi|f|T}{2\alpha}-\frac{\pi(1-\alpha)}{4\alpha}\right) \\
&= 0 && \text{otherwise}
\end{aligned}
\tag{C.2}
$$

The origin of the apellation "raised cosine" is apparent in the third line of Equation C.2.

The parameter α in Equations C.1 and C.2 is referred to as an excess-bandwidth parameter. If $\alpha = 0$, the spectrum defined in Equation C.2 is exactly equal to the theoretical minimum bandwidth $1/2T$ for signaling rate $1/T$. As α increases from 0 to 1, the excess spectrum width increases to 100%. Raised-cosine channel spectra are illustrated in Figure C.2 for several values of α.

Practical systems are typically designed for excess bandwidths of 30% or more for several reasons. First, "brick wall" filters needed to produce the infinite attenuation slopes implied by $\alpha = 0$ are physically unrealizable. Second, as shown in Figure C.3, the time-domain pulse response for small values of α exhibits large amounts of ringing. Slight errors in the sample times cause significant degradations in performance due to intersymbol interference. Third a slight deviation in the signaling rate from the design rate also produces significant intersymbol interference.

It must be emphasized that Equation C.2 defines the desired spectrum at the output of the channel (the input to the decision circuit). Thus the desired response results from a combination of the input pulse spectrum and the channel filter responses. Often, the input pulse spectrum arises from a square pulse of duration T:

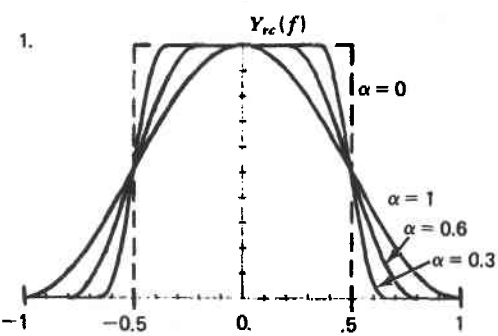


Figure C.2. Raised-cosine spectrums for various values of α.

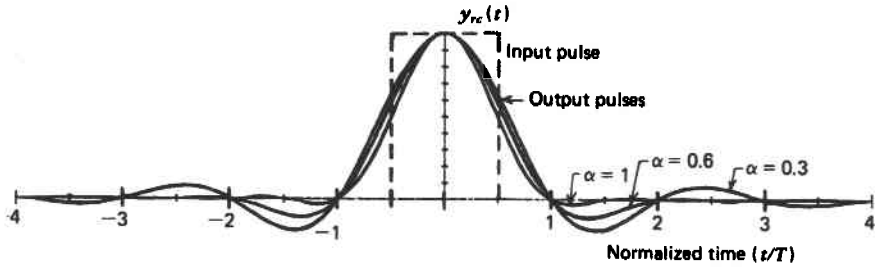


Figure C.3. Raised-cosine pulse responses for various values of α.

$$x_s(t) = \begin{cases} 1 & |t| \le \frac{1}{2}T \\ 0 & \text{otherwise} \end{cases} \tag{C.3}$$

The frequency spectrum corresponding to $x_s(t)$ in Equation C.3 is the "sin(x)/x" spectrum also referred to as a "sine" function:

$$X_s(f) = \frac{(T)\sin(\pi f T)}{\pi f T}$$

$$= T\,\mathrm{sinc}(\pi f T) \tag{C.4}$$

When the channel input spectrum is as defined in Equation C.4, the filter function of a channel to produce a raised-cosine output is determined as

$$H(f) = \frac{Y_{rc}(f)}{X_s(f)} \tag{C.5}$$

The channel filter functions implied in Equation C.5 are shown in Figure C.4 for the same values of α shown in Figure C.2. Figure C.2 displays channel output spectra while Figure C.4 displays corresponding frequency-domain transfer functions for the channel.

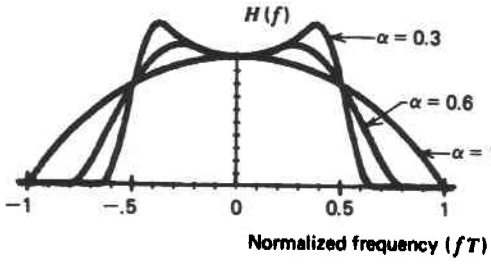


Figure C.4. Channel filter responses needed to produce raised-cosine outputs when excited by $\sin(\omega T/2)/(\omega T/2)$ pulses.

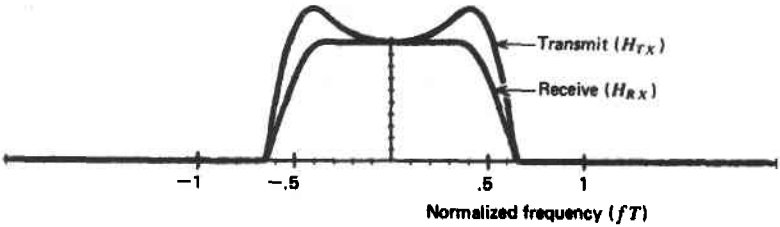


Figure C.5. "Optimum" transmit and receive filter functions for raised-cosine response with $\alpha = 0.3$ and $\sin(\omega T/2)/(\omega T/2)$ excitation.

As mentioned in Chapter 3, the design of a smoothing filter in a digital voice decoder sometimes requires a modification like that defined in Equation C.5. In the case mentioned, the ideal "flat" filter for reconstructing speech from narrow impulselike samples should be modified by dividing the flat response by the spectrum of the finite-width samples (Equation C.4). When the samples are narrower than the signaling interval, the $\sin(x)/x$ response is essentially flat across the bandwidth of the filter. However, when the samples are made to last for the entire duration of the sample interval, the filter response should be "peaked" to compensate for $\sin(x)/x$ roll-off.

As discussed in Chapter 6, a channel filter function is usually partitioned between the transmitter and the receiver. The so-called "optimum" filter partitioning occurs when the receive filter response is the square root of the desired output response $Y_{rc}(f)$, and the transmit filter response is whatever is necessary to transform the channel input spectrum into the complex conjugate of the receive filter response. (The output spectrum of the transmitter is also the square root of the desired channel response.) When the channel input is a pulse of duration T, the "optimum" filter functions are defined as

$$H_{RX}(f) = \{Y^{rc}(f)\}^{1/2} \tag{C.6}$$

$$H_{TX}(f) = \frac{[Y_{rc}(f)]^{1/2}}{X_s(f)} \tag{C.7}$$

Notice that the transmit spectrum resulting from $H_{TX}(f)$ in Equation C.7 is equal to the square root of the desired output response, no matter what the input pulse shape is. Hence, when optimum partitioning is used, the transmit spectrum and the receiver design are independent of the channel excitation.

The transmit and receive filter functions for square-wave excitations (Equations C.6 and C.7) are shown in Figure C.5.* Notice that the transmit filter function has peaking at frequencies other than direct current. The implication for midband attenuation is one of the reasons why "optimum" partitioning may not be optimum in a system with source power limitations.

*This discussion assumes baseband transmission. The principles are easily extended to carrier-based systems by translating the filter functions to the carrier frequency.

C.2.1 Optimum Filtering for Minimum Shift Keying

As discussed in Chapter 6, minimum shift keyed (MSK) modulation can be represented as quadrature channel modulation with baseband excitation defined as cosine pulse shapes:

$$x_c(t) = \begin{cases} \cos\left(\dfrac{\pi t}{T}\right) & |t| \le \frac{1}{2}T \\ 0 & \text{otherwise} \end{cases} \tag{C.8}$$

The transform of $x_c(t)$ is

$$X_c(f) = \left(\frac{\pi}{2T}\right)\frac{\cos \pi f T}{1-(2fT)^2} \tag{C.9}$$

When a raised-cosine output response and optimum filter partitioning is desired, the MSK filter functions are

$$H_{RX}(f) = [Y_{rc}(f)]^{1/2} \tag{C.10}$$

$$H'_{TX}(f) = \frac{[Y_{rc}(f)]^{1/2}}{X_c(f)} \tag{C.11}$$

The transmit spectrum and the receiver of an optimally partitioned MSK system are identical to counterparts in an optimally partitioned, offset-keyed 4-PSK system!

C.2.2 Partial-Response Systems

As another digital transmission system design example, consider a $1 + D$ partial-response system. The desired time response of the channel is defined in Equation C.12 and the associated frequency spectrum in Equation C.13. The pulse response of a $1 + D$ partial-response channel is shown in Figure C.6. Notice that a single pulse contributes equally to the response at two successive sample times but crosses zero at all other sample times:

$$y_c(t) = \frac{4\cos(\pi t/T)}{\pi[1-(2t/T)^2]} \tag{C.12}$$

$$Y_c(f) = \begin{cases} \cos \pi f T & |f| \le \dfrac{1}{2T} \\ 0 & \text{otherwise} \end{cases} \tag{C.13}$$

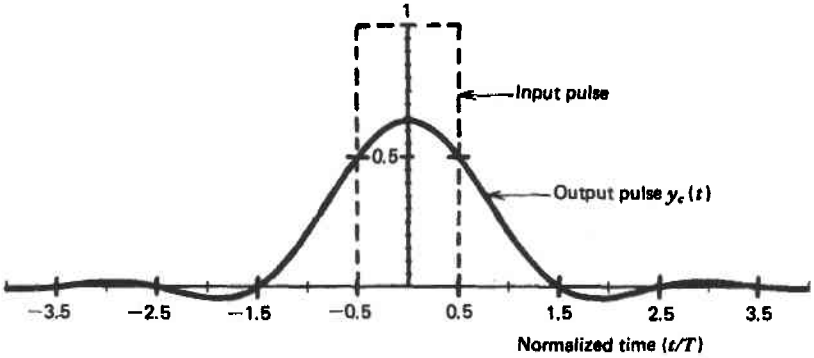


Figure C.6. Output pulse of $1 + D$ partial-response channel.

Optimum filter partitioning is again achieved with a receive filter having an amplitude response equal to the square root of the desired output response $Y_c(f)$. If the channel is excited by square pulses of duration T, as defined in Equation C.3, the optimum filter functions are

$$H_{RX}(f) = [Y_c(f)]^{1/2} \qquad |f| \le \frac{1}{2T} \tag{C.14}$$

$$H_{TX}(f) = \frac{[Y_c(f)]^{1/2}}{X_s(f)} \qquad |f| \le \frac{1}{2T} \tag{C.15}$$

where $X_s(f)$ is defined in Equation C.4.

The optimum filter functions defined in Equations C.14 and C.15 are shown in Figure C.7 along with the desired output response of the channel. Notice that, unlike the full-response (raised-cosine) systems, optimum partitioning of a partial-response system does not require peaking of the transmit filter.

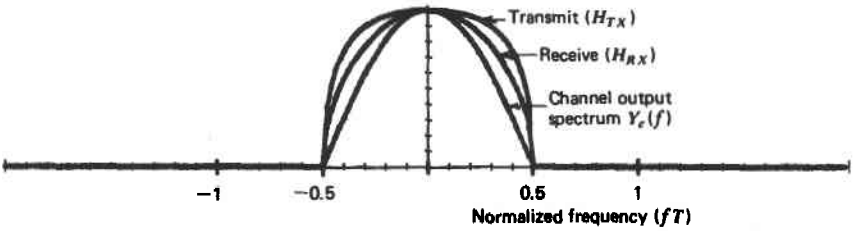


Figure C.7. Spectra of $1 + D$ PRS channel and "optimum" filter responses for $\sin(\omega T/2)/(\omega T/2)$ excitation.

C.3 ERROR RATE ANALYSES: BASEBAND SYSTEMS

C.3.1 Binary Transmission

Consider the receiver model of the digital transmission system shown in Figure C.8. The receiver consists of two parts: signal processing circuitry and a data detector (decision circuit). For the time being, assume that the output of the signal processing circuit produces a pulse of amplitude $+V$ when a 1 is transmitted and a pulse of amplitude $-V$ when a 0 is transmitted. Obviously, the detector merely examines the polarity of its input at the sample times defined by the sample clock. A decision error occurs if noise at the sample times has an amplitude greater than V and a polarity opposite to the transmitted pulse.

The most commonly analyzed type of noise is assumed to have a Gaussian or normal probability distribution. Thus the probability of error P_E can be determined as

$$P_E = \frac{1}{\sqrt{2\pi}\,\sigma} \int_V^{\infty} e^{-t^2/2\sigma^2}\, dt \tag{C.16}$$

where σ^2 is the rms noise power at the detector. Using the error function,

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt \tag{C.17}$$

Equation C.16 is sometimes rewritten as

$$P_E = \tfrac{1}{2}(1 - \operatorname{erf} z) \tag{C.18}$$

where $z = V/\sqrt{2}\,\sigma$. The error probability P_E can also be expressed in terms of the complimentary error function:

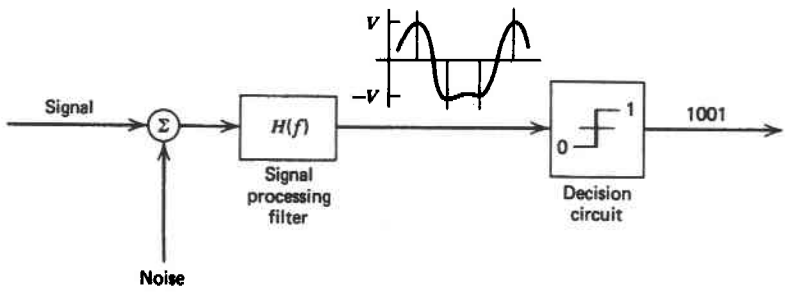


Figure C.8. Digital receiver model.

$$P_E = \tfrac{1}{2}\,\mathrm{erfc}\, z \tag{C.19}$$

where

$$\mathrm{erfc}\, z = 1 - \mathrm{erf}\, z \qquad z = \frac{V}{\sqrt{2}\,\sigma}$$

In lieu of evaluating the integral in Equation C.17 or C.19 (which has no closed form solution), the error functions can be approximated as

$$\mathrm{erfc}\, z \doteq \frac{e^{-z^2}}{(z\sqrt{\pi})} \qquad (z >> 1) \tag{C.20}$$

C.3.2 Multilevel Transmission

The error rate of a multilevel baseband system is easily determined by an appropriate reduction in the error distance. If the maximum amplitude is V, the error distance d between equally spaced levels at the detector is

$$d = \frac{V}{L-1} \tag{C.21}$$

where L is the number of levels. Adjusting the error distance V of a binary system to that defined in Equation C.21 provides the error rate of a multilevel system as

$$P_E = \left(\frac{1}{\log_2 L}\right)\left(\frac{L-1}{L}\right)\mathrm{erfc}\left(\frac{V}{(L-1)\sqrt{2}\,\sigma}\right) \tag{C.22}$$

where the factor $(L-1)/L$ reflects the fact that interior signal levels are vulnerable to both positive and negative noise and the factor $1/\log_2 L$ arises because the multilevel system is assumed to be coded so symbol errors produce single-bit errors ($\log_2 L$ is the number of bits per symbol).

Equation C.22 reveals that, with respect to peak signal-to-noise ratios at the detector, L-level transmission incurs a penalty of $20 \log_{10}(L-1)$ decibels. If V is increased by a factor $L-1$, the error rate of the L-level system is identical to the error rate of the binary system [except for the factors $1/\log_2 L$ and $(L-1)/L$, which typically represent only a few tenths of a decibel].

Equation C.22 relates error rate to the peak signal power V^2. To determine the error rate with respect to average power, the average power of an L-level system is determined by averaging the power associated with the various pulse amplitude levels:

$$(V^2)_{\text{avg}} = \frac{2}{L}\left[\left(\frac{V}{L-1}\right)^2 + \left(\frac{3V}{L-1}\right)^2 + \cdots + V^2\right]$$

$$= \frac{2V^2}{L(L-1)^2}\sum_{i=1}^{L/2}(2i-1)^2 \tag{C.23}$$

where the levels

$$\frac{V}{L-1}\{\pm 1, \pm 3, \ldots, \pm(L-1)\}$$

are assumed to be equally likely.

C.3.3 Energy-per-Bit-to-Noise-Density Ratios

The foregoing error rate equations relate P_E to the signal energy at the sample times and the rms noise power at the detector. When comparing various digital modulation formats, multilevel systems in particular, it is more relevant to relate error performance to signal power and noise power at the input to the receiver (in front of the signal processing circuitry). As a first step in developing an error rate equation based on signal-to-noise ratios at the receiver input, the noise power at the detector is determined.

The variance of the noise σ^2 used in the previous equations is exactly equal to the rms power that would be measured at the detector in the absence of a signal. This noise power can be determined analytically as

$$\sigma^2 = \int_{-\infty}^{\infty} |H(f)(\tfrac{1}{2}N_0)|^2\, df$$

$$= N_0 \int_{0}^{\infty} |H(f)|^2\, df \tag{C.24}$$

$$= N_0 \cdot B_N \tag{C.25}$$

where N_0 is the one-sided noise power spectral density in watts per hertz and $B_N = \int_0^\infty = |H(f)|^2\, df$ is the noise-equivalent bandwidth or simply the noise bandwidth of the receiver filter function $H(f)$.

In Equation C.25, the noise source is assumed to be white. That is, a uniform spectral density exists across the entire band of interest. This noise may exist in the trans-

mission medium itself, or it may occur in the "front-end" amplifier of the receiver. If the rms noise power passing through a square (ideal) filter of bandwidth B is measured, a reading of $(N_0)(B)$ watts would be obtained. Thus B_N represents the bandwidth of a perfectly square filter that passes the same amount of noise as the receiver filter $H(f)$ [$H(f)$ may be decidedly nonsquare in its amplitude response].

The receiver function $H(f)$ necessarily provides a compromise between two conflicting objectives. First, it must minimize the amount of noise passed to the detector (i.e., minimize B_N). Second, the difference between sample values ($+V$ and $-V$) must be maximized. Obviously, the signal-to-noise ratio at the detector (V^2/σ^2) should be maximized to minimize the error probability. A classical result of digital communication theory states that V/σ is maximized when $H(f)$ is "matched" to the received signal.

When viewed in the time domain, a "matched filter" is implemented by correlating (multiplying) the received signal with each of the receivable (noise-free) pulse shapes. The outputs of the correlators are integrated across a signal interval to determine the overall average correlation during the interval. The integrator with the maximum output indicates the most likely symbol to have been transmitted.

In most systems all signals or pulses have the same shape but differ only in amplitude and polarity. Thus a single matched filter can be used. Detection merely involves comparing the matched filter output to appropriate decision levels. The output of a single correlator $h(t)$ and its integrator is expressed as

$$V = \int_0^T s(t)h(t)\,dt$$

$$= \int_0^T |s(t)|^2\,dt \tag{C.26}$$

where $s(t)$ is the signal or pulse shape being measured. Notice that V is, in essence, a measure of the energy in the signal over the signaling interval 7.

When analyzed in the frequency domain, a matched filter response $H(f)$ is the complex conjugate of the channel pulse spectrum $S(f)$. Thus the matched filter output $Y(f)$ can be expressed in the frequency domain as

$$Y(f) = H(f) \cdot S(f)$$

$$= S^*(f) \cdot S(f) \tag{C.27}$$

Frequency-domain representations are most convenient when the transmit signal is strictly bandlimited, implying that the duration of the pulse response is theoretically unlimited. In this case the energy in a pulse (e.g., raised cosine) is directly proportional

to the detector voltage at the optimum sample time. Hence optimum detection is achieved by merely sampling the output of the receive filter at the proper time.

Using the relationship of Equation C.25 and the parameter E_S to represent symbol energy, we express the binary error rate Equation C.19 as

$$P_E = \tfrac{1}{2}\,\mathrm{erfc}(z) \tag{C.28}$$

where $z^2 = E_S/(N_0 B_N)$.

Notice that for a given system (fixed B_N), the error rate is dependent only on the ratio of the symbol energy E_S and the noise density N_0. This ratio is commonly referred to as a signal-to-noise ratio, although it is not a signal-power-to-noise-power measurement. Equation C.28 is the preferred form of the error rate equation for comparing different modulation schemes. In a binary scheme, the symbol energy E_S is equal to the energy per bit E_b.

As an example of a specific system, consider a baseband raised-cosine channel with optimum partitioning. The output spectrum of the transmitter is the square root of the raised-cosine spectrum $Y_{rc}(f)$ defined in Equation C.2. The matched receiver filter also has a square root of a raised-cosine response (Equation C.6). Hence the noise bandwidth of the receiver can be determined as

$$\begin{aligned} B_N &= \int_0^{\infty} |H(f)|^2\, df \\ &= \int_0^{(1+\alpha)/2T} |Y_{rc}(f)|\, df \\ &= \frac{1}{2T} \qquad \text{(independent of } \alpha) \end{aligned} \tag{C.29}$$

As defined in Equation C.1 and shown in Figure C.3, the (normalized) peak sample value at the detector of a raised-cosine pulse is 1. Using unnormalized pulses of amplitude E_S, we can determine the error rate of a binary ($+E_S$, $-E_S$) raised-cosine channel as

$$P_E = \tfrac{1}{2}\,\mathrm{erfc}(z) \tag{C.30}$$

where $z^2 = (E_S/N_0)T$ and T is the duration of a signal interval.

Although Equation C.30 was derived for a raised-cosine channel, it is more general in that it is applicable to any binary system using antipodal signaling. Thus Equation C.30 is plotted in Figure 4.23 as the best performance achievable by any digital trans-

mission system detecting one pulse at a time. (Lower error rates are possible when redundant signals or error-correcting codes are used.)

The error rate performance presented in Equation C.22 for multilevel systems is based on noise power at the detector. As long as the signal bandwidth is identical for all systems, Equation C.22 is valid since the noise bandwidth of the receiver is independent of the number of levels. When the signaling rate is held constant, however, the bit rate increases with the number of levels. To compare multilevel systems on the basis of a given data rate, the signaling interval T and, hence, the noise bandwidths must be adjusted accordingly.

If T is the signaling interval for a two-level system, the signaling interval T_L for an L-level system providing the same data rate is determined as

$$T_L = T \log_2 L \tag{C.31}$$

Using the noise bandwidth of a raised-cosine filter in Equation C.29, we extend Equation C.22 to multilevel systems as

$$P_E = \left(\frac{1}{\log_2 L}\right)\left(\frac{L-1}{L}\right)\operatorname{erfc}(z) \tag{C.32}$$

where

$$z = \frac{V/(L-1)}{(N_0/T_L)^{1/2}}$$

Equation C.32 can be simplified and presented in a more customary form by using the relationship that energy per symbol $E_s = E_b \log_2 L = V^2 T_L$, where E_b is the energy per bit:

$$P_E = \left(\frac{1}{\log_2 L}\right)\left(\frac{L-1}{L}\right)\operatorname{erfc}(z) \tag{C.33}$$

where

$$z = \frac{(\log_2 L)^{1/2}}{L-1}\left(\frac{E_b}{N_0}\right)^{1/2}$$

Equation C.33 is plotted in Figure 4.26 for 2, 4, 8, and 16 levels. These curves represent the ideal relative performances of multilevel baseband systems at a constant data rate. The bandwidth requirements of the higher-level systems decrease in proportion to $\log_2 L$.

Equation C.33 and Figure 4.26 represent the performance of multilevel systems with respect to E_b/N_0 (common data rate but different bandwidths). The following relationship can be used to determine error rates with respect to signal-to-noise ratios at the decision circuit (different data rates but common bandwidth):

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}}$$

$$= \frac{(E_b)(\log_2 L)(1/T_L)}{(N_0)(1/2T_L)}$$

$$= (2)(\log_2 L)\left(\frac{E_b}{N_0}\right) \tag{C.34}$$

where $1/2T_L$ is the minimum "Nyquist" bandwidth of the signal.

The SNR obtained in Equation C.34 represents the ratio of signal power at the sample time to noise at the detector. This ratio is sometimes referred to as a postdetection SNR because it is the signal-to-noise ratio at the output of the signal processing circuitry.

Some communications theorists use predetection signal-to-noise ratios in determining error rates. Since predetection SNRs are measured prior to bandlimiting the noise, a noise bandwidth must be hypothesized to establish a finite noise power. Commonly, a bit rate bandwidth ($1/T$) or a Nyquist bandwidth ($1/2T$) is specified. The latter specification produces SNRs identical to that in Equation C.34. Exceptions occur with double-sideband modulation using coherent demodulation (e.g., 2-PSK) where the predetection SNR is 3 dB higher than the postdetection SNR. (All signal power is coherent to the demodulator carrier reference, but only half of the noise is "coherent.")

C.3.4 Partial-Response Systems

The error rate equation for a $1 + D$ partial-response system is obtained by incorporating the following modifications to a full-response system:

1. The error distance is exactly one-half the error distance of a corresponding full-response system (see Figure C.6).
2. The noise bandwidth of the receiver (assuming square-root partitioning) is obtained by integrating H_{RX} of Equation C.14:

$$B_N = \int_0^{1/2T} \cos \pi f T \, df$$

$$= \frac{1}{\pi T} \tag{C.35}$$

Thus the noise bandwidth of a $1 + D$ partial-response system is 2 dB lower than the noise bandwidth of a raised-cosine (full-response) system obtained in Equation C.29. Since the error distance is reduced by 6 dB, partial-response filtering incurs a net penalty of 4 db—with respect to unfiltered signal power at the source. In terms of channel powers, the partial-response system incurs a smaller penalty owing to greater spectrum truncation in the transmit filter.

The difference in channel powers of the two systems can be obtained by integrating the respective channel power spectra. If square-root partitioning is used, the channel power of the partial-response system is 2 dB below the channel power of the corresponding full-response systems. In this case, the difference in channel powers is exactly equal to the difference in noise bandwidths because the receiver filter responses are matched to the respective channel spectra. That is, integration of channel spectra has essentially been accomplished in Equations C.29 and C.35.

When spectrum truncation loss in a transmitter is considered, a square-root partitioned partial-response system is only 2 dB worse than that of a corresponding full-response system. Of course, the partial-response system requires less bandwidth than the full-response system. For completeness, the error rate equation of a square-root partitioned $1 + D$ partial-response system is

$$P_E = \left(\frac{1}{\log_2 L}\right)\left(\frac{L-1}{L}\right)\operatorname{erfc}(z) \tag{C.36}$$

where $z = (\pi/4)(E_b/N_0)^{1/2}(\log_2 L)^{1/2}/(L-1)$ and E_b is the energy per bit on the channel.

Equation C.36 is identical to Equation C.33 except for the factor $\pi/4$ resulting from the lower error distance, the lower noise bandwidth, and the lower channel power of the partial-response system.

C.4 CARRIER TRANSMISSION SYSTEMS

C.4.1 Filter Design

Except for a few relatively uncommon frequency-modulated systems, digitally modulated carrier systems can be designed and analyzed with baseband-equivalent channels. Carrier-based filters are derived by translating the baseband filters to bandpass filters centered at the carrier frequency. The output pulse response of the channel is determined by the composite of the baseband-equivalent filters. Thus pulse shaping can be achieved by filtering the baseband signals or the modulated signals. Partitioning of the channel filter function is dependent on the application. In all cases the composite channel response is identical to that defined in Equation C.2 for raised-cosine channels or Equation C.13 for partial-response systems.

C.4.2 Error Rate Analysis

Error rate analyses of baseband systems can be directly applied to carrier systems—under one important condition: Coherent modulation and demodulation must be used. For example, coherent demodulation of a 2-PSK signal $y(t) = \cos[\omega t + \phi(t)]$ involves implementing the following equations:

$$\begin{aligned}
\phi(t) &= \text{lowpass}\{\cos[\omega t + \phi(t)]\, 2\cos(\omega t)\} \\
&= \text{lowpass}\{[(\cos\phi(t)\cos\omega t - \sin\phi(t)\sin\omega t]2\cos\omega t\} \\
&= \text{lowpass}\{\cos\phi(t)(1+\cos\omega t) - \sin\phi(t)\sin 2\omega t\} \\
&= \cos\phi(t) \\
&= \begin{cases} 1 & \text{for } \phi(t) = 0 \\ -1 & \text{for } \phi(t) = \pi \end{cases}
\end{aligned} \tag{C.37}$$

Notice that coherent demodulation involves multiplying the received signal by a local carrier that is exactly in phase with the respective incoming signal. Hence coherent demodulation is closely related to matched filter detection as presented in Equations C.26 and C.27. To complete optimum detection of a digitally modulated signal, the baseband equivalent of the receiver filter function must also be matched to the carrier envelope: $\cos\phi(t)$. When coherent demodulation (also called coherent detection) is used, the error rate performance is identical to the analogous baseband system. Hence the error rate of a coherently detected 2-PSK system is provided in Equation C.19 or C.28.

The correspondence of coherent demodulation to matched filter detection is illustrated in Figure C.9. For convenience, digital biphase (diphase) is compared to an NRZ baseband signal. The basic principle also applies to sine wave carriers at any frequency. The important point to notice in Figure C.9 is that the output of the coherent demodulator (or equivalently, the matched filter) is identical to the NRZ signal. Fur-

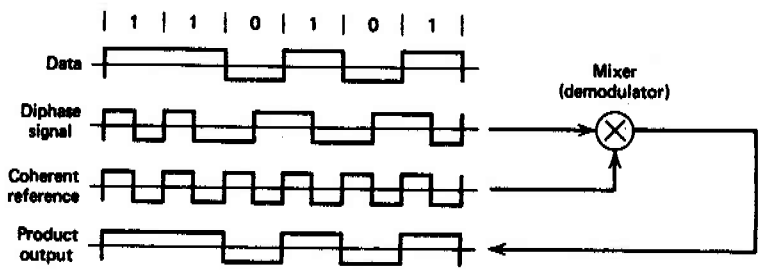


Figure C.9. Coherent demodulation (detection) of digital biphase signal.

thermore, the noise power density coming out of the coherent demodulator is identical to the noise power density of the baseband system. (Positive-weighted white noise is statistically no different than negative-weighted white noise.)

Figure C.9 demonstrates that coherently demodulated carrier-based signals produce the same signal-to-noise ratios at the detector as baseband systems, despite the fact that (double-sideband) carrier systems require twice as much bandwidth. Coherent demodulation leads to a receiver noise bandwidth equal to the baseband-equivalent noise bandwidth because only one-half of the noise power in the carrier signal bandwidth is passed by the coherent demodulator. (Noise in the carrier signal bandwidth that is out of phase with respect to the coherent reference is translated to twice the carrier frequency and therefore eliminated by the low-pass filter, $\sin \omega t \cdot \cos \omega t = \frac{1}{2} \sin 2\omega t$.)

C.4.3 QAM Error Rates

The error rate equation of a coherently detected QAM system is identical to the error rate of the corresponding multilevel baseband system applied individually to each quadrature channel. Thus the error performance of a 16-QAM system is provided in Equation C.33 for $L = 4$ levels. In a QAM system coherent demodulation causes one-half of the noise power in the carrier signal bandwidth to show up at the I-channel detector and one-half shows up at the Q-channel detector. Of course, the total signal power is divided in two so the signal-to-noise ratio at the individual detectors is identical to the carrier signal-to-noise ratio (i.e., predetection SNR is equal to postdetection SNR). The error rate performance of QAM systems is plotted in Figure 6.20 in terms of E_b/N_0. To relate those results to signal-to-noise ratios, use Equation C.34 using L as the number of levels on each quadrature channel.

C.4.4 PSK Error Rates

The error rate of a multilevel PSK system is derived most easily by using quadrature channel representations for the signals. For example, Figure C.10 displays the regions of decision errors for a representative signal phase in an 8-PSK signal. The received signal is processed by two orthogonal phase detectors to produce quadrature signals $Y_I(t)$ and $Y_Q(t)$ in Equations 6.10 and 6.11, respectively.

If the transmitted phase is $\pi/8$ (corresponding to data values 011), a decision error results if noise causes $Y_Q(t)$ to go positive at the sample time [$Y_Q(t)$ is positive downward to represent the Q (sine) channel as lagging the I (cosine) channel]. The normalized error distance is $\sin(\pi/8)$. A decision error also results if noise causes $-Y_Q(t)$ to exceed $Y_I(t)$, indicating the phase is greater than $-\pi/4$. This latter condition can be represented as a negative value for the transformed signal $Y_B(t) = 0.707Y_Q(t) + 0.707Y_Q(t)$. [Here, $Y_B(t)$ is a projection of the received signal onto a $-\pi/4$ basis vector (see Figure 6.13)]. Examination of Figure C.10 reveals that the error distance for this second type of error is also $\sin(\pi/8)$. Since the noise variance $Y_B(t)$ is identical to the

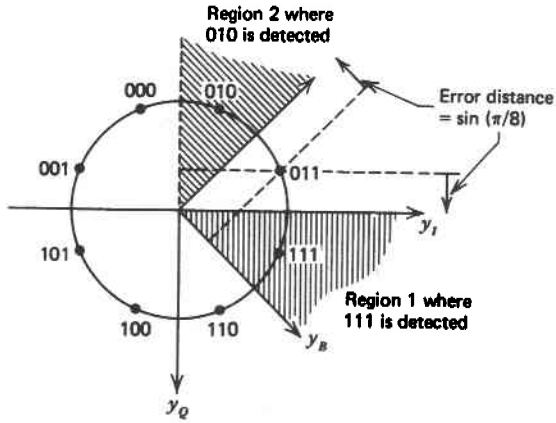


Figure C.10. Regions of decision error for 8-PSK signal at $\pi/8$ (011).

noise variance of $Y_I(t)$ (by virtue of the 0.707 multipliers), both types of errors are equally likely.

In general terms, the error distance of a PSK system with N phases is $V \cdot \sin(\pi/N)$, where V is the signal amplitude at a detector (i.e., the radius of a PSK signal constellation). A detection error occurs if noise of the proper polarity is present at the output of either of two phase detectors. A detection error, however, is assumed to produce only a single-bit error. The general expression for the theoretical error rate for PSK modulation is now determined by modifying Equation C.22 as

$$P_E = \left(\frac{1}{\log_2 N}\right) \operatorname{erfc}(z)$$

where

$$z = \frac{\sin(\pi/N)V}{\sqrt{2}\,\sigma} \tag{C.38}$$

The signal amplitude V can be expressed as

$$V = \left[E_b(\log_2 N)\left(\frac{1}{T}\right)\right]^{1/2} \tag{C.39}$$

and the rms noise voltage σ as

$$\sigma = \left[N_0\left(\frac{1}{2T}\right)\right]^{1/2} \tag{C.40}$$

for noise in a Nyquist bandwidth.

Combining Equations C.38, C.39, and C.40 relates PSK error rates to energy per bit to noise density on the channel:

$$P_E = \left(\frac{1}{\log_2 N}\right)\mathrm{erf}(z) \tag{C.41}$$

where

$$z = \sin\left(\frac{\pi}{N}\right)(\log_2 N)^{1/2}\left(\frac{E_b}{N_0}\right)^{1/2}$$

Equation C.41 is plotted in Figure 6.16 for PSK systems with various numbers of phases.* To determine error rates with respect to signal-power-to-noise-power ratios, use the following:

$$\mathrm{SNR} = \log_2 N\left(\frac{E_b}{N_0}\right) \qquad (N > 2) \tag{C.42}$$

*For 2-PSK systems, the error rates as specified in Equation C.38 or C.41 should be divided by 2 because only one phase detector is needed and it produces errors for one polarity of noise only.

APPENDIX D

TRAFFIC TABLES

Table D.1 is a table of maximum offered loads A for various blocking probabilities B and number of servers N. The blocking probabilities are for infinite sources with lost calls cleared (Erlang-B, Equation 10.8).

Table D.2 is a table of maximum offered loads A for various blocking probabilities B, number of servers N, and finite number of sources M. The offered load A is determined as $M\rho$, where ρ is the average source activity assuming no calls are cleared. The blocking probability for finite sources is determined from Equation 12.13.

Note: The following tables were obtained from *Telephone Traffic Theory, Tables, and Charts*, Siemens Aktiengesellschaft, Munich, 1970.

TABLE D.1 Maximum Offered Load Versus B and N[a]

N/B	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30	40
1	.0001	.0005	.001	.005	.010	.020	.053	.111	.176	.250	.429	.667
2	.014	.032	.046	.105	.153	.223	.381	.595	.796	1.00	1.45	2.00
3	.087	.152	.194	.340	.455	.602	.899	1.27	1.60	1.93	2.63	3.48
4	.235	.362	.439	.701	.869	1.09	1.62	2.05	2.50	2.95	3.89	5.02
5	.452	.649	.762	1.13	1.36	1.66	2.22	2.88	3.45	4.01	5.10	6.60
6	.728	.996	1.15	1.62	1.91	2.28	2.96	3.76	4.44	5.11	6.51	8.19
7	1.05	1.39	1.58	2.16	2.50	2.94	3.74	4.67	5.46	6.23	7.86	9.80
8	1.42	1.83	2.05	2.73	3.13	3.63	4.54	5.60	6.50	7.37	9.21	11.4
9	1.83	2.30	2.56	3.33	3.78	4.34	5.37	6.55	7.55	8.52	10.6	13.0
10	2.26	2.80	3.09	3.96	4.46	5.08	6.22	7.51	8.62	9.68	12.0	14.7
11	2.72	3.33	3.65	4.61	5.16	5.84	7.08	8.49	9.69	10.9	13.3	16.3
12	3.21	3.88	4.23	5.28	5.88	6.61	7.95	9.47	10.8	12.0	14.7	18.0
13	3.71	4.45	4.83	5.96	6.61	7.40	8.83	10.5	11.9	13.2	16.1	19.6
14	4.24	5.03	5.45	6.66	7.35	8.20	9.73	11.5	13.0	14.4	17.5	21.2
15	4.78	5.63	6.08	7.38	8.11	9.01	10.6	12.5	14.1	15.6	18.9	22.9
16	5.34	6.25	6.72	8.10	8.88	9.83	11.5	13.5	15.2	16.8	20.3	24.5
17	5.91	6.88	7.38	8.83	9.65	10.7	12.5	14.5	16.3	18.0	21.7	26.2
18	6.50	7.52	8.05	9.58	10.4	11.5	13.4	15.5	17.4	19.2	23.1	27.8
19	7.09	8.17	8.72	10.3	11.2	12.3	14.3	16.6	18.5	20.4	24.5	29.5
20	7.70	8.83	9.41	11.1	12.0	13.2	15.2	17.6	19.6	21.6	25.9	31.2

TABLE D.1 ***(Continued)***

N/B	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30	40
21	8.32	9.50	10.1	11.9	12.8	14.0	16.2	18.7	20.8	22.8	27.3	32.8
22	8.95	10.2	10.8	12.6	13.7	14.9	17.1	19.7	21.9	24.1	28.7	34.5
23	9.58	10.9	11.5	13.4	14.5	15.8	18.1	20.7	23.0	25.3	30.1	36.1
24	10.2	11.6	12.2	14.2	15.3	16.6	19.0	21.8	24.2	26.5	31.6	37.8
25	10.9	12.3	13.0	15.0	16.1	17.5	20.0	22.8	25.3	27.7	33.0	39.4
26	11.5	13.0	13.7	15.8	17.0	18.4	20.9	23.9	26.4	28.9	34.4	41.1
27	12.2	13.7	14.4	16.6	17.8	19.3	21.9	24.9	27.6	30.2	35.8	42.8
28	12.9	14.4	15.2	17.4	18.6	20.2	22.9	26.0	28.7	31.4	37.2	44.4
29	13.6	15.1	15.9	18.2	19.5	21.0	23.8	27.1	29.9	32.6	38.6	46.1
30	14.2	15.9	16.7	19.0	20.3	21.9	24.8	28.1	31.0	33.8	40.0	47.7
31	14.9	16.6	17.4	19.9	21.2	22.8	25.8	29.2	32.1	35.1	41.5	49.4
32	15.6	17.3	18.2	20.7	22.0	23.7	26.7	30.2	33.3	36.3	42.9	51.1
33	16.3	18.1	19.0	21.5	22.9	24.6	27.7	31.3	34.4	37.5	44.3	52.7
34	17.0	18.8	19.7	22.3	23.8	25.5	28.7	32.4	35.6	38.8	45.7	54.4
35	17.8	19.6	20.5	23.2	24.6	26.4	29.7	33.4	36.7	40.0	47.1	56.0
36	18.5	20.3	21.3	24.0	25.5	27.3	30.7	34.5	37.9	41.2	48.6	57.7
37	19.2	21.1	22.1	24.8	26.4	28.3	31.6	35.6	39.0	42.4	50.0	59.4
38	19.9	21.9	22.9	25.7	27.3	29.2	32.6	36.6	40.2	43.7	51.4	61.0
39	20.6	22.6	23.7	26.5	28.1	30.1	33.6	37.7	41.3	44.9	52.8	62.7
40	21.4	23.4	24.4	27.4	29.0	31.0	34.6	38.8	42.5	46.1	54.2	64.4
41	22.1	24.2	25.2	28.2	29.9	31.9	35.6	39.9	43.6	47.4	55.7	66.0
42	22.8	25.0	26.0	29.1	30.8	32.8	36.6	40.9	44.8	48.6	57.1	67.7
43	23.6	25.7	26.8	29.9	31.7	33.8	37.6	42.0	45.9	49.9	58.5	69.3
44	24.3	26.5	27.6	30.8	32.5	34.7	38.6	43.1	47.1	51.1	59.9	71.0
45	25.1	27.3	28.4	31.7	33.4	35.6	39.6	44.2	48.2	52.3	61.3	72.7
46	25.8	28.1	29.3	32.5	34.3	36.5	40.5	45.2	49.4	53.6	62.8	74.3
47	26.6	28.9	30.1	33.4	35.2	37.5	41.5	46.3	50.6	54.8	64.2	76.0
48	27.3	29.7	30.9	34.2	36.1	38.4	42.5	47.4	51.7	56.0	65.6	77.7
49	28.1	30.5	31.7	35.1	37.0	39.3	43.5	48.5	52.9	57.3	67.0	79.3
50	28.9	31.3	32.5	36.0	37.9	40.3	44.5	49.6	54.0	58.3	68.5	81.0
51	29.6	32.1	33.3	36.9	38.8	41.2	45.5	50.6	55.2	59.7	69.9	82.7
52	30.4	32.9	34.2	37.7	39.7	42.1	46.5	51.7	56.3	61.0	71.3	84.3
53	31.2	33.7	35.0	38.6	40.6	43.1	47.5	52.8	57.5	62.2	72.7	86.0
54	31.9	34.5	35.8	39.5	41.5	44.0	48.5	53.9	58.7	63.5	74.2	87.6
55	32.7	35.3	36.6	40.4	42.4	44.9	49.5	55.0	59.8	64.7	75.6	89.3
56	33.5	36.1	37.5	41.2	43.3	45.9	50.5	56.1	61.0	65.9	77.0	91.0
57	34.3	36.9	38.3	42.1	44.2	46.8	51.5	57.1	62.1	67.7	78.4	92.6
58	35.1	37.8	39.1	43.0	45.1	47.8	52.6	58.2	63.3	68.4	79.8	94.3
59	35.8	38.6	40.0	43.9	46.0	48.7	53.6	59.3	64.5	69.7	81.3	96.0
60	36.6	39.4	40.8	44.8	46.9	49.6	54.6	60.4	65.6	70.9	82.7	97.6
61	37.4	40.2	41.6	45.6	47.9	50.6	55.6	61.5	66.8	72.1	84.1	99.3
62	38.2	41.0	42.5	46.5	48.8	51.5	56.6	62.6	68.0	73.4	85.5	101.
63	39.0	41.9	43.3	47.4	49.7	52.5	57.6	63.7	69.1	74.6	87.0	103.
64	39.8	42.7	44.2	48.3	50.6	53.4	58.6	64.8	70.3	75.9	88.4	104.
65	40.6	43.5	45.0	49.2	51.5	54.4	59.6	65.8	71.4	77.1	89.8	106.

TABLE D.1 ***(Continued)***

N/B	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30	40
66	41.4	44.4	45.8	50.1	52.4	55.3	60.6	66.9	72.6	78.3	91.2	108.
67	42.2	45.2	46.7	51.0	53.4	56.3	61.6	68.0	73.8	79.6	92.7	109.
68	43.0	46.0	47.5	51.9	54.3	57.2	62.6	69.1	74.9	80.8	94.1	111.
69	43.8	46.8	48.4	52.8	55.2	58.2	63.7	70.2	76.1	82.1	95.5	113.
70	44.6	47.7	49.2	53.7	56.1	59.1	64.7	71.3	77.3	83.3	96.9	114.
71	45.4	48.5	50.1	54.6	57.0	60.1	65.7	72.4	78.4	84.6	98.4	116.
72	46.2	49.4	50.9	55.5	58.0	61.0	66.7	73.5	79.6	85.8	99.8	118.
73	47.0	50.2	51.8	56.4	58.9	62.0	67.7	74.6	80.8	87.0	101.	119.
74	47.8	51.0	52.7	57.3	59.8	62.9	68.7	75.6	81.9	88.3	103.	121.
75	48.6	51.9	53.6	58.2	60.7	63.9	69.7	76.7	83.1	89.5	104.	123.
76	49.4	52.7	54.4	59.1	61.7	64.9	70.8	77.8	84.2	90.8	105.	124.
77	50.2	53.6	55.2	60.0	62.6	65.8	71.8	78.9	85.4	92.0	107.	126.
78	51.1	54.4	56.1	60.9	63.5	66.8	72.8	80.0	86.6	93.3	108.	128.
79	51.9	55.3	56.9	61.8	64.4	67.7	73.8	81.1	87.7	94.5	110.	129.
80	52.7	56.1	57.8	62.7	65.4	68.7	74.8	82.2	88.9	95.7	111.	131.
81	53.5	56.9	58.7	63.6	66.3	69.6	75.8	83.3	90.1	97.0	113.	133.
82	54.3	57.8	59.5	64.5	67.2	70.6	76.9	84.4	91.2	98.2	114.	134.
83	55.1	58.6	60.4	65.4	68.2	71.6	77.9	85.5	92.4	99.5	115.	136.
84	56.0	59.5	61.3	66.3	69.1	72.5	78.9	86.6	93.6	101.	117.	138.
85	56.8	60.4	62.1	67.2	70.0	73.5	79.9	87.7	94.7	102.	118.	139.
86	57.6	61.2	63.0	68.1	70.9	74.5	80.9	88.8	95.9	103.	120.	141.
87	58.4	62.1	63.9	69.0	71.9	75.4	82.0	89.9	97.1	104.	121.	143.
88	59.3	62.9	64.7	69.9	72.8	76.4	83.0	91.0	98.2	106.	123.	144.
89	60.1	63.8	65.6	70.8	73.7	77.3	84.0	92.1	99.4	107.	124.	146.
90	60.9	64.6	66.5	71.8	74.7	78.3	85.0	93.1	101.	108.	126.	148.
91	61.8	65.5	67.4	72.7	75.6	79.3	86.0	94.2	102.	109.	127.	149.
92	62.6	66.3	68.2	73.6	76.6	80.2	87.1	95.3	103.	111.	128.	151.
93	63.4	67.2	69.1	74.5	77.5	81.2	88.1	96.4	104.	112.	130.	153.
94	64.2	68.1	70.0	75.4	78.4	82.2	89.1	97.5	105.	113.	131.	154.
95	65.1	68.9	70.9	76.3	79.4	83.1	90.1	98.6	106.	114.	133.	156.
96	65.9	69.8	71.7	77.2	80.3	84.1	91.1	99.7	108.	116.	134.	158.
97	66.8	70.7	72.6	78.2	81.2	85.1	92.2	101.	109.	117.	135.	159.
98	67.6	71.5	73.5	79.1	82.2	86.0	93.2	102.	110.	118.	137.	161.
99	68.4	72.4	74.4	80.0	83.1	87.0	94.2	103.	111.	119.	138.	163.
100	69.3	73.2	75.2	80.9	84.1	88.0	95.2	104.	112.	121.	140.	164.

[a]*N* is the number of servers. The numerical column headings indicate blocking probability (%).

TABLE D.2 Maximum Offered Load Versus *B*, *N*, and Finite Sources *M*[a]

N	*M*	.01	.05	0.1	0.5	1.0	2	5	10	15	20	30
1	2	.0002	.0010	.0020	.0100	.0200	.0400	.100	.202	.307	.417	.659
	3	.0002	.0008	.0015	.0075	.0151	.0303	.077	.159	.246	.341	.559
	4	.0001	.0007	.0013	.0067	.0134	.0270	.069	.143	.224	.312	.519
	5	.0001	.0006	.0013	.0063	.0126	.0254	.065	.136	.213	.298	.498
2	3	.030	.067	.095	.212	.300	.425	.678	.980	1.23	1.47	1.97
	4	.023	.052	.074	.167	.238	.342	.560	.832	1.07	1.30	1.78
	5	.021	.046	.065	.149	.213	.308	.510	.767	.997	1.22	1.70
	6	.019	.043	.061	.139	.200	.289	.482	.731	.955	1.18	1.65
	7	.018	.041	.058	.133	.191	.277	.464	.707	.928	1.15	1.62
3	4	.186	.317	.400	.685	.864	1.09	1.50	1.95	2.31	2.65	3.35
	5	.148	.254	.322	.561	.715	.918	1.30	1.72	2.08	2.42	3.13
	6	.131	.227	.288	.505	.648	.837	1.20	1.62	1.97	2.31	3.02
	7	.122	.211	.268	.473	.609	.790	1.14	1.55	1.90	2.24	2.95
	8	.116	.201	.255	.452	.583	.759	1.10	1.51	1.85	2.19	2.90
	9	.111	.194	.246	.437	.565	.737	1.07	1.47	1.82	2.16	2.86
	10	.108	.188	.240	.426	.551	.720	1.05	1.45	1.79	2.13	2.84
	15	.100	.174	.222	.396	.514	.675	.994	1.38	1.72	2.06	2.76
4	5	.500	.748	.889	1.33	1.59	1.89	2.24	2.98	3.43	3.86	4.76
	6	.408	.617	.737	1.12	1.36	1.64	2.15	2.71	3.16	3.60	4.51
	7	.365	.554	.665	1.02	1.24	1.52	2.01	2.56	3.02	3.46	4.38
	8	.340	.517	.621	.963	1.17	1.44	1.92	2.47	2.93	3.37	4.30
	9	.323	.492	.592	.922	1.13	1.39	1.86	2.41	2.87	3.31	4.24
	10	.310	.474	.571	.892	1.09	1.35	1.82	2.36	2.82	3.27	4.20
	15	.280	.429	.518	.816	1.00	1.25	1.71	2.24	2.70	3.14	4.08
5	6	.951	1.31	1.51	2.08	2.40	2.77	3.39	4.04	4.58	5.09	6.17
	7	.794	1.11	1.28	1.80	2.10	2.45	3.07	3.73	4.28	4.80	5.91
	8	.716	1.00	1.16	1.66	1.94	2.29	2.90	3.56	4.12	4.65	5.77
	9	.668	.940	1.04	1.56	1.84	2.18	2.78	3.45	4.01	4.55	5.68
	10	.635	.896	1.09	1.50	1.77	2.11	2.71	3.37	3.94	4.48	5.61
	12	.592	.839	.979	1.42	1.68	2.01	2.60	3.27	3.84	4.38	5.52
	15	.556	.791	.924	1.35	1.60	1.92	2.51	3.18	3.74	4.29	5.44
	20	.525	.748	.876	1.28	1.53	1.84	2.42	3.09	3.66	4.21	5.37
6	7	1.51	1.97	2.21	2.90	3.26	3.69	4.38	5.12	5.73	6.32	7.59
	8	1.28	1.70	1.91	2.55	2.90	3.32	4.02	4.78	5.41	6.02	7.32
	9	1.17	1.55	1.76	2.37	2.71	3.12	3.82	4.59	5.24	5.85	7.17
	10	1.09	1.46	1.65	2.25	2.58	2.98	3.69	4.47	5.12	5.74	7.07
	15	.926	1.25	1.43	1.97	2.29	2.68	3.38	4.17	4.84	5.48	6.84
	20	.865	1.17	1.34	1.87	2.17	2.56	3.26	4.05	4.72	5.37	6.74
	30	.813	1.10	1.27	1.77	2.07	2.45	3.15	3.94	4.62	5.28	6.66
7	8	2.15	2.70	2.98	3.76	4.17	4.63	5.39	6.20	6.89	7.56	9.01
	9	1.85	2.36	2.62	3.36	3.75	4.22	5.00	5.85	6.56	7.25	8.74
	10	1.70	2.17	2.42	3.13	3.52	3.99	4.78	5.64	6.37	7.07	8.58
	15	1.39	1.80	2.02	2.68	3.05	3.51	4.31	5.21	5.97	6.70	8.25
	20	1.28	1.67	1.88	2.52	2.88	3.33	4.14	5.05	5.81	6.56	8.13
	30	1.19	1.56	1.76	2.38	2.74	3.18	3.99	4.90	5.68	6.44	8.03

TABLE D.2 (*Continued*)

N	M	.01	.05	0.1	0.5	1.0	2	5	10	15	20	30
	9	2.85	3.48	3.80	4.65	5.09	5.59	6.41	7.30	8.05	8.80	10.4
	10	2.49	3.08	3.37	4.20	4.64	5.14	6.00	6.92	7.71	8.48	10.2
	11	2.29	2.85	3.14	3.94	4.38	4.89	5.76	6.71	7.51	8.29	9.99
8	12	2.16	2.70	2.97	3.77	4.20	4.17	5.59	6.56	7.37	8.17	9.88
	15	1.93	2.43	2.70	3.46	3.89	4.40	5.29	6.29	7.13	7.94	9.69
	20	1.76	2.24	2.49	3.23	3.65	4.16	5.06	6.07	6.93	7.76	9.54
	30	1.63	2.08	2.32	3.03	3.45	3.95	4.86	5.89	6.77	7.61	9.41
	10	3.59	4.30	4.64	5.57	6.03	6.56	7.44	8.39	9.22	10.0	11.9
	11	3.18	3.84	4.17	5.07	5.54	6.09	7.00	8.01	8.87	9.72	11.6
	12	2.94	3.57	3.89	4.79	5.26	5.81	6.75	7.78	8.66	9.52	11.4
	13	2.78	3.39	3.71	4.59	5.06	5.61	6.57	7.62	8.51	9.39	11.3
9	14	2.66	3.26	3.57	4.44	4.91	5.47	6.43	7.50	8.41	9.29	11.2
	16	2.50	3.08	3.38	4.23	4.70	5.27	6.24	7.33	8.25	9.15	11.1
	18	2.39	2.95	3.25	4.09	4.56	5.13	6.11	7.21	8.15	9.06	11.0
	20	2.31	2.87	3.15	3.99	4.46	5.02	6.02	7.13	8.07	8.99	11.0
	30	2.12	2.64	2.91	3.73	4.19	4.76	5.77	6.90	7.87	8.81	10.8
	11	4.38	5.15	5.52	6.49	6.98	7.54	8.47	9.49	10.4	11.3	13.3
	12	3.91	4.64	5.00	5.97	6.47	7.04	8.02	9.09	10.0	11.0	13.0
	13	3.63	4.34	4.69	5.65	6.16	6.74	7.75	8.86	9.81	10.8	12.8
	14	3.45	4.13	4.47	5.43	5.94	6.54	7.56	8.69	9.66	10.6	12.7
10	15	3.31	3.98	4.32	5.27	5.78	6.38	7.41	8.56	9.55	10.5	12.6
	16	3.20	3.86	4.19	5.14	5.65	6.25	7.30	8.46	9.46	10.4	12.6
	18	3.04	3.68	4.01	4.95	5.46	6.07	7.13	8.31	9.33	10.3	12.5
	20	2.93	3.56	3.88	4.81	5.32	5.93	7.01	8.21	9.23	10.2	12.4
	25	2.75	3.36	3.68	4.59	5.10	5.72	6.81	8.04	9.08	10.1	12.3
	30	2.65	3.25	3.56	4.47	4.98	5.59	6.69	7.93	8.99	10.0	12.2
	12	5.19	6.01	6.41	7.44	7.95	8.53	9.50	10.6	11.6	12.5	14.7
	13	4.68	5.46	5.85	6.88	7.40	8.01	9.04	10.2	11.2	12.2	14.4
	14	4.37	5.13	5.51	6.54	7.07	7.69	8.76	9.94	11.0	12.0	14.2
	15	4.15	4.90	5.27	6.30	6.84	7.47	8.56	9.77	10.8	11.9	14.1
11	16	3.99	4.72	5.09	6.12	6.66	7.30	8.40	9.63	10.7	11.7	14.0
	17	3.86	4.59	4.95	5.98	6.52	7.17	8.28	9.53	10.6	11.7	14.0
	18	3.76	4.48	4.84	5.86	6.41	7.06	8.18	9.44	10.5	11.6	13.9
	20	3.60	4.31	4.66	5.68	6.23	6.88	8.03	9.31	10.4	11.5	13.8
	25	3.36	4.04	4.39	5.40	5.95	6.62	7.79	9.10	10.2	11.3	13.7
	30	3.32	3.90	4.24	5.24	5.79	6.46	7.64	8.98	10.1	11.2	13.6
	13	6.03	6.90	7.31	8.39	8.92	9.52	10.5	11.7	12.7	13.8	16.1
	14	5.47	6.31	6.72	7.80	8.35	8.98	10.1	11.3	12.4	13.4	15.8
	15	5.13	5.95	6.35	7.44	8.00	8.65	9.77	11.0	12.1	13.2	15.7
	16	4.88	5.69	6.09	7.18	7.75	8.42	9.56	10.8	12.0	13.1	15.5
12	17	4.70	5.50	5.90	6.99	7.56	8.24	9.40	10.7	11.8	13.0	15.5
	18	4.56	5.35	5.74	6.83	7.41	8.09	9.27	10.6	11.8	12.9	15.4
	20	4.34	5.12	5.51	6.60	7.18	7.87	9.08	10.4	11.6	12.8	15.3
	25	4.03	4.77	5.16	6.24	6.83	7.54	8.78	10.2	11.4	12.6	15.1
	30	3.85	4.58	4.96	6.04	6.63	7.35	8.61	10.0	11.3	12.5	15.0

[a]The numerical column headings indicate blocking probability (%).

GLOSSARY

2B1Q. Four-level line code used in ISDN basic rate access lines.

99% Power bandwidth. Bandwidth containing 99% of the energy of a signal.

Abbreviated dialing. *See* Speed dialing.

Access tandem. "Switching system within a LATA used as an access point by long-distance carriers."

Add–drop multiplexer (ADM). Network element that accesses only a portion of a higher level digital signal to extract and insert local traffic at an intermediate point of a route.

Added-channel framing. TDM framing format utilizing an additional channel with the same rate as the message channels for the express purpose of defining frame boundaries.

Added-digit framing. TDM framing format utilizing an additional bit in every frame with a prescribed repetitive sequence for the express purpose of defining frame boundaries.

Advanced mobile phone service (AMPS). Analog cellular mobile telephone standard of North America.

Alternate mark inversion (AMI) signal. Three-level signal with values $+V$, 0, $-V$. A space (logic 0) is encoded with 0 V. A mark (logic 1) is encoded with alternating values of $+V$ and $-V$, which are normally only 50% duty cycle pulses. Also referred to as a bipolar signal.

Alternate mark inversion violation. In AMI coding, the occurrence of two successive marks with the same polarity. Also referred to as a bipolar violation.

Amplitude distortion. Distortion of a transmission signal caused by nonuniform passband attenuation as a function of frequency.

Amplitude modulation (AM). Modulation of the amplitude of a carrier wave with the amplitude of a baseband signal.

Antipodal signaling. Technique of encoding binary signals so that the symbol for a 1 is the exact negative of the symbol for a 0. Antipodal signaling provides optimum error performance in terms of the signal-to-noise ratio.

Asynchronous network. A network in which the clocks of the transmission links and switching systems are not synchronized to each other.

Asynchronous transfer mode (ATM). Packet-switched method of time division multiplexing (labeled multiplexing) utilizing fixed packet (cell) sizes of 53 bytes.

Asynchronous transmission. Mode of communication characterized by start–stop transmissions with undefined time intervals between transmissions. Each transmission burst generally contains a single word or byte of information.

Automatic call distributor (ACD). Switching system used to evenly distribute incoming calls to a number of stations on a first-come, first-served basis. Applications include operator assistance and airline reservations.

Automatic number identification (ANI). Process of identifying and forwarding a calling number to network call control services.

Automatic repeat request (ARQ). Error correction process involving the use of redundant check bits to detect corrupted blocks of data and trigger requests for retransmission of the same.

Availability. (1) With respect to switching systems: the number of outlets available from a particular inlet. (2) With respect to equipment in general: the percentage of time the equipment is providing acceptable service.

Balanced code. Line code with equally occurring positive and negative energy to preclude a dc component in the frequency spectrum.

Baseband. Literally, the frequency band of an unmodulated signal. A baseband signal is an information-bearing signal that is either transmitted as is or used to modulate a carrier.

Baud rate. Unit of signaling speed (symbols per second). For binary signaling the data rate is the same as the baud rate. For multilevel transmission the data rate is equal to the baud rate times $\log_2(L)$, where L is the number of levels.

Bipolar coding. *See* Alternate mark inversion.

Bis. Second version. Ter means three. For example, V.27bis and V.27ter are the second and third versions of the V.27 standard, respectively.

Bit leaking. Process of converting byte-sized timing adjustments into multiple bit-sized or fractional bit-sized timing adjustments.

Blocked calls cleared (BCC). Service discipline in which unserviceable requests are rejected by the system without service. Also called lost calls cleared (LCC).

Blocked calls held (BCH). Service discipline in which unserviceable requests stay "in the system" without being serviced but having a portion of their desired service time elapse until service begins. Also called lost calls held (LCH).

Blocking. Inability of a calling party to be connected to a called party because either all circuits are busy or there is internal (matching loss) blocking within a switch matrix.

Bridged tap. Extra pair of wires connected in shunt to a main cable pair. The extra pair is normally open circuited but may be used at a future time to connect the main

pair to a new customer. Short-bridged taps do not affect voice frequency signals but can be extremely detrimental to higher frequency digital signals.

Building integrated timing supply (BITS). Master timing supply for all equipment in a building.

Busy hour. The 60-minute period of a day (and sometimes of a week) in which the average offered traffic load is highest.

Call admission control (CAC). ATM control process responsible for negotiating quality-of-service parameters for new connections dependent on the traffic loads of the network at the time.

Call congestion. Blocking probability of a trunk group.

Call forwarding. Process of signaling a switching system to reroute calls intended for a particular number to some other number on a temporary basis.

Call waiting. Process of inserting a short tone into the voice path of an active user informing the user that another call is waiting to be connected.

Calling number identification (CNI). Service whereby a telephone network provides the identity of a calling number (or person) with incoming ringing.

Carrier sense multiple access/collision detection (CSMA/CD). Access protocol for a common bus or radio link in which sources first try to detect the presence of a carrier and then begin transmitting when the facility is idle. If another source begins transmitting during a period of time corresponding to the maximum transmission delay, a collision is detected and all active nodes cease transmission for a random amount of time.

Carrierless amplitude/phase (CAP) modulation. QAM implementation technique using direct (DSP) generation of filtered quadrature signal waveforms followed by DSP summation for the composite signal.

CCS. Measure of traffic intensity expressed as so many hundred call seconds per hour: 36 CCS = 1 erlang.

Cell delay variation (CDV). An ATM QoS parameter that specifies the maximum range of cell propagation delay through a network.

Cell transfer delay (CTD). An ATM QoS parameter specifying the average propagation time for a cell to be transferred from its source to its destination.

Central office. Public network switching office (and particularly the switching machine within it) to which subscriber lines are connected. Also referred to as an end office or class 5 office. In a more general sense, any switch in the public network is sometimes referred to as a central office.

Centrex service. Private switching service provided by a local exchange carrier to an organization utilizing hardware and software partitions of a central office switch.

CEPT-1. Term used to designate the 2.048-Mbps first-level digital signal of the ITU-T digital hierarchy. (More often referred to as an E1 signal.)

Channel-associated signaling (CAS). Use of separate signaling channels for each message channel. The individual signaling channels may be separate from the

message channels (as in E1) or derived within the message channels themselves (as in T1).

Channel bank. Equipment that converts multiple analog interfaces to and from a time-division-multiplexed digital bit stream (24 analog to one DS1 or 30 analog to one E1). Channel banks are also used for FDM applications.

Channel service unit (CSU). Equipment installed on customer premises at the interface to phone company lines to terminate a DDS or T1 circuit to provide network protection and diagnostic capabilities. Also called a customer service unit.

Circuit Emulation Service (CES). An ATM service in which constant-bit-rate virtual circuits use AAL1 adaptation to emulate an end-to-end physical circuit.

Circuit switching. The principle of establishing an end-to-end connection between users of a network. The associated facilities are dedicated to the particular connection and held for the duration of the call.

Clear-channel capability. Ability to provide a transparent 64-kbps channel through a North American network (using B8ZS and common-channel signaling).

Clock. Time base used to control the transfer of digital information.

C-message weighing. Selective attenuation of voiceband noise in accordance with the subjective effects as a function of frequency (noise weighting filter characteristic used in North America).

Codec (coder–decoder). Integrated circuit providing analog-to-digital and digital-to-analog conversion of voice signals.

Coded mark inversion (CMI). Two-level line code in which a binary 0 is coded with a positive level and an equal-magnitude negative level for half a unit time interval each. A binary 1 is coded with a full-period positive or negative level in an alternating manner to maintain dc balance.

Coherent demodulation. Demodulation using a carrier reference that is synchronized in frequency and phase to the carrier used in the modulation process.

Common-channel signaling. Use of a separate channel dedicated to transport of signaling information pertaining to a multiple number of message channels.

Community dial office (CDO). Small, normally unattended switching system that is used in small communities and is controlled from a larger central office.

Companding. Process of compressing a signal at the source and expanding it at the destination to maintain a given end-to-end dynamic range while reducing the dynamic range between the compressor–expander.

Concentration. Process of switching some number of lightly used channels or sources onto a smaller number of more heavily used channels.

Conference call. Call in which three or more stations are interconnected in a manner that allows all parties to talk and be heard by all other parties.

Connection admission control (CAC). An ATM function that determines whether a virtual circuit (VC) connection request should be accepted or rejected.

Constant (or continuous) bit rate (CBR). An ATM class of service that supports transmission of a continuous data rate.

Constant envelope. A (frequency or phase) modulated carrier with no variation in the carrier amplitude.

Controlled slip. Slip involving the repetition or deletion of an integral number of TDM frames.

Crossbar switch. Rectangular array of crosspoints arranged so that an input line connected to a row of crosspoint inputs can be switched to any output line connected to a column of crosspoint outputs.

Crosstalk. Unwanted signal transfer from one circuit into another.

Custom Local Area Signaling Services (CLASS). Signaling service for analog telephone lines provided between the first two rings of an incoming telephone call. Principal features enabled by CLASS services are Caller ID, Call Return, Call Screening, Preferred Call Forwarding, Priority Ringing, and Repeat Dialing.

Datagram. Single packet message (from the point of view of a packet network).

Dataphone Digital Service (DDS). Service offering of common-carrier telephone companies providing digital channels at 2.4, 4.8, 9.6, 56, or 1544 kbps.

dBm. Power level in decibels relative to 1 mW.

dBrnC. Power level of noise with C-message weighting expressed in decibels relative to reference noise. Reference noise power is -90 dBm $= 10^{-12}$ W.

dBrnC0. Noise power measured in dBrnC but referenced to the zero-level transmission level point.

Decision feedback equalization (DFE). Equalization technique involving the use of detected data values to produce simulated channel responses in a receiver that are then subtracted from subsequent received signal samples to cancel channel distortion.

Delta modulation. Technique for digitally encoding an analog signal by continuously measuring and transmitting only the polarity of the signal slope. The analog signal is reconstructed as a staircase approximation of the input.

Demand traffic. First attempt traffic. (Offered traffic not including retries.)

Desynchronizer. Circuitry that extracts a tributary digital signal from a composite, higher level signal and derives the source clock frequency of the tributary.

Digital advanced mobile phone service (D-AMPS). Digital mobile telephone standard of North America, also known as IS-136 TDMA.

Digital circuit multiplication (DCM). Process of increasing the number of voice signals carried on a digital transmission link through the use of speech compression or silence removal or both.

Digital loop carrier (DLC). A TDM system for transporting multiple digital signals to and from multiple subscriber lines.

Digital pad. Digital circuit that changes PCM sample values for the purpose of adjusting (attenuating) the power level of the encoded analog signal.

Digital satellite system (DSS). Direct broadcast satellite (DBS) system developed by Hughes Electronics Corporation.

Digital speech interpolation (DSI). Digital equivalent of TASI wherein digital speech signals are monitored and connected to a transmission channel only when voice activity is present (a form of DCM).

Digital video broadcasting (DVB) group. European digital TV standard for TV, audio, and data. DVB can be broadcast via satellite or terrestrial systems.

Direct broadcast satellite (DBS). One-way TV broadcast service from a satellite to a small 18-in. dish antenna.

Direct inward dialing (DID). Process of a public telephone network providing PBX extension numbers with incoming calls.

Double-sideband modulation. Modulation technique in which a baseband signal with no dc energy directly modulates a carrier to produce both upper and lower sidebands but no carrier energy.

Dual-tone multifrequency (DTMF) signaling. Generic name for push-button telephone signaling equivalent to the Bell System's TOUCH-TONE.

Dynamic range. Range of power levels (minimum to maximum) achievable by a signal or specified for equipment operation.

E&M lead signaling. Interface between a switching system and a transmission system utilizing pairs of wires for signaling that are separate from the voice pairs. Outgoing control signals from each facility are carried by respective M-leads to E-leads of the other facility.

Echo canceler. Device that removes talker echo in the return branch of a four-wire circuit by subtracting a delayed version of the signal transmitted in the forward path.

Echo suppressor. Device that is activated by voice energy in one path of a four-wire circuit to insert a high amount of signal loss in the other path for the purpose of blocking an echo.

Echo. Reflected and delayed signal. Common sources of echo in the telephone network are electrical reflections at four-wire to two-wire conversions and acoustic reflections with speakerphones.

Edge switch. An ATM switch providing gateway interfaces to non-ATM communication links.

Elastic store. First-in, first-out data buffer that accepts data under control of one clock and outputs data under control of another clock.

Electronic automatic exchange (EAX). Designation of stored-program control switching machines developed by General Telephone and Electronics.

Electronic switching system (ESS). Designation of stored-program control switching machines developed by AT&T.

End office. Class 5 switching office. Also referred to as a central office.

Envelope delay. Derivative of channel phase response with respect to frequency. Ideally, the phase response should be linear, indicating that all frequencies are delayed equally.

Envelope detector. Method of detecting amplitude modulation by tracking only the peak values of a carrier wave.

Equalization. Practice of compensating for transmission distortions with fixed or adaptive circuitry.

Erlang. Measure of traffic intensity. Basically, a measure of the utilization of a resource (e.g., the average number of busy circuits in a trunk group, or the ratio of time an individual circuit is busy).

Error second. A 1-sec interval containing at least 1 bit error.

Exchange area. Contiguous area of service defined for administrative purposes that typically comprises an entire town or city and includes the immediate countryside and suburbs. An exchange area may have one end office or many end offices interconnected by trunks and tandem offices.

Expansion. Switching of a number of input channels onto a larger number of output channels.

Extended superframe format (ESF). A T1 framing format that embeds diagnostic capabilities and a facility data link into the 8-kbps DS1 framing channel.

Failed seconds (FS). An (ESF) transmission link failure defined as 10 consecutive severely errored seconds.

Far-end crosstalk (FEXT). Unwanted energy coupled from one channel or circuit into another circuit and appearing at the far end of the transmission link.

Fiber distributed data interface (FDDI). High-bandwidth fiber transmission system for interconnecting local area networks.

FIFO. First-in, first-out service discipline for a queue.

Flow control (in data network). Procedure for controlling the rate of transfer of packets from one node to another.

Foreign exchange circuit (FX). Extension of service from one switching office to a subscriber normally serviced by a different switching office.

Forward error correction (FEC). Error correction technique with sufficient redundancy embedded into the signal for correcting errors without retransmission.

Four-wire circuit. Circuit using two separate channels for each direction of transmission. When wireline transmission is involved, each direction of transmission is provided by a separate pair of wires.

Fractional speech loss. Fraction of speech that gets clipped in a TASI or DSI system because all channels are in use when talk spurts begin.

Frame alignment. Identification of the beginning and end of a TDM frame using a framing data pattern.

Frame. Set of consecutive time slots in a TDM format in which the position of individual time slots can be identified by reference to a frame alignment signal.

Frequency diversity. In radio systems the use of one or more back-up transmitters, channels, and receivers to protect against atmospheric (multipath) fading.

Frequency division multiplex (FDM). Partitioning the bandwidth of a transmission link into separate channels of lesser bandwidth into which separate signals are inserted and extracted.

Frequency modulation (FM). Modulation process that varies the frequency of a carrier signal in accordance with a baseband information signal.

Frequency shift keying (FSK). Digital modulation technique in which data are represented by discrete frequencies.

Fricative. Speech sound produced with continuous air flow through one or more vocal tract restrictions to cause air turbulence (friction). Example fricative sounds are s, f, t, x, or j.

Full-duplex. Transmission in two directions simultaneously (also referred to as simply a duplex operation).

Gateway network element (GNE). A SONET node with an interface to unsynchronized tributary signals.

Gaussian noise. Background noise with a uniform frequency spectrum across a band of interest and having amplitude sample values that follow a normal (gaussian) probability distribution.

Glare. Simultaneous seizure of both ends of a two-way trunk by two switching machines trying to set up two separate connections.

Global positioning system (GPS). Global satellite navigation system commonly used by telecommunications systems as a time and frequency synchronization source.

Half-duplex transmission. Transmission in both directions but only in one direction at a time.

HDB3. Modified AMI (bipolar) line code in which strings of four 0's are encoded with an AMI violation in the last bit.

Head-of-line blocking. Blocking of cells or packets in a FIFO queue because the next entry cannot be serviced while other entries in the queue could be serviced.

Highway. A TDM path inside a digital (TDM) switching machine.

Hook flash. Momentary depression of the switchhook to alert an operator or equipment, but not so long as to be interpreted as a disconnect.

Hot standby. Redundant equipment kept in an operational mode as backup for primary equipment. Usually, automatic switching to the standby equipment occurs when the primary equipment fails.

Huffman coding. Process of assigning variable-length digital codewords to a message set wherein the length of a codeword is inversely proportional to the probability of occurrence of respective messages. Thus a frequently transmitted message requires fewer bits than the average length of all messages.

Hybrid fiber coax (HFC). Cable television distribution system involving the use of fiber in the feeder portion of the network from a head-end to fiber nodes followed by coax drops to residences in neighborhoods.

Hybrid. Device used to connect a two-wire, bidirectional circuit to a four-wire circuit.

Idle channel noise. Noise occurring during silence intervals of speech.

Idle character. Control character inserted into a synchronous transmission link when there is no information to be sent.

Impulse noise. Short-duration spike of noise that is much larger than randomly distributed (Gaussian) noise.

Inband signaling. Signaling transmitted within the same channel and band of frequencies used for message traffic.

Integrated digital network (IDN). Network in which digital TDM transmission links are directly interfaced into digital TDM switching matrices.

Integrated services digital network (ISDN). Integrated digital network in which subscriber interfaces are provided with direct access of a signaling channel and one or more digital 64-kbps channels.

Intercept. Process of diverting a call from an intended station to an operator or recorded announcement.

International Telecommunications Union (ITU). International standards organization that subsumed the CCITT in March of 1993 (renamed ITU-T).

Internet. Connected set of networks using TCP/IP suite.

Intersymbol interference (ISI). Interference in a transmission system caused by a symbol in one signaling internal being spread out and overlapping the sample time of a symbol in another signal interval.

Jitter. Short-term variations of the significant instants of a digital signal from their ideal positions in time. Short-term variations are often defined to correspond to frequencies above 10 Hz.

Joint photographic experts group (JPEG). ISO standards group that defines how to compress still pictures.

Justification (pulse stuffing). Process of adapting the rate of digital signal by adding non-information-carrying bits at a source interface and extracting the same bits at a destination interface.

Justification ratio. Ratio of the justification (pulse-stuffing) rate used to synchronize a tributary signal to the maximum possible justification rate allowed by a higher level multiplex format. (A justification ratio of $\frac{1}{3}$ implies one-third of the rate adjustment opportunities are null bits and two-thirds of the opportunities are tributary data bits.)

Key system. Small, customer premises telephone switching system that allows users to directly select telephone lines (and receive dial tone from a central office).

LAN Emulation. An ATM service offering that emulates Ethernet or token ring LANs.

Line administration. *See* Load balancing.

Line code. Set of electrical (or optical) waveforms chosen to represent data for the purpose of transmission.

Load balancing. Adjusting the assignment of very active customer lines so that all groups of customer lines in a multistage switch receive approximately the same amount of traffic.

Loading coils. Lumped-element inductors inserted at periodic points in cable pairs to flatten out their voice frequency response. Although loading coils improve voice frequency transmission, they severely attenuate higher frequencies as required for digital transmission.

Local access and transport area (LATA). Service area of a local exchange carrier (local operating company).

Longitudinal current. Electrical current passing along a pair of wires in a common direction. *Contrast* Metallic current.

Loop start. Off-hook signaling protocol involving the flow of dc current between the tip and the ring of a subscriber loop.

Loop timing. Synchronizing the transmit timing of a bidirectional link to the received timing of the same link.

Lost calls cleared (also loss system). Mode of operation in which blocked calls are rejected by the network and may or may not return.

Lost calls held. Mode of operation that holds blocked call requests until a channel becomes available. The portion of a call that gets blocked is lost.

M13 Multiplexer. Asynchronous network multiplexing equipment using bit stuffing for 28 individual 1.544-Mbps DS1 signals to create a 44.736-Mbps DS3 signal.

Main distributing frame (MDF). Framework used to cross connect outside plant cable pairs to central office switching equipment. The MDF provides protection and test access to the outside plant cable pairs.

Management information base (MIB). Collection of software defined objects that can be accessed via a network management protocol (e.g., SNMP or TMN).

Master clock. Particularly accurate and stable frequency source from which various nodes in a network extract their operating frequency (clock).

Master frame. Set of consecutive TDM frames that are identified by a master frame alignment signal (MAS).

Matching loss. Blocking within a multistage switch matrix resulting from at least one link in all suitable paths through the matrix being busy.

Maximum time interval error (MTIE). Largest time interval error occurring in all possible measurement intervals of length S within a measurement period T.

Message switching. Practice of transporting complete messages from a source to a destination in non–real time and without interaction between source and destination, usually in a store-and-forward fashion.

Message unit accounting. Activity-based billing as opposed to flat-rate billing, which is independent of usage.

Metallic current. Current flowing in opposite directions in a wire pair as a result of a difference signal propagating along the pair. *Contrast* Longitudinal current.

Minimum shift keying (MSK). Digital frequency shift modulation wherein the magnitude of a frequency shift is the minimum amount required to produce a 180° of phase shift in a symbol interval.

Modem. Contraction of the terms modulation and demodulation. A device used to generate "voicelike" data signals for transmission over telephone lines. A modem is referred to as a "data set" in older Bell System terminology.

Multifrequency (MF) signaling. Signaling method used for interoffice applications. MF signaling uses two of six possible tones to encode 10 digits and five special auxiliary signals.

Multipath. Transmission phenomenon in which a single transmit signal follows two or more separate paths to a receiver with differing delays.

Multiplexing. Process of combining multiple signals into a composite signal for transmission over common facilities.

Mutually synchronized network. Technique for synchronizing a network in which all nodes derive their operating frequency as the average of their own free-running clock frequency and the frequencies of some number of other nodes in the network.

Near-end crosstalk (NEXT). Unwanted energy coupled from a transmitter in one circuit into a receiver of another circuit at the same location (near end).

Network element (NE). Internal node of a SONET network whose interfaces are signals of the SONET hierarchy.

Network management. Network management is the function of supervising a communications network to ensure maximum utilization of the network under all conditions. Supervision requires monitoring, measuring, and, when necessary, action to control the flow of traffic.

Next generation digital loop carrier (NGDLC). A DLC system capable of using SONET transmission links and providing basic compliance to the Bellcore GR 303 standard.

Nonblocking. Switching network that always has a free path from any idle incoming trunk or line to any idle outgoing trunk or line.

Nonreturn to zero (NRZ). Line code that switches directly from one level to another. Each level is held for the duration of a signal interval.

Nyquist rate. (1) Minimum sampling rate required to extract all information in an analog waveform. The Nyquist rate is equal to twice the bandwidth of the signal being sampled. (2) In a baseband channel with a bandwidth B, the maximum rate ($2B$) that pulses can be transmitted without intersymbol interference.

Off hook. State or condition that a telephone receiver is requesting service or in use. Also a supervisory signal to indicate active status of a telephone or line.

Offered traffic. Amount of traffic carried by a system assuming the system has infinite capacity and therefore does not block or delay any calls.

On hook. Inactive status of a telephone or line.

One-way trunk. Trunk circuit that can be seized at only one end.

Open systems interconnection (OSI). Communications protocol reference model introduced by ISO consisting of seven layers (physical, link, network, transport, session, presentation, application).

Out-of-band signaling. Signaling technique that uses the same path as message traffic but a portion of the channel bandwidth above or below that used for voice.

Pair-gain system. Subscriber transmission system that serves a number of subscribers with a smaller number of wire pairs using concentration, multiplexing, or both.

Parity. Process of adding a redundant bit to a group of information bits to maintain either odd or even numbers of 1's in the composite group of bits. A parity error results if an odd number of 1's is detected when even parity is transmitted or vice versa.

Partial-response signaling (PRS). Use of controlled intersymbol interference to increase the signaling rate in a given bandwidth.

Permanent virtual circuit (PVC). Virtual circuit (X.25), virtual connection (Frame Relay), or virtual channel connection (ATM) that emulates a leased-line connection.

Personal communications system (PCS). Digital mobile cellular telephone system operating in 1900-MHz frequency bands in North America (1850–1910 MHz for mobile to base station 1930–1990 MHz for base station to mobile).

Per-trunk signaling. Method of signaling in which the control information pertaining to a particular call is transmitted over the same circuit (channel) that carries the call.

Phase distortion. Signal distortion resulting from nonuniform delay of frequencies within the passband.

Phase reversal keying (PRK). Special case of phase shift keying involving only two phases 180° apart.

Phase shift keying (PSK). Form of digital modulation that uses 2^n distinct phases to represent n bits of information in each signal interval.

Plesiochronous digital hierarchy (PDH). Designation of the original digital multiplexing hierarchy utilizing free-running clocks at all levels of the multiplexing hierarchy.

Plesiochronous. Method of network synchronization involving the use of independent (unsynchronized but highly accurate) clocks at the same nominal rate.

Pointer burst. In SONET, the occurrence of multiple pointer justification events (PJEs) of one polarity within the time constant of a desynchronizer circuit.

Pointer justification event (PJE). Change in a pointer value to accommodate phase drift.

Polar signaling. Two-level line coding for binary data using balanced (symmetric) positive and negative levels.

Power spectral density. Distribution of signal power as a function of frequency.

Primary reference source (PRS). Top level (Stratum 1) clock source with an accuracy of $\pm 1 \times 10^{-11}$.

Priority ringing. Feature for incoming call discrimination in which a telephone rings with a special ringing pattern for calls from a select set of phone numbers.

Private branch exchange (PBX). Switching equipment used by a company or organization to provide in-house switching and access to the public network.

Protection switching. Practice of transferring message traffic from one transmission link to a spare transmission link when the former is degraded in some fashion or for maintenance tests or equipment upgrades.

Provisioning. Allocating or deallocating facilities in support of a particular service offering (turning a service on or off).

Psophometric weighting. Noise-weighting filter recommended by CCITT.

Pulse amplitude modulation (PAM). Process of representing a continuous analog waveform with a series of discrete-time samples. The amplitudes of the samples are continuous and therefore analog in nature.

Pulse code modulation (PCM). Process of converting PAM samples into discrete levels represented by digital codewords.

Pulse stuffing. *See* Justification.

QPRS. Quadrature channel modulation using partial-response signaling on each channel.

Quadrature amplitude modulation (QAM). Independent amplitude modulation of two orthogonal channels using the same carrier frequency.

Quantization noise. Difference between the discrete sample value represented by a digital code and the original analog sample value.

Raised-cosine channel. Digital transmission channel with a particular type of pulse response that produces no intersymbol interference at the sample times of adjacent signaling intervals. The designation "raised cosine" is derived from the form of the analytical frequency spectrum (1 + cosine or cosine squared).

Recirculation. Passing a cell back to an input stage of a multiple-stage ATM switch when the cell would otherwise be discarded because it encountered blocking in an output stage.

Reframing time. Amount of time it takes to detect and synchronize to a frame alignment signal.

Regeneration. Process of recognizing and restoring a digital signal before perturbations of the signal accumulate to the point that bit errors occur.

Regenerative repeater. Device used to detect, amplify, reshape, and retransmit a digital bit stream.

Return loss. Difference in decibels between reflected and incident energy at a signal reflection point.

Ringback. Signaling tone returned by switching equipment to a caller indicating that a called party telephone is being alerted (ringing).

Ringing voltage. Low-frequency ac signal used to activate the ringer of a telephone (typically, a 20-Hz signal at 90 V rms).

Robbed bit signaling. North American practice of using the least significant PCM bit of every sixth frame for channel-associated signaling.

Router. Store-and-forward packet-switching element that processes packet headers and determines an appropriate output link for each packet using routing tables and routing protocols.

Severely errored seconds (SES). An (ESF) transmission link performance measure defined as the occurrence of 320 or more errored seconds.

Sidetone. Portion of a talker's signal that is purposely fed back to the earpiece so that the talker hears his or her own speech.

Signaling. Exchange of electrical information (other than by speech) specifically concerned with the establishment and control of connections and management in a communication network.

Simple network management protocol (SNMP). Standard protocol for performing network management functions (created by the Internet community).

Simplex transmission. Mode of operation involving transmission in one direction only.

Singing. Audible oscillation of a telephone circuit caused by a net amount of gain in a four-wire segment of the circuit.

Single mode fiber. Fiber that is narrow enough to preclude transmission of an optical signal along more than one path.

Single-frequency (SF) signaling. Method of conveying dial-pulse and supervision signals from one end of a trunk or line to the other using the presence or absence of a single specified frequency. A 2600-Hz tone is commonly used.

Slip. Deletion or repetition of data in a data stream caused by an overflow or an underflow of an elastic store due to variations in the write and read rates. A controlled slip is a deletion or repetition of an entire frame of data so the position of the framing bits is undisturbed.

Slope overload. Saturation of the range of difference signals in a differential encoder caused by a signal with a large, rapid change in amplitude.

SONET island. Collection of one or more SONET GNEs and NEs that create a continuous SONET path for a digital signal.

Space diversity. In radio systems, the use of two receiving antennas and possibly two separate receivers to provide protection against atmosphere-induced signal attenuation (fading).

Space expansion. Use of more center stages in a multistage switch than there are inputs to a first-stage array or outputs from a last-stage array.

Span line. Repeatered T1 line from end to end but not including channel banks.

Speed dialing. Use of a short address code to represent an often-called telephone number. The common control computer in a PBX or end office providing such a service translates the short code into the desired number.

Spread spectrum. Process of distributing an information-bearing signal across a bandwidth that is much wider than the inherent bandwidth of the signal.

Star network. Network with a single node to which all other nodes are connected.

Statistical multiplexor (Stat-mux). Multiplexer that assigns transmission capacity to a tributary only when the source of the tributary is actively transmitting information.

Step-by-step (SXS) switch. Electromechanical switching system using step-by-step switch elements to progressively set up a connection as each digit is dialed.

Stored-program control (SPC). Computer-controlled switching.

Stratum clock. Definition of free-run accuracy of a clock depending on stratum level in synchronization network.

Stuff Ratio. *See* Justification ratio.

Supervisory signal. Signal used to indicate the status of a line or to control equipment on the line.

Switched multimegabit data service (SMDS). Public, packet-switched wide-area data communications service offered by common carriers.

Switchhook. Hook or buttons upon which a telephone handset rests when it is not being used.

Symbol rate. *See* Baud rate.

Synchronizer. Circuitry that inserts a tributary digital signal into a higher level multiplexed signal and performs justification (pulse stuffing) in the process.

Synchronous network. Network in which the clocks of separate nodes operate at identical rates.

Synchronous transmission. Mode of digital transmission in which discrete signal elements (symbols) are transmitted at a fixed and continuous rate.

Synchronous. Mode of operation in which two or more pieces of equipment of network perform their operations at precisely the same rate for an indefinite amount of time. The rate of operation is derived by distribution of a timing reference.

Tag switching. Fast packet-switching process that integrates network layer (layer 3) routing and data link layer (layer 2) switching to simplify address and protocol processing.

Talker echo. Portion of a talker's voice that is returned to the talker and heard by the talker. The amount of annoyance to a talker is dependent on both the amount of delay and the amplitude of the returning signal.

Tandem office. In general, any intermediate switch used to establish a connection. In specific terminology, a tandem office is a switch used to interconnect end offices in an exchange area.

TASI. Time assignment speech interpolation, the practice of concentrating a group of voice signals onto a smaller group of channels by dynamically switching active voice signals to idle channels.

Ternary coding. Use of all states of a three-level code to send more than 1 bit of information in a single symbol. This is in contrast to bipolar coding, which uses three levels, but only one of two in any particular interval. One method of interfacing binary data to a ternary line code is to encode 4 bits with three ternary symbols (4B3T).

Tie line. Dedicated circuit connecting two private branch exchanges.

Time compression multiplex. Transmitting in two directions on a single transmission link using alternating directions of transmission ("ping-pong" transmission).

Time congestion. Ratio of time that all facilities of a system are busy (congested). Time congestion refers to the status of the system and does not necessarily imply that blocking occurs.

Time deviation (TDEV). Square root of the time variance (TVAR).

Time division multiplexing (TDM). Sharing a transmission link among multiple users by assigning time intervals to individual users during which they have the entire bandwidth of a system.

Time expansion. Use of more time slots on internal links of a switch than exist on external links.

Time interval error (TIE). Variation in time delay of a given signal relative to an ideal timing signal over a particular time period.

Time Variance (TVAR). Measure of the expected time variation of a signal as a function of time separation.

Traffic carried. Amount of traffic carried (by a group of circuits or a group of switches) during any period is the sum of the holding times expressed in hours (CCITT).

Traffic engineering. Network management activity that analyzes average and peak traffic loads to determine how to design or reconfigure a network to maximize the traffic capacity of the network and to ensure an acceptable value of blocking probability.

Traffic flow. Amount of traffic (in erlangs or CCS) carried or offered to a transmission link or switching system.

Transhybrid loss. Amount of isolation (in decibels) between go and return paths on the four-wire side of a four-wire to two-wire hybrid.

Transmission level point (TLP). Specification, in decibels, of the signal power at a point in a transmission system relative to the power of the same signal at a (hypothetical) zero transmission level point (0-TLP).

Transmultiplexer (transmux). Equipment that converts FDM voice signals to TDM voice signals and vice versa.

Transversal equalizer. Time-domain equalizer utilizing a tapped delay line and weighting coefficients at each of the taps to remove intersymbol interference.

Traveling class mark. Code that accompanies connection setup messages indicating the nature of the service request and any special provisions that may be desired.

Trunk. Circuit or channel between two switching systems.

Two-way trunk. Trunk circuit that can be seized at either end of the circuit.

Two-wire circuit. Circuit consisting of a single pair of wires and capable of simultaneously carrying two signals in opposite directions.

Unipolar. Binary line code using single polarity pulses and zero voltage for the two coding levels.

Unit interval (UI). Nominal difference in time between pulse positions at a specified data rate.

Unspecified bit rate (UBR). Class of service of ATM without specific quality assurances sometimes referred to as "best effort" service.

User datagram protocol (UDP). An Internet protocol without network delivery guarantees. (Lost or errored packets are not retransmitted.)

Variable bit rate (VBR). Class of service of ATM with quality assurances as long as the source adheres to preestablished average and peak data rates.

Vector quantization. Choosing one codeword representative of one discrete waveform of a set of discrete waveforms such that the selected waveform has the best match to a segment of the signal being encoded.

Vestigial sideband transmission. Form of single-sideband transmission that includes a vestige of the deleted sideband and a small amount of carrier energy.

Virtual circuit. Path through a packet-switched network through which all packets associated with a particular "connection" flow.

Waiting time jitter. Timing jitter caused by waiting for a timing adjustment opportunity beyond the time when the need for an adjustment arises.

Wander. Long-term variations of the significant instants of a digital signal from their ideal positions in time, where long term implies phase oscillations of frequency less than or equal to a demarcation point (typically 10 Hz) that is specified for each interface rate.

Wavelength division multiplexing (WDM). Use of multiple optical wavelengths to carry multiple optical signals.

White noise. Noise with a flat frequency spectrum across a band of interest in which samples taken at twice the bandwidth or lower have zero correlation.

Wide area telecommunications services (WATS). Service that permits customers to make (OUTWATS) or receive (INWATS or 800 Service) long-distance calls and to have them billed on a bulk basis rather than individually.

X.25. International protocol standard for accessing packet-switched networks.

ANSWERS TO SELECTED PROBLEMS

1.1 2000 pWC, $2(10)^{-6}$ mWC
1.3 500 pWC
1.5 250 pW at –3 dB TLP

3.1 1 kHz, 2 kHz, 3 kHz
3.2 2.43 dB degradation
3.4 49.4 Mbps
3.6 38.5 dB
3.8 0/110/0110
3.10 23.5 dB
3.14 (a) 13; (b) 33
(c)

Sample	Quantization Noise	Noise Power	SQR (dB)
30.2	.2	.04	43.6
123.2	–.8	.64	43.7
–2336.4	.4	.16	75.3
8080.9	.9	.81	79.1

3.17 (a) 3 dB (b) 1.76 dB
3.19 SQR increases by 20.6 dB
3.21 0/110/0001, 0/110/0001, 1/110/0001, and 1/110/0001 (noninverted)

4.1 37.5 bits
4.3 1.76 dB
4.5 (a) Average frame time = .097 seconds.
(b) Maximum average reframe time = 0.193 seconds.
4.7 141 bits
4.9 94 ms
4.11 (a) 1.99 dB (b) 3.78 dB
4.15 (a) 81 (b) 19 (c) 18

4.17

		+1	−3	+1	−1	+3	+3	−3
(a)	1+D		−2	−2	0	+2	+6	0
(b)	1−D		−4	+4	−2	+4	0	−6
(c)	$1-D^2$			0	+2	+2	+4	−6

5.1 250
5.2 (a) 621
5.3 (a) .11 (b) .25
5.5 (a) 55,296
5.7 (a) 41 (b) 10,828
5.9 (a) Total number of memory bits = 48,000.
(b) Complexity = 1504.
5.11 900 bits

6.1 (a) .5 dB degradation
6.2 New BER is $1(10)^{-8}$
6.4 1600 Hz
6.5 (a) 17% of peak signal power.
(b) 15.3 dB reduced error distance.
6.7 (a) 7.14 dB advantage for 32-QAM.
(b) 4.78 dB advantage for 32-QAM.
6.9 11.2 dB degradation in error distance
6.11 $1.5(10)^{-3}$
6.13 3 dB

7.1 195 Bits if the receiver knows the nominal data rate.
370 Bits to accommodate peak-to-peak shifts.
7.3 ±1.3 UI with 99% probability
7.5 $5.37(10)^{-5}$ misframes/sec (once every 5.2 hours)
7.7 25.6 dB relative to 1 UI^2
7.9 1.2%
7.11 (a) TIE = 44.736 UI (b) MTIE = 44.736 UI.

8.1 .188 dB/km
8.3 16 km
8.5 $\Delta\lambda$ at 1300 nm is 5 times larger than $\Delta\lambda$ at 1550 nm.
8.7 (a) 2m for mB1P (b) m + 1 for mB1C
8.9 (a) Minimum DS1 rate = 1.542 Mbps.
(b) Maximum DS1 rate = 1.546 Mbps.
8.11 (a) Minimum rate = 44.712 Mbps.
(b) Maximum rate = 44.784 Mbps.
8.13 68.7 μsec

9.1 6500 kbps.
9.3 11.4 kbps
9.5 Channel 1 output = +7; channel 2 output = +9
9.7 SIR = 12 dB

10.1 (a) 2.5 kbits/message (b) 100 bits/message
10.3 Total number of bits is 86 bits.
10.5 (a) 0.445 (b) delay is increased by 30 milliseconds

11.1 1250 meters
11.3 56 to 62.67 kbps

12.1 (a) 40 calls/hour (b) .989 (c) 9.5% (d) .19 E (e) .453
12.3 1.4 E
12.5 (a) 15 ports (14 ports with finite source analysis) (b) B = 9%
12.7 (a) 2.2% (b) 1%
12.8 (a) B = 30% (b) add 6 channels
12.10 (a) $A_{transit}$ = 32 E.
(b) 46 circuits required for B < .5%.
12.13 (a) B = 10% (b) B = .78%
12.14 2 WATS lines
12.16 (a) 83.3% (b) .833 sec (c) 30.6% (d) .833

INDEX