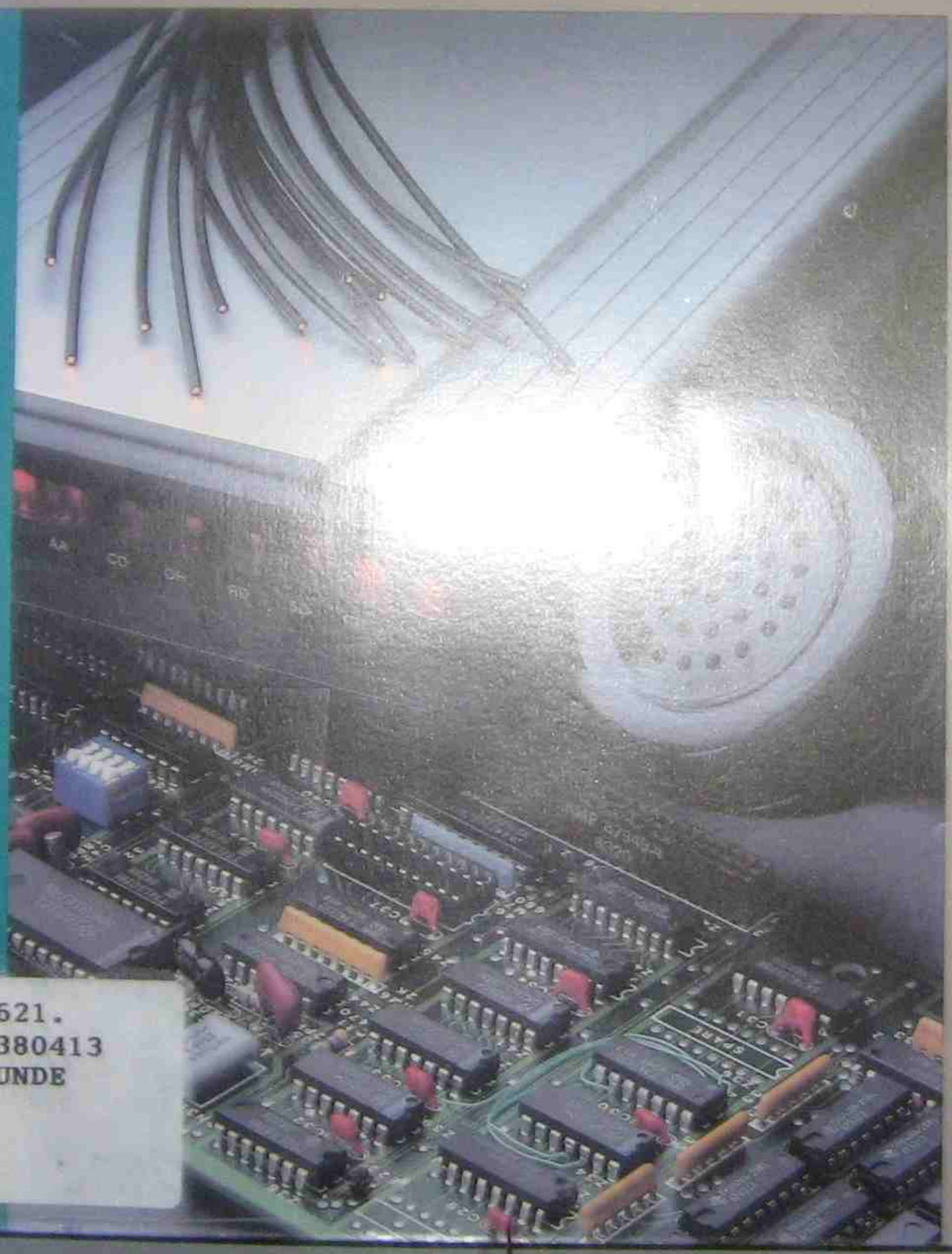
27270

Sams Understanding Series

Understanding Data Communications

TAFE NSW LIBRARY NETWORK
3 5555 07073679 0



Second Edition

Data Communications — the transmission of words or symbols from a source to a destination — is no longer exclusive to the business world. Its pervasive impact can be felt everywhere. Learn the basic principles, and a whole lot more, in this fully illustrated, easy-to-understand format that's ideal for self-paced, individualized learning.

By
George E. Friend
John L. Fike
H. Charles Baker
John C. Bellamy

Revised by:
Gilbert Held

621.
380413
UNDE

© 1988 by Howard W. Sams & Company
A Division of Macmillan Inc.

About the cover:

Represented on the cover are the components of modern day data communications: a telephone, a modem, optical fiber cables, integrated circuits, and the interconnecting printed wiring boards used to assemble the equipment.

SECOND EDITION
THIRD PRINTING—1989

All rights reserved. No part of this book shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher. No patent liability is assumed with respect to the use of the information contained herein. While every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions. Neither is any liability assumed for damages resulting from the use of the information contained herein.

International Standard Book Number: 0-672-27270-9
Library of Congress Catalog Card Number: 88-60044

Acquisitions Editor: *James S. Hill*
Manuscript Editor: *Katherine Stuart Ewing*
Compositor: *Shepard Poorman Communications, Indianapolis*

Printed in the United States of America

Table of Contents

Preface iv

Chapter		Page
1	*An Overview of Data Communications*	1
	Quiz	24
2	*Data Terminals*	25
	Quiz	53
3	*Messages and Transmission Channels*	55
	Quiz	74
4	*Asynchronous Modems and Interfaces*	75
	Quiz	114
5	*Synchronous Modems and Digital Transmission*	115
	Quiz	144
6	*Fiber Optic and Satellite Communications*	145
	Quiz	167
7	*Protocols and Error Control*	169
	Quiz	196
8	*Alternatives in Local Area Networks*	197
	Quiz	225
9	*Architectures and Packet Networks*	227
	Quiz	247
10	*Network Design and Management*	249
	Quiz	270
11	*ISDN*	271
	Quiz	278
	Glossary	279
	Bibliography	283
	Index	284
	Answers to Quizzes	291

Preface

The communication of information of all kinds by means of binary signals – the ones and zeros that are used by computers – has gained such an important place in American society that it is fair to say that successful operation of the U.S. economy would be impossible without it. Information sent using data communications controls a major portion of the long distance telephone network, enables the rapid authorization of credit purchases and cashing of checks, and provides for the inventory and ordering of goods in stores of all kinds from fast-food franchises to photo finishers. Data communications technology allows the simultaneous printing in widely separated cities of national magazines and newspapers, whose entire text and picture copy is transmitted from a central source to the city in which they are to be printed on a daily basis.

A basic knowledge of such a pervasive force in our lives is as necessary for all of us as some knowledge of the workings of another transportation system – the automobile. We have attempted to cover the basic principles of data communications and to explain those areas of application which have a daily impact: communications between terminals and computers, including local area networks and packet networks; communication of telephone conversations and control of the telephone network; and the use of light rays and satellites for carrying the ever-increasing volume of data helping to make our lives more comfortable and sometimes a great deal more confusing.

This book is arranged somewhat like a textbook. Each chapter starts by saying what it covers, ends by saying what has been covered, and provides a short multiple-choice quiz on the chapter contents. Like other books in the series, this book builds understanding step-by-step. Try to master each chapter before going on to the next one.

If your objective is to obtain general information about the subject, the more detailed portions of the text and the quiz can be bypassed. This will allow you to absorb the information more easily and, hopefully, enjoy the style of presentation. If you need more basic information about the telephone system and telephone electronics, a companion book in the series titled *Understanding Telephone Electronics* should be helpful. For more detailed and more advanced information about data communications, refer to the books listed in the bibliography.

G.F.
J.F.
C.B.
J.B.

An Overview of Data Communications

ABOUT THIS BOOK

This book is intended to explain how data communications systems and their various hardware and software components work. It is not intended to tell you how to build the equipment, nor will it tell you how to write computer programs to transmit data between two computers. What can be learned is a basis for understanding data communications systems in general, and some tips for setting up your own system to communicate between your personal or professional computer and another personal computer, a data base service, or an electronic bulletin board.

To understand data communications, it is necessary to have some understanding of the telephone channel because it is the medium used by most data communications systems to move information from one place to another. Therefore, part of this book is devoted to describing the public telephone network and the equipment used to interface between it and a computer.

ABOUT THIS CHAPTER

This chapter presents some history of data communications and a general description of a data communications system. Explanations of bits, bytes, two-state communications systems, and codes help lay the foundation for further discussions in the remainder of the book.

WHAT IS DATA COMMUNICATIONS?

Data communications is often referred to as computer communications due to the ever increasing use of computers and their support equipment.

Data communications is the process of communicating information in binary form between two points. Data communications is sometimes called computer communications because most of the information interchanged today is between computers, or between computers and their terminals, printers or other peripheral devices. The data may be as elementary as the binary symbols 1 and 0, or as complex as the characters represented by the keys on a typewriter keyboard. In any case, the characters or symbols represent information.

WHY IS DATA COMMUNICATIONS IMPORTANT?

It is important for us to understand data communications because of its significance in today's world. Data communications is commonly used in the world of business and it is being used more and more in homes as well. Whether it is the transmission of bank account information from a central computer to a convenient electronic teller machine, or the downloading of a video game from a computer bulletin board to a home computer's memory, data communications is becoming an integral part of our daily activities.

THE FIRST DATA COMMUNICATIONS SYSTEMS

Modern data communications involves the use of electrical or electronic apparatus to communicate information in the form of symbols and characters between two points. Since electricity, radio waves, and light waves are all forms of electromagnetic energy, it is stretching things only a bit to say that early forms of communication, such as the puffs of smoke from the Indian's signal fire (*Figure 1-1*), or reflections of sunlight from a hand-held mirror, also were forms of the same type of data communications. To carry this idea further, think of the puffs of smoke as discrete symbols, just as the symbols used in today's communications systems are discrete.

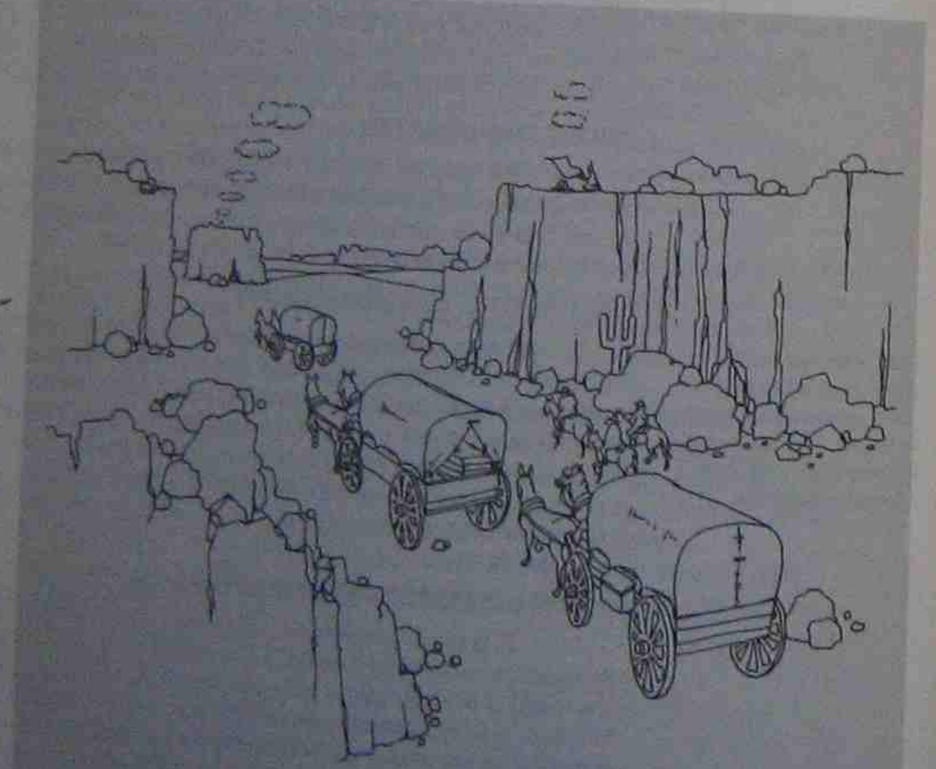
Early Uses of Electricity

Originally, data communications depended on codes transmitted by visual systems such as mirrors, flags, and smoke. Electrical data communication systems transmit codes by switching electrical current.

The discovery and harnessing of electricity introduced many new possibilities for communications codes beyond the smoke signals, mirrors, signal flags and lanterns in use in the eighteenth and nineteenth centuries.

One of the early proposals, submitted to a Scottish magazine in 1753, was simple but had profound implications for hardware. This idea was to run 26 parallel wires from town to town, one wire for each letter of the alphabet. A Swiss inventor built a prototype system based on this 26-wire principle, but the technology of wire making eliminated that idea from serious use.

Figure 1-1. Looks Like Trouble Ahead



The first data communications did not rely on electricity.

In 1833, Carl Friedrich Gauss used a code based on a 5 by 5 matrix of 25 letters (I and J were combined) to send messages by deflecting a needle from one to five times, right or left. The first set of deflections indicated the row; the second the column.

The Telegraph

Samuel F.B. Morse perfected the telegraph, the first mass data communications system based on electric power.

The first notable development in data communications occurred in the nineteenth century when an American, Samuel F. B. Morse, invented the electric telegraph. Although other inventors had worked on the idea of using electricity to communicate, Morse's invention was by far the most important because he coupled the human mind (intelligence) with the communications equipment.

A basic telegraph system is diagrammed in *Figure 1-2*. When the telegraph key at station A was depressed, current flowed through the system and the armature at station B was attracted to the coil, clicking as it struck the stop. When the key was released, it opened the electrical circuit and the armature of the sounder was forced to its open position by a spring, striking the other stop with a slightly different click. Thus, the telegraph sounder had two distinctive clicks. If the time between successive clicks of the sounder was short, it represented a dot; if longer, a dash. Morse developed a code, similar to the one shown in *Figure 1-3*, to represent characters by a series of these dots and dashes. The transmitting operator converted the characters in the words of a message to be sent into a series of dots and dashes. The receiving operator interpreted those dots and dashes as characters; thus, the information was transmitted from point A to point B.

Morse first developed the telegraph in 1832, but it wasn't until much later that he successfully demonstrated its use. The best-known demonstration took place in 1844, when Morse transmitted over a wire from Washington to Baltimore the message, "What hath God wrought!"

The telegraph system was the first electrically based communications system to connect the east and west coasts of the United States, and both sides of the Atlantic.

Since mail delivery by the Pony Express was the typical means of communication before the telegraph, the telegraph quickly became a success because of its much greater speed. The equipment was simple and rugged; the key and the sounder each contained only one moving part. Both the system's strength and weakness (and its only real complexity) was the human mind — the transmitting and receiving operators. By the time of the Civil War, a telegraph line spanned the continent, crossing the prairies and deserts to connect California with the rest of the United States. It was from this historic technological breakthrough that the Western Union Telegraph Company made its legacy and indeed was named; their line connected the *West* with the *Union*. By the time the telephone was invented some thirty years later, the telegraph industry was large and prosperous, with many companies providing service to almost every city and town in the United States. In 1866, the telegraph connected the nations of the world with the laying of the trans-Atlantic cable between the United States and France.

Figure 1-2.
Basic Telegraph System

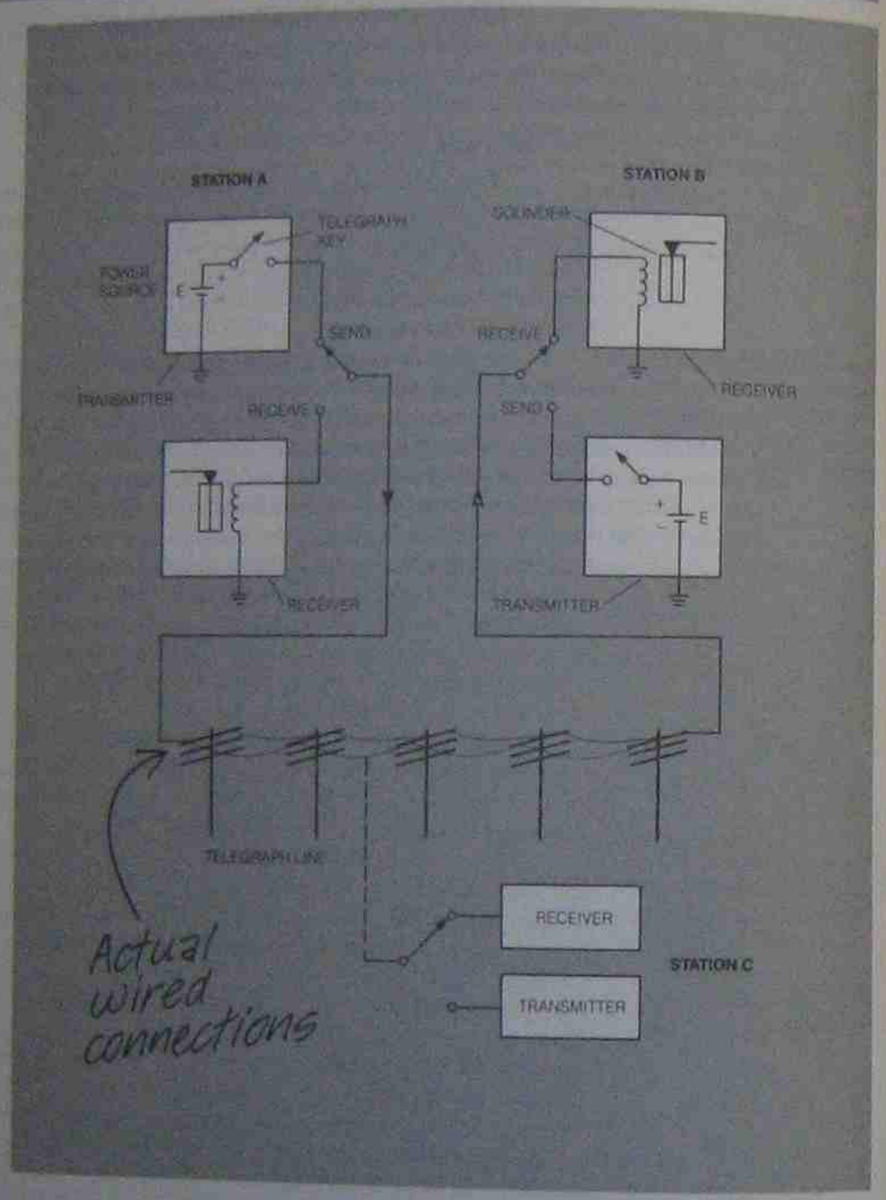
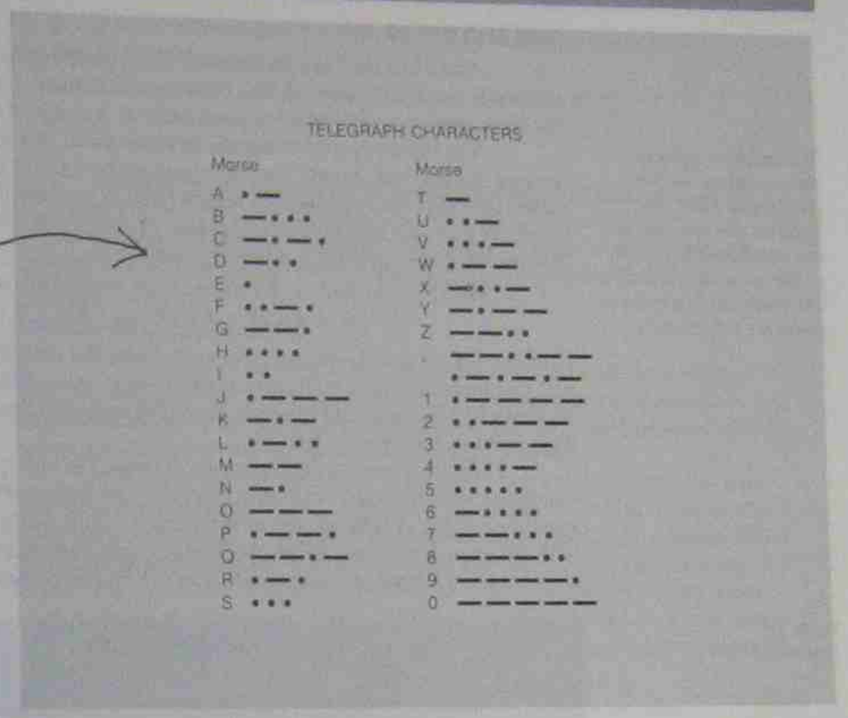


Figure 1-3.
International Morse Code



Much of Morse terminology is still in use today.

TWO-STATE COMMUNICATIONS SYSTEMS

Much of the telegraph's terminology and principles of operation are still used today. The most important of these is the two-state communications system.

The importance of the Morse telegraph is not just historical; it illustrates the simplicity of a complete data communications system. Much of the terminology that developed around the Morse system still is in use today. For example, consider the terms "mark" and "space." If a device were arranged so that paper continually moved under a pen attached to the telegraph sounder armature, then a mark would be made on the paper when the armature was attracted to the coil. Thus, we could refer to the state of current flow in the line as the marking state and the state of no current flow in the line as the spacing state. Worldwide standards for data communications today still use the terms mark and space, with the idle condition on the transmission channel called the marking condition.

The wire ("telegraph channel") between the operators, therefore, is in one of two states; either current is flowing, or it is not. This illustrates a simple idea which has been repeated over and over again in the development of data communications systems. A two-state communications system is the simplest, the easiest to build, and the most reliable. The two states can be On and Off (as in the telegraph), or Plus and Minus (with current flowing in opposite directions), or Light and Dark (like turning a flashlight on and off to send code), or 1 and 0 (the concept used in the computer), or some other design with only two possible values. A two-state or two-valued system is referred to as a binary system.

BITS AND BYTES

Transmission codes in the binary system are made up of digits called bits. Each bit can have one of two possible states (high or low, on or off). Several bits combined in a uniform group are called a byte.

The 0 and the 1 are the symbols of the binary system. A binary digit is commonly referred to as a bit. The individual line changes in digital data (like the mark and space) are called bits and each bit is assigned a value of 0 or 1.

The binary system uses positional notation, just as the familiar decimal system does, except that each position has only two possible values, rather than ten. For example, the number 345.27 in decimal means three hundreds plus four tens plus five ones plus two tenths plus seven hundredths. The *weight* of each position is a power of ten, thus:

$$345.27 = (3\times100) + (4\times10) + (5\times1) + (2\times1/10) + (7\times1/100)$$
$$= 3\times10^{2} + 4\times10^{1} + 5\times10^{0} + 2\times10^{-1} + 7\times10^{-2}$$

Note the use of positive and negative exponents, and remember the definition that *any* number raised to the zero power is equal to 1. In the decimal system, as shown by the preceding example, ten is the *base*, and the weight of each digit position is a power of ten.

Similarly, we can define powers of two in the binary positional system, where the weight of each position is two times the weight of the one to the immediate right. A table of powers of two can be quite useful when dealing with binary numbers because most of us don't want to memorize them. *Table 1-1* is such a table for powers of two up to 2^8 which will be sufficient for this book.

Table 1-1. Powers of Two

Power of Two	Positional Weight	9-Bit Binary Number
2^0	1	000000001
2^1	2	000000010
2^2	4	000000100
2^3	8	000001000
2^4	16	000010000
2^5	32	000100000
2^6	64	001000000
2^7	128	010000000
2^8	256	100000000

Using this table makes it easy to convert a binary number to its decimal equivalent. For example;

$$1101001 = 1\times2^{6} + 1\times2^{5} + 0\times2^{4} + 1\times2^{3} + 0\times2^{2} + 0\times2^{1} + 1\times2^{0}$$
$$= 64 + 32 + 0 + 8 + 0 + 0 + 1$$
$$= 105 \text{ decimal}$$

Computers are more efficient if their internal paths and registers are some power of two in length (two bits, four bits, eight bits, sixteen bits, etc.). A commonly used grouping is eight bits and this 8-bit group is called a byte or an octet. From *Table 1-1*, you can see that a byte can represent 2^8 or 256 unique sequences of 1s and 0s.

COMMUNICATIONS CODES

The use of simplified and standardized binary codes allowed information to be encoded and decoded by mechanical or electrical means, and made it possible to automate data communications.

A common characteristic of data communications systems is the use of an intelligent device to convert a character or symbol into coded form and vice versa. In the Morse system, the intelligent "devices" were the telegraph operators who converted the characters into dots and dashes.

Skilled telegraph operators were always in short supply and the work was difficult and exhausting, so an electrical or mechanical means of coding the characters was needed. However, it was essentially impossible to automate the transmitting and receiving operations because of the varying duration of the dots and dashes in the Morse code, and the fact that the codes for the characters were made up of different quantities of dots and dashes. Therefore, a code that had an equal number of equal duration signaling elements for each character was needed.

Some Definitions

A code is a previously agreed upon set of meanings that define a given set of symbols and characters; in data communications these codes are built into the equipment.

Characters are the letters, signs and symbols on an input device's keyboard; some of these do not print, but are used to control the system.

Before we go any further, let's establish the following definitions in order to discuss communication codes:

Codes: Standard (agreed-upon-in-advance) meanings between signaling elements and characters. The key idea is *standard*. The codes used in data communications systems are already defined and the code set is built into the equipment. About the only time the user might need to deal with codes is when interfacing two machines (computers, printers, etc.) from different manufacturers.

Characters: The letters, numerals, space, punctuation marks, and other signs and symbols on a keyboard. (Remember that the space character is just as important as any other even though we tend to think of it as "nothing" or "blank.") For example, A 7# is a sequence of four characters. Communication systems also use control characters which do not print, but they also must be coded. Some of these (such as Carriage Return or Tab) may be on a keyboard, but many are not.

Symbols are the representations of characters that are transmitted over transmission lines. It is easier to design and build machines to recognize symbols rather than characters.

Signaling Elements: Something that is sent over a transmission channel and used to represent a character. The dots and dashes (or marks and spaces) of the Morse code are signaling elements, as are the ones and zeros in this sequence:

0100000101 0000001011 0111011011 0110001011

This is the way A 7# might look when transmitted between a personal computer and another computer or printer. You will see later that the code is ASCII, with even parity and one start and one stop bit.

The definitions for characters and signaling elements illustrate why machines and people need different ways to represent information. People quickly and reliably recognize printed characters by their distinctive shapes, but that is difficult and expensive for a machine to do. On the other hand, machines can easily handle long strings of two-valued signaling elements such as marks and spaces or ones and zeros, but that is hard for a person to accomplish with any accuracy.

Baudot Code

Emile Baudot developed one of the more successful codes suited for machine encoding and decoding. However, it was limited because it used only five symbols.

As stated above, the Morse code was unsuitable for machine encoding and decoding because of the problems caused by the varying lengths of the codes for the characters. When, early in the twentieth century, interest developed in replacing human telegraph operators with machines, several suitable codes already existed. The most prominent of these had been invented in the 1870s by a Frenchman named Emile Baudot. Since Baudot's code, similar to the one shown in *Figure 1-4*, used the same number of signaling elements (marks and spaces) to represent each character, it was better suited to machine encoding and decoding. Unfortunately, the number of signaling elements was limited to five by problems in timing the electromechanical devices. The 5-bit code could generate only 32 possible combinations, fewer than was needed to represent the 26 characters of the alphabet, the 10 decimal digits, the punctuation marks, and the space character. In addition, Baudot used two shift-control characters — the letters (LTRS) shift and the figures (FIGS) shift — to permit the code set to represent all the characters which seemed necessary at the time. The shift codes do not represent printable characters, but select one of two character sets, each composed of 26 to 28 characters. Receipt of the letters shift code causes all following codes to be interpreted as letters of the alphabet; receipt of the figures shift code causes all the following characters to be interpreted as numerals and punctuation marks. Notice that the LTRS and FIGS codes, as well as the other control codes and the space character, always have the same interpretation no matter which shift mode the machine is currently in. Although Baudot's invention did not immediately revolutionize telegraphy (because of the difficulty that human operators had in sending equal length codes), it did provide a basis for the later development of the teleprinter.

Figure 1-4. Five-Level Teleprinter Code

Code Signals • Denotes positive current							LTRS Shift	FIGS Shift	
Start	1	2	3	4	5	Stop		CCITT Standard International Telegraph Alphabet No. 2 Used for Telex	North American Teletype Commercial Keyboard
	•	•				•	A	–	–
	•			•	•	•	B	?	?
		•	•	•		•	C		
	•			•		•	D	Who are you?	$
	•					•	E	3	3
	•		•	•		•	F	Note 1	!
		•		•	•	•	G	Note 1	&
			•		•	•	H	Note 1	#
		•	•			•	I	8	8
	•	•		•		•	J	Bell	Bell
	•	•	•	•		•	K	(	(
		•			•	•	L	)	)
			•	•	•	•	M	.	.
			•	•		•	N	,	,
				•	•	•	O	9	9
		•	•		•	•	P	0	0
	•	•	•		•	•	Q	1	1
		•		•		•	R	4	4
	•		•			•	S	'	'
					•	•	T	5	5
	•	•	•			•	U	7	7
		•	•	•	•	•	V	=	;
	•	•			•	•	W	2	2
	•		•	•	•	•	X	/	/
	•		•		•	•	Y	6	6
	•				•	•	Z	+	"
						•	Blank		
	•	•	•	•	•	•	Letters shift (LTRS)		
	•	•		•	•	•	Figures shift (FIGS)		
			•			•	Space		
				•		•	Carriage return		
		•				•	Line feed		

Notes:
1. Not allocated internationally; available to each county for internal use.

There was a need for new codes that could represent all characters, be able to check for errors, and leave room for further expansion.

Modern Codes

The Baudot code and variations of it were the backbone of communications for almost half a century, but they clearly left much to be desired. The newspaper industry found the lack of differentiation between upper and lower case letters to be a problem and devised a six-level code to designate the difference between upper and lower case letters. This was just one example of a general need: modern communications required a code that could represent all printable characters and still leave room for error checking. The code had to permit decoding without reliance on correct reception of previous transmissions, and also had to permit decoding by machine. Perhaps most important of all, the new code needed to be expandable.

During the 1960s, a number of data transmission codes were developed. Most of these have fallen by the wayside, leaving three predominant codes: a single 5-bit code (CCITT International Alphabet No. 2) still used for telex transmission; the Extended Binary-Coded-Decimal Interchange Code (EBCDIC, pronounced "eb-see-dik") developed by IBM and primarily used for synchronous communication in systems attached to large mainframe computers; and the American Standard Code for Information Interchange (ASCII, pronounced "as-key") code defined by the American National Standards Institute (ANSI) in the United States and by the International Standards Organization (ISO) worldwide.

EBCDIC is a modern code using 8-bits to represent 256 characters. It was developed by IBM to provide a standard code for its own products.

EBCDIC

When a clear need arises for standardization, standards can come into existence in two ways. A single manufacturer (especially a dominant one) can define a standard for its own products, and the rest of the industry may follow. This is what IBM® did. It created the EBCDIC 8-bit code, allowing 256 characters to be represented. The world would probably be better off if EBCDIC had become the standard, because it has enough unique characters to allow almost any representation. However, only IBM and firms who build IBM-compatible equipment adopted EBCDIC. Since EBCDIC is primarily used in large IBM-compatible computing systems, it will not be used in most of the examples in this book.

The ANSI developed the 7-bit ASCII data communications code that is in general use today. It can represent 128 characters.

ASCII

The other, more common method of creating a standard is through a committee, which serves as a forum for examination of needs, discussion between interested parties, and compromise. This process produced the ASCII 7-bit code, formally known as ANSI standard X3.4-1977. ASCII can represent 128 characters, but not all of them represent printed symbols. Included in the character set are all the letters of the English alphabet (both uppercase and lowercase), the numerals 0-9, punctuation marks, and many symbols. This standard code set is used in virtually all small computers and their peripherals, as well as in large computers in most of the world.

®IBM is a trademark of International Business Machines Corporation.

Figure 1-5 shows the ASCII characters and their associated codes. Compare this chart to the Morse code in *Figure 1-3* and the 5-bit teleprinter code in *Figure 1-4*. The Morse code had a varying number of elements (dots and dashes) for each character, and was quite restricted — only letters, numerals, and a few punctuation marks. The 5-bit code had a constant number of elements and a few more special characters, but still couldn't distinguish between upper case and lower case letters.

ASCII not only has upper and lower case, but also has a great deal of regularity that may not be readily apparent. For example, to convert any upper case alphabetic character (A through Z) to lower case, it is only necessary to change bit 6 from zero to one! Another feature is that bits 4 thorough 1 of the numeric characters (0, 1, 2, 3, 4, 5, 6, 7, 8 and 9) are the binary-coded-decimal (BCD) value of the character. Another advantage of ASCII is that 128 different characters can be represented by the seven bits used in the code, rather than only 32 characters from the 5-bit code.

The ASCII format is arranged so that lower case letters can be changed to upper case by changing only one bit. Bits 4 through 1 of the numeric characters also are the BCD value of the number.

Figure 1-5. American Standard Code for Information Interchange (ASCII)

Bit Positions: 7	0	0	0	0	1	1	1	1
6	0	0	1	1	0	0	1	1
5	0	1	0	1	0	1	0	1
4 3 2 1								
0 0 0 0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	STX	DC2	"	2	B	R	b	r
0 0 1 1	ETX	DC3	#	3	C	S	c	s
0 1 0 0	EOT	DC4	$	4	D	T	d	t
0 1 0 1	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	ACK	SYN	&	6	F	V	f	v
0 1 1 1	BEL	ETB	'	7	G	W	g	w
1 0 0 0	BS	CAN	(	8	H	X	h	x
1 0 0 1	HT	EM	)	9	I	Y	i	y
1 0 1 0	LF	SUB	*	:	J	Z	j	z
1 0 1 1	VT	ESC	+	;	K	[	k	{
1 1 0 0	FF	FS	,	<	L	\	l	|
1 1 0 1	CR	GS	-	=	M	]	m	}
1 1 1 0	SO	RS	.	>	N	^	n	~
1 1 1 1	SI	US	/	?	O	_	o	DEL

Non-Printable Control Characters

Sample of Control Characters (Bold)
STX = Start of text
EOT = End of transmission
CR = Carriage return
HT = Horizontal tabulation

Examples:
1000011 = C
0110011 = 3
1010000 = P
0110000 = 0 (Zero)
0100000 = SP (space)

ASCII also has several control codes which, in addition to their defined control functions, have been used by some manufacturers to represent specialized functions.

The two left-most character columns in the chart of *Figure 1-5* represent the nonprinting control characters which may be used to control the operation of the receiving device. For example, the control codes for carriage return (CR) and linefeed (LF), which are commonly used on a typewriter, are shown. Other control codes include formfeed, bell, horizontal tab and vertical tab. These control codes were designed for printing or display devices, although some manufacturers have used the control codes for all manner of special functions. Also, some codes control how a receiving device will interpret subsequent codes in a multiple character function or command. Two shift characters called Shift In (SI) and Shift Out (SO) are used to shift between ASCII and character sets other than those used in English. ANSI standards X3.41-1974 and X3.64-1979 expand the definition of the escape (ESC) control code for even greater flexibility. Other control codes delimit text, such as start of text (STX) and end of text (ETX). These codes are used primarily in block or synchronous data transmission, and you'll see more of them in later chapters.

Escape Character

Escape sequences are code sequences made up of non-control characters that are to be interpreted as control codes.

The escape (ESC) character designates that the codes that follow have special meaning. Characters received in an escape sequence are not interpreted as printing characters, but as control information to extend the range of the "standard" character set by allowing other definitions. The escape character has the effect of making all character codes available for control of a device. Graphics characters, foreign language character sets, and special applications sets have been developed which are accessible via escape character sequences; thus, they permit a much richer variety of displayed symbols than is possible with any single code.

The CRT terminal probably has benefited most by the escape sequence. The serial communications link to these terminals is the same as for a teleprinter, and ordinarily, any characters received via this channel are displayed on the terminal screen as expected. But the people who developed the ASCII standard did not forsee (thus did not make provision for) capabilities for character and line deletion, and display enhancements such as inverse video, underlining, and blinking that are available on the CRT terminal. Unfortunately, little standardization of these sequences existed until the ANSI X3.64 standard came out in 1979. Before then, and without standardization, designers felt free to exercise their creativity. For example, one major feature now found on most video display terminals is absolute cursor positioning. The computer can send a command to the terminal that will place the cursor anywhere on the screen. This capability is important for many types of form-filling operations. Unfortunately, there are almost as many escape sequences to do cursor positioning as there are terminal manufacturers. Even different models in a manufacturer's line may use different escape sequences to do the same thing.

The result of this "creativity" with the character set is that some manufacturers' equipment will not operate correctly with that of most of the rest of the world. An example is a printer that automatically inserts a linefeed after receiving a carriage return. Since most computers send both a carriage return and linefeed in response to only a carriage return input, the printout will not be spaced as desired because of the extra linefeed supplied by the printer.

The technology and techniques used in teleprinting have provided the basis for asynchronous data communication.

Teleprinters

The teleprinter was the next major step after the telegraph in data communications. Teleprinter equipment has been the backbone of non-voice business communications for over half a century. As recently as the mid-1970s, a teletypewriter (TTY), which is a teleprinter with a keyboard for input, was the standard terminal for small- and medium-sized computers. Many companies have nationwide and even worldwide private teleprinter networks. Two nationwide public teleprinter networks are the TWX and TELEX services. Several common carriers will transmit telex messages to any teleprinter in the United States, and companies called "international record carriers" will deliver them to any teleprinter in the world. A company's telex number can be found in a directory similar to a telephone directory.

Like the telegraph, teleprinters are important not only in their own right because they were the principal data communication method for almost fifty years, but also because most of the standards and terminology for low-speed or asynchronous data communications came from the world of teleprinters. We'll discuss asynchronous transmission more thoroughly in Chapter 2; for now, it is sufficient to realize that the transmission of each of the coded characters begins with a "start" symbol and ends with a "stop" symbol. Although this technique permitted synchronization of teleprinter equipment thousands of miles apart, it also added to the already high overhead (in terms of extra time) required for transmission of each character.

Teleprinter equipment was inherently limited to slow transmission speeds because it was electromechanical: solenoids had to be energized to attract armatures; motors had to start, turn and stop; and clutches had to engage and disengage. All of these mechanical operations limited the maximum speed at which telepinters could operate reliably to less than thirty characters per second. International telex operates even slower—less than ten characters per second! Either speed is much too slow for a computer to communicate efficiently with another computer, or even with a modern printer or CRT terminal.

DATA COMMUNICATIONS IN COMPUTING

The introduction of the electronic computer provided greater speed and capabilities to the data communications system.

Further improvements in data communications were necessitated by the widespread use of electronic computers which were introduced in the early 1950s. These computer systems were capable of storing large quantities of information and of processing it quickly. Input and output equipment were much improved over the teletypewriter and teleprinter so they could operate much faster. As more computers were used, it became necessary for computers to communicate with each other. Since this communication did not need any electromechanical equipment, data transfer theoretically could occur at extremely fast rates. Practically, however, data transfer rates are limited by the transmission medium — primarily the public telephone network.

The 1950s

The first computer systems had little need for data communication because they were located close together, I/O was usually direct, and information was processed in discrete batches.

The typical computer system of the 1950s, shown in *Figure 1-6*, used punched cards for input, printers for output, and reels of magnetic tape for "permanent" mass storage. There was little or no data communications in these systems because the input devices, output devices and computer were all located close together and were directly connected by short cables. Information was processed on a "one job at a time" or "batch" basis.



Figure 1-6. Typical Data Processing System of the 1950s
(Courtesy of International Business Machines Corporation)

For example, a wholesale grocery company would receive orders by mail and telephone during the day. At the end of the day, these would be collected and punched into cards which were then assembled as a batch. The data on the cards were read into the computer for processing at night. During the processing, the computer would do such things as check the customer's credit. If satisfactory, the computer would generate a pick list for the warehouse so the order could be filled and shipped the next day. The quantity ordered for the different types of items would be deducted from the quantity on hand, and the invoices and various warehouse tickets would be printed. This system worked well except that the person taking the order from the customer during the day, whether on the phone or at the customer's premises, was never certain whether the ordered items were available in the warehouse. Although an updated inventory report was available each morning, a running inventory was not kept during the day as orders were received.

The 1960s

In the 1960s, the use of outlying, on-line terminals required communications links to connect them to the main computer.

Batch processing systems such as the one just described are the most efficient in the use of computer time and equipment; however, they are relatively inefficient in the use of the order clerks, sales people and other resources of the company. While the costs of people were increasing, the solid-state revolution caused the cost of computer logic and memory to decrease sharply. This meant that companies could invest in more computer equipment to reduce the costs of personnel, as well as to make themselves more competitive in business.

In the 1960s, batch processing was largely replaced by on-line processing (*Figure 1-7*). In the example of the wholesale grocery company, the clerks took the orders over the phone and used on-line terminals to enter the orders into the computer at the same time. Some of these terminals were directly connected to the computer by parallel communication channels, but many of the terminals were teleprinter devices which used serial data communications over dedicated private cables to the computer.

As the order was entered, the computer checked the inventory to see if the goods were on hand. If so, the customer was given a delivery date; if not, he was told that the goods were out of stock, and was given the option of placing a back order or selecting another item. While more expensive than batch processing in terms of computer time and equipment, this method of doing business made the wholesale grocery company more competitive because it gave the customer better service.



Figure 1-7.
Typical Data Processing System of the 1960s
(Courtesy of Control Data Corporation Archives)

he 1970s, the introduc-
of the minicomputer
the very portable mi-
omputer required that
e units have increased
munications with their
nframe computer.

The 1970s

The 1970s changed this process still further (*Figure 1-8*). Other computers, sometimes called minicomputers, could communicate with the larger computer in the wholesale grocer's central office. The on-line order terminals were not necessarily in the room next to the computer; they might have been on a different floor, in a different building, or even in a different city. Toward the end of the decade, even smaller computers, called microcomputers or personal computers, began to be seen not only in the homes of some of the employees, but also on their desks at the office. Sometimes the employees would communicate with the main computer through these smaller computers. The company salesmen began carrying portable computer terminals with them on their sales calls. They could enter their orders from the customers' office directly into the main computer over a standard telephone line.

True data communications was everywhere in the system. There was more equipment around than anyone would have thought possibe in the 1950s — more computers, more terminals, more communications channels — and businesses were run more efficiently because of it.

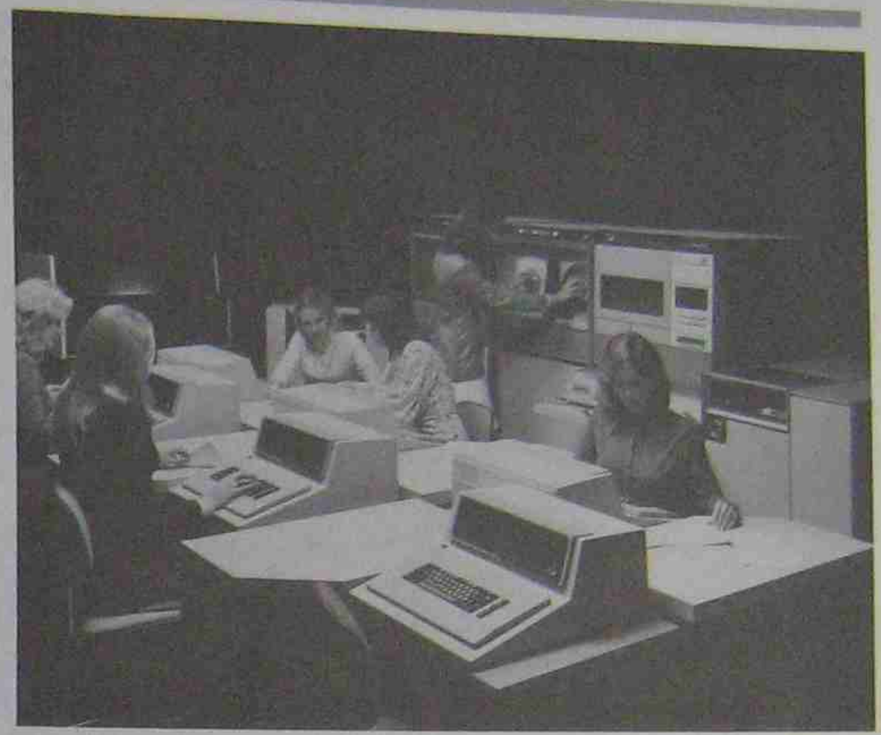


Figure 1-8.
Typical Data Processing System of the 1970s
(Courtesy of Mohawk Data Sciences Corporation)

Advances in technology increased the capabilities, reduced the size and power requirements of computers, and provided a new communications link — the satellite.

The 1980s

As the wholesale grocery company grew, they decided to produce some of the food products that they sold, so they set up producing operations in California, Texas and Florida. Because their wholesale business is primarily in the northeast, they have major warehouse facilities in New Jersey, Pennsylvania and Ohio, and sales offices in Chicago, Philadelphia and New York City. The company headquarters is also in New York City.

With the expansion to the south and west, the company's data communications system also needed expansion. They decided to use the relatively new satellite communications to tie the New York office with the Texas and California operations. (Satellite communications are discussed in later chapters.) The satellite could be used for other points as well, but the company (for various reasons) uses leased terrestrial circuits to the Florida, Chicago and Philadelphia sites. The smaller sales offices in the states where the company operates are tied together with a network of low-speed teletypewriter-like terminals that allow order entry and some communication with the home office. The company's salesmen still carry portable terminals, but advances in technology have reduced the size of the terminal while increasing its capabilities. A typical one is shown in *Figure 1-9*.

Figure 1-9.
TI 703KSR Teleprinter



As low-cost personal computers became more widely available during the mid-1980s, many managers acquired them for departmental use. Although the original intention was to use the local processing capability of personal computers for such tasks as budgeting with spreadsheet programs and word processing, corporate management quickly began to integrate personal computers into their company communications networks. Some personal computers replaced CRT terminals and served as both terminals and local processing devices. Other PCs were linked together into a local area network, which enabled users to share peripheral devices and provided the capability of sending electronic messages to other users on the network.

Introduced in mid-1981, more than ten million IBM PC and compatible computers had been manufactured by 1987, when the Personal System/2 Model 30 computer replaced the IBM PC. *Figure 1-10* illustrates the major components of an IBM PC computer system. Today, the personal computer is as prevalent in many organizations as is the office typewriter.

Figure 1-10.
IBM PC System
(Courtesy of IBM Corporation.)



Besides the technological revolution, the breakup of the Bell System changed the way business used and paid for communications.

CHANGES IN THE INDUSTRIES

The hardware and software were not the only things that were changing. Both the computer industry and the communications industry were changing in ways that would have been unimaginable twenty years earlier. The rise of the minicomputer, and later the microprocessor in personal computers, and programmable controllers for equipment, created a situation where computing power was rather inexpensive. The suppliers of large main frame computers, such as IBM, UNISYS, Control Data and Honeywell, were joined by companies such as Texas Instruments, Apple, Radio Shack and Commodore, that had not previously made computers. With many more computers used by many more people to create and use much more data, interest in data communications increased. In communications, the advent of satellites followed by optical fiber transmission offered the promise of far greater bandwidths, thus, faster data transmission speeds.

The revolution in computing and data transmission was paralleled by an upheaval in the structure of the communications industry. The telephone companies' monopoly of almost a century disappeared forever, beginning in the late 1960s and accelerating until the Bell System was broken up in 1984. For the first time, other firms not only could compete with the telephone companies in offering long distance service; but they also could sell equipment, including data communications devices, to be connected to the telephone network.

These two forces, the advent of the microprocessor-based personal computer and the increase in competition in both the computing and communications industries, brought about dramatic changes and accelerated technological advances in the communications field. They also brought communications and computing closer together. In the example wholesale grocery company of the 1980s, it is hard to say where communications stops and computing begins. Is the salesman's portable terminal or the manager's

personal computer a communications device or a computer? The answer, of course, is that they are both, because they do whatever combination of tasks is required for the job at hand.

In modern computer-communications systems, we can interconnect equipment of many different sizes and capabilities from a variety of suppliers. One of the purposes of this book is to explain how to interconnect, or interface, these different kinds of equipment so it all works together in a system.

GENERAL DESCRIPTION OF A DATA COMMUNICATIONS SYSTEM

A data communications system can be described simply in terms of three components: the transmitter (also called the "source"), the transmission path (usually called the channel, but sometimes the line), and the receiver (occasionally called the "sink"). In two-way communications, however, the source and the sink may interchange roles; that is, the same piece of equipment may transmit and receive data simultaneously. Therefore, it is easier to think of a data communications system between point A and point B *(Figure 1-11)* in terms of the Universal Seven-Part Data Circuit which consists of:

The most accurate way to separate a data communications system's parts is in terms of the Universal Seven-Part Data Circuit.

1. The data terminal equipment (DTE) at point A.
2. The interface between the DTE and the data circuit-terminating equipment (or data communications equipment) (DCE) at point A.
3. The DCE at point A.
4. The transmission channel between point A and point B.
5. The DCE at point B.
6. The DCE-DTE interface at point B.
7. The DTE at point B.

The function of the various parts of the data communications system, both hardware and software, now can be described more easily. The DTE is the source, the sink, or both in the system. It transmits and/or receives data by utilizing the DCE and data transmission channel. Don't be misled by the name "data terminal equipment"; these could indeed be CRT or teletype terminals, but they also may be personal computers, printers, front-end processors for large mainframe computers, or any other device that can transmit or receive data. The whole purpose of the data communications system is to transmit useful information between point A and point B; the information may be used directly by the DTE, or the DTE may process and display the information for use by human operators.

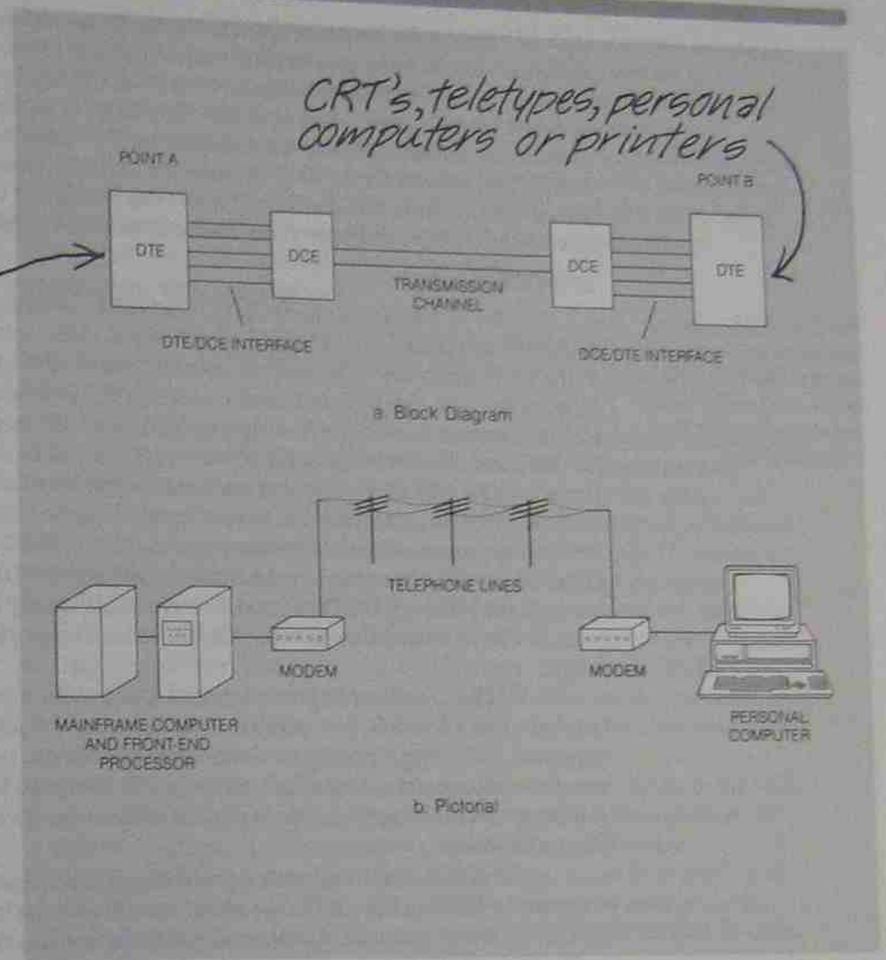
Form and Content of Information

The data communications system is only concerned with the correct transmission of data, in its coded form, between two points.

The DCE and the transmission channel perform the function of moving the data from point A to point B. In general, they neither know nor care about the content of the information transmitted; it could be stock market quotations, a display for a video game, or recipes for Aunt Martha's fudge.

This brings us to an important point about data communications — the difference between the *form* and *content* of the information transmitted. The form of the information might be English language text represented by code for a telegram; the content might advise you that a rich uncle had just left you a million dollars. Clearly, the average user of data communications is much

Figure 1-11. The Universal Seven-Part Data Circuit



more interested in the content of the information, and really is not concerned about the mechanics of the communication process so long as the information is received correctly. It would be rather disappointing if, in fact, your uncle had left you only $100.00, but because of errors in transmission, the telegram you received stated that you had inherited $1,000,000.00.

The data communications system itself is concerned only with the correct transmission between points A and B of the information given it; the system does not operate on the content of the information at all. This means that where we talk about the "correctness" of the transmitted information in this book, we mean only that the information received has the identical form as the information transmitted. If someone gave a message to the telegraph operator stating that your uncle had left you a million dollars, when he had left you only a hundred dollars, the inaccuracy of that information would not be the fault of the telegraph operator or the data communications equipment.

Protocol is the name given to the hardware and software rules and procedures for making sure that any transmission errors are detected. These may be as simple as transmitting an extra bit of information in each character to detect errors, as used with personal computers, or as complex as some systems use for satellite data communications. The satellite system not only sends extra information to allow the receiver to detect transmission errors, but also to correct the errors and to make the receiver appear to be in the room next to the DTE, rather than about 50,000 transmission miles away.

DTE-DCE Interface

The DTE-DCE interface consists of the input/output circuitry in the interface, as well as the cables and connectors that link DCE and DTE.

We've mentioned the DTE-DCE interface several times as though it is something special. It is. The interface consists of the input/output circuitry in the DCE and in the DTE, and the connectors and cables that connect them. In most systems, this interface conforms to the RS-232C standard as published by the Electronics Industry Association (EIA) in the United States. (The RS-232C and other standard DTE-DCE interfaces will be discussed later in this book.) The RS-232C interface, and other serial interfaces which use some of the RS-232C specification, but depart from it in some important way, are by far the most commonly-used interfaces in data communications. The RS-232C standard specifies the rules by which data is moved across the interface between the DTE and the DCE, and, therefore, ultimately from point A to point B. The word serial means that the bits cross the interface one at a time in series.

DTEs are a very important part of the process of moving data between points A and B. Not only are basic input and output capabilities important, but today's intelligent electronic terminals can perform many complex software-driven functions whose goal is to ensure better performance and accuracy of the data transfer. We'll look at these functions in several of the chapters to follow.

Similarly, while we devote a great deal of discussion to the interface between the DTE and the DCE, we spend less time on the interface between the DCE and the communications channel. This is because the latter interface is quite simple (either two wires or four wires, rather than two to twenty-four as in the RS-232C interface), and because there is not a problem with the sequencing of the electrical signals across this interface.

We won't spend a great deal of time discussing the transmission channel either, even though it is obviously crucial to the data communications system. The various electrical characteristics of the channel itself (such as its bandwidth, relative delay at various frequencies, and other parameters) usually conform to published specifications, and these specifications are often contained in the tariffs published by the telephone companies or other carriers. We have very little control over the transmission channel itself (except to hang up and redial), and the equipment used for data communications is tolerant of a wide range of channel characteristics. We will discuss some of these characteristics and some of the problems, but unless the data transmission rate is at a very high speed, the communications channel usually does not cause any trouble.

So we really don't care much about how the channel works, so long as a standard voice-grade channel between points A and B is available. A standard voice-grade channel is a telephone line which normally would be used for voice communication, that is, a telephone conversation. Thus, a fundamental problem in data communications is that the transmission of digital data must be accomplished over facilities that were designed for voice (analog) communications. The modem, a type of DCE which is used to convert digital data to an analog form (like speech or tones) for transmission over the telephone channel, is described later in this book.

WHAT HAVE WE LEARNED?

Now that we have revealed the fundamental secrets of data communications, let's briefly review the problem, the terms and the parts of the system:

1. Problem: To encode information using some kind of standard code, convert it to a form that can be transmitted over the existing telephone network, transmit it between point A and point B without introducing errors, and reverse the process at the receiving end to recover the original information.
2. Channel: A transmission facility connecting points A and B. It's usually a voice-grade telephone channel provided by the telephone company.
3. Data Terminal Equipment (DTE): At one end, the source of the data to be transmitted; at the other end, the sink which receives the transmitted data.
4. Data Circuit-terminating (or Communications) Equipment (DCE): The conversion equipment between the DTE and the transmission channel. One type of DCE, the modem, converts the data into tones for transmission over the voice channel.
5. RS-232C Interface: The interface wiring and electronics between the DTE and the DCE equipment at either end of the system. Although there are other important interface standards, this is the most common.
6. Information: The data to be transmitted between A and B. It could be as simple as a command to turn on an indicator light, or as complex as the commands for drawing a multicolor illustration on a CRT display. In either case, the communications system itself is concerned only about moving the data between A and B without any error. The data communications system does not respond to, nor act on, the content of the data transmitted.
7. Protocol: The rules for transmission of information between two points. They include rules for handling such questions as what to do if a transmission error occurs, or how to determine if the receiver is ready to receive the transmission.

Quiz for Chapter 1

1. Which is *not* an example of data communications?
   a. A teletype printing news bulletins.
   b. A computer transmitting files to another computer.
   c. An automatic teller machine checking account balances with the bank's computer.
   d. A salesman telephoning orders to the office.
2. Two-state (binary) communications systems are better because:
   a. They can interface directly with the analog telephone network.
   b. The components are simpler, less costly, and more reliable.
   c. People think better in binary.
   d. Interstate calls are less costly.
3. Which is *not* a positional notation system?
   a. Roman (MCMXXXVII)
   b. Binary (01111110)
   c. Decimal (1492)
   d. Hexadecimal (3A2B)
4. Codes are always:
   a. Eight bits per character.
   b. Either seven or eight bits per character.
   c. Agreed upon in advance between sender and receiver.
   d. The same in all modern computers.
5. The Baudot code:
   a. Was invented by the Baudot brothers, Mark and Space.
   b. Requires the escape character to print numbers.
   c. Requires shift characters to provide sufficient combinations.
   d. Was invented by Emile's sister, Bridgette.
6. The ASCII:
   a. Is version II of the ASC standard.
   b. Has 128 characters, including 32 control characters.
   c. Is a subset of the 8-bit EBCDIC code.
   d. Is used only in the U.S. and Canada.
7. Escape sequences:
   a. Use the ESC character to indicate the start of a special control sequence.
   b. Are used to switch (escape) between ASCII and EBCDIC codes.
   c. Are a popular daydream for inmates.
8. The principal difference between batch processing and on-line processing is:
   a. Computer resources are used more efficiently for on-line processing.
   b. Teleprinters are used for batch processing; CRTs are used for on-line processing.
   c. Transactions are grouped for batch processing; transactions are processed as needed for on-line processing.
9. DCE and DTE:
   a. Mean "digital communications equipment" and "digital termination equipment".
   b. Are connected by either two or four wires.
   c. Refer to the modem and the computer or terminal, respectively.
10. The correctness and accuracy of the transmitted message content is:
    a. Verified by the modem.
    b. Determined by the sender and receiver, not by the communications system.
    c. Ensured by use of digital techniques.

Data Terminals

ABOUT THIS CHAPTER

Having defined data communications in an overall sense, we now begin describing the various equipment and software which make up the system. We'll start with the part of the system that is most visible to the user — the terminal. Terminals are one of the most dynamic areas of the computer industry; they are associated with convenience, speed, and glamour. The great majority of data terminals in use today fall into two categories: teleprinter terminals and CRT terminals. Growing additions to this category are small business computers and personal computers used as terminals. We'll discuss both the older teleprinter terminals and the newer CRTs, as well as their rivals.

TELEPRINTERS

The teleprinter, the next step in data communications after the Morse telegrapher, is a low-speed printer that has a serial communications interface.

The device that was invented to replace the Morse telegrapher was called the teleprinter. It was actually a form of specialized telegraph which used the 5-bit character code. An early teleprinter is shown in *Figure 2-1*. The largest manufacturer of teleprinters in the United States was the Teletype Corporation, and because of this, the word "teletype" has been commonly used to refer to any teletypewriter or teleprinter. This usage is improper, however, because "Teletype" is a trademark of AT&T Teletype Corporation.

A teleprinter may be defined as any device that combines a low-speed printer with a serial communications interface. Although a keyboard is an important component of most teleprinters, it is not required to fit this definition (which is used in the commercial marketplace). Thus, the common line printer qualifies if it has a serial interface.

Teleprinter Communications

Teleprinters transmit characters using a simple but slow asynchronous transmission method that either reverses current flow or turns it on and off.

Teleprinters use start-stop asynchronous transmission which is the most widely used method of data communications for a serial interface. It is the simplest technique, but it's also the least efficient (the price paid for simplicity).

Teleprinters transmit at a speed much lower than the capacity of a telephone voice grade line. Common teleprinter speeds in North America are 75 and 110 bits per second (bps). Common speeds elsewhere in the world are 100 and 200 bps. The international teleprinter (telex) network and many European lines operate at 50 bps. Teleprinter signals are formed simply by switching an electrical current on and off or by reversing its direction of flow. The current is



Figure 2-1.
An Early Teleprinter
(Courtesy of National Broadcast Museum, Dallas, Texas)

either 20 or 60 milliamperes, and the equipment is called current-loop equipment. (Most teleprinters today use RS-232C as the external interface, but current-loop circuitry still may be used internally.) The on and off pulses form a code which is interpreted by the receiving device. Single-current teleprinter signaling, also known as neutral or unipolar signaling, is common in the United States; but double-current signaling, also called polar or bipolar signaling, is more common in Europe. We'll discuss current-loop signaling more in a later chapter.

Teleprinter Terminals

The speed of the teleprinter's printing mechanism is very important in applications where most of its time will be spent in receiving and printing incoming data.

Two types of teleprinters are available: those with typewriter-style keyboards (called keyboard send-receive or KSR), and those without keyboards (called receive-only or RO). The typical teleprinter is capable of printing from 10 to 30 characters per second (cps) or from 100 to 300 words per minute (wpm) which is much faster than most typists. However, because much of the terminal's time may be devoted to printing output from a computer rather than entering data, it is usually desirable to have the fastest possible printing mechanism. Thus, teleprinters are available with printing speeds of up to 120 cps.

In the usual operation of a teleprinter, a user enters inputs to the computer via the keyboard. The inputs are transmitted back (called echo) from the computer character-by-character to the printer and printed to verify receipt. Output originating in the computer or another source is printed as it is

received, thus, a "hard copy" of both the inputs and the outputs is available. While many teleprinters are utilized in this way, some are strictly on-line printing devices; thus, they don't require a keyboard.

Teleprinters that use microprocessors can be programmed for a variety of different applications. This has reduced their cost and extended the machine's life.

The classic and perhaps best application for teleprinter terminals is in interactive sessions with a computer, such as timesharing or bulletin boards. Here the comparative slowness of the teleprinter is not restrictive and the printed copy is often quite useful. The slow speed of operation is not as noticeable as in other applications because much of the operator's time is spent in thinking, rather than in waiting for output.

Modern teleprinters, such as the one illustrated in *Figure 2-2*, utilize a variety of printing techniques and operate at a wide range of print speeds. They provide features such as programmable format control, adjustable forms control, upper and lower case printing, interchangeable character styles (fonts), bidirectional printing and paper feeding, selectable character and line spacing, status indicators, portability, and additional keys such as a numeric keypad.

Manufacturers have incorporated the microprocessor into teleprinters to reduce design, development and production costs, while providing a variety of applications that can be implemented by either the vendor or user. The microprocessor delays the time when a particular teleprinter design will become obsolete because future applications can be accommodated by reprogramming the microprocessor to do other jobs. From the user's point of view, microprocessor technology also offers the major advantage of lower cost because the manufacturer's cost savings resulting from using microprocessors are passed on to the customer.

Figure 2-2.
TI 820 KSR Teleprinter



Compared to total terminal sales, few new teleprinter units are sold outside of the traditional record services markets (telex, twx, and private teleprinter network). Even most of these are specialized. An example is the portable printing terminal often used by a traveling business person whose needs are satisfied by a small, light-weight (13 to 18 pounds), hand-carried terminal. Such terminals are available from Texas Instruments and other vendors.

The rapid decrease in the price of portable computers means that they can be used now for applications that previously were reserved for portable terminals. Not only can portable personal computers communicate with large corporate computers, but these portables permit the traveling business person to compose memorandums and letters and to analyze business decisions with spreadsheet programs. Although portable terminals communicating with corporate computers could perform such functions, the personal computer can complete these tasks in a local mode of operation, which eliminates many long distance communication sessions.

Although many vendors call their personal computers portable, the word transportable may be more appropriate in certain cases. Personal computers that weigh between 20 and 30 pounds are best categorized as transportable, whereas truly lightweight (weighing under 12 pounds) personal computers can be considered portable, especially when you carry them through an airport.

Teleprinter Versus CRT Terminals

A teleprinter's disadvantages are: slower speed, little or no variability in format, and limited reliability due to mechanical wear.

The drawbacks to teleprinters are speed, formatting, and reliability. They are far slower in operation than CRT terminals; even their highest print speed doesn't compare to the 300 or 1200 characters per second on the CRT. Because teleprinters are designed primarily for message communications, they do not provide sophisticated capabilities for data editing or formatting. It is quite difficult to create a form on a teleprinter, and of course such attributes as blinking, highlighting, and similar enhancements that are available on a CRT terminal are impossible on a teleprinter. While teleprinters are quite reliable, they are, nevertheless, electromechanical devices that are subject to wear and misalignment. CRT terminals on the other hand, generally have higher reliability because of far fewer parts, most of which do not move.

Serial Printers

Teleprinter terminal serial printers, either impact or non-impact, print one character at a time sequentially.

Teleprinter terminals utilize serial (or character) printers, so named because they print one character at a time. Such printers may be classified into two categories: impact printers that mechanically strike the paper to produce a printed image, and non-impact printers that produce an image by some other means. Impact printers are further divided into two sub-categories: those that produce a "fully-formed" character (where the character is created by impact from a single piece of type), and "dot matrix" printers, where the character image on the paper is formed by a matrix of dots. Non-impact teleprinters form characters by the dot matrix method only, using either an electrothermal or ink-jet printing method.

Bookmark

HOWARD W. SAMS & COMPANY

DEAR VALUED CUSTOMER:

Howard W. Sams & Company is dedicated to bringing you timely and authoritative books for your personal and professional library. Our goal is to provide you with excellent technical books written by the most qualified authors. You can assist us in this endeavor by checking the box next to your particular areas of interest.

We appreciate your comments and will use the information to provide you with a more comprehensive selection of titles.

Thank you,

Vice President, Book Publishing
Howard W. Sams & Company

COMPUTER TITLES:

Hardware
☐ Apple 140 ☐ Macintosh 101
☐ Commodore 110
☐ IBM & Compatibles 114

Business Applications
☐ Word Processing J01
☐ Data Base J04
☐ Spreadsheets J02

Operating Systems
☐ MS-DOS K05 ☐ OS/2 K10
☐ CP/M K01 ☐ UNIX K03

Programming Languages
☐ C L03 ☐ Pascal L05
☐ Prolog L12 ☐ Assembly L01
☐ BASIC L02 ☐ HyperTalk L14

Troubleshooting & Repair
☐ Computers S05
☐ Peripherals S10

Other
☐ Communications/Networking M03
☐ AI/Expert Systems T18

ELECTRONICS TITLES:
☐ Amateur Radio T01
☐ Audio T03
☐ Basic Electronics T20
☐ Basic Electricity T21
☐ Electronics Design T12
☐ Electronics Projects T04
☐ Satellites T09
☐ Instrumentation T05
☐ Digital Electronics T11

Troubleshooting & Repair
☐ Audio S11 ☐ Television S04
☐ VCR S01 ☐ Compact Disc S02
☐ Automotive S06
☐ Microwave Oven S03

Other interests or comments: _____

Name _____
Title _____
Company _____
Address _____
City _____
State/Zip _____
Daytime Telephone No. _____

A Division of Macmillan, Inc.
4300 West 62nd Street Indianapolis, Indiana 46268

27270

Bookmark

BUSINESS REPLY CARD
FIRST CLASS PERMIT NO. 1076 INDIANAPOLIS, IND.

POSTAGE WILL BE PAID BY ADDRESSEE

HOWARD W. SAMS & CO.
ATTN: Public Relations Department
P.O. BOX 7092
Indianapolis, IN 46209-9921

NO POSTAGE NECESSARY IF MAILED IN THE UNITED STATES

HOWARD W. SAMS & COMPANY

Impact Printers

Impact printers are those that use mechanical action to strike a ribbon against the paper to print the desired character, either completely formed, or in a dot matrix.

Impact printing mechanisms include replaceable type-ball elements (as used on the IBM Selectric[1] typewriters), daisy wheels (a flat disk with petal-like projections for the character elements), elements shaped like a cup with finger-like projections, moving type belts with the characters engraved on the belt, a rotating cylinder containing the characters, and a type block with characters embedded in the block. All of these share the same advantages of fully-formed characters in almost any style and shape, and the ability to make carbon copies. They also share the same disadvantages of low speed and mechanical complexity. No matter how the character is mounted or engraved on the type element, it must be positioned in front of the platen and then struck against an inked ribbon and paper to produce an image. This limits the speed of the printer to a maximum of 120 cps (but it's usually much lower), as well as creating a large amount of noise in the process.

This speed limitation was an incentive for printer manufacturers to seek a different approach that would extend the upper limit of printing speed for serial impact printers. Their search led to the development of the dot matrix printer. Dot matrix printers are a compromise (and often a successful one) between decreased character quality and substantially higher print speeds of 180 cps or more. The printed image is formed from a rectangular matrix of dots, typically 9 dots high by 7 dots wide. Printing is performed by moving a print head containing a column of 9 pins across the paper and selectively actuating the pins at 7 successive intervals to form each character. Though they contain comparatively few moving parts, matrix printers are subject to a large amount of wear and heat within the print head as a result of the succession of pin movements required to create each character.

Dot matrix teleprinters are typically less expensive than similarly featured fully-formed character teleprinters. With the improved print quality now available with higher-resolution dot matrix printing, most teleprinter users are satisfied with dot matrix printers.

Another development that has improved speed in newer teleprinters is the bidirectional print head with logic-seeking control. The print head can print while moving in either direction (left to right or right to left). The logic-seeking control circuitry determines the shortest horizontal distance to move the print head to print the next line. Thus, on full pages of text, one line is printed left to right and the next is printed right to left, etc. This feature, which is available on both fully-formed character printers and dot matrix printers, eliminates the time required for a full carriage return to the left side for every line.

Non-Impact Printers

Non-impact printers use electrothermal or ink-jet printing techniques. They are more reliable, quieter, and faster than impact printers.

The non-impact types of teleprinters employ various electronic and chemical techniques to produce printed images. All non-impact teleprinters currently on the market utilize dot matrix character formation. Some of the printing techniques have been developed from xerography and facsimile communications techniques. Others were specifically developed for use in high-speed printing applications where print speeds of more than 2000 lines (that's

[1]IBM and Selectric are trademarks of International Business Machines Corporation.

ST. GEORGE TECHNICAL COLLEGE LIBRARY

lines, not characters) per minute are not uncommon. Still others were designed to meet a specific goal, such as quiet operation.

The electrothermal printing technique is the most commonly used non-impact technique. The print head applies heat through the matrix pins to chemically coated paper. The heat causes a chemical reaction which produces a matrix of colored dots that forms the character. It is quiet and efficient, but some people object to having to supply special paper.

The ink-jet technique uses a stream of electrically charged ink droplets sprayed onto ordinary paper to produce the characters as shown in *Figure 2-3*. Dot placement is accomplished by electrostatic deflection plates that control the direction of the charged ink droplets in much the same manner as electron beam position is controlled within a cathode ray tube. By electronic control of switching between reservoirs of different colored ink, instant color changes can be made as the printer operates. Ink-jet printers were very expensive, but prices have been reduced substantially in the last few years.

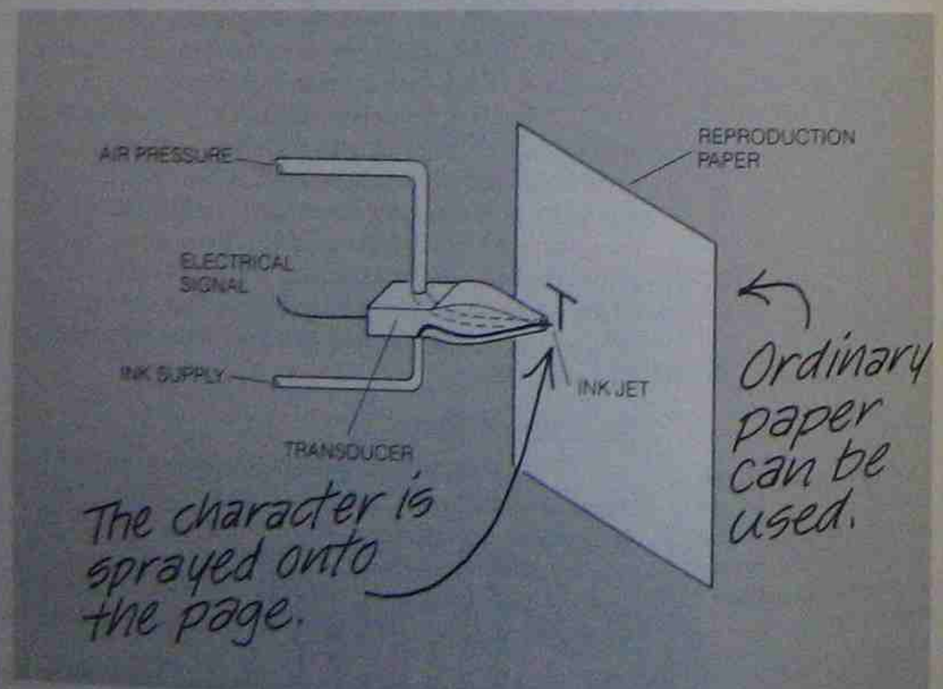
Reliability of non-impact printers is high compared to impact printers because they have few mechanical parts. Their inability to produce more than one copy at a time (no carbon copies) might sometimes be a disadvantage, but their quiet and reliable operation is certainly an advantage.

CRT TERMINALS

In many applications, the CRT terminal has replaced the teleprinter terminal as the interface between computers and people.

The cathode ray tube (CRT) terminal, also called a video display terminal (VDT), is currently the most common interface between people and computers. Originally developed as an alternative to teleprinter terminals, its introduction in 1965 revolutionized the data communications environment where the teleprinter had been the only interface device. Although many companies entered the marketplace, prices remained high due to low-volume

Figure 2-3.
Basic Ink-Jet Printing Technique



production and the cost (price and assembly labor) of discrete components. The advent of the integrated circuit, followed by large-scale integration, brought prices down, although they were still out of reach of the average individual. Then came the microprocessor, which replaced much of the integrated circuit logic and drove prices still lower. The result is high-quality, high-performance and relatively low-cost CRT terminals that even the hobbyist can afford. A modern CRT terminal designed for businesses is shown in *Figure 2-4*.

Figure 2-4.
TI 931 CRT Terminal



ASCII Terminals

A CRT terminal that can operate without a separate controller, and can transmit and receive asynchronous data using the ASCII is called an ASCII terminal.

In the computer industry, the term "ASCII display terminal" usually refers to a CRT terminal that operates without a separate controller, uses the ASCII code, and transmits and receives asynchronous data. This term originated with CRT terminals (the glass Teletype) designed to replace Teletype ASR 33/35 teleprinter terminals; and ASCII terminals are still described as being "Teletype-compatible". ASCII terminals are available in a wide range, from low-end dumb units that do little more than emulate a teleprinter (or which emulate another CRT terminal, such as the ADM-3, which emulates a teleprinter), up to 132-column split-screen editing terminals with several pages of electronic memory storage.

CRT terminals have three capabilities in common: 1) a keyboard that can generate a full alphanumeric character code set, 2) a CRT monitor that can display the characters of that code set, and 3) the capability to send and receive

data via communications lines to a remote host computer. CRT terminals fall into one of three general categories: dumb, smart, and user-programmable. Of course, there is some overlap between categories, but we'll define them as follows:

CRT terminals can be classed in one of three general categories: dumb terminals, smart terminals, and user-programmable terminals.

Dumb terminals are Teletype-compatible and offer a limited number of functions; perhaps none beyond keyboarding and display. They are the lowest-cost, but often the difference in price between them and the smart terminals is small compared to the difference in performance and features.

Smart terminals offer extended functions, such as editing and formatted data entry. The user may be able to tailor the terminal to fit his or her application through limited programming, such as format creation, parameter definition, or input checking.

User-programmable (or intelligent) terminals feature software support. The manufacturer provides an operating system; an assembler, compiler, or interpretive programming language; subroutines; I/O utilities; one or more protocol emulators; and one or two application programs, such as data entry and text editing. They provide some form of program storage and more memory for text.

Dumb Versus Smart Terminals

The description of the capabilities of teleprinters also describes the capability of the dumb CRT terminals except for print speed. These essentially consist of a CRT, video (display) memory, and interface electronics. Few truly dumb terminals are still manufactured; virtually all CRT terminals currently manufactured are microprocessor-controlled, making them smart terminals according to our definition. The microprocessor chips cost only a few dollars, which is less than the cost of the logic they replace.

Microprocessor control has given CRT smart terminals the flexibility and capabilities to perform many operations beyond simple display and keyboarding.

The microprocessor programs (firmware) reside in read-only memory (ROM) or programmable ROM (PROM). ROM-resident programs control those features which are "permanent" while PROM-resident programs are typically produced in smaller quantities and implement customized or modifiable features. Either type can be replaced by a different type with a different program by simply removing the old chip and putting in a new one, although this feature is more talked about than actually used. In addition to controlling basic terminal functions, the microprocessor can provide protocol emulation, define the character code sets to be generated by the keyboard, implement special features, set control parameters, etc. Moreover, the microprocessor delays obsolescence, since future needs can be implemented via reprogramming.

In addition to the ability to simply display alphabetic and numeric characters, microprocessor control provides the capability to highlight characters by means of underscoring, reverse video, blinking, differing levels of brightness, and combinations of these attributes. One of the most impressive capabilities of these terminals is the ability to display a predefined form and let the operator "fill in the blanks" while checking the entries as they are made.

Many CRT terminals can display double size characters and most have a graphics character set for creating forms and formats. Some also support business graphics; for example, bar, column, and pie charts to

graphically show sales, income and expense, inventory levels, etc. Interactive or engineering graphics, on the other hand, is a highly specialized area usually requiring a completely different, high-resolution graphics terminal.

Because of the need for increased requirements, the emphasis will shift to using microcomputers for terminal functions.

User-Programmable Terminals Versus Personal Computers

Taken separately, the words programmable and terminal are precisely definable. But when combined to describe one class of equipment, their meaning becomes vague and difficult to define with precision. In 1979, a popular periodical defined a user-programmable terminal as a terminal which permits local file updating and accommodates user-written applications programs. At that time, this definition stirred up quite a bit of controversy because such a terminal would have to provide for the following in addition to the basic terminal functions:

1. The availability of, and the support and documentation of, at least one general-purpose programming language for development of programs by customer personnel.
2. Entry of programs via the terminal keyboard.
3. The ability to access and retrieve programs stored in the host computer memory, and the ability to execute (use) those programs at the terminal.

Much of the value of user-programmable terminals is their off-line processing capability provided by the above listed items. All user-programmable terminals have built-in main memory that the user can use to store and execute programs, but the capacity varies from as little as 1K byte (which will accommodate only a very small program) to as much as 512K bytes (which will accommodate a large program and much data). Most (though not all) user-programmable terminals also provide for local mass storage using floppy, hard, or cartridge disks; or cassette or cartridge magnetic tape.

With disks, tapes, memory, an operating system, and a programming language, the user-programmable terminal sounds a lot like a personal or professional computer, doesn't it? This is the revolution in the terminal market. Dumb terminals and low-end smart terminals are being squeezed aside by personal and professional computers running terminal programs, which relegates terminals to specialized tasks. Many office workers now have a personal or professional computer on their desks, which can be used either stand-alone as a computer or placed on-line to the large host computer. Files can be up- and down-loaded and processed wherever it makes the most sense to do so. All because of data communications and the microprocessor!

Emulation

When the personal computer operates a communications program that enables the computer to duplicate the attributes of a terminal, the software is known as a terminal emulator program. Such programs permit the personal computer to communicate with main frame application programs developed to work with predefined types of terminals.

Depending on the terminal emulation software marketed for the personal computer, you may be able to replace several types of terminals

with one computer. For an example, consider the organization that has both Hewlett Packard (HP) and Digital Equipment Corporation (DEC) main frame computers. With HP 2645 and DEC VT100 terminal emulation software, one can use a personal computer to emulate common Hewlett Packard and Digital Equipment Corporation terminal devices and thereby obtain full screen access to two different computer systems that are designed to operate with two specific types of terminals.

Non-ASCII Terminals

This section describes CRT terminals that do not fit the ASCII label, together with some specialized terminals that use other displays. Of these, the most important are the batch or cluster type.

Batch Terminals

Communications between a terminal and its computer tends to fall into two basic categories, interactive and batch, which are distinguished by the types of functions to be performed. Terminals and software designed for interactive communications between the terminal and the host support conversational inquiry/response, interactive data entry, and other such applications. Batch communications, on the other hand, supports either remote job entry or remote batch data communications. Remote job entry (RJE), in which the terminal acts as a remote console for the central computer, allows local users to load and execute applications programs. For remote batch data communications, the terminal acts as a remote input/output device for an application program being performed in the host.

The reasons for these distinctions are efficiency and accuracy. An interactive terminal designed to communicate in asynchronous character-by-character mode is not a very efficient method for transmitting large blocks of data as required in the batch mode, nor is the error checking (if any) adequate to ensure correct transmission. On the other hand, it would be wasteful to tie up a complete remote batch terminal just to allow one user to use the keyboard to ask questions about particular data stored in the host computer with the answers sent back by the computer.

Clustered Terminals

When a remote location generates a large amount of data via key entry, the use of a clustered system may be indicated. This is typical in large-scale data processing installations where it is common practice to control many peripheral units, such as CRT terminals and key-to-disk data input devices, at one location with a single control unit. Data entered through the keyboard devices are stored in a buffer for transmission or transfer to magnetic tape or disk. Then the output of many operators is transmitted in one block to the main computer at much higher speed than that used by teleprinters and ASCII CRT terminals. In addition to the key entry devices, storage, and controller; the cluster may include printers and mass storage. This type of architecture is fundamental to modern data processing and a widely-used example is the IBM[2] 3270 series.

[2]IBM and IBM PC are trademarks of International Business Machines Corporation.

IBM 3270 Series

The IBM 3270 is not a terminal per se, but is a designation for a family of devices designed for use in clusters such as described above. The 3270 series has had a strong impact on the synchronous terminal market since its introduction in 1971. The first generation, which was discontinued in 1982, included the 3271/3272 control units, 3275 display station, 3277 display, and 3284/3286/3288 printers. In 1977, the product line was expanded and updated with the 3274 control unit, 3276 control/display, 3278 display, and 3287/3289 printers. In 1979, color displays and printers were added. In 1983, IBM made some long-awaited changes and enhancements to the 3270 product line. Added were the 3178 display station, a new 3274 control unit, the 3290 information panel (a gas plasma display) and the 3299 terminal multiplexer.

The 3299 multiplexer is normally used to reduce the amount of cable required to connect terminals to a control unit when the distance between the location of the terminals and the control unit exceeds a few hundred feet. Up to eight terminals can be connected to a 3299, with only one cable required to connect the 3299 to a control unit.

Newer terminals in the 3270 series include the 3179 color display station, the 3279G color graphics display station, the 3180 and 3191 monochrome display stations, and the 3194 color display stations. The latter two are most notable for their design, which minimizes the desk space that they require compared with earlier terminals in the 3270 series.

In the area of personal computer connectivity to the 3270 series, IBM has introduced many products. Two of the more significant products are an adapter card that is installed into a system expansion slot of a computer in the IBM PC series and a new control unit. The adapter card permits direct connection of the personal computer to a control unit, which enables the PC to function as a 3270 terminal. With this configuration, personal computer users also can transfer files and do local processing, both of which are beyond the capability of conventional 3270 terminals. In 1986, IBM introduced the 3174 control unit, which features an optional token-ring gateway. This feature permits one member of the IBM PC series to be cabled to the 3174 as a gateway device on a token-ring local area network. Other personal computers are then connected to the gateway personal computer, enabling many personal computers on the network to obtain access via one personal computer.

The IBM 3270 family of clustered synchronous terminals is receiving strong competition from microcomputers.

In providing the PC interface, IBM finds itself in an interesting quandary. The 3270 line has been quite profitable over the years; now it faces competition from personal computers, especially the one manufactured by IBM! As personal computing becomes the rule instead of the exception in most major corporations, the terminal industry is on the verge of the biggest change since the introduction of the CRT.

The many other firms which offer 3270-compatible products must likewise compete with personal computers. In order to remain competitive, these companies have traditionally offered some combination of lower prices, improved price-performance, and shorter delivery times in order to penetrate the IBM "plug-compatible" market. In addition to the 3270-compatible

vendors, some ASCII terminal manufacturers have invaded the 3270 market through protocol conversion. On a 3270 network, synchronous terminals can be replaced with asynchronous terminals coupled with "black boxes" that convert the asynchronous data to synchronous form with the proper headers, check bits, etc. These devices allow an ASCII terminal to support the functional characteristics of the 3270 terminal. The advantage of this strategy is that ASCII terminals are considerably less expensive than their 3270 counterparts.

Other Types of Terminals

The majority of terminals manufactured today employ a keyboard and a CRT. The popularity of this device stems from its flexibility, high character capacity, and relatively low cost. However, specialized needs have led to the development of a broad range of other terminals which include optical bar code readers, voice response units, portable terminals, and the ordinary tone dialing telephone. In addition, such things as supermarket cash registers, portable communicating terminals for inventory management, and a host of other types of equipment now employ data communications. By the DTE/DCE definition, all of these are terminals.

PARTS OF A TERMINAL

Having discussed the various types of terminals, let's now discuss the components of a typical CRT terminal. The average user is concerned only with the components with which he or she interacts — the keyboard and the display. But first, let's talk about one of the current buzz words in the terminal business — ergonomics.

Ergonomics

ANSI standards define ergonomics as: "A multi-disciplinary activity dealing with the interactions between man and his total working environment, plus such traditional environmental aspects as atmosphere, heat, light, and sound, as well as the tools and equipment of the workplace."

Terminal manufacturers have become increasingly aware of the need to consider ergonomics in the design of their equipment. The trend toward making CRT terminals more "operator-friendly" began in Europe, and most people agree that European manufacturers still lead in this area. Recent developments in terminals, such as sloped keyboards and green or amber displays, have come about because of ergonomic considerations.

Keyboards

The majority of display terminals now have keyboards that are detached or detachable. Most of these are connected to the console via a coiled cable, usually three to six feet long, which allows the operator to place the keyboard in a comfortable position while using the terminal. A more recent development is the "light-link" or cordless keyboard, where an infrared transmitter/receiver replaces the cable to allow even more freedom of movement.

Another design factor is the slope and thickness of the keyboard. Most keyboards today are either sloped or stepped, with the angle being 5 to 15 degrees. Most recent terminals have sculptured key caps instead of flat key

caps, which speeds data entry and improves operator comfort. Audible keyboard feedback may be provided by a beep or key click from a speaker, and tactile feedback may be provided by a unique "feel" when the stroke is registered. Like most other ergonomic considerations, the above features are not, in themselves, the most important criteria for choosing one terminal over another; what they do indicate is the manufacturer's commitment to providing equipment that is up-to-date and easy to use.

Keyboard Layout

The layout of the keyboard is a primary concern of ergonomics. The de facto standard typewriter key layout is that of the IBM Selectric typewriter illustrated in *Figure 2-5*. A keyboard that has this arrangement of the alphabetic keys is called a QWERTY (pronounced quir-t) keyboard because these are the first six letters of the top row of alphabetic keys. The QWERTY is far from the best layout. In fact, the QWERTY keyboard layout was designed for early mechanical typewriters to arrange the key hammers to reduce key hangups. This arrangement also slowed the typist, so that he or she could not key faster than the early typewriters could operate! In the 1890s, there were over a hundred different typewriter keyboards because almost every company that made a typewriter used a different keyboard. Some of these keyboard types were: the Crandall (ZPRCHMI), the American (CJPFUBL), the Hall (KBFGNIA), and the Morris (XVGWSLZ).



Figure 2-5.
IBM Selectric Typewriter Keyboard Layout

Various attempts have been made to improve keyboard layout. Some had the letters of frequently used words close together. One proposed by Dr. August Dvorak put the five vowels under one hand and the five most common consonants under the other hand. There were circular keyboards, semicircular keyboards, keyboards with six rows, and keyboards with one long row. The idea of a shift key didn't come along until 1875; prior to that, upper and lower case letters were on different keys.

The ability to program the keyboard provides the opportunity to rearrange the layout and/or key functions to suit personal taste.

Programmable Keyboards

You may be thinking that the operator has no choice but to adapt since the keyboard is built-in and the key arrangement is fixed. Ten years ago that was true, but now it may not always be. Alternate keyboard arrangements (including the Dvorak) are available from some manufacturers, either as standard equipment, or as add-on devices for several terminals and computers. More importantly, many keyboards are programmable. The processor in the terminal considers each key depression simply as a contact closure in a matrix of contacts. The matrix location represented by that row and column is used by the processor to look up the character in the corresponding location in a table in memory. Thus, the keyboard layout may be changed by simply reprogramming the table; that is, the "C" key could be defined in the table so that a "?" is printed instead of "C". This suggests that the terminal of the future might have several keyboards which can be changed by plugging in the new one and telling the program which table to use.

Numeric Keypads

Most terminals have some form of numeric keypad with the numerals 0 through 9 grouped like those on a ten-key adding machine to provide faster entry of numeric data. Other keys such as +, −, and . also may be included. The numeric keypad may be entirely separate from the main keyboard or it may be implemented through an alternate definition of alphabetic keys (usually the "IOP", "KL;", and "M,." groups). Using the main keyboard's number keys to enter numbers requires the use of both hands, thus, number entry is faster with a numeric keypad because all of the numerals are under the fingers of the right hand.

cial keys provide access to functions unique to puter input. While e are a large number ese keys, no uniform- xists as to their ioning on the oard.

Other Keys

In addition to the alphabetic and numeric keys, terminal keyboards offer a variety of special keys that are unique to computer input. Unfortunately, there is no standard arrangement of these keys. These may include the CTL (control) key (which allows input of some or all of the ASCII control characters), TAB (usually programmable), ESC, arrow keys (which control cursor movement), and programmable F (function) keys.

For example, on the Wyse WY-50 terminal, the LINE DELETE key sends ESC R. Some keys send different codes in the shifted and unshifted positions. Thus, the unshifted PAGE key means scroll one page forward, and the shifted PAGE key means scroll one page backward. The eight function keys are capable of producing 16 code sequences by using the SHIFT key. When the terminal is powered up, these are set to Ctrl A@ through Ctrl-AO, but they

may be changed by the computer program, which sends the terminal a series of escape sequences. For example, the sequence ESC z A DIR B: CR DEL will reprogram the F2 key to send DIR B: followed by a carriage return. Each function key may be programmed with up to eight characters. Like most modern terminals, the WY-50 has the ability to display function key labels on the bottom line of the CRT screen.

Those users who feel that the more keys the merrier should be very merry indeed because many terminals now have over one hundred keys. However, the keys sometimes are not laid out in the most useful way. For example, the cursor arrow keys may be in a row, rather than in a diamond pattern where their direction can be felt. These additional keys may be packed close around the regular keyboard, or arranged in logical groups and slightly separated from the main keyboard. When keys have unusual or unexpected placement, the operator is likely to hit them by mistake. The result can be devastating.

Other Input Devices

Alternatives to keyboard input, which can make entry of specialized data easier, include light pens, touch-sensitive screens, graphic tablets, a "mouse" (a cursor positioning device), and even speech input. Proponents of all of these devices, and more, suggest that they will eventually replace the keyboard for most input. Don't bet on it. Buy the most comfortable, best designed, easiest-to-use keyboard that you can find; it and the display are your window into the computer. The other devices can be quite useful (especially the mouse), but most input probably will be through the keyboard for a long time to come.

Display

The display provides the visualization of the terminal inputs and outputs. Most use dot matrix character formation with 24 text lines having 80 characters per line.

There has been more discussion recently about CRT displays than any other part of the terminal. Some researchers have noted eyestrain, headaches, dizziness, back pains, nausea, and nervous symptoms in workers whose jobs require that they operate CRT terminals for long periods of time. Eyestrain and fatigue are considerations which must be dealt with when designing a CRT display screen. Most CRT tubes are etched or contain a bonded faceplate to reduce glare; tilt and swivel adjustments are also popular for the same reason. The latter also allow the operator to place the screen at the most comfortable viewing angle. The phosphor color and character size is important, too. White or green phosphors are generally used in the United States, with green becoming increasingly popular. Amber phosphors are used in Europe and are becoming available in the U.S.A.

Most display terminals use the dot matrix technique to form characters. The more dots that are contained in the character cell, the better defined the character will appear on the screen. For years, 5 by 7 dot characters were standard, but today 7 by 7 and 7 by 9 characters are more common. Some manufacturers use high screen refresh rates and non-interlaced scanning to reduce image flicker and improve legibility. The size of the characters generated depends on the size of the screen and the display format used. Characters will be larger on 15-inch screens than on 12-inch screens; likewise, characters will be larger in an 80 character-per-line format than in a

132 character-per-line format. Display enhancements such as double-height and double-width characters can make characters larger, but these are intended for highlighting important data, not for routine use.

Most CRT terminals display 24 text lines of 80 characters each. Many also provide a twenty-fifth line which is used for labels, status, or other terminal or program related information. The status line sometimes is displayed in reverse video.

Memory-Mapped Displays

The majority of ASCII terminals use a technique called memory-mapped video. An advantage of the memory-mapped technique is that the microcomputer can write to and read from the video memory in the same way as ordinary memory because a portion of the ordinary memory is used for the video memory. The difference is that whatever characters are stored in video memory are displayed on the CRT when addressed by the microprocessor.

An 8-bit microprocessor uses two bytes (16 bits) for a memory address; thus, it can directly address 2^{16} or 65,536 memory locations. If each memory location can hold one character (which is the usual arrangement), and if the display size is 80 characters by 24 lines, then 1920 character locations are needed in video memory. If the program memory is restricted to addresses below 63,615, then memory locations from 63,615 through 65,535 could be used for video memory to store an array of characters for the display.

Assume that the upper left-hand corner of the CRT display receives the character from the first location (63,615) in video memory, the next horizontal position on the CRT receives the character from the second location (63,616), etc. Thus, if the first video memory location contains "J", the second "O", the third "H", and the fourth "N"; the upper left-hand corner of the CRT will display "JOHN".

Memory-mapped displays are fast and inexpensive. Because the character set is translated into dots within the dot matrix using a table lookup technique, it is also possible to redefine the character set by simply changing the table entries (if the table is user-accessible). Line insertion and deletion; scrolling, and other effects are easily accomplished by moving data around in video memory, or even simpler, by changing starting addresses. The principal disadvantage of memory-mapped video of the type described is that it takes up space in the memory used to store programs and data. This usually is not a problem for terminals, but for computers, video memory is now usually separate. Also, computer video memory now is often "bit-mapped" rather than "character-mapped" as discussed above. With bit-mapping, each dot on the screen has a separate memory location. This feature allows high-resolution graphic displays, but it requires much more video memory.

Other Types of Displays

Although the CRT is the dominant display device today, there are other devices for displaying information. These include solid-state devices such as light-emitting diodes and liquid crystal displays, and plasma (gas discharge) matrices such as used in the IBM 3290. These alternate types of displays provide extremely sharp images, but the dominance of the CRT remains unchallenged for now.

The Wyse WY-100 is an asynchronous terminal that also is capable of block-mode transmission. Special functions are available to display the dot matrix characters.

Features of a Typical Low-Cost Terminal

The Wyse WY-50 (*Figure 2-6*) is fairly typical of current terminal technology; thus, an examination of its operating features will illustrate some of the previous points. When power is turned on, the WY-50 performs self-tests on the microprocessor, memory, communications and printer ports, keyboard, and display. The terminal halts all further operation if a fault is detected and an error message pointing to the fault is displayed. The main display is 24 lines of 80 characters, plus a top line for local and host messages, and a bottom line for function-key labels. The main 24 by 80 display can be split horizontally into two screens. An escape sequence initiates this feature. For example, ESC x 1 + will split the screen horizontally, with the lower screen starting at line 12.

Characters are formed by an 8 by 10 dot matrix in a 10 by 11 cell which provides very good character definition. The terminal can display the 128 ASCII characters and a number of graphics characters. Normal and reverse video, underscore, dim, blink, and blank attributes may be assigned in any combination on a line-by-line basis; for example, ESC A1T creates a reverse video, half-intensity function-key labeling line. Actual labels are entered with another escape sequence; for example, ESC z 3 DEL CHAR will enter DEL CHAR into the label field for the fourth function key (keys are numbered 0 through 7). Attributes and text are entered into the local and host message fields in a similar manner.



Figure 2-6.
Wyse WY-50 CRT Terminal
(Courtesy of Wyse Technology)

In addition to asynchronous character-by-character transmission, the WY-50 is capable of block-mode transmission. When in the block mode, the terminal performs editing and validation for data entered through the keyboard. This function requires a second page or screen of video memory. One 24 by 80 page is used as the data "form" and the second page is used to define the data-validation parameters. Local editing features include line insert and delete, character insert and delete, and automatic word wrapping.

DATA TRANSMISSION

Serial and Parallel Transmission

Data are commonly transferred between computers and terminals by changes in the current or voltage on a wire or channel. Such transfers are called parallel if a group of bits move over several lines at the same time, or serial if the bits move one-by-one over a single line. *Figure 2-7* illustrates parallel and serial data transmission.

parallel data trans-
ssion, each bit of a
aracter is sent over its
wire. In serial trans-
sion, all bits travel over
same wire one after
ther.

In parallel transmission, each bit of a character travels on its own wire. A signal, called the strobe or clock, on an additional wire indicates to the receiver when all of the bits are present on their respective wires so that the values can be sampled. Computers and other digital systems that are located near one another (within a few feet) usually use parallel transmission because it is much faster. As the distance between equipment increases, the multiple wires not only become more costly, but also the complexity of the line drivers and receivers increases because of the difficulty of transmitting and receiving pulse signals on long wires.

Serial transmission is used for transmission over long distances. The conversion from parallel to serial and vice versa is accomplished with shift registers. Transmission of serial data is called synchronous if the exact sending or receiving time of each bit is determined before it is transmitted or received, or asynchronous if the timing of the bits in a character are not determined by the timing of a previous character.

Timing

ming is critical to the
ste transmission and
tion of data.

Figure 2-8 illustrates a series of bits transmitted serially over a line. Note that while both the sender and receiver use the same nominal clock rate, the receiver must somehow determine a more exact clock for decoding the data. This is one of the fundamental problems of data communications. In this case, the receiver retimes its clock on the negative-going edge of the transition (change) from one to zero of the start bit, then uses its new timing to find the middle of the start bit as shown in *Figure 2-8a*. Notice in *Figure 2-8b* that although the receiver clock is slightly fast, it doesn't cause an error because the sample strobe still occurs within each bit time.

Figure 2-7.
Parallel and Serial Data Transfer

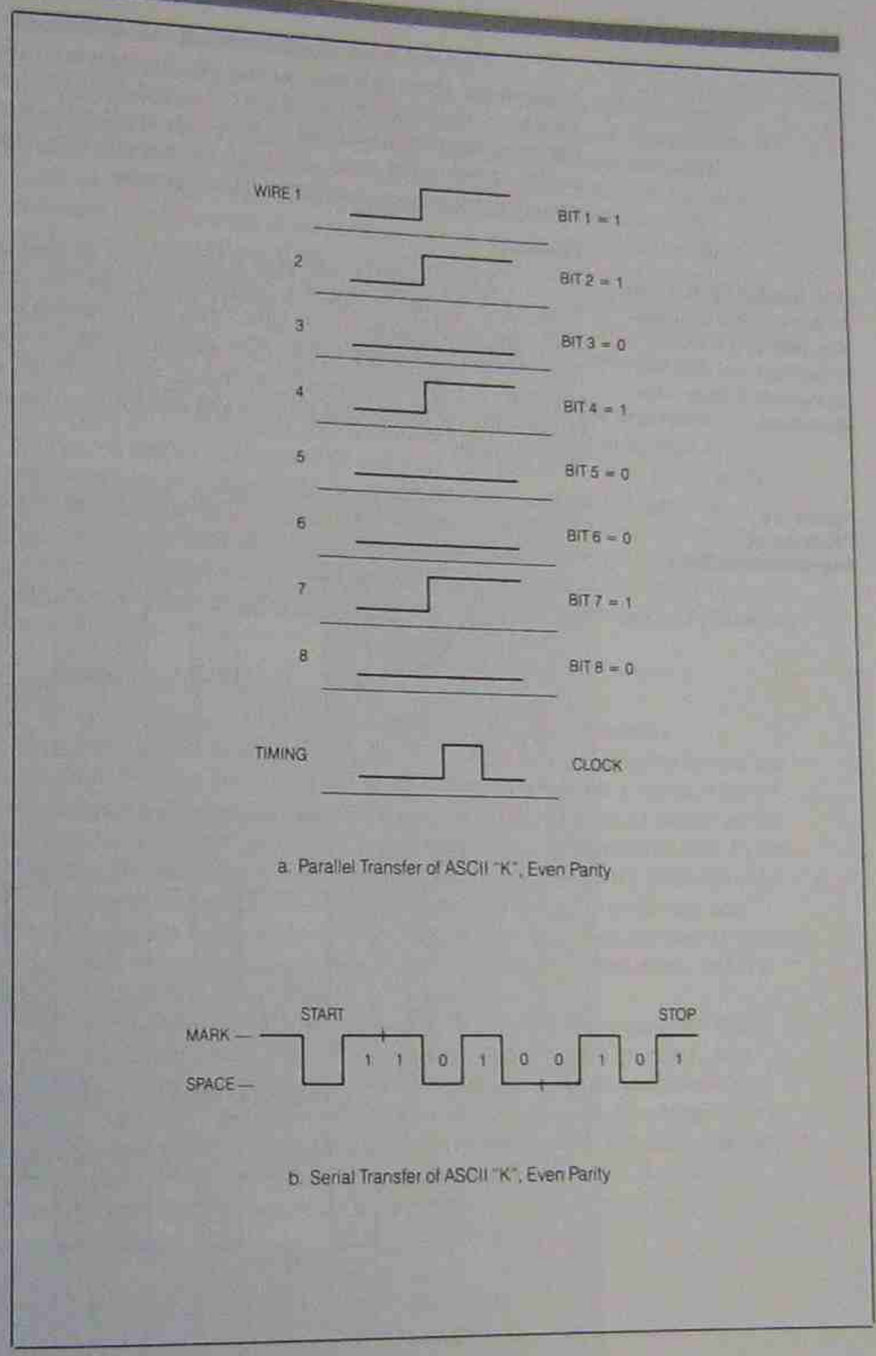


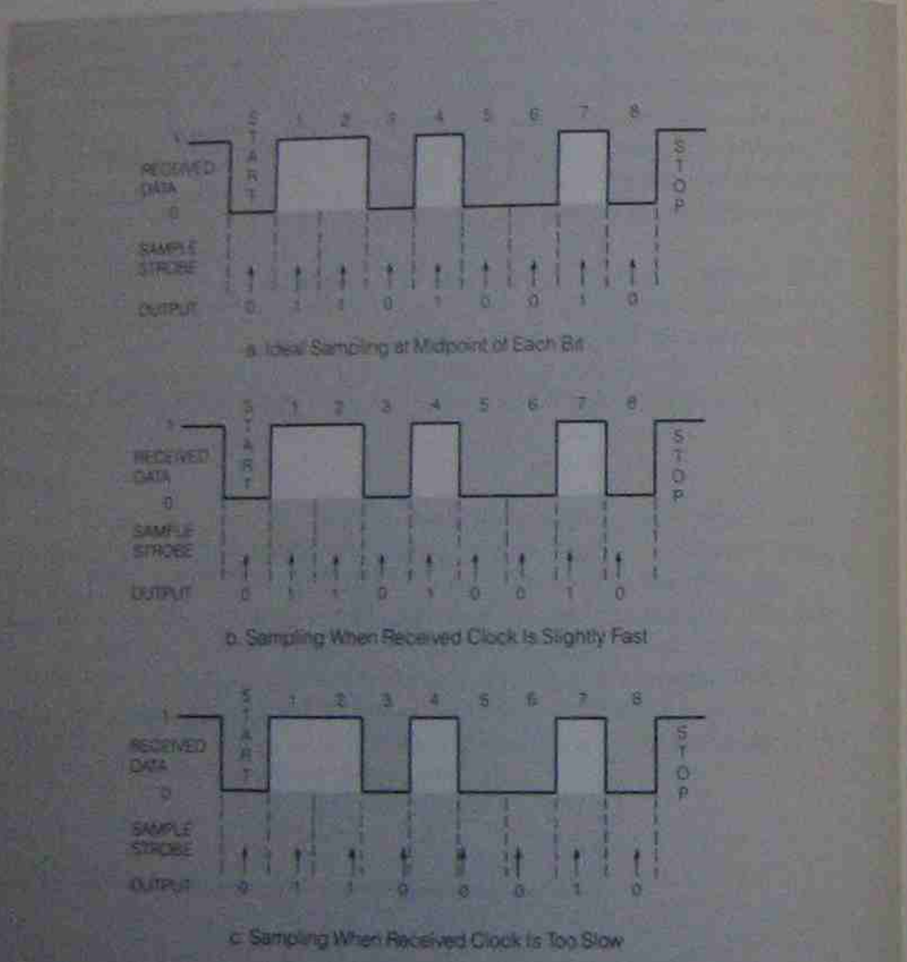
Figure 2-8c shows a different situation. Here, the receive clock is so much slower than the transmit clock that the sample strobe does not sample bit 4 at all, thus, an error occurs in the output. This illustrates the fact that the communications system must somehow ensure that the receive clock is timed so that errors do not occur. Although bit timing is very important, it alone isn't enough to complete the communications process.

Framing

Framing is the next step in timing after bit timing. In the example of *Figure 2-8*, it could be called character timing, since the start and stop bits frame one character for the receiver. This is the framing used in asynchronous data transmission which will be discussed in association with teleprinter operation in the next section. Synchronous data communications systems utilize different methods of framing and these will be discussed in later chapters.

raming method used
chronous transmis-
efines a character by
start and stop bits
arate it from other
ters.

2-8.
ng of
hronous Data



The first asynchronous communication systems used electric motors driving rotating contacts at each end. The rotation of the motors had to be synchronized.

Asynchronous Transmission

Asynchronous communications systems evolved before electronic systems, therefore, electromechanical systems had to be used. The problem in those systems was to synchronize the operation of two electromechanical devices (motors), one at each end of the line, whose speed could not be easily adjusted. The timing and framing problem was solved by retiming each character using the edge of the start bit, which was required to be the first bit on the line. Nevertheless, if the difference in motor speeds was too great, the bit timing would drift with each successive bit until the sampling of the last received data bit in the character could be incorrect. Thus, five bits became the standard for the low speed teleprinter code because the chance of error increased significantly if more bits were used.

Imagine an operator keying a teleprinter keyboard to produce characters for transmission. The framed character generated by each key depression begins with a start bit and ends with one or two stop bits. The stop condition is a mark or positive voltage on the line (negative voltage with double-current signaling). The positive voltage remains on the line until the next character starts because the line idle condition is internationally defined as current flow or mark, rather than no current or space. As soon as a space (start bit) is detected, the receiving device starts and will be in synchronization with the transmitting device.

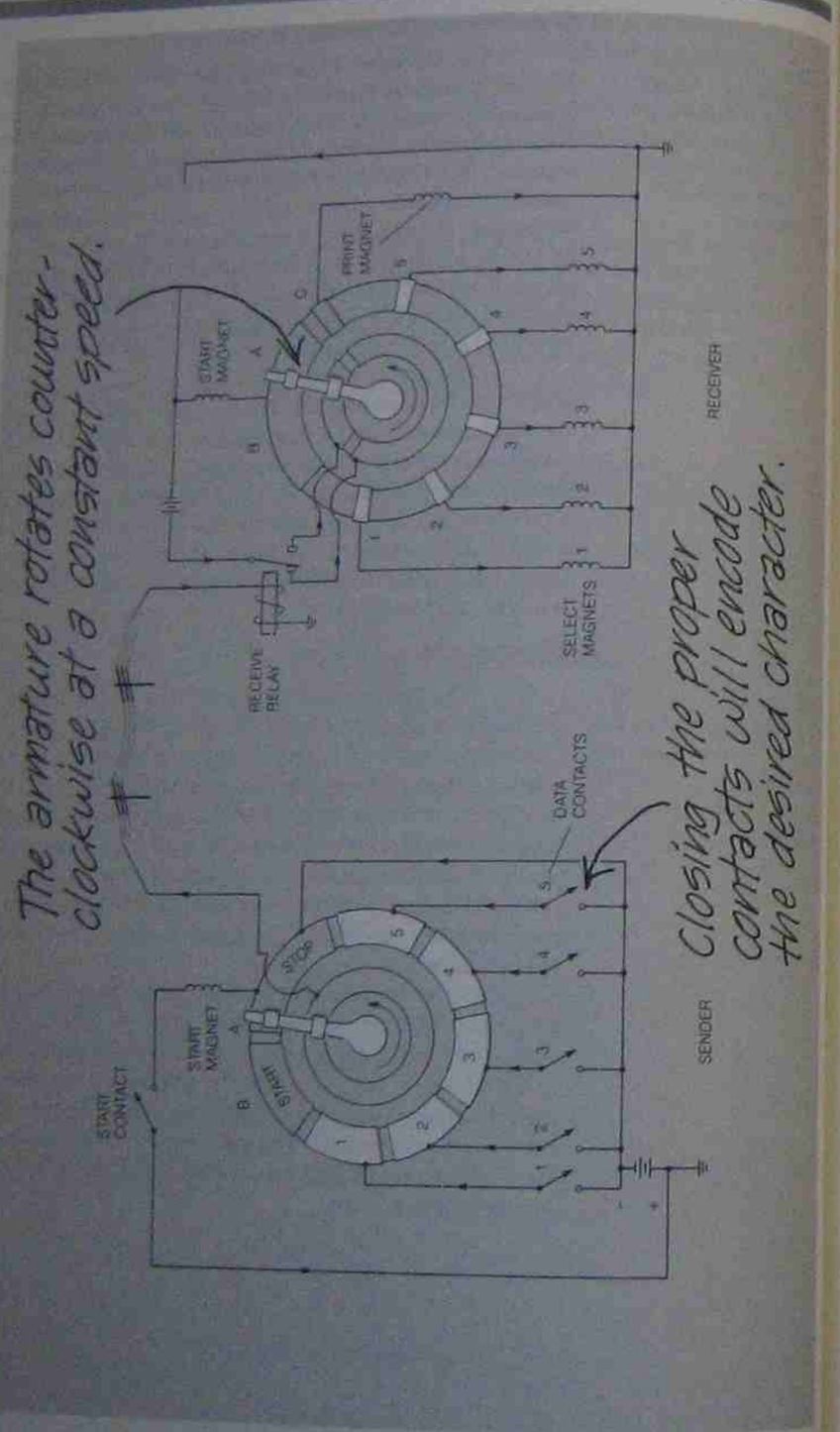
Electromechanical Method

Current pulses, placed on the line by the sender's rotating armature, select the character to be printed through the synchronized receiver contacts.

The diagram in *Figure 2-9* illustrates the operation of an electromechanical start-stop teleprinter machine. The sending device and the receiving device each have an armature, A, which rotates at a constant speed when a clutch connects it to an electric motor in the machine. Contacts on the armature connect an outer ring of contacts, B, to the transmission line. In the diagram, the armature is in its stop position. Current from the battery in the sender flows through the contact labeled stop, through the armature, and through the line to the receiver. This current causes the receive relay to remain energized, keeping the clutch disengaged, the armature stationary, and the teleprinter in the idle condition.

Now suppose that an operator at the sending device presses the H key on the keyboard. In accordance with the 5-bit code in *Figure 1-4*, data contacts 3 and 5 close to encode the H, and the start contact at the sending machine closes. This energizes the start magnet to engage the clutch to turn the armature of the sending device counterclockwise. The armature connects the contacts on the outer ring to the line in sequence: start, 1, 2, 3, 4, 5, stop. Notice that the start contact and data contacts 1, 2, and 4 do not conduct current.

Figure 2-9. Electromechanical Start-Stop Coding and Decoding



Bit timing *within* a character must be precise to prevent errors. However, the timing *between* characters is of no importance for accuracy.

As soon as the armature of the sender travels to its start contact, the current through the line and the receive relay ceases. The receive relay is de-energized, and current flows through the other set of relay contacts and through the start magnet of the receiver. This engages the clutch and the receiver armature begins turning counterclockwise at approximately the same speed as that of the sender armature. When the sender armature is over data contact 3, current flows through the line and the receive relay. The receive relay is energized, and as the armature of the receiver passes over data contact 3 on the outer ring, current flows through select magnet 3. Select magnet 5 is operated similarly. The operation of these two magnets cause the letter H to be selected on the type mechanism of the receiver. When the sender armature reaches its stop contact, the receiver armature passes over contact C, current flows to the print magnet, and the letter H is printed. Both armatures then come to rest in the position shown in the figure, unless the sender is ready to transmit another character immediately.

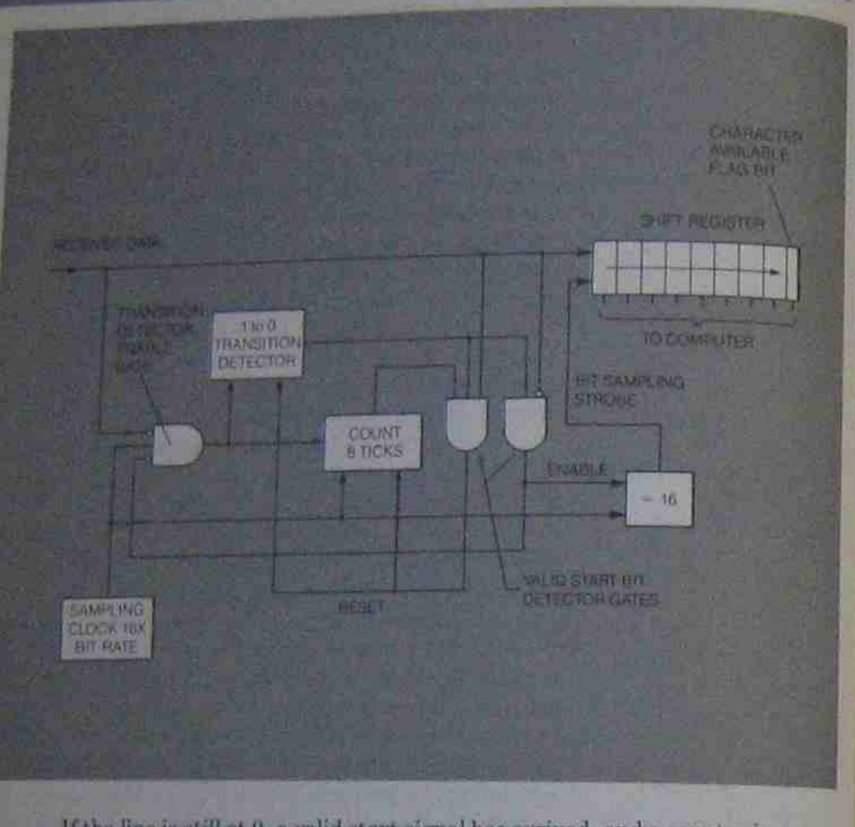
Note that the start and stop elements have those exact functions; that is, they start and stop the armature rotation. As you can visualize, if there were much variation in the rotational speed of the sending and receiving armatures, the armatures would not be in synchronization and the character would be decoded incorrectly. As discussed previously, this is one reason that the Baudot code was limited to five symbols.

With start-stop transmission, a new character can begin at any time after the stop bits of the preceding character have been received. The time between one character and the next is indeterminate, but within a character, the timing is precisely defined. For that reason, it might be better to call this technique "self-synchronized" or "internally synchronized", rather than asynchronous, meaning "not synchronized". (The lack of a continuous synchronous agreement between the transmitter and the receiver — specifically, the lack of a clocking signal within or accompanying the data channel — is the reason for the name "asynchronous".)

Electronic Receiving Circuit

The electromechanical arrangements of the early teleprinters have been replaced with electronic circuits. Since these can be synchronized within closer tolerances, an 8-bit code can be used. An electronic circuit for receiving asynchronous serial data in an 8-bit code is shown in *Figure 2-10*. It utilizes a clock which runs at 16 times the symbol rate of the incoming data. This rapid rate is used in order to detect the 1 to 0 transition (when the start bit begins) as soon as possible after it occurs. The circuit which detects the 1 to 0 transition enables a spike detection circuit. Eight "ticks" of the 16X clock (one-half a bit time) are counted, then the line is checked to see if it is still in the 0 state. If it is not, it is assumed that the initial 1 to 0 transition was due to noise on the line, the spike detection circuit is reset, and no further action is taken.

Figure 2-10. Asynchronous Receiver
(Source: McNamara, John E., Technical Aspects of Data Communications, 2nd Ed., Digital Press, © 1982)



If the line is still at 0, a valid start signal has arrived, and a counter is enabled which divides the 16X clock by 16 to produce a sampling clock which ticks once per bit time for the shift register. This tick occurs roughly at the center of the bit being sampled. The off-center error can be made smaller by sampling at 32 times the bit rate, and even further reduced by sampling at 64 times the bit rate. However, when higher sampling rates are used, the counter in the spike detection circuit and the counter in the bit sampler circuit must count proportionately higher.

The advent of the more accurate electronic circuit has made possible the use of the 8-bit data communication code.

The bit sampler circuit strobes the shift register eight times to sample the state of the line to get the eight bits into the serial-to-parallel shift register. Then a signal, called a flag, is sent to the computer or controller with which it is associated to announce that a character has been received. The computer then signals the shift register to transfer the eight bits in parallel into the processing circuits.

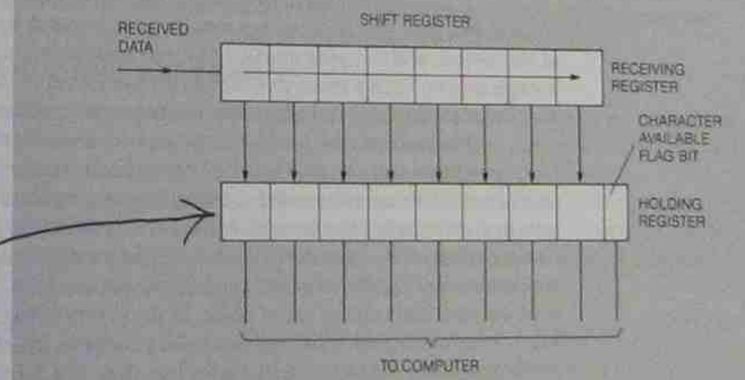
A single-buffered interface requires the receiving computer to read each character as soon as it is received with no break before the next one begins.

A problem with using only one shift register (called a single-buffered interface) is that, when characters are arriving continuously, the computer has only the duration of the stop bit to read the received character before the next character begins entering the register. A simple improvement is to provide a holding register into which the received character can be parallel transferred as soon as the eighth bit has been sampled. A character available flag is sent to the computer when the parallel transfer occurs, and the receiving register becomes available for the next character. This arrangement, called a double-buffered interface, is shown in *Figure 2-11*. In either case, the arrival of a character which cannot be handled because the previous character has not been read is called a data overflow. If overflow occurs, most receiver circuits overwrite the old character with the new one (the old one is lost), and place an error signal on a separate lead to the computer.

Stop Bits and Timing Errors

With a double-buffered interface, the stop bit arrival time is no longer the only time available to the computer for reading the character, so there is time for circuitry to check the stop bit (sometimes called the ninth bit) to confirm that it is a 1. If it is not a 1, one of the following conditions exists: the communications channel is broken, receiver timing is confused, or the transmitting station is sending a special signal. The existence of one of these conditions typically provides an error signal on a lead in the asychronous receiver.

Figure 2-11. Double-Buffered Interface
(Source: McNamara, John E., Technical Aspects of Data Communications, 2nd Ed., Digital Press © 1982)



Allows time to read the character while another one is being received.

A double-buffered interface provides a holding register to hold the character for reading, while the next character received fills the receiving register.

rge data communica-
ns systems normally do
use asynchronous
nsmission because the
rt-stop codes take too
ch time, and error de-
tion and correction
hods are impractical.

Such an error occurs when the receiver has lost track of which zeros in the transmission are the start bits and which are just zeros in the data. If, for some reason, the receiver treats a data bit as a start bit, it will assemble the next eight bits as a character. Since these eight bits are really parts of two characters (the end of one and the beginning of another) the ninth bit to arrive will not be the stop bit, but rather a data bit from the second character. If it is a zero, the error checking circuit will detect an error. Since framing is the process of deciding which groups of eight bits constitute characters, and this error is due to a failure in that process, it is called a framing error.

Failure in the framing process generally can be avoided. The idle line condition is a mark condition, and any amount of idle time more than a character time in length will correct this condition. Even if characters are sent continuously, the receiver will eventually become realigned to the correct start elements, no matter which zero bit is chosen as the start bit. Of course, the characters received while the framing is not in alignment will be decoded incorrectly.

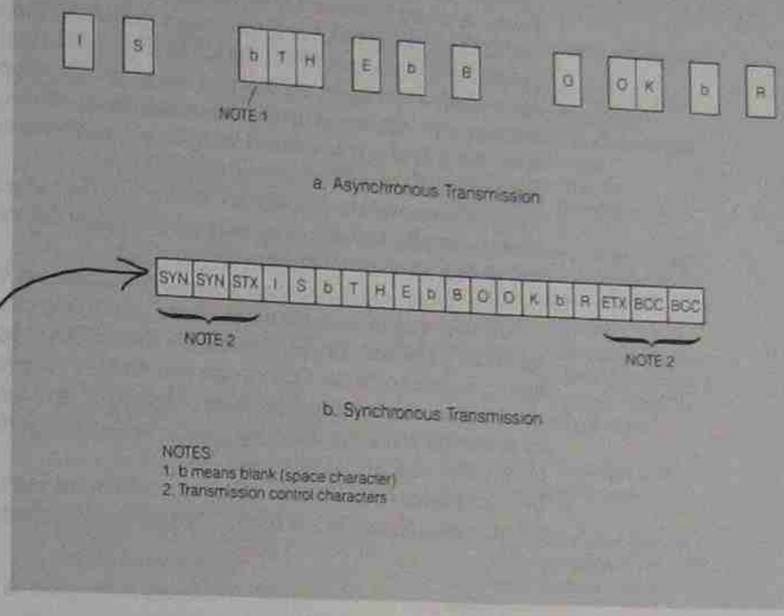
Other Types of Transmission

Asynchronous transmission is the most common in data communications simply because there are more low-speed terminals and small computer applications in which it is used. Large systems and computer networks usually utilize other methods in addition to asynchronous. This is because of the large overhead penalty of 20% associated with start-stop codes; that is, two (the start and stop bits) out of the ten bits transmitted are for control rather than information. This is not a problem in conversational timesharing, where more time is spent in looking at the screen and thinking than in transmission, but it would be a heavy price to pay when transferring a file of 500,000 characters (4 million bits) at 9600 bps.

A second problem in such large transfers is error checking. The timesharing user checks his or her own input and output for errors by looking at the screen, and rekeying or asking for retransmission of portions that contain errors. Such a procedure is clearly impractical for long file transfers that occur at fast rates and often without an operator present.

The method used to solve these and other problems of large-volume, high-speed data transmission is called synchronous transmission. In synchronous transmission, start and stop bits are not used. Characters are sent in groups, called blocks, with special synchronization characters placed at the beginning of the block and within it, to ensure that enough 0 to 1 or 1 to 0 transitions occur for the receiver clock to remain accurate. Error checking is done automatically on the entire block. If any errors occur, then the entire block is retransmitted. This technique also carries an overhead penalty (nothing is free), but the overhead is far less than 20% for blocks of more than a few dozen characters. *Figure 2-12* compares asynchronous and synchronous transmission of a sequence of characters.

Figure 2-12. Comparison of Asynchronous and Synchronous Data Transmission



Characters sent in blocks.

In addition to asynchronous and synchronous transmission methods to achieve receiver timing, there is now a third method called isochronous transmission. This technique is used in most modern computer networks. It involves timing that is controlled by the network rather than the DCE or DTE, as is true for synchronous and asynchronous transmission. Synchronous and isochronous data transmission will be discussed later in this book.

WHAT HAVE WE LEARNED?

1. Compared to CRT terminals, teleprinters are much slower in printing, have limited formatting capability, don't have enhancements such as blinking and highlighting, and are less reliable in operation.
2. A serial or character printer prints one character at a time. The printer may be either an impact or a non-impact printer. An impact printer may produce fully-formed characters or characters formed by a dot matrix. A non-impact printer usually uses a thermal or ink-jet method.

3. CRT terminals may be classified as dumb, smart, and user-programmable. A dumb terminal is essentially the same as a teleprinter except for printing speed. A smart terminal has capabilities such as editing, formatting, highlighting, and blinking. A user-programmable terminal has a built-in programmming language and the capability to store and execute programs.
4. As applied to terminals, ergonomics has to do with the color, character size and non-glare surface of the display, and with the shape, position, and layout of the keys. It also is concerned with the physical position of both the display and the keyboard with respect to the operator.
5. In serial transmission, the bits are transmitted one at a time over a single wire. In parallel transmission, each bit of a group moves over its own wire and all bits of the group move at the same time.
6. Both bit timing and framing are necessary to maintain synchronization for proper decoding in data communications. Bit timing usually is accomplished by clocking the bits. Framing usually is accomplished by inserting special bits or characters in the data stream just for that purpose.
7. In asynchronous transmission, each character is sent by itself and is framed by a start bit and a stop bit. This is the transmission method used by teleprinters and ASCII terminals.
8. In synchronous transmission, many characters are sent together in groups called blocks. Each block is framed by special synchronization characters.

Quiz for Chapter 2

1. Teleprinters:
   a. Are only for printing at remote locations, not for input.
   b. Offer both high speed operation and a variety of formatting controls.
   c. Have a printer for output and may have a keyboard for input.
2. Impact printers:
   a. Strike a ribbon against the paper to produce character images.
   b. Include ink-jet and thermal devices.
   c. Are rapidly becoming obsolete.
3. Glass teletypes:
   a. Are among the most recent developments in CRT terminals.
   b. Were so-named because they had the same interface as a teleprinter.
   c. Are teleprinters designed to interface with fiber-optic transmission systems.
4. Electromechanical teleprinters:
   a. Use a complex mechanical buffer to match speed between transmitting and receiving machines.
   b. Use start/stop codes to synchronize sending and receiving equipment.
   c. Are rarely used today.
5. ASCII terminals are generally defined as:
   a. Terminals using synchronous transmission in EBCDIC.
   b. Terminals using synchronous transmission in ASCII.
   c. Terminals using asynchronous transmission in ASCII.
   d. Any terminal having an American (dollar-sign) keyboard.
6. "3270" terminals refer to:
   a. Asynchronous terminals made by IBM.
   b. Any terminal which is painted blue.
   c. Synchronous terminals that interface with an IBM-type cluster controller.
7. The difference between timing and framing is:
   a. Timing is concerned with the individual bits; framing is concerned with the boundaries between characters.
   b. Timing refers to serial transmission; framing refers to parallel.
   c. Timing is concerned primarily with asynchronous systems; framing is concerned with synchronous systems.
8. Escape codes are so called because:
   a. In effect, they provide a means to temporarily "escape" from the standard meanings of the character set.
   b. They initiate operation of the escapement mechanism in teleprinters.
   c. They cause the cursor to escape from the boundaries of the CRT screen and roam around in memory.
9. The QWERTY keyboard:
   a. Is still considered to be the layout allowing the greatest typing speed.
   b. Is the most popular keyboard, but not necessarily the best.
   c. Is a key layout that is rarely used.
10. Memory-mapped displays:
    a. Are associated with electromechanical teleprinters.
    b. Have the advantage that they do not take up memory space.
    c. Allow direct addressing of display locations by the processor.
11. Serial printers
    a. Are used to transmit grain prices.
    b. Are faster than CRT terminals, and offer more flexibility.
    c. Print one character at a time.
    d. Usually use serial interfaces.
12. Non-impact printers:
    a. Are normally quieter than impact printers.
    b. Generate carbon copies easily.
    c. Produce fully-formed characters.

13. CTR terminals:
    a. Are the most widely-used hard-copy terminals.
    b. Offer high-speed display and formatting flexibility.
    c. Do not normally utilize microprocessors.
14. User-programmable terminals:
    a. Are replacing personal and professional computers.
    b. Offer more flexibility at lower cost.
    c. Are being replaced by personal and professional computers.
15. Ergonomics:
    a. Involves the interface between people and machines, such as terminals.
    b. Is the application of ergo-economics to communications.
    c. Utilizes three-level ergo-coding for transmission over certain channels.
16. Serial and parallel transmission:
    a. Differ in how many bits are transferred per character.
    b. Are used in synchronous and asynchronous systems, respectively.
    c. Differ in whether the bits are on separate wires or all on one.
17. Memory-mapped displays:
    a. Are utilized for high-resolution graphics, such as maps.
    b. Uses ordinary memory to store the display data in character form.
    c. Stores the display data as individual bits.
18. Asynchronous transmission:
    a. Is less efficient than synchronous, but simpler.
    b. Is much faster than synchronous transmission.
    c. Is another name for isochronous transmission.
19. Single-buffering:
    a. Is more efficient than double-buffering.
    b. Is less efficient than no buffering.
    c. Provides very little time to unload the incoming character from the register.
20. Most terminal keyboards:
    a. Provide numerous additional specialized keys.
    b. Are strictly typewriter-style, with few extra keys.
    c. Use the Dvorak layout.

Messages and Transmission Channels

ABOUT THIS CHAPTER

Useful communication requires four elements as shown in *Figure 3-1:* a message (information) to be communicated, a sender of the message, a medium or channel over which the message can be sent, and a receiver. Messages usually are information that is useful to people, but the sender and receiver may or may not be human. The medium must be one suitable to convey the type of message. In this chapter, we will examine in more detail the types of messages and the media which carry them.

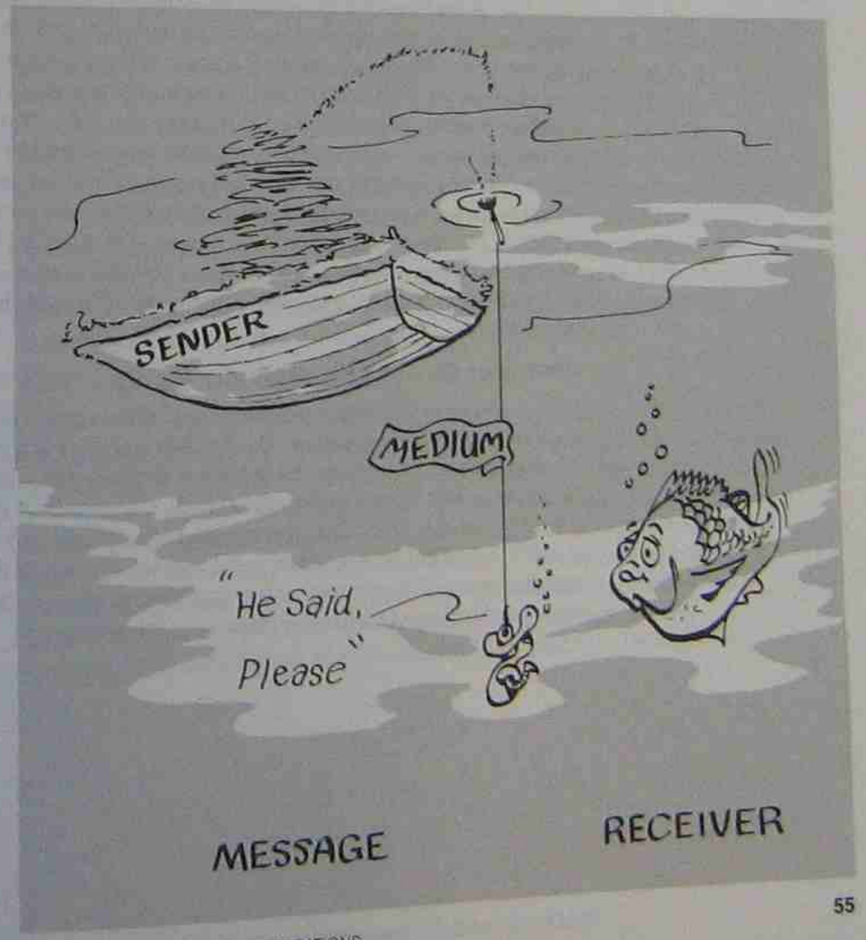


Figure 3-1. The Elements of Communication

WHAT IS INFORMATION?

Information as a Quantity

The unit of information used in data communications is the bit which has only two possible values, 1 and 0. Seven bits are needed to select any one of 128 different symbols in ASCII.

"Information," according to one definition, "is the communication or reception of knowledge or intelligence." Information is also "a numerical quantity that measures the uncertainty in the outcome of an experiment to be performed." This second definition has an application in the sending of messages. For example, suppose we have a machine that can send only two symbols, A1 and A2. We can then say that the "experiment" is the accurate recognition of the two symbols (A1 and A2) being sent from one machine to another. As far as the receiving machine is concerned, it is just as likely to receive one symbol as the other. So we can say that the "numerical quantity" in this experiment is a unit of information that will allow a selection between two equally likely choices. This quantity, or unit of information is usually called a bit (a contraction of binary digit), and has two possible values, 0 and 1. If these two values are used to represent A1 and A2, then A1 could be represented by the bit value 0, and A2 by the bit value 1. The number of bits per symbol is 1, but we still need a way of selecting which symbol (bit) we want to use. A machine which needs only two symbols only needs a 1-bit select code (0 and 1).

A machine that uses only two symbols is not of much use for communication, but suppose the machine used 128 symbols (like the ASCII character set). Then the number of equally likely choices to be handled is 128, and the number of bits (information) required to represent each of those 128 symbols is seven (refer to *Table 1-1*). You can see, then, that if the knowledge (or intelligence) to be communicated can be represented by a set of equally likely symbols, then the amount of information required per symbol to communicate it is necessarily dependent upon the total number of bits of information. The ASCII is particularly useful for selecting the information to be communicated since it can select one of 128 ASCII symbols with only one 8-bit byte, a common bit grouping in computers (the eighth bit is not used in this case).

Information Content of Symbols

In many information systems, not every symbol is equally likely to be used in a given communication. The English language is a good example. In a message written in English, the letter e is 12 times more likely to occur than the letter s. This uneven distribution is also characteristic of particular groups of letters and of words. What this means, then, is that each of the 128 symbols in ASCII is not likely to occur an equal number of times in any given communication. For example, notice that the use of the letters e and g, and the letter combinations th and er, is unequal in this paragraph.

Redundancy is the measurement of the probability of occurrence of any particular character in a character set.

In 1949, Claude Shannon published a book titled *The Mathematical Theory of Communication*. In this book, he discusses the uncertainty or amount of disorder of a system, which he called entropy. The entropy of a set of equally likely symbols (like the digits 0-9 in a table of random numbers) is the logarithm to the base 2 of the number of symbols in the set. The entropy of the English *alphabet*, which contains 26 letters and a space, is then $\log_2(27) = 4.76$ bits per symbol. Because of the uneven use of letters in the English *language*, however, its entropy was estimated by Shannon as 1.3 bits per symbol. This means that the language is about 70% redundant, and that it should be possible to reconstruct English text accurately if every other letter is lost or changed due to noise or distortion. Obviously, redundancy is desirable in order to raise the chances of receiving a good message when the medium is noisy. (The words noise and noisy in this book refer to electrical noise; that is, an electrical signal that is not supposed to be present.)

Redundancy can be used to reconstruct a message even if part of the message is lost during transmission. It is very important in verifying data accuracy.

Using Redundancy in Communications

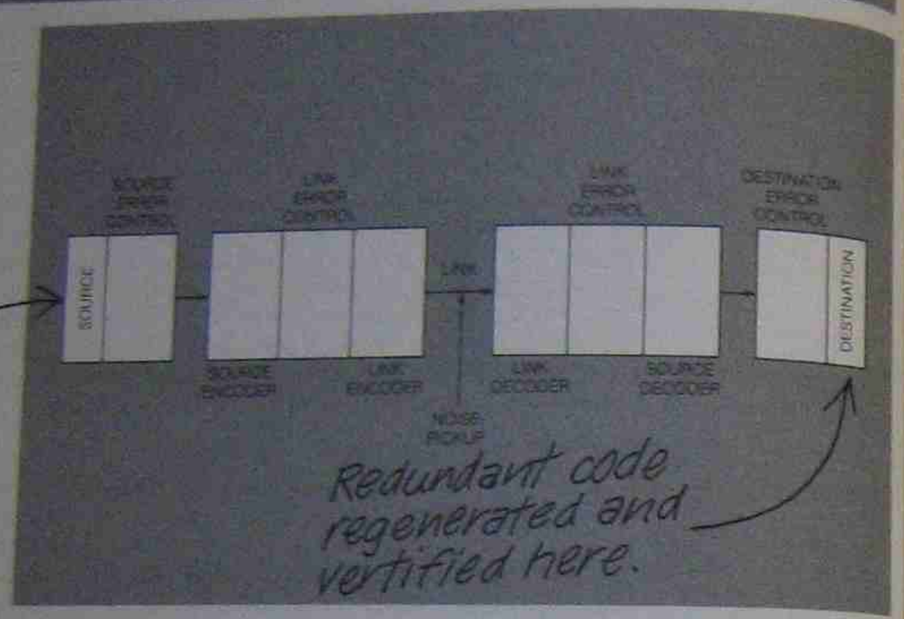
So, you wonder, what does all this have to do with the real world of data communications? Quite a bit, since almost every scheme in current use for sending data uses redundancy in an attempt to verify that the data have been received exactly as sent; that is, no errors have been introduced by the sending mechanism, the transmission medium or the receiver. The redundant information may consist simply of a retransmission of the entire original message. Although simple to implement, retransmission is not efficient, therefore, special techniques are used to generate redundant information that is related to the message in a way that is known to both the sender and receiver. The sender generates the redundant information during transmission and sends it along with the message. The receiver regenerates and checks the redundant information when the message is received. This scheme is represented by *Figure 3-2*. Verification usually occurs at the end of each link in the chain making up the transmission path. The details of this process and various methods in current use are described in a later chapter.

WHAT CARRIES THE MESSAGES?

Bounded data channels confine the signals within wires, coaxial cable or optical fiber cable. Unbounded channels permit the signal to radiate in all directions.

The physical channels (the media) that carry data are of two types; bounded and unbounded. In a bounded medium, the signals are confined to the medium and do not leave it (except for small leakage amounts). A pair of wires, coaxial cable, waveguide, and optical fiber cable are examples of a bounded media. The atmosphere, the ocean, and outer space are examples of unbounded media, where electromagnetic signals originated by the source radiate freely into the medium and spread throughout the medium. The unbounded media are utilized by various radio frequency transmitting schemes to carry messages. The main feature of unbounded media is that once the signal is

Figure 3-2.
Error Checking Points



radiated from the transmitter, it radiates equally in all directions (unless restricted) and continues forever onward. As it moves farther from the source, the energy is spread over a larger area so the level continually gets weaker at greater distances. As the wave moves through the medium, it is affected by natural disturbances which can interfere with the signal.

Wire Pairs

Wire pairs are the simplest type of bounded carrier. The two wires which provide a go and return signal path, when balanced to ground, provide the most protection from induced noise.

The simplest type of bounded medium is a pair of wires providing a go and return path for electrical signals. Early telegraph systems used the earth itself instead of a wire for one of the paths as shown in *Figure 3-3a* with repeaters inserted along the line to reduce the effects of noise and attenuation (loss of signal strength). However, this scheme did not work well because the earth is not always a good conductor and the path was susceptible to large noise currents induced by lightning. Losses were reduced by using two wires as shown in *Figure 3-3b*, but the line was still unbalanced to ground so it was subject to picking up noise from almost every noise-producing device. Finally, the balanced two-wire line shown in *Figure 3-3c* was used to greatly reduce noise pickup.

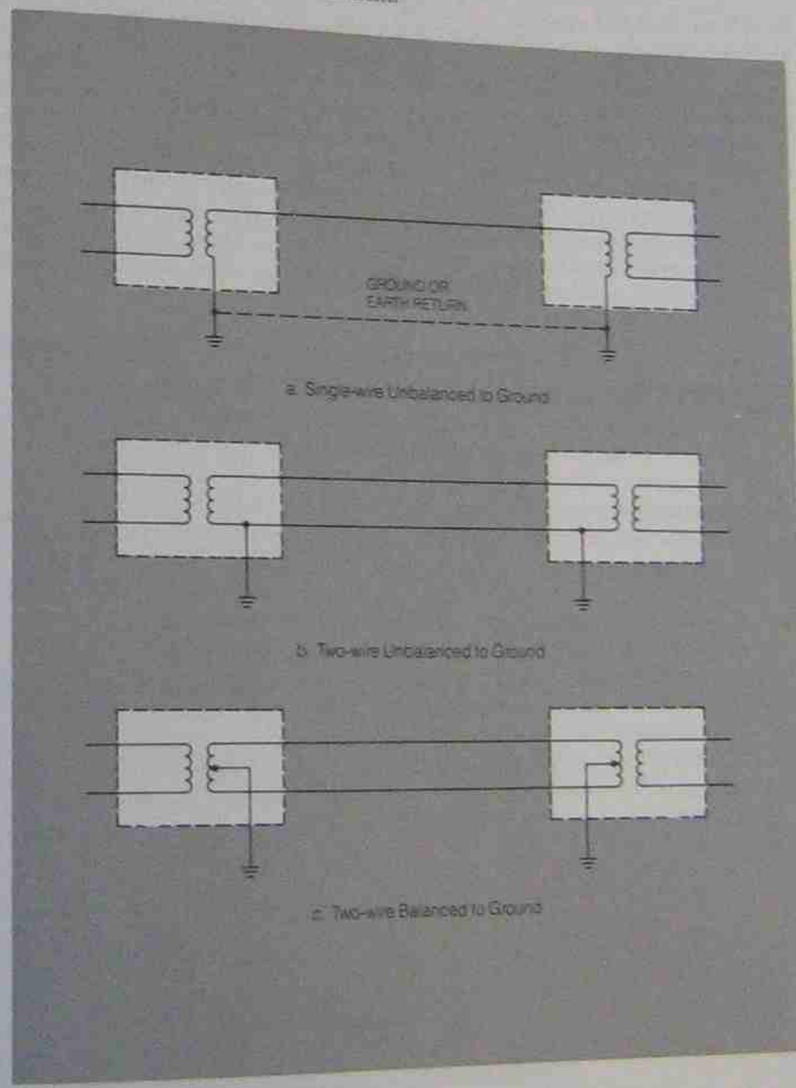
The most common type of bounded medium consists of wire pairs twisted together and made up into cables of from 4 to 3000 pairs. The size of the wire used varies from 16 AWG (American Wire Gauge) with a wire diameter of 0.05082 inch to 26 AWG with a diameter of 0.01594 inch. In modern cables, each wire is insulated with a polyethelene or a polyvinyl chloride jacket; however, there is still a large quantity of older cable in use in which the insulation for each wire is paper.

The size of the wires used and the distance between them affect the attenuation. Twisted wire pairs are limited to a maximum frequency of about 1 MHz.

Open-wire lines have low attenuation at voice frequencies due to the large size of the wire and the relatively large distance between the two wires when mounted on a crossarm of a utility pole. A typical value of attenuation for 104 mil (0.104 inch) diameter open wire lines is 0.07 decibel (dB) per mile; while 19 gauge (0.03589 inch diameter) twisted wire pairs in a multipair cable have a voice frequency attenuation of about 1 dB per mile.

The attenuation of twisted wire pairs rises rapidly with increasing frequency, and the amount of crosstalk between adjacent pairs also increases with frequency. The maximum usable frequency for wire pairs in cables is around 1 MHz without special treatment.

Figure 3-3.
Types of Transmission Circuits



The Effect of Inductance

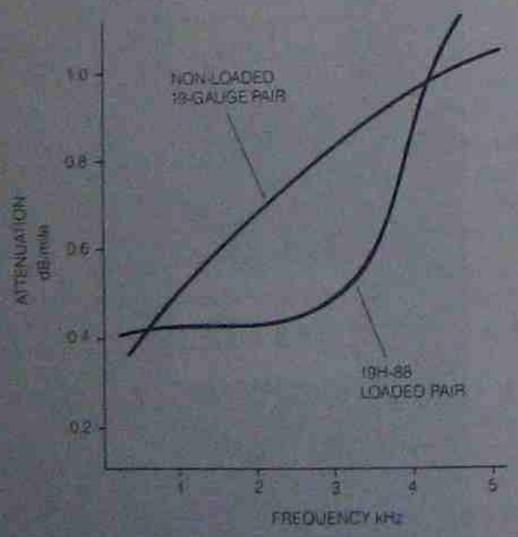
A concept that may not be obvious about paired wire circuits is that the addition of inductance in the line can help reduce attentuation at voice frequencies. The line impedance (ac resistance) is increased, so that a given amount of power can be transmitted with less current, but at a higher voltage. The result is a reduction in the series losses and an increase in the shunt losses. Since the series losses are usually the most severe, there is a net reduction in attentuation until the inductance rises to the point where series and shunt losses are equal.

By loading a circuit with added inductance, attenuation can be reduced at higher frequencies so the circuit response is relatively constant over the voice band.

Adding inductance to wire pairs is called loading, and a circuit to which inductance has been added is called a loaded line or loaded circuit. The effect of loading is illustrated in *Figure 3-4*. The typical frequency versus attenuation performance is shown for a nonloaded 19 gauge cable pair and a 19 gauge cable pair loaded with 88 millihenrys of inductance every 6000 feet. (The standard notation for this is 19H-88 loaded pair.) *Figure 3-4* shows that the attentuation of the loaded circuit is less than that of the unloaded one, and that it changes very little with increasing frequency up to a certain point above 3 kHz. This point is called the cutoff point or cutoff frequency.

Loading was introduced around 1900 on long distance open wire lines to reduce losses due to attenuation, since there were no amplifiers for the signals.

Figure 3-4.
Effect of Inductance Loading



Electronic Amplifiers

Electronic amplifiers eliminate the requirement to load long distance circuits because amplification can compensate for the line losses.

In 1883, Thomas Edison discovered the rectifying properties of the thermionic vacuum tube. A thermionic vacuum tube is one in which a stream of electrons is emitted by an incandescent substance. Edison's tube was a two-element tube (called a diode) consisting of a cathode (the incandescent substance) and an anode.

However, the principle of thermionic vacuum tubes lay unused until 1904 when Sir John Ambrose Fleming, an English physicist and engineer, adapted the diode for use as a demodulator (detector) of radiotelegraph signals for the Marconi Wireless Telegraph Co. In 1906, Lee DeForest introduced a third element, called a control grid, to the diode and created the triode. By 1912, the triode and its associated circuits were developed for use as an amplifier.

When the use of the DeForest triode as an amplifier began in 1914, it was no longer necessary to load long distance circuits because the losses in the line could be compensated for by amplification. However, loading is still used on longer local loops (from the telephone office to the customer), since it is cheaper than adding active components for amplification. The presence of loading on local circuits has a considerable effect on their ability to carry high-frequency data signals which causes problems for some new types of telephone service.

Coaxial Cable

To put more data transmissions onto one channel at the same time, the transmissions must be at frequencies higher than those practical for wire pairs. Coaxial cables can be used for signals with frequencies of 10,000 MHz.

In order to make telephone service economical, a way had to be found to put more than one conversation onto a channel. Indeed, the invention of the telephone arose out of Alexander Bell's experiments on a "harmonic telegraph," an attempt to put more than one telegraph signal on a channel. Putting more conversations or more data on a single channel requires a larger bandwidth (ability to carry more frequencies), which, as a practical matter, means higher frequencies. Since the practical limit for wire pairs is around 1 MHz, some other method had to be developed.

Some significant and interesting effects occur in the vicinity of a wire carrying an alternating current signal. One of these effects is that both an electric field and a magnetic field are created around the conductor. The magnetic field can induce the signal it is carrying into adjacent conductors. (In communications, the induced and unwanted signal is called crosstalk.) However, if one conductor of the pair is the ground side of the circuit, and is made to surround the other conductor, both the radiated electric field and the magnetic field can be confined within the tube formed by the outer conductor as illustrated in *Figure 3-5*.

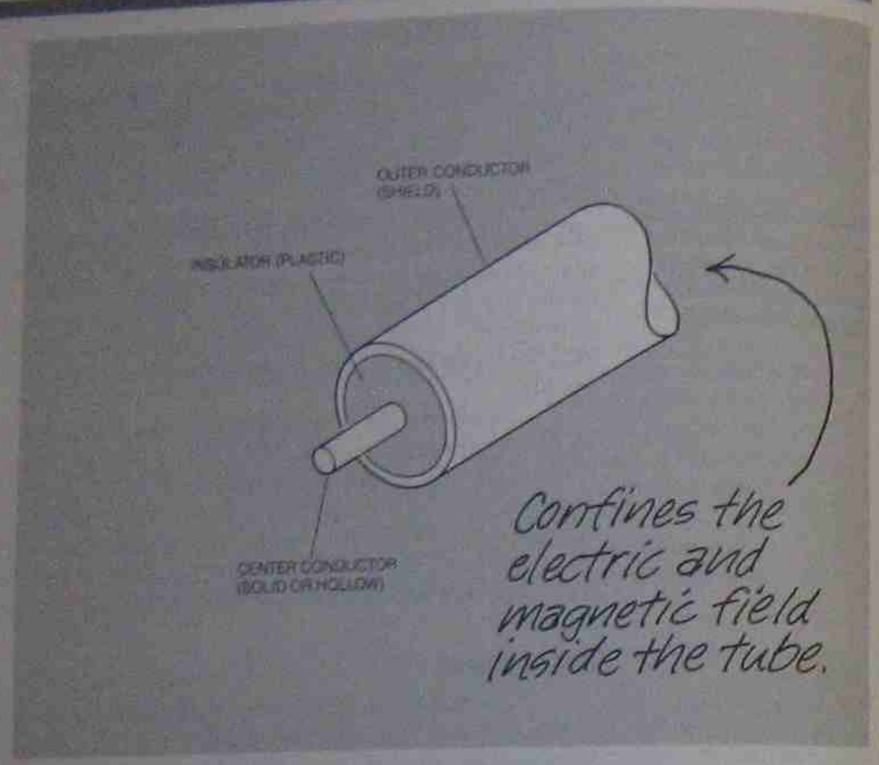


Figure 3-5.
Structure of Coaxial Cable

This medium is called a coaxial cable, because the two conductors have a common axis. The self-shielding works well at frequencies above about 100 kHz, but at low frequencies the "skin depth" of the current is comparable to the thickness of the outer conductor and the shielding becomes ineffective. The resistive loss of coaxial cable increases as the square root of the frequency, making coaxial cable generally usable at frequencies of up to 2000 MHz, although some types may be used up to 10,000 MHz.

Waveguide

Waveguides can be used to frequencies of 2,000 MHz.

If the frequency of transmission is high enough, the electric and magnetic components of a signal can travel through free space, requiring no solid conductor at all. However, to avoid interference and losses due to signal spreading, and to be able to route the signal as desired, it is sometimes useful to confine these waves to another bounded medium called a waveguide.

Waveguides are commonly used at frequencies from 2000 MHz up to 110,000 MHz to connect microwave transmitters and receivers to their antennas. Waveguides are pressurized with dry air or nitrogen to drive out moisture from inside the waveguide because moisture attenuates the microwaves. Older waveguides were constructed with a rectangular cross section, but common practice today is to make the guides circular as shown in *Figure 3-6*. Waveguides remain in use as a conductor of high-power, high-frequency signals, but optical fiber cables are being used in newer systems.

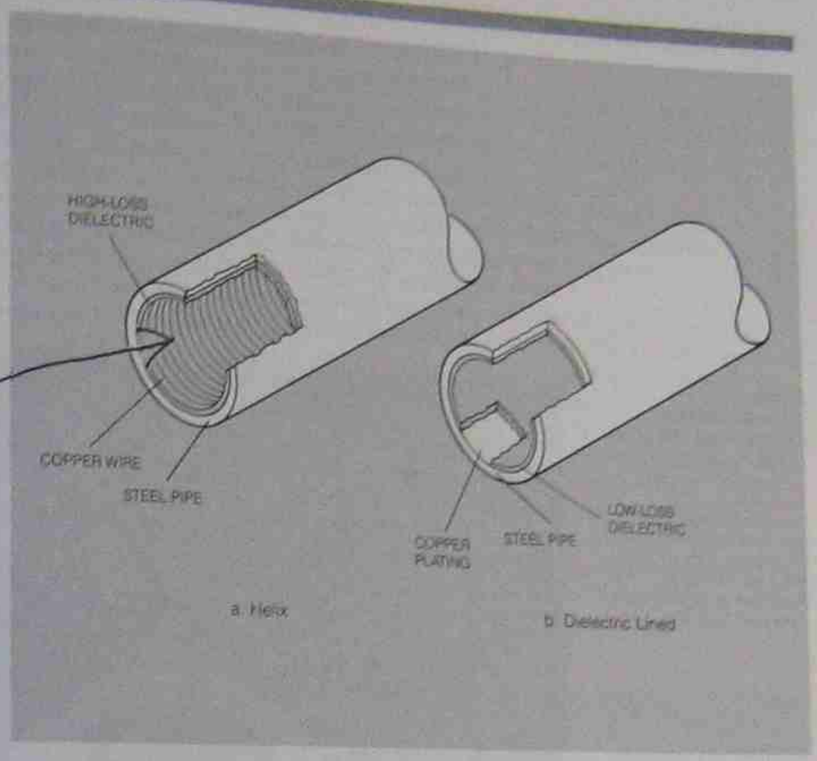


Figure 3-6.
Circular Waveguides
(Source: Members of the Technical Staff, Bell Telephone Laboratories, Engineering and Operations of the Bell System. © 1977, AT&T Bell Telephone Laboratories, Inc., used by permission.)

Fiber optic systems can handle high capacity, high frequency transmissions economically because they have low loss, are free of electromagnetic interference, are non-conductive, non-inductive, and do not radiate energy.

Fiber Optic Systems

The capacity of a transmission system is a direct function of the highest frequency it can carry, therefore, progress in transmission technology has been measured by the bandwidth of the media available to carry signals. Recent developments in the use of glass fibers to carry binary signals have shown these systems to be extremely well-suited to high data rate applications. Fiber optic systems are attractive for several reasons:

1. The low transmission loss, as compared with wire pairs or coaxial cable, allows much greater separation between repeaters. A fiber optic system with no repeaters has been demonstrated which transmits 420 megabits per second (Mbps) over a span of 75 miles with an error rate lower than high-quality coaxial cable systems.
2. Since the optical fibers carry light rays, the frequency of operation is that of light. The transmission wavelength used for current single-mode fibers is 1.2 micrometers, equivalent to a frequency of around 800 terahertz (800,000,000,000,000 Hz). Such frequencies allow data transmission rates of 20,000 Mbps over short distances.
3. Optical fiber cables do not radiate energy, do not conduct electricity, and are noninductive. They are essentially free from crosstalk and the effects of lightning-induced interference, and present no security problem from an inductively coupled "wire tap."

4. Optical fiber cables are smaller, lighter and cheaper than metallic cables of the same capacity. It is economically feasible to provide several unused fibers in a cable for spares and for future growth. A cross section of a typical optical fiber cable is shown in *Figure 3-7*.

One standard fiber optic system currently in service is the AT&T FT3 lightwave system which can carry 80,000 two-way voice conversations at the same time. The cable for the this system is one-half inch in diameter and contains 144 fibers. Each fiber pair operates at 90 Mbps for a total data rate of about 6000 Mbps. The system provides one spare fiber for every operating one, and switchover to a spare is automatic upon loss of signal in a fiber. A more detailed discussion of fiber optic systems is given in a later chapter.

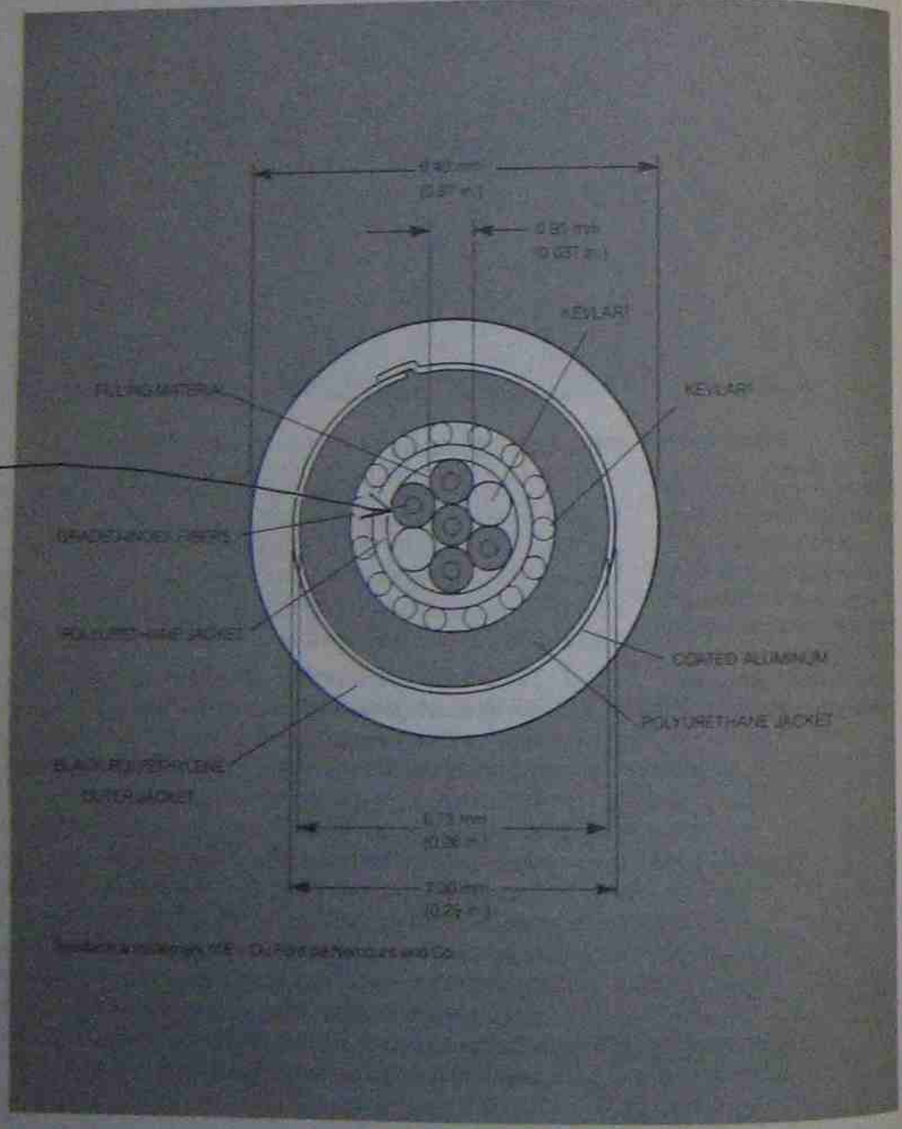


Figure 3-7.
Typical Five-Fiber Optical Fiber Cable for Direct Burial
(Source: R.L. Freeman, Telecommunication Transmission Handbook, 2nd Ed., John Wiley & Sons, 1981. Copyright © 1981 by John Wiley and Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

Carry the light waves.

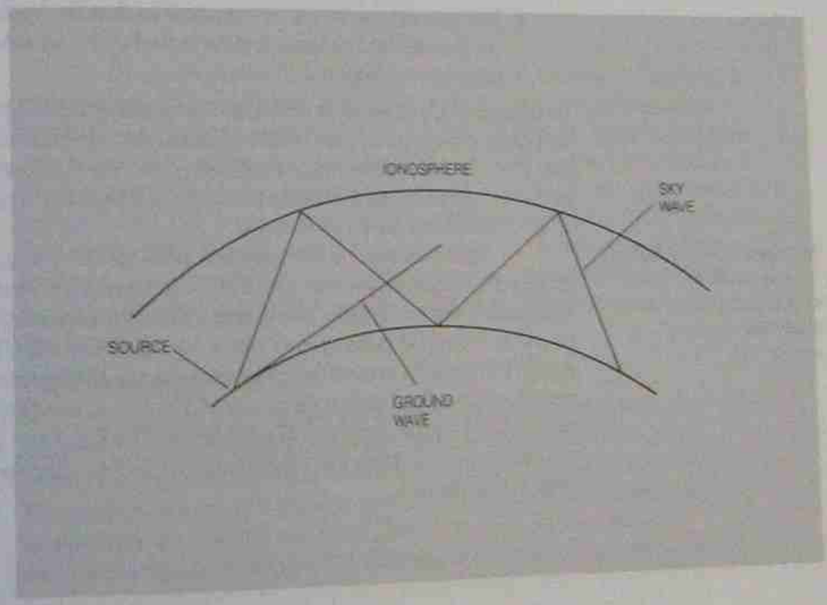
High-Frequency Radiotelephone

By convention, radio transmission in the frequency band between 3 MHz and 30 MHz is called high-frequency (HF) radio. Frequency bands within the HF spectrum are allocated by international treaty for specific services, such as mobile (aeronautical, maritime, and land), broadcasting, radio nagivation, amateur radio, space communications, and radio astronomy. HF radio has properties of propagation that make it less reliable than some other frequencies; however, HF radio does allow communications over great distances with small amounts of radiated power.

High frequency (HF) signals radiate from an antenna over two paths: a ground wave following the earth's surface and a skywave that bounces between the earth and ionosphere.

HF radio waves transmitted from antennas on the earth follow two paths as indicated in *Figure 3-8* when they leave the antenna. The groundwave follows the earth's surface and the skywave bounces back and forth between the earth's surface and various layers of the earth's ionosphere. The groundwave is useful for communications up to about 400 miles, and works particularly well over water. The skywave propagates signals for up to 4000 miles with a path reliability of about 90%. Data signals are carrried on HF radio systems as continuous wave (CW) radio telegraphy at about 15 bits per second (bps), and frequency shift keyed (FSK) single sideband signals are carried on HF at 75 bps. Higher data bit rates (up to 4800 bps) are converted to standard 3 kHz voice channel analog signals by modems and these analog signals are transmitted on voice frequency (VF) carrier systems using HF radio.

Figure 3-8.
Paths of Radio Waves



Microwave Radio

Microwave radio is used by line-of-sight carrier systems to carry large quantities of voice and data signals. It is affected adversely by atmospheric conditions and solid objects.

The tall towers with large horns and/or dish antennas that you see while driving through the country are the repeater stations for line-of-sight (LOS) microwave radio systems (sometimes called radiolink systems). Such systems have become popular for carrying large quantities of voice and data traffic for several reasons:

1. They require no right-of-way acquisition between towers.
2. They can carry very large quantities of information per radio system, due to their high operating frequency.
3. They require the purchase or lease of only a small area of ground for installation of each tower.
4. Because the wavelength of the transmitted signal is short, an antenna of reasonable size can focus the transmitted signal into a beam. This provides a much greater signal strength at the receiver without increasing transmitter power.

Radiolink systems are subject to transmission impairments which limit the distance between repeater points and cause other problems. The microwave signals are:

1. Attenuated by solid objects (including the earth) and, in addition, the higher frequencies are attenuated by rain, snow and fog.
2. Reflected from flat conductive surfaces (water, metal structures, etc.).
3. Diffracted (split) around solid objects.
4. Refracted (bent) by the atmosphere so that the beam may travel beyond the line-of-sight distance and be picked up by an antenna that is not supposed to receive it.

In spite of all these possible problems, radiolink systems are highly successful and carry a substantial part of all telephone, data and television traffic in the U.S. The microwave range of radio frequencies is allocated for various purposes by international treaty. Some of the frequency assignments for the United States are shown in *Table 3-1*.

Although there are microwave systems that send data in digital form, most send data as analog signals.

Most common carrier radiolink systems carry analog signals, principally frequency modulation (FM). There are a few systems, however, which carry digital signals. Two examples in extensive service in the U.S. are the AT&T 3A-RDS radio system which operates in the 11 GHz band, and the AT&T DR-18 radio system which operates in the 18 GHz band. The 3A-RDS system carries DS-3 digital signals at 44.736 Mbps, and the DR-18 system carries DS-4 digital signals at 274.176 Mbps. The DS-3 and DS-4 signals, which are made up of several lower bit-rate signals, will be discussed in more detail later in this chapter.

Table 3-1. Frequency Assignments for Microwave Radiolink Systems

Assigned by international treaty.

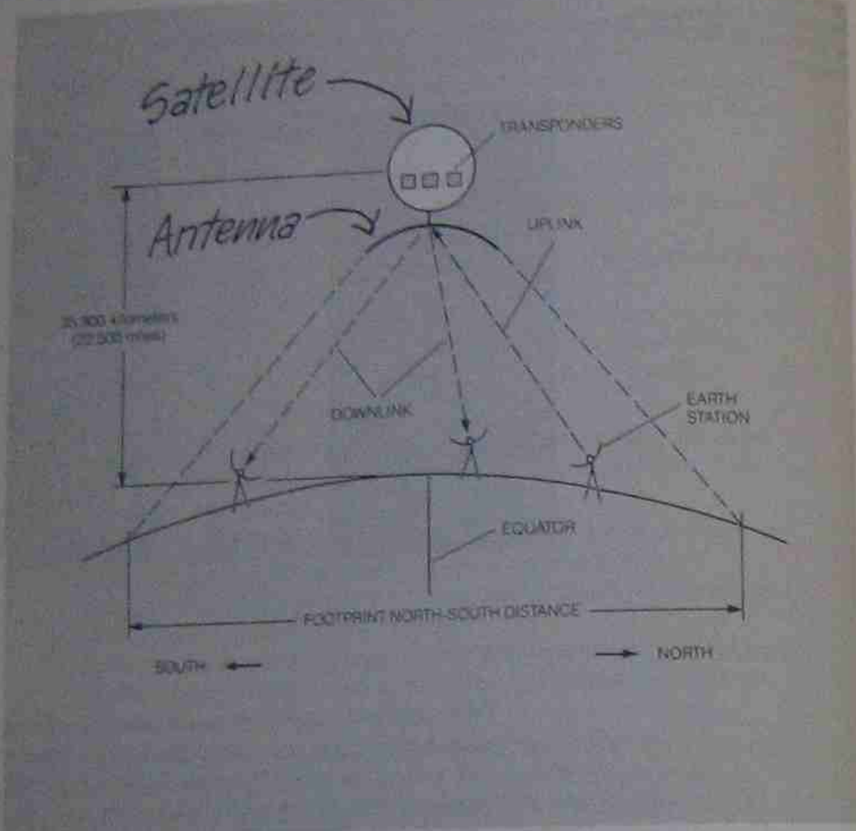
Service	Frequency, GHz	Service	Frequency, GHz
Military	1.710-1.850	Common Carrier and Satellite (uplink)	5.925-6.425
Operational Fixed	1.850-1.990	Operational Fixed	6.575-6.875
Studio Transmitter Link	1.990-2.110	Studio Transmitter Link	6.875-7.125
Common Carrier	2.110-2.130	Common Carrier and Satellite (downlink)	7.250-7.750
Operational Fixed	2.130-2.150	Common Carrier and Satellite (uplink)	7.900-8.400
Common Carrier	2.160-2.180	Common Carrier	10.7-11.7
Operational Fixed	2.180-2.200	Operational Fixed	12.2-12.7
Operational Fixed (TV)	2.500-2.690	CATV Studio Links	12.7-12.95
Common Carrier and Satellite (downlink)	3.700-4.200	Studio Transmitter Link	12.95-13.2
Military	4.400-4.990	Military	14.4-15.25
Military	5.250-5.350	Common Carrier	17.7-19.3

Terrestrial radiolink systems are point-to-point; that is, the signal is transmitted in a beam from a source microwave antenna across the earth's surface to the antenna at which it is aimed. The width of the beam transmitted by a microwave antenna varies between 1 degree and 5 degrees as a function of the frequency of transmission and antenna size. As a result, the transmission is highly directional, which is desirable if the information is intended for only one destination (for example, a telephone conversation). For many applications, however, the information has multiple destinations (for example, television broadcasts), which makes the satellite radiolink system more practical and desirable.

Satellite Radiolink Systems

Figure 3-9 is a simple model of a satellite radiolink system. The satellite contains several receiver/amplifier/transmitter sections, called transponders, each operating at a slightly different frequency. Each of the 12 transponders on each satellite has a bandwidth of 36 MHz. Individual transmitter sites, called uplink earth stations, send narrow beams of microwave signals to the satellite. The satellite acts as a relay station. A transponder receives the signal from a single transmitter, then amplifies it and retransmits it toward earth on a different frequency. Note that the transmitting earth station sends to only one transponder on a single satellite. The satellite, however, sends to any and all downlink receiving earth stations in its area of coverage, called its footprint.

Figure 3-9.
Satellite Radio Link System



A satellite radio link consists of an uplink transmitter signal to the satellite and a down link retransmitted signal to receiving stations on earth.

Commercial communications satellites are launched into geostationary orbit at an altitude of 35,900 kilometers (22,300 miles) above the equator. This means that the geostationary satellite is orbiting the earth at a constant speed and in the same direction as the earth's rotation about its axis. The orbiting speed is such that it causes the satellite to have a fixed location with respect to the earth. The earth station antennas, therefore, can be fixed in position and do not have to track a moving target in the sky. The angle of view for a geostationary satellite is almost 120 degrees wide. In principle, three such satellites equally spaced around the equator could cover the earth from 60 degrees north latitude to 60 degrees south latitude. In practice, the coverage angle is restricted to less than 110 degrees because the earth station's antenna must be elevated above the local horizon by more than five degrees.

Several types of signals are carried by satellite systems. For example, 6 MHz bandwidth standard televisions programs, multiplexed 64 kilobits per second (kbps) telephone channels, and high-speed data all can be carried simultaneously. One privately operated system, the Satellite Business System, is all digital with each of the 10 transponders per satellite capable of carrying 43 Mbps of digital data.

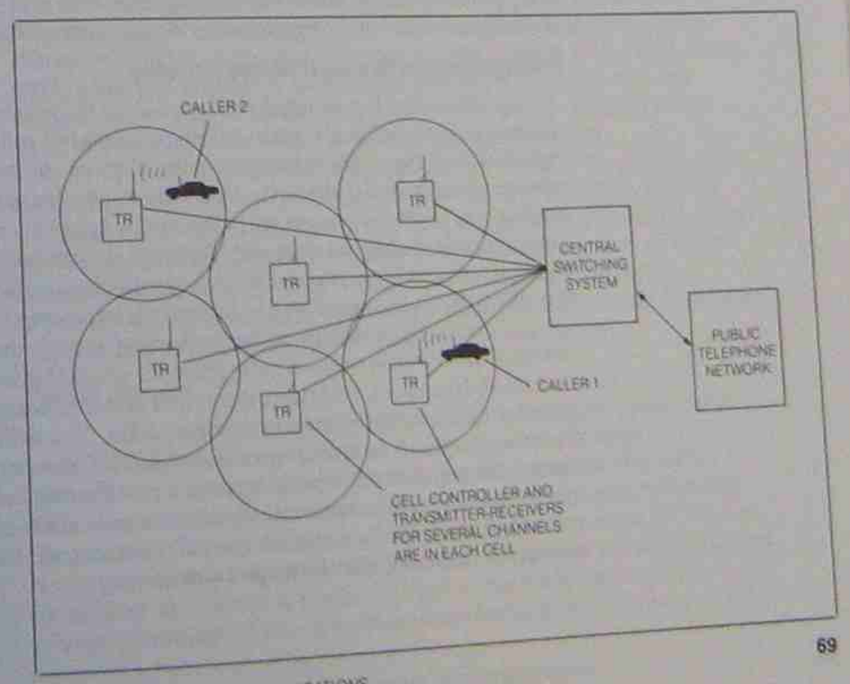
Cellular Radio Systems

The introduction of cellular radio systems effectively increased the number of radio channels available to mobile telephone systems.

Americans have demonstrated an insatiable desire to communicate with each other, anywhere, at any time. It seems that no location is too private, too noisy (audible noise), or too busy to exclude the installation of a telephone or a data terminal. Since Americans spend a lot of time in their cars, mobile telephones and data terminals are in great demand. Each telephone conversation requires a separate radio channel, and since only a limited number of such channels were available in the past, the demand for mobile telephone channels far outstripped the radio frequencies available to provide them. However, in 1982, a system allowing the reuse of channels within a metropolitan area, called the cellular radio system, began trial operation in Chicago. This system provides many more mobile telephone channels.

A diagram of a simple cellular system is shown in *Figure 3-10*. A metropolitan area is divided into several cells, each of which is served by a low-powered transmitter and associated receiver. The radio channels are suitable for data transmission up to 4.8 kbps as well as voice. The number of radio channels assigned to each cell is sufficient for the predicted number of users in that cell at any one time. When a caller makes a call, his mobile unit automatically seizes a free channel in his current cell. When the caller moves out of the cell, the cell controller automatically switches control of the call from the cell being left to the one being entered. Even a different radio channel may be used, but the caller doesn't have to do anything and is never aware of anything happening. The call is linked from the cell controller to a central switching system. The central switching system can link the caller via radio to another mobile or can access the public telephone network for connection to any fixed telephone.

Figure 3-10.
Cellular Radio System



EFFECTS OF BANDWIDTH ON A TRANSMISSION CHANNEL

Theoretically, a channel's data rate can be increased as desired as long as the signal-to-noise ratio increases. The practical maximum data rate limit was defined by Nyquist's theory to be two times the bandwidth.

All transmission channels of any practical interest are of limited frequency bandwidth. The limitations arise from the physical properties of the channel or from deliberate limitations on the bandwidth to prevent interference from other sources. For primarily economic reasons, most data communications systems seek to maximize the amount of data that can be sent on a channel.

Theoretical Information Handling Capacity of a Channel

Claude Shannon proved that the maximum capacity of an ideal channel whose only impairments are finite bandwidth and noise randomly distributed over that finite bandwidth is:

$$C = W \times \text{Log}_2[1 + (P/N)] \text{ bits per second}$$

where: P is the power in watts of the signal through the channel
N is the power in watts of the noise out of the channel
W is the bandwidth of the channel in hertz

Neglecting all other impairments, some typical values for a voice-grade analog circuit used for data are: W = 3000 hertz, P = 0.0001 watts (−10 dBm), N = 0.0000004 watts (−34 dBm). According to Shannon's Law, the value of C is:

$$3000 \times \text{Log}_2[1 + 250] = \text{about } 24{,}000 \text{ bits per second}$$

Shannon's value of C is never achievable because there are numerous impairments in every real channel besides those taken into account in Shannon's Law. Also, there are no ideal modems. However, Shannon's Law provides an upper theoretical limit to a binary channel. It is important to note that, due to the nature of the function Log_2, the value of C in the formula can be increased more readily by increasing W than by increasing (P/N).

Intersymbol Interference Reduces Capacity

One of the factors which tends to reduce the achievable capacity of a channel below the value of C in the formula is a problem called intersymbol (or interbit) interference. If a rectangular pulse like that shown in *Figure 3-11* is input to a bandlimited channel, the bandwidth limitation of the channel rounds the "corners" of the pulse as shown in the output waveform of *Figure 3-11* and causes an undesired signal to appear. The "tail" or overshoot part of this new signal interferes with previous and subsequent pulses, adding uncertainty to the signal; that is, the signal may be incorrectly interpreted at the destination.

Harry Nyquist analyzed the problem of intersymbol interference and developed an ideal rounded pulse shape for which that impairment is minimized. Nyquist also did much theoretical research dealing with sampling of analog signals for representation in binary form. Nyquist's Sampling Theorem says that if an analog signal is sampled 2f times per second, then the samples may be used to perfectly reconstruct the original signal over a spectrum f hertz. For example, if a signal is sampled at the rate of 8000 times per second, then those samples can be used to reconstruct the original signal with perfect accuracy over the range 0–4000 hertz.

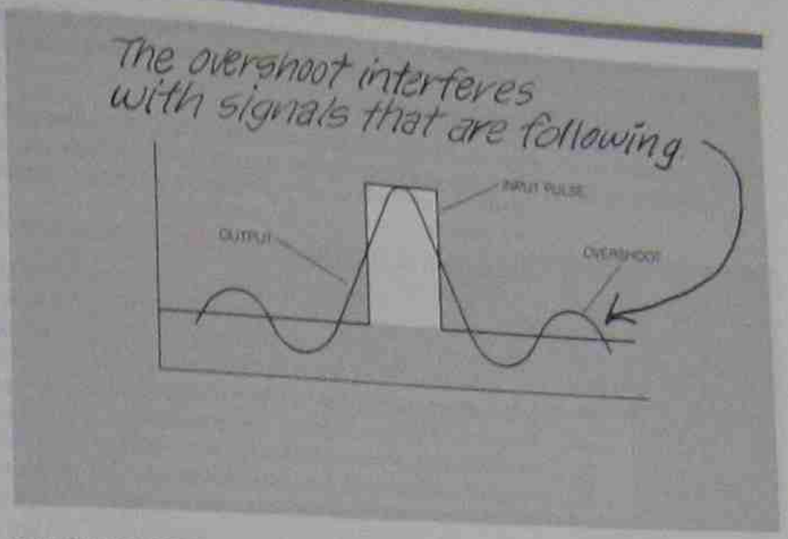


Figure 3-11. Pulse Response Through a Band-Limited Channel

BANDWIDTH REQUIREMENTS FOR VARIOUS SIGNALS

Analog Signals

As improved analog signal processing techniques continue to evolve, a given bandwidth can be used more efficiently.

Transmission rates for data communications seem to follow a corollary of Parkinson's Law; that is, data rates increase to fill the bandwidth available. A good example is the introduction of low-cost modems operating at 1200 and 2400 bps for personal computers, which are four and eight times faster than the rate of 300 bps that was used for a long time. Large scale integrated circuits made possible the remarkable increases in performance and decreases in price of these analog modems. In 1974, a 300 bps full-duplex modem cost $600 to $800. Now, a 1200 bps modem can be bought for under $100, whereas 2400 bps modems in the $175 to $250 range are readily available from many vendors.

The 300 bps full-duplex signals use two bands of frequencies, each occupying about 300 Hz. Thus, the total 600 Hz used out of the 3000 Hz available bandwidth is inefficient. The 1200 bps modems also are full-duplex and use most of the available bandwidth. Therefore, four times as much information can be sent in the same channel in a given time period. Through sophisticated signal processing techniques, modems can even carry up to 19,200 bps using an analog signal over a voice channel.

Digital Signals

Digital signals require a greater bandwidth than an equivalent analog signal. The advantages offered by transmitting signals in digital form more than offset this disadvantage.

Transmission of signals in binary form can require considerably more bandwidth than an equivalent analog signal. For example, the transmission of 24 analog voice channels requires about 95 kHz (24 x 4 kHz). Transmission of these same 24 voice channels in digital form using the standard T1 time division multiplex format requires about 776 kHz, or about 8 times as much bandwidth (776/96). However, the advantages gained by sending the signals as binary data more than offset the requirement for greater bandwidth. (For a discussion of the tradeoffs and advantages, refer

to another book in this series: J. L. Pike and G. E. Friend, *Understanding Telephone Electronics*, Texas Instruments Incorporated, © 1983.)

CARRIER SYSTEMS

In general, carrier systems are systems that provide a means to send signals from more than one source over a single physical channel. The bandwidth available to carry signals in a particular medium can be allocated in two ways: by frequency or by time intervals.

Analog Carrier Systems

In FDM, the total available channel bandwidth is divided into smaller bandwidth portions (sub-channels), each with its own signal source.

The frequency spectrum represented by the available bandwidth of a channel can be divided into smaller bandwidth portions, with each of several signal sources assigned to each portion. This is the principle of frequency division multiplexing (FDM). FDM is still used in some simple data communications systems, and is extensively used in the long-haul part of the public telephone network. One common example of a frequency division multiplexing system is a standard low-speed (300 bps) modem, which divides the spectrum available in a voice channel into two portions, one for transmit and one for receive (this is discussed more and illustrated in a later chapter).

The electronic systems that implement FDM are called analog carrier systems. The carrier in an analog carrier system is a signal generated by the system and the carrier is modulated by the signal containing the information to be transmitted. Table 3-2 shows the standard analog carrier systems in use in the public telephone network.

Table 3-2
Analog Carrier Systems

Multiplex Level	No. of Voice Circuits	Frequency Band, KHz
Voice Channel	1	0-4
Group	12	60-108
Supergroup	60	312-552
Mastergroup	600	564-3,084
Jumbogroup	3600	564-17,548

Digital Carrier Systems

TDM is used in digital carrier systems because many bits can be packed into the very narrow time blanks allowed to each of several signal sources.

The second method of dividing the capacity of a transmission channel among several separate signal sources is to allocate a very short period of time on the channel in a repeating pattern to each signal. This technique is called time division multiplexing (TDM). It is well-suited to binary signals consisting of pulses representing a one or a zero. These pulses can be made of very short duration and still convey the desired information; therefore, many of them can be squeezed into the time available on a digital carrier channel.

The original signal may be an analog wave which is converted to binary form for transmitting as in the case of speech signals in the telephone network, or the original signal may already be in binary form as in the case of a business machine. The electronic systems that perform this TDM process are called digital carrier systems. As with the analog carrier systems, there is a standard hierarchy of digital carrier systems in the public telephone network, as shown in Table 3-3.

Table 3-3
Digital Carrier Systems

Digital Signal No.	No. Voice Circuits	Bit Rate, MBPS
DS-1	24	1.544
DS-2	96	6.312
DS-3	672	44.736
DS-4	4032	274.176

WHAT HAVE WE LEARNED?

1. Information is a numerical quantity that measures the randomness of a system.
2. Symbols such as letters have an information content. Not every symbol is equally likely to occur.
3. The redundancy of a system measures how likely symbols are to be repeated.
4. Data communications systems use redundancy to detect and correct errors in transmission.
5. Signals can travel through guided and unguided transmission media.
6. Adding inductance to pairs of wires is called loading. Loading is used to reduce high-frequency attenuation over the wire pair.
7. Fiber optic transmission systems send data signals as light rays. These systems have much higher bandwidth and have immunity from external interference.
8. Transmission rates over channels are limited by the bandwidth of the channel, the signal-to-noise ratio, and the amount of intersymbol interference in the transmitted waveform.

Quiz for Chapter 3

1. The amount of uncertainty in a system of symbols is also called:
   a. Bandwidth.
   b. Loss.
   c. Entropy.
   d. Quantum.
2. Redundancy measures:
   a. Transmission rate of a system.
   b. How likely symbols are to be repeated.
   c. Time between failure.
   d. System cost.
3. An example of a bounded medium is:
   a. Coaxial cable.
   b. Waveguide.
   c. Fiber optic cable.
   d. All of the above.
4. Loading refers to the addition of:
   a. Resistors.
   b. Capacitors.
   c. Bullets.
   d. Inductance.
5. Coaxial cable has conductors with:
   a. The same diameter.
   b. A common axis.
   c. Equal resistance.
   d. None of the above.
6. Fiber optic cables operate at frequencies near:
   a. 20 MHz.
   b. 200 MHz.
   c. 2 GHz.
   d. 800 THz.
7. HF radio waves follow how many basic paths on leaving the transmitter?
   a. Two.
   b. Four.
   c. One.
   d. Many.
8. The area of coverage of a satellite radio beam is called its:
   a. Beamwidth.
   b. Circular polarization.
   c. Footprint.
   d. Identity.
9. Transmission of binary signals requires:
   a. Less bandwidth than analog.
   b. More bandwidth than analog.
   c. The same bandwidth as analog.
   d. A license from the FAA.
10. The standard first-level digital multiplex system in the U.S. operates at:
    a. 2.048 Mbps.
    b. 44.736 Mbps.
    c. 1.544 Mbps.
    d. 9600 bps.

Asynchronous Modems and Interfaces

ABOUT THIS CHAPTER

Having laid the groundwork by describing the form of the data (codes), the source of the data to be transmitted (terminals), and the transmission media; we now can discuss *how* the data are transmitted. This chapter explains how modems work in terms of frequency, bandwidth and modulation; discusses multiplexing; and describes the data terminal equipment/data communication equipment (DTE/DCE) interface.

WHY DATA CAN'T BE TRANSMITTED DIRECTLY

Digital information cannot be transmitted directly over the switched telephone network because the portion of the network beyond the central office cannot carry direct current.

Although it may not have been obvious at the time, the reasons that modems are necessary were discussed in Chapter 3. Remember that the switched telephone network utilizes a variety of transmission methods from wire pairs to microwave, but only in the local area (primarily within the area served by a Central Office) does an actual metallic (wire) path exist from one telephone to another. This means that we cannot transmit data in its original form as a series of pulses much farther than from our telephone to the Central Office (CO). Actually, if loading coils are used in the local loop, we can't even transmit it that far. As discussed in Chapter 3, the bandwidth (the passband) of the telephone channel beyond the CO after it has been filtered and amplified is approximately 300 to 3400 Hz as shown in *Figure 4-1*. Since direct current (zero Hz) is below 300 Hz, it is not within the passband; thus, data in its original pulsed dc or "baseband" form cannot be transmitted over this channel.

SOLVING THE PROBLEM WITH MODEMS

A transmitting modem changes the digital signals produced by computers to an analog signal with a frequency bandwidth that can be transmitted over the telephone network.

Because we are interested in transmitting data beyond the CO using the existing telephone network, we need to change the data pulses to another form that can be transmitted over the telephone channel. Since the telephone network is designed and optimized for transmission of analog signals in the voice band, why not make the data look like these analog signals for transmission? This is exactly the function of a modem. The ones and zeros of the data stream from the DTE are converted to tones (or analog waveforms resembling tones) having frequencies within the 300 to 3400 Hz range. Thus, the modem is nothing more than a rather complex interface device.

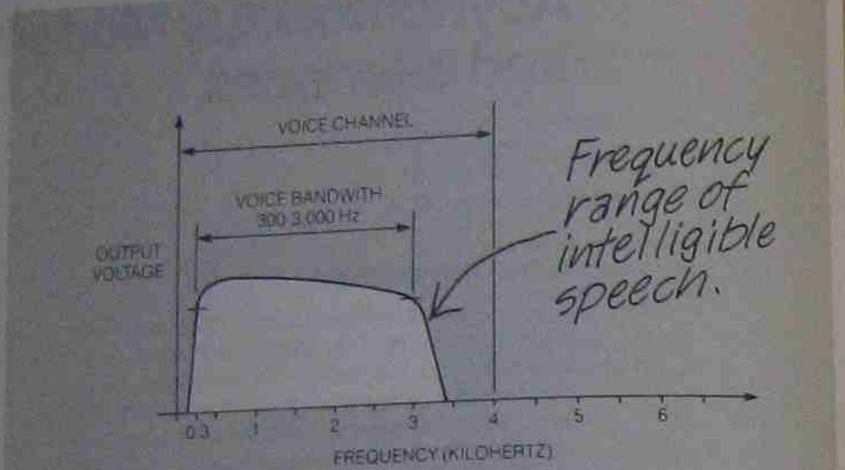


Figure 4-1. Telephone Channel Bandwidth
(Source: J.L. Fike and G.E. Friend, Understanding Telephone Electronics, © 1983 Texas Instruments Incorporated)

A receiving modem converts the analog signal back to its original digital form so that the receiving computer can use the data.

At the transmitting end, incoming pulses from the DTE are converted to tones and transmitted over the telephone channel; at the receiving end, the tones are converted back to pulses which are passed to the DTE. In other words, the transmitting modem modulates an analog signal, called a carrier, with the data (more about modulation in the next section) and uses the analog signal to carry data to the other end of the telephone circuit. The receiving modem demodulates the analog signal to recover the data in their original pulse form and passes it to the receiving DTE. In fact, the name "modem" is a contraction of the words modulator-demodulator. One important fact to remember is this: A modem does not operate on the content of the data; it merely changes the form for transmission.

Telephone Channel Restrictions on Modems

In addition to the bandwidth limits, another restriction of the telephone channel which affects modem design is inherent in any analog transmission facility; that is, the transmission is best at frequencies near the center of the passband and poorer for frequencies toward the upper and lower limits of the passband. High-speed modems use almost all of the voice band for one channel; therefore, most high-speed modems in North America utilize a carrier frequency of 1700 to 1800 Hz because these frequencies are very near the middle of the voice band. Low-speed modems, because of their narrower bandwidth requirements, can use more than one carrier frequency within the voice band and still operate in the "good" portion of the band.

Modems cannot use the frequencies in the transmission channels that are used for telephone network signaling.

Still another restriction of the telephone channel is that certain frequencies cannot be used. The telephone network uses the transmission channel for passing information and control signals between the switching offices. This process, called in-band or in-channel interoffice signaling, utilizes tones at frequencies within the voice band. A modem cannot use these same frequencies because the network might interpret them as control tones with disastrous results to the call.

Modem Interfaces

A modem has an interface to the telephone network and an interface to the DTE. The one to the telephone network is the simpler because that interface consists of only two wires called the "tip and ring." As long as the modem adheres to the voltage, current, power and frequency rules of the telephone company, the telephone channel is really just a pipe to move analog tones from one place to another.

A protocol establishes the procedures which are used by the transmitting and receiving equipment in establishing and maintaining communications. "Handshaking" assures the equipment is ready to go.

The interface wiring between the modem and the DTE is more complex and is governed by standards which will be discussed later in this chapter. This interface also requires that certain procedures (called a protocol) be observed in establishing communications between the two ends. First, the DTE and the modem at the transmitting end must establish communication with one another. The DTE indicates to the modem that it wishes to transmit, and this modem signals to the modem at the other end of the circuit to see if it is ready to receive. Since modems do not store data, the receiving modem must contact its DTE to see if it is ready to receive. (This communication between the equipment often is referred to as "handshaking.") After the transmitting modem knows that the receiving modem and DTE are listening on the line, it notifies the transmitting DTE which then begins passing data to the transmitting modem for modulation and transmission. On the receiving end, the receiving modem demodulates the incoming signal and passes the received data to the receiving DTE.

In half-duplex transmission (only one direction at a time), when the transmitter is finished and wants a reply from the other end, the channel must be "turned around." To do this, much of the handshaking must be done again to establish transmission in the opposite direction, and this turn-around handshaking must occur each time the direction is changed. In full-duplex transmission (both directions simultaneously), the transmission uses two different carrier frequencies, thus, the handshaking is necessary only for the initial set up.

A sine wave signal often is used as the carrier in an analog modulation process. A sine wave can be easily generated by electromechanical generators and electronic oscillators.

frequency, amplitude, and phase are the characteristics of a sine wave that can be varied to achieve modulation.

ANALOG MODULATION

Modulation is the process of using some medium as a carrier to carry information between two points. For example, we could send Morse code by turning a flashlight on and off to modulate the light beam. Reflecting on this process, we see that we are changing some property of the carrier to represent the data; in this instance it is the intensity of the beam. (This might be more easily understood if we imagine that we use a bright light for a mark and a dim light for a space. In this case, we are modulating the brightness or "amplitude" of the light beam to send data.)

A sine wave that might be used as a carrier in a modem is illustrated in *Figure 4-2*. Sine waves are not only mathematical functions; they also are fundamental phenomena of the physical universe. Sine waves are generated by devices such as electromechanical generators and electronic oscillators.

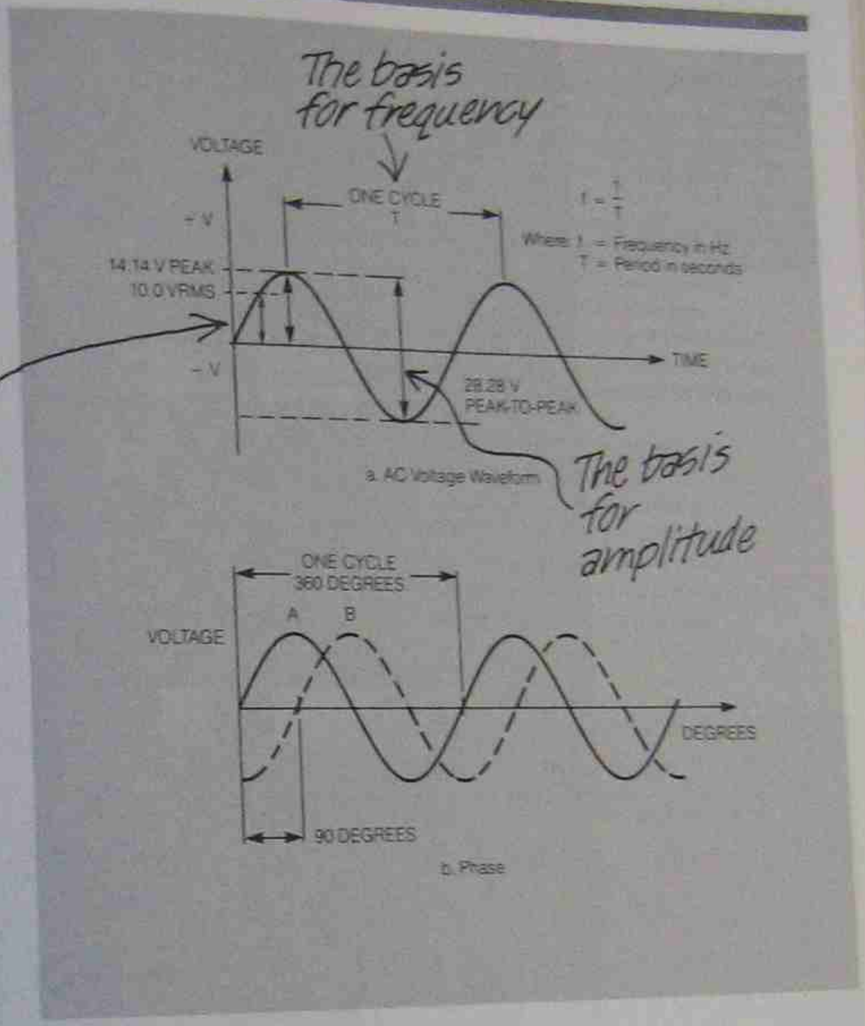
A man named Fourier discovered that any series of pulses, sounds, voltages, or similar waves may be broken down into a series of sine waves of varying frequencies and amplitudes. This means that human speech, in the air or on a telephone wire, can be analyzed in terms of the sine waves it contains. That is not to say that the human larnyx creates speech by combining sound waves at different frequencies; it simply means that Fourier analysis is a fundamental and useful tool for analyzing the characteristics of any transmission channel or device, whether it be a modem or the speakers of a high fidelity sound system.

As illustrated in *Figure 4-2a*, a sine wave can be defined by its frequency and amplitude. Frequency is measured in cycles per second with the unit of Hz. Amplitude can be measured in units of volts peak-to-peak, volts peak, and volts rms as indicated in the figure.

Another parameter of a sine wave is the phase of the wave, but it only has meaning in reference to another wave of the same frequency. Two different sine waves having the same frequency can be compared by the amount by which one leads or lags the other. Since one complete cycle of a sine wave occurs in 360 degrees, we can consider the difference between the two waves in terms of degrees as indicated in *Figure 4-2b*. The amplitude of the two waves so compared does not have to be the same, but the frequency has to be exactly the same.

Since the frequency, amplitude, and phase completely characterize a sine wave, these are the only parameters of the carrier sine wave that can be changed to modulate the carrier for transmitting information from one modem to another. All modems, therefore, utilize either amplitude, frequency or phase modulation, or some combination of these.

Figure 4-2.
Sine Wave Fundamentals
(Source: D.L. Cannon, Understanding Digital Troubleshooting, © 1983, Texas Instruments Incorporated)



Sine wave is defined by frequency and amplitude.

The modem converts the 1s and 0s of the data stream into an analog signal that is within the telephone voice band so the data can be transmitted over the telephone network.

Figure 4-3 illustrates amplitude, frequency, and phase modulation. Notice that the same sequence of 1s and 0s of the data stream affect the carrier in different ways. After modulation, the pulses are represented by an ac signal having frequencies within the voice band, thus, the information can be transmitted over telephone channels. Actually, simple amplitude modulation is not used for data communication because it is very susceptible to electrical noise interference which can cause errors in the received data. Low-speed modems use frequency modulation; higher speed modems use phase modulation; and the very highest speed modems for voice-band transmission use a combination of phase and amplitude modulation.

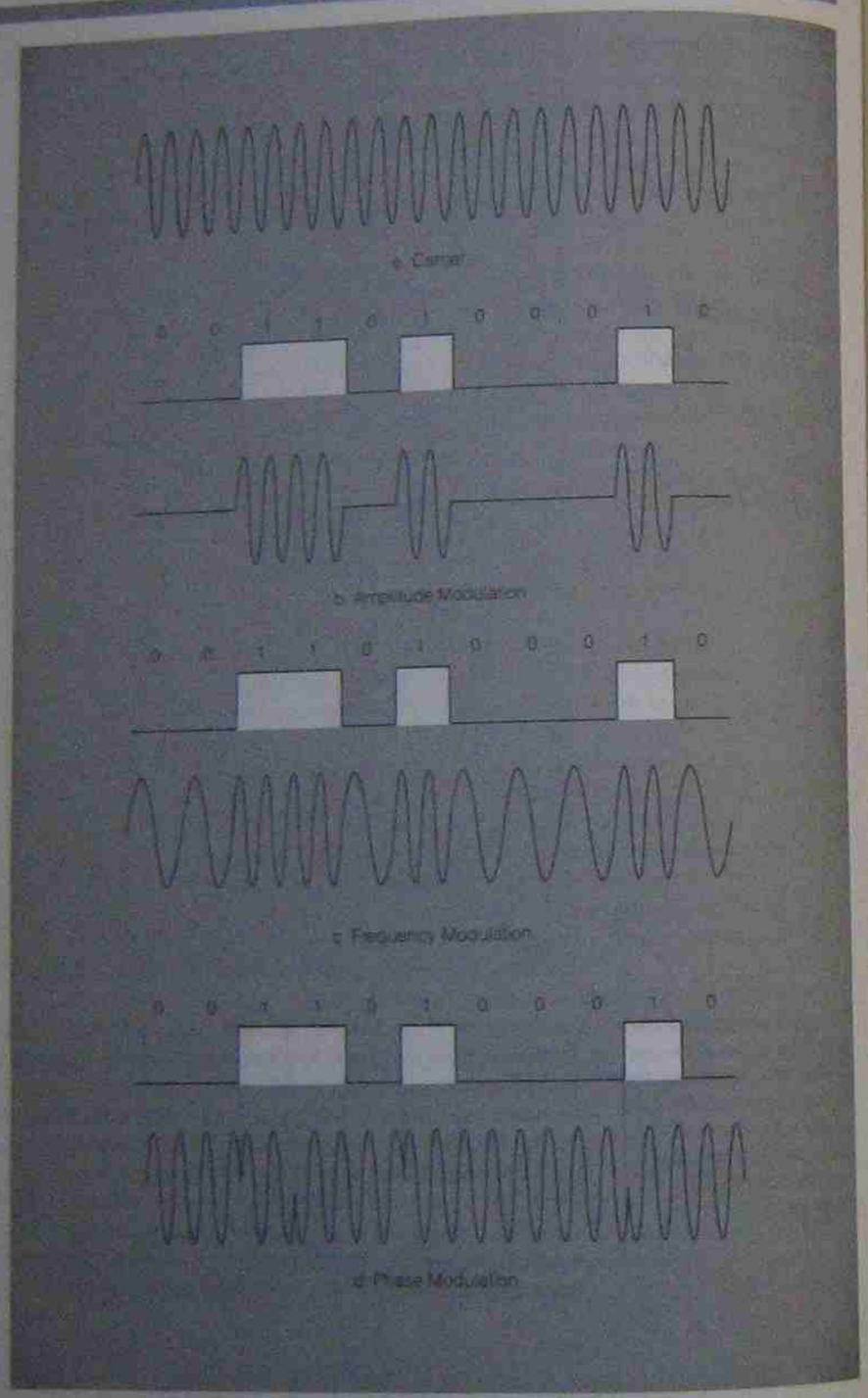


Figure 4-3. Types of Analog Modulation
(Source: D. Doll, Data Communications, Facilities, Networks and System Design, John Wiley & Sons, 1978. Copyright © 1978 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

LOW-SPEED MODEM OPERATION

With this background, we can describe how a standard 0-300 baud low-speed modem works. *Figure 4-4* shows the voice band divided into two sub-bands for transmission in both directions simultaneously. Recall that such two-way transmission is called full-duplex, meaning that the same bandwidth is available in both directions.

For some people, the process of transmitting data over one channel in two directions at the same time may be hard to visualize. One could think of it as being similar to two people talking on the same local call simultaneously. Both people can talk and hear at the same time, although they may have trouble understanding each other because the entire bandwidth of the channel is available to both parties in both directions. However, the modems completely separate the data into the two bands shown, one for each direction, so that each can understand the other.

The separation is accomplished when one modem is set to the "originate" mode, and the other to the "answer" mode, by a switch on each modem. These terms come from the use of low-speed modems in dial-up computer applications where a user calls a computer. The calling modem is usually in the originate mode, and the called modem in the answer mode (makes sense, doesn't it?). By doing things this way, the marking tone of the answer modem (the high-pitched "whistle" heard when it answers) also disables any echo suppressors and companders (which interfere with data transmission) that may be in the circuit.

By separating the total channel bandwidth into two bands, modems can transmit and receive data simultaneously over the same channel. The calling modem is set to the originate mode and the answering modem to the answer mode.

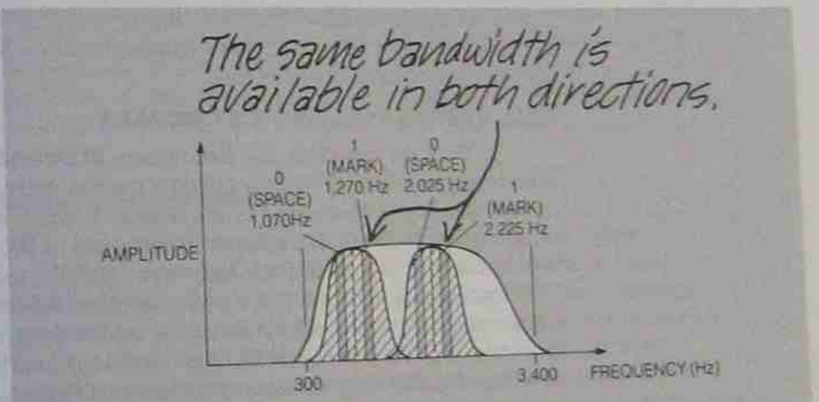


Figure 4-4. Frequencies Used for 300 Baud Full-Duplex Transmission
(Source: J.L. Fike and G.E. Friend, Understanding Telephone Electronics, © 1983, Texas Instruments Incorporated)

In *Figure 4-4*, notice that the originate modem transmits 0s (spaces) at 1070 Hz and 1s (marks) at 1270 Hz. The answer modem also transmits spaces and marks, but at 2025 and 2225 Hz, respectively. This type of modulation is called frequency-shift keying (FSK). (The term "keying" means turning a transmitter on and off.) The appropriate mark or space frequency is simply turned on and off by the transmitting modem as it wishes to send a one or zero. If you listen on a telephone line which is carrying low-speed FSK modem transmissions, you may hear a warbling sound as the frequencies shift back and forth.

The FSK modem just described is more commonly known as a Bell System 113 type device. Originally, the Bell System supplied all modems that could be attached to the switched telephone network. By the time other vendors were permitted to sell modems for use on the switched network, the Bell System's monopoly had created a large base of telephone company devices, which encouraged vendors to manufacture modems compatible with Bell System products.

FSK is a straightforward and economical modulation method that works well over telephone channels, so the logical question is: "Why not use it for higher speed modems?" The answer is "between the lines" (no pun intended) of *Figure 4-4*. Higher data rates require more bandwidth; thus, if we wanted to transmit at higher speeds, then the mark and space frequencies for each sub-band would have to be farther apart, and the two sub-bands would have to be separated further to provide enough bandwidth. Soon we would be outside of the telephone channel; in other words, we would require more bandwidth than is available. Therefore, higher speed modems use some form of phase modulation because phase modulation requires the least bandwidth of the three analog modulation methods.

BELL SYSTEM 212A AND V.22 MODEMS

Until the mid-1980s, the Bell System 212A modem was the most popular upgrade for Bell System 113 type modem users. The 212A modem can be considered two modems in one because it employs FSK modulation using the Bell System 113 frequency assignments at 300 bps. At its higher 2000 bps rate, it uses dibit phase shift keyed (DPSK) modulation. Under DPSK modulation, two bits at a time are encoded into one phase shift or signal change. Thus, the signal rate of the modem (commonly known as its baud rate) is 600 or one-half of its 1200 bps data rate. The phase shift encoding of a 212A type modem is illustrated in *Table 4-1*.

Dibit	Phase Shift (Degrees)
00	90
01	0
10	180
11	270

When a 212A modem operates at 1200 bps, it can transmit data either asynchronously or synchronously. One key advantage to using 212A type modems is that when connected to a computer they can receive transmission from terminal devices operating at either 300 or 1200 bps.

The V.22 standard is similar to the 212A at 1200 bps. However, at the 212A's lower data rate, the two modems are incompatible. The V.22 modem's second data rate is 600 bps compared with the 212A's lower data rate of 300 bps. A second major difference between the lower speed of the two modems is in the method of modulation used. The 212A uses FSK modulation at 300 bps, whereas the V.22 uses 2-phase PSK at 600 bps. Fortunately, the V.22 modem is used primarily in Europe, whereas the 212A type modem is used primarily in North America.

V.22 bis

The CCITT V.22 bis (bis meaning second in Latin) has received widespread acceptance throughout the United States and Europe during the last two years and represents one of the few modems that can be used for worldwide communications. The V.22 bis recommendation governs modems designed for asynchronous data transmission at 2400 bps over the switched telephone network, with a V.22 fallback method of operation. When the V.22 bis modem operates in its fallback mode, incompatibility problems can arise because of the different methods used to manufacture this modem.

In the United States, most if not all V.22 bis modems follow the Bell System 212A specifications for fallback operations—DPSK at 1200 bps and FSK at 300 bps. In Europe, most V.22 bis modems follow the V.22 specifications for fallback operations, which includes 2-phase PSK at 600 bps. Thus, at their lowest fallback rate, V.22 bis modems manufactured for use in North America are incompatible with those manufactured for use in Europe.

NONSTANDARD MODEMS

The requirements of personal computer users for higher transmission rates for file transfer and interactive full screen display operations resulted in several vendors designing proprietary operating modems to achieve data rates that would have been beyond belief several years ago. To achieve data rates up to 19.2 Kbps over the switched telephone network, some modems incorporate data compression and decompression algorithms, whereas other modems attempt to transmit data on up to 512 distinct frequencies over the communications link bandwidth.

Modems incorporating compression really operate at variable data rates because their throughput depends on the susceptibility of the data to the compression algorithms in the modem. As an example of how a modem can compress data, consider the sequence ACCOUNT ␣␣␣␣␣␣␣ AMOUNT where ␣ indicates a blank or space character. This sequence could represent the column headings of a report and is similar to most accounting reports that contain columns of data with each column separated by spaces from the next column. If the modem uses a special character denoted as S_c to show space compression, the sequence is transmitted as ACCOUNT S_c7

AMOUNT between modems. At the receiving modem, the character S_c indicates the occurrence of space compression, and the numeric 7 indicates the number of space characters compressed. This information provides the receiving modem with the ability to decompress the data into its original form.

To prevent the natural occurrence in the data stream of a character indicating space compression and thus falsely affecting the modem at the opposite end of the data link, the originating modem is programmed to stuff an extra space compression indicating character into the transmission sequence. Thus, if the data stream into the modem is XYZS$_c$ABC, the modem transmits the sequence XYZS$_c$S$_c$ABC. At the receiving modem, the occurrence of the first S_c character causes the modem to examine the next character. If the receiving modem finds a second S_c character following the first, it will disregard the second S_c character, restoring the data stream to its original form while preventing a false decompression of the data.

Most compression modems use a mixture of compression algorithms. Although the actual throughput is variable, the net effect of compression is roughly double the data transfer rate of the device. For further information covering data compression techniques and applications, the reader is referred to the book *Data Compression 2nd Edition* by Gilbert Held. Published by John Wiley & Sons in 1987, this book explains eleven compression algorithms and contains BASIC language programs that can be used to compress and decompress data.

PACKETIZED ENSEMBLE PROTOCOL MODEM

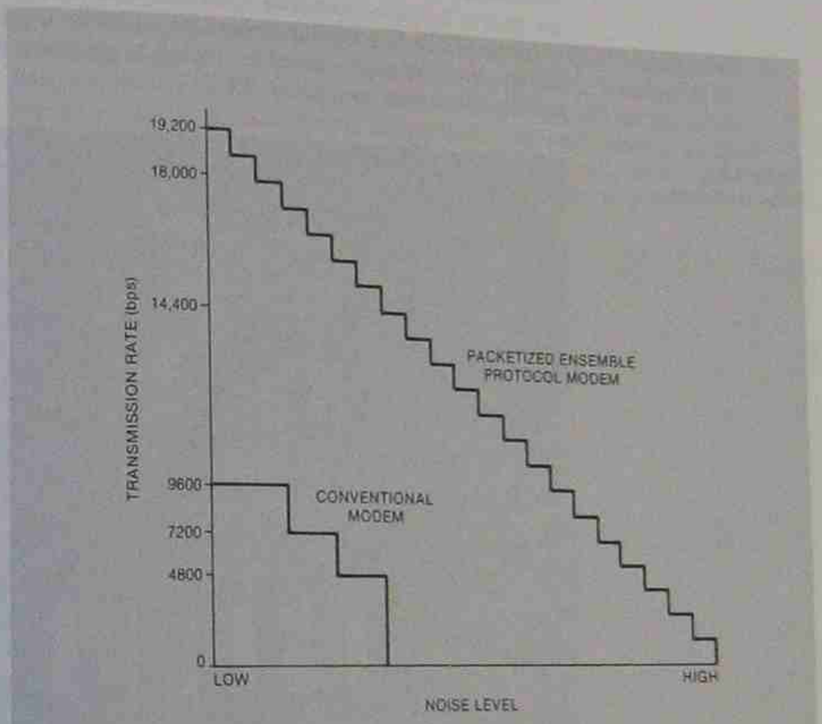
A second type of nonstandard modem gaining widespread acceptance is known formally as a Packetized Ensemble Protocol Modem. This modem incorporates a high-speed microprocessor and roughly 70,000 lines of instructions built into read-only memory (ROM) chips to obtain a revolutionary advance in modem technology.

Under the Packetized Ensemble Protocol mode of operation, the originating modem simultaneously transmits 512 tones onto the line. The receiving modem determines which tones are usable and reports back to the originating device the usable frequencies over which data can be transmitted. The originating modem then selects a transmission format most suitable to each tone, forms a group of bits into a packet for transmission on each tone and adds an error checking group of bits that covers the data to be transmitted before actually transmitting the data over the full bandwidth. The receiving modem performs a check based on the same algorithm used by the first modem to develop the error checking group of bits. If the transmitted and locally generated check group of bits do not match, the receiving modem requests the transmitting modem to retransmit the data, which results in error corrections by retransmission that is transparent to the modem user.

One of the prime advantages of a Packetized Ensemble Protocol Modem is its capability of adjusting automatically to the usable frequencies on a line. This capability enables the modem to lower its fallback data rate in small increments to correspond to the loss of a few tones at one time.

Figure 4-5 illustrates the transmission rate of a Packetized Ensemble Protocol Modem compared with a conventional modem as the noise level on a circuit increases. The Packetized Ensemble Protocol Modem loses its capability of transmitting on one or a few tones as the noise level increases, resulting in a gradual decrease in the data rate of that modem. In comparison, most conventional modems, such as a 9600 bps device, are designed to fall back to a predefined fraction of their main data rate, such as 7200 or 4800 bps.

Figure 4-5. Transmission Rate Versus Noise Level



The original Packetized Ensemble Protocol Modem was developed by Telebit Corporation and is marketed as the Trailblazer. Other vendors, including GTE, Digital Communications Corporation, and Racal-Vadic, have entered into various marketing agreements with Telebit that enable them to market this modem.

DIGITAL MODULATION

Sine waves are all we have to work with in transmitting over the analog telephone channel because it doesn't transmit pulses. Digital transmission systems *will* transmit pulses, and with them we can encode either analog or digital information by modulating pulses. A modem as described above is not required in these systems, but other conversion

equipment is. However, this process is beyond the scope of this book so we will only briefly describe the types of digital modulation here.

Digital modulation methods include: PAM, PWM, PPM, and PCM.

Figure 4-6 shows three ways to modulate a series of pulses to carry data. When the amplitude of the pulses is varied to represent analog information, the method is called pulse amplitude modulation (PAM). This method is very susceptible to electrical noise interference. In the second method, called pulse width modulation (PWM), the information is represented by varying the width of the pulses. Both of these techniques are used in telephone switching equipment, such as a private branch exchange (PBX). The third technique, which varies the position of pulses within a group of pulses (called the frame) to represent information, is called pulse position modulation (PPM).

Figure 4-6.
Pulse Modulation



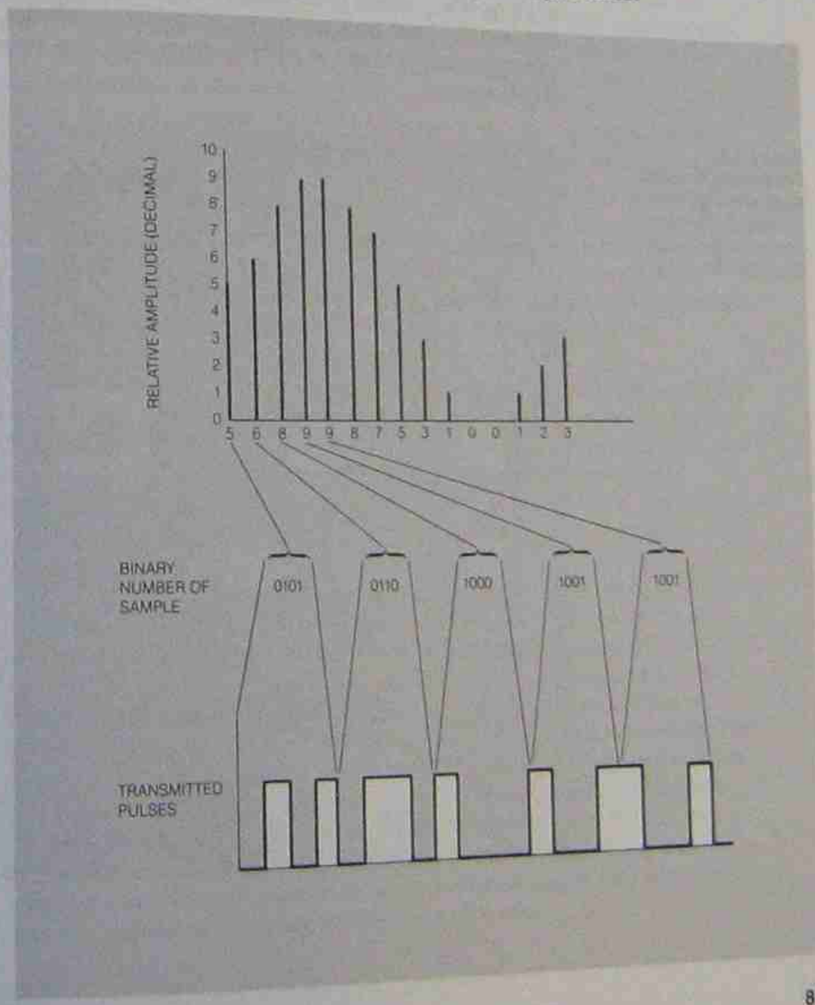
Figure 4-7 shows the process of sampling an analog signal as in PAM, but the amplitudes of the samples are encoded into binary numbers represented by constant amplitude pulses that are transmitted. This process, called pulse code modulation (PCM) overcomes the noise interference problem of PAM. Most long-distance telephone calls are transmitted by PCM.

MULTIPLEXING

Sharing a Channel

Just as modulation may be understood using the idea of sending code with a flashlight, so multiplexing has a simple example. Let's imagine that we have several letters to take to the post office. We could get into the car and take one letter, mail it, come back home and get the second letter, take it to the post office and mail it, come back to get the third . . . silly, right? Why not take all of the letters in the same trip, since they're all going to the post office?

Figure 4-7.
Pulse Code Modulation



This is what multiplexing amounts to: using a resource (in this case, the car going to the post office) to carry more than one message at a time. In the systems we are concerned with, a microwave system is a transmission facility (like the car), and a telephone call is the message (like the letter). The transmission facility is divided in either of two ways in order to share it; by frequency or by time.

Frequency Division Multiplexing

FDM translates the frequencies within each of several voice-band channels to different frequency channels that are combined and transmitted together.

Figure 4-8 shows the use of frequency division multiplexing (FDM) to carry more than one telephone conversation over a transmission channel. In effect, the frequencies in each call are changed so that they can be placed side-by-side in a wideband channel and be transmitted as a group. At the other end, the frequencies in each call are changed back to the original frequencies. FDM has been the mainstay of telephone transmission for many years; it is more efficient in terms of bandwidth than digital systems. The problem is that noise is amplified along with the voice. This fact, and the great decrease in cost of digital electronics, has led to the widespread replacement of FDM systems with time division multiplexing (TDM) systems.

Individual calls are changed in frequency and transmitted as a group.

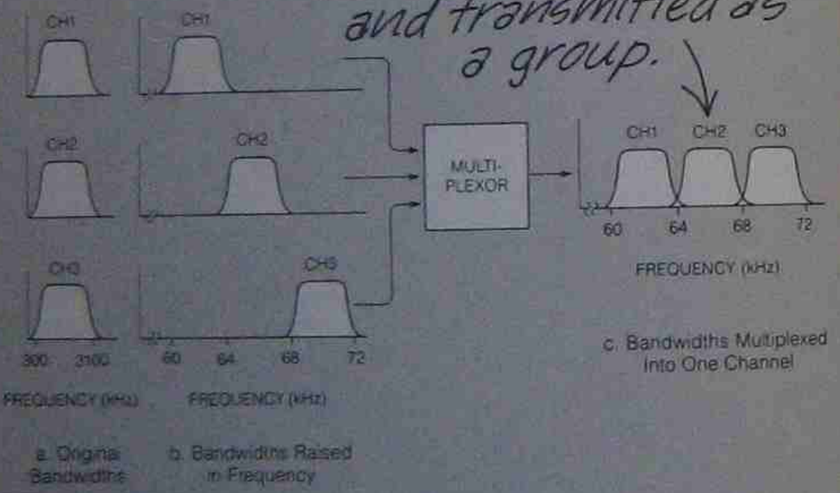


Figure 4-8. Frequency Division Multiplexing *(Source: Andrew S. Tannebaum, Computer Networks, Prentice-Hall, Inc.)*

In time division multiplexing, at transmission, the digital modulation techniques of PAM and PCM are used to convert and encode samples of an analog signal into pulses. The pulses are decoded and converted upon receiving.

Time Division Multiplexing

Figure 4-9 illustrates a TDM system carrying three telephone conversations simultaneously. The analog speech signals are sampled and the samples are transmitted in series over the same channel, one at a time. At the destination, the demodulation process is synchronized so that each sample of each channel is routed to the proper channel. Thus, this is multiplexing because the same transmission facility is used for more than one call at a time, and it is TDM because the available time is divided among three calls.

Most communications networks include several TDM systems that are used primarily because of economics. When one circuit and a pair of multiplexers replace two or more data circuits, the savings that accrue from the elimination of data lines usually can pay for the required equipment in less than a year.

To understand the rationale for using multiplexers, consider an example where eight terminals located in San Francisco must communicate with a main frame computer located in Chicago. If each terminal requires four hours of computer access per day and the cost of one hour of communications on the switched telephone network is $30, conventional dialing by each terminal operator to Chicago costs the organization $960 per day. If a month has 22 business days, the monthly cost for accessing the Chicago computer would slightly exceed $21,000, exclusive of the cost of modems.

When a number of terminal devices at one location require access to a computer at a second location, multiplexing should be considered. *Figure 4-10* illustrates a network configuration that could be used to satisfy the San Francisco to Chicago communications requirements of the organization.

Assuming the cost of each multiplexer is $3,000 and the leased line costs $1,000 monthly, during the first year, the multiplexing method would cost $18,000, exclusive of modems. In the second year of operation, the cost of multiplexing would be reduced to $12,000 because we assume the equipment was purchased the first year. Thus, in this example, the two-year cost of multiplexing is $30,000, resulting in a savings of roughly $10,000.

A second reason for the use of multiplexers is the relative fixed cost associated with their use. Multiplexer systems are connected to one another via leased lines that have rates which may increase slightly from year to year because of tariff changes. The cost of the multiplexer system is rather stable and predictable because the cost of the line is billed on a monthly basis based on the distance between locations and is not dependent on usage. In comparison, the cost of using the switched telephone network depends on many factors, including the duration of the call, the distance between calling and called parties, the day and time of day the call was originated and whether operator assistance was required. If only a few terminal users increase their connect time 15 minutes per day, the monthly cost of communications could substantially increase.

Figure 4-9. Simplified Diagram of TDM
(Courtesy of GTE Communication System Corporation)

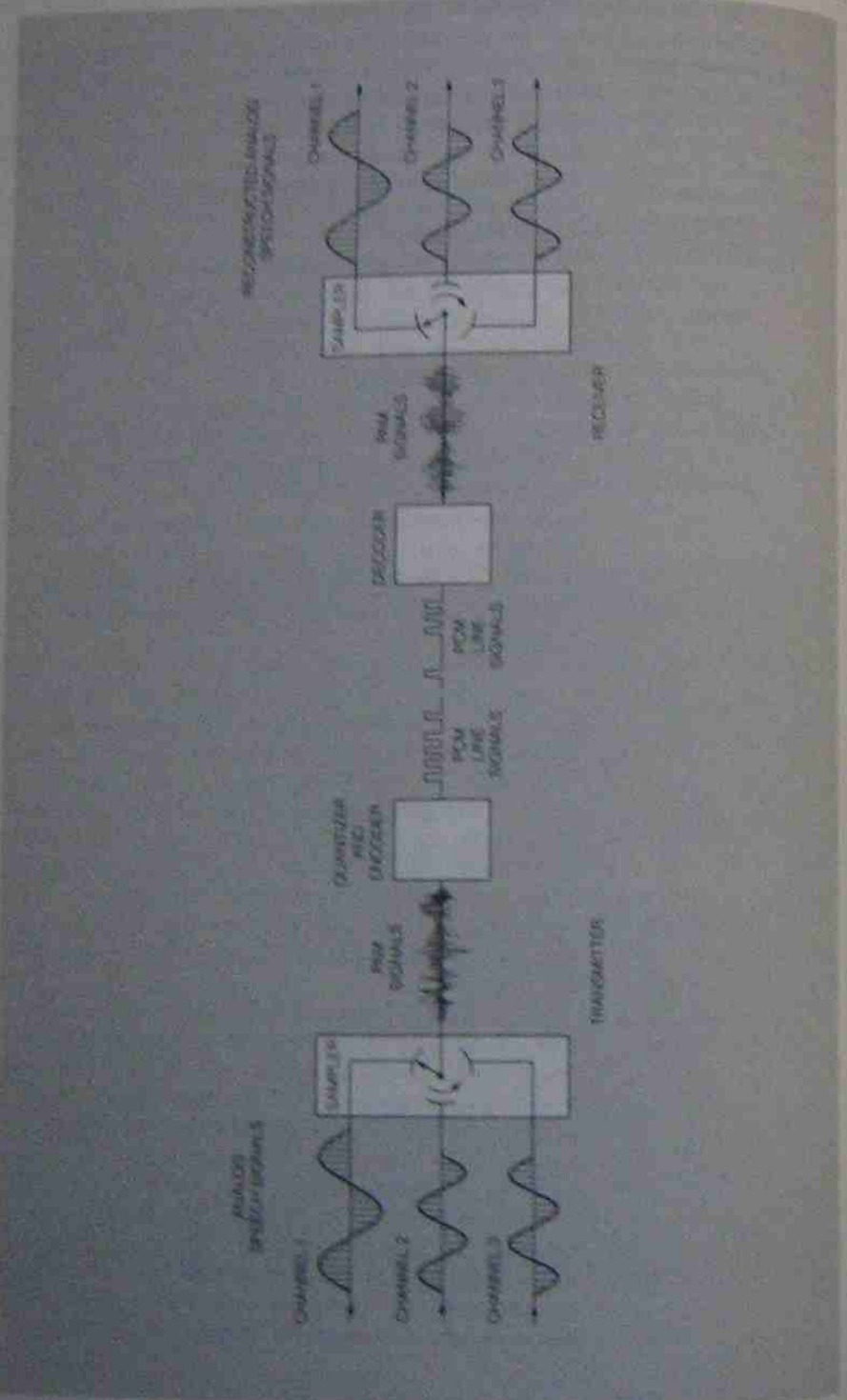
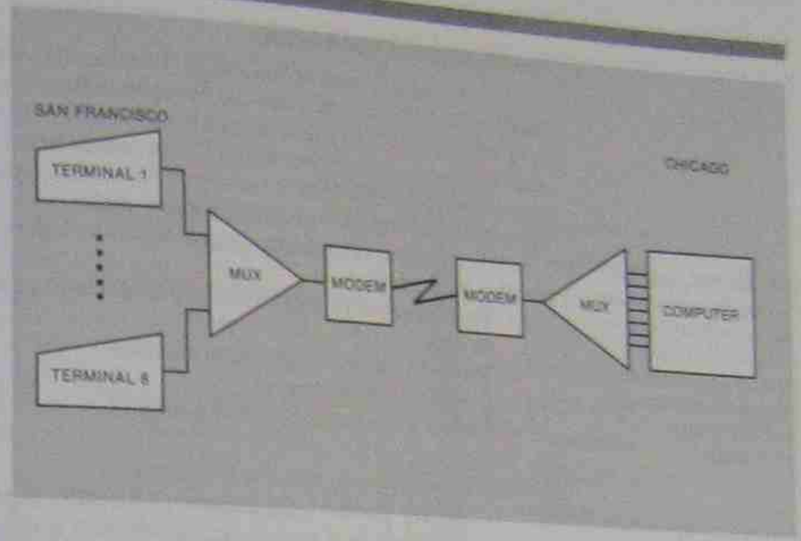


Figure 4-10. Using Multiplexer to Reduce Communications Costs



Correct interfacing between equipment is of utmost importance to achieve error-free data communications. The most common interface standard is RS-232C.

INTERFACE AND SIGNALING STANDARDS

There are many "standards" and "recommended practices" used to define data communications interfaces and signaling; in fact, the entire problem of data communications may be looked at as the task of passing information through a series of interfaces and transmission channels without loss of meaning.

The interface that is of the greatest interest is the one between the equipment which originates and/or receives the data, the Data Terminal Equipment (DTE), and the equipment which handles the problem of transmitting them from place to place, the Data Communications Equipment (DCE). Computers are called Data Terminal Equipment (whether or not they look or act like a CRT or other terminal). Modems are called Data Communications Equipment (also called data circuit-terminating equipment or data sets), because their function is to communicate the data rather than consider, compute, or change them in any way.

Terms like RS-232C, V.24, RS-422, RS-423, RS-449, X.21, X.25, current loop, and several others are the designations of the various standards and recommendations designed to make the task of connecting computers, terminals, modems and networks easier. As we'll see, manufacturers have in many cases used the standardized interfaces for functions that were never intended; this helps by not increasing the number of interfaces, but adds to the confusion of the user.

In terms of the present computer and communications world, the Electronic Industry Association (EIA) RS-232C is *the* standard; therefore, we'll describe it and use it as a way to explain interfaces in general. We'll consider some of the limitations of RS-232C, and we'll discuss some of the newer developments which eventually may replace RS-232C.

The RS-232C and V.24 Interface

The proper name of RS-232C is "Interface Between Data Terminal Equipment and Data Communication Equipment Employing Serial Binary Data Interchange." The latest revision is C, thus, we refer to it as RS-232C. What we say here also applies to CCITT Recommendation V.24, which is almost identical with RS-232C; however, the electrical signal characteristics are specified separately in CCITT Recommendation V.28.

The RS-232C standard covers the mechanical and signal interface between data terminal equipment and data communication equipment employing serial binary data interchange.

In addition to explanatory notes and a short glossary, the RS-232C standard covers four areas:

1. The mechanical characteristics of the interface (which has some surprises).
2. The electrical signals across the interface.
3. The function of each signal.
4. Subsets of signals for certain applications.

Mechanical Interface

Like any standard, the purpose of RS-232C is primarily as a reference for designers of equipment; therefore, it is not tutorial, and not particularly easy reading. (Perhaps that explains why it seems that many people who write about RS-232C have never read it.) Almost anyone in the computer or communications field, including authors, will tell you that the DB-25 connector is defined in the standard. That is simply not true. All that the mechanical section covers is the assignment of signals to the connector pins (discussed in the next section), which piece of equipment has the female connector (the DCE), the recommended maximum cable length (50 feet), and the maximum cable capacitance (2500 picofarads). Because the DB-25 connector *has become* almost universally associated with RS-232C, its pin arrangement is shown in *Figure 4-11*. However, it is not defined in the standard, and some manufacturers (notably IBM) use a different connector on much of their equipment.

Connector shape and pin arrangement are not covered in the standard, but the assignment of signals to connector pins, the connector gender, and cable length and capacity are.

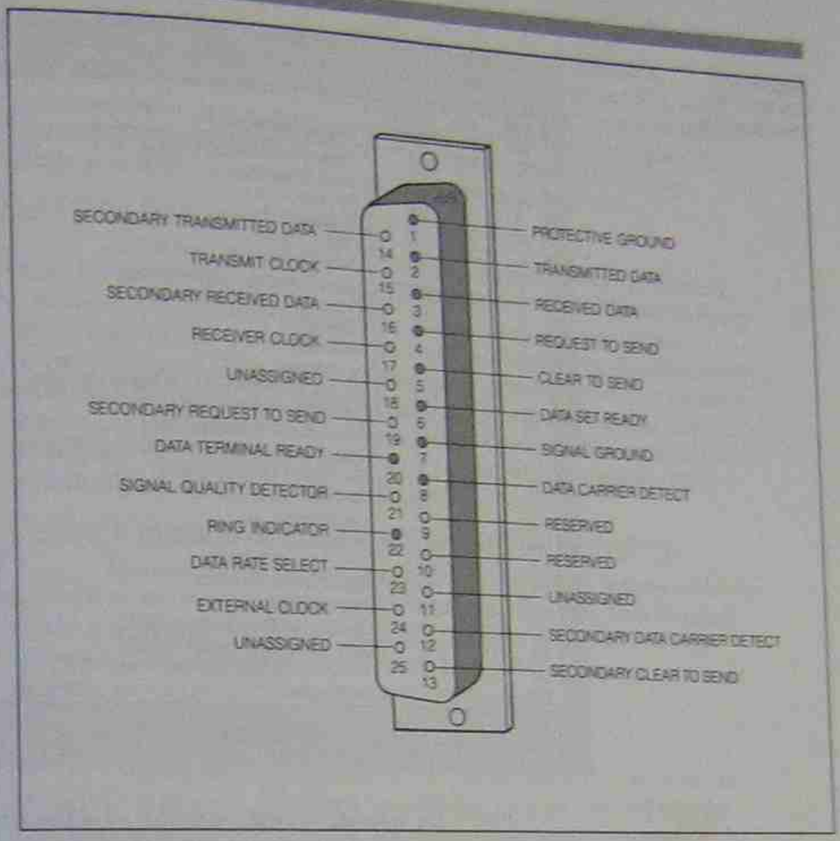
Electrical and Functional Interfaces

Although it doesn't make much sense to say that one part of a standard is more important than another, the heart of RS-232C is certainly the electrical and functional sections. The electrical portion covers the all important voltage and current specifications for each pin, along with such features as the requirement that the equipment not be damaged if *any* two pins are shorted together. (We didn't say the equipment would still work, but it's not supposed to burn up.) The functional portion, perhaps the most important of all, defines the sequencing of the signals and the action taken by the DTE and DCE in response.

The standard covers the specifications for current and voltage, special electrical considerations, and the sequencing of signals across the interface.

Although the RS-232C standard defines the procedure for automatic answering by the modem and for reversing the transmission direction in half-duplex communications, it does not define automatic dialing. This is covered in EIA Standard RS-366. As we will see later, many newer modems incorporate automatic dialing in ways that the authors of RS-232C and RS-366 did not imagine.

Figure 4-11.
Typical RS-232C Female Connector



A Pin by Pin Tour of RS-232C

In *Figure 4-12*, the signals are numbered and named according to three standard systems, plus a fourth that isn't a standard but which many people use. The first is by pin number, which is the way that people who deal regularly with the interface think of the signals: pin 2, pin 3, pin 5, etc. The second is the EIA designation: BA, BB, CB, etc. The third is the CCITT designation: 103, 104, 106, etc. The fourth is an abbreviation of the signal description: TD for Transmitted Data, RD for Received Data, CTS for Clear To Send, etc. A given signal often has more than one of these abbreviations because they aren't standard. Let's now take a pin-by-pin tour and look at the signals by category and function.

Pins 1 and 7, Protective Ground (Gnd) and Signal Ground (SG)

Two grounds are provided. Chassis (frame) ground (Pin 1) for the system, and a reference signal ground (Pin 7) for the signals on all the other pins.

If provided, Pin 1 is connected to the chassis of the equipment and is intended to connect one end of the shield if shielded cable is used. Shielded cable can be used to minimize interference in high-noise environments. Never connect the shield at both ends of the cable. Pin 7 is the common reference for all signals, including data, timing, and control signals. Pin 7 must be connected at both ends in order for the DTE and DCE to work properly across the serial interface.

Figure 4-12. RS-232C Pin Designations

PIN NO.	EIA CKT.	CCITT CKT.	Signal Description	Common Abbrev.	From DCE	To DCE
1	AA	101	Protective (Chassis) Ground	GND		
2	BA	103	Transmitted Data	TD		X
3	BB	104	Received Data	RD	X	
4	CA	105	Request to Send	RTS		X
5	CB	106	Clear to Send	CTS	X	
6	CC	107	Data Set Ready	DSR	X	
7	AB	102	Signal Ground/Common Return	SG	X	X
8	CF	109	Received Line Signal Detector	DCD	X	
9			Reserved			
10			Reserved			
11			Unassigned			
12	SCF	122	Secondary Received Line Signal Detector		X	
13	SCB	121	Secondary Clear to Send		X	
14	SBA	118	Secondary Transmitted Data			X
15	DB	114	Transmitter Signal Element Timing (DCE)		X	
16	SBB	119	Secondary Received Data		X	
17	DD	115	Receiver Signal Element Timing		X	
18			Unassigned			
19	SCA	120	Secondary Request to Send			X
20	CD	108/2	Data Terminal Ready	DTR		X
21	CG	110	Signal Quality Detector	SQ	X	
22	CE	125	Ring Indicator	RI	X	
23	CH	111	Data Signal Rate Selector (DTE)			X
23	CI	112	Data Signal Rate Selector (DCE)		X	
24	DA	113	Transmitter Signal Element Timing (DTE)			X
25			Unassigned			

Pin 7 is the reference signal ground for all of the other pins, thus, it is very important. The interface will not work without it because none of the signal circuits would be completed. One difficulty of RS-232C is the use of two separate grounding wires; grounding of distributed analog systems is often difficult, and having two ground paths doesn't help any. For example, suppose there is an RS-232C cable between two pieces of equipment where Protective Ground is open (or not connected), but where Signal Ground is connected to the power outlet ground at both ends. However, due to problems in the building ground circuit, the two ends are at different ground potentials. This causes a current to flow through the Signal Ground wire, and the resistance in the wire causes a difference in potential to develop between pin 7 at one end and pin 7 at the other end. If this potential is large enough, it could cause the data to be received incorrectly.

Pins 2 and 3, Transmitted Data (TD) and Received Data (RD)

The voltage polarities transmitted and received on the data lines (Pin 2 and 3) range from 5 to 25 volts for logic 0 and 1, and are reversed on the control lines.

At last we're getting down to the nitty-gritty. These are the pins that count; if it weren't for the data that passes through them, all the rest would be unnecessary. One important point to remember is that all signal names in the RS-232C standard are as viewed from the DTE. Thus, the DTE transmits on pin 2 and receives on pin 3, but the DCE transmits on pin 3 and receives on pin 2. If you imagine yourself as a computer (easier than imagining yourself as a modem), then the names are easier to understand.

For people accustomed to more modern electronic interfaces, RS-232C signal levels may be a bit surprising because they are not TTL levels. In RS-232C, a positive voltage between 5 and 15 volts on pin 2 or 3 with respect to pin 7 represents a logic 0 level, and a negative voltage between -5 and -15 volts on either pin represents a logic 1. These are the levels for data; the voltage polarities are reversed for logic 0 and 1 on the control lines.

The signals on pins 4, 5, 6, and 20 are the handshaking signals which establish the communications link.

Pins 4 and 5, Request to Send (RTS) and Clear to Send (CTS)

Terminals may not transmit until Clear to Send is received from the DCE. For private-line transmission, CTS is usually linked to Request to Send; in such cases, the DTE may use RTS to turn on the modem transmission carrier if the modem is so optioned. In this type of application, the relationship between RTS and CTS is usually a simple count-down timer which is optioned when the DCE is installed. The time delay will be set to such a value as to allow time for the carrier to turn on and become stable before CTS is returned to the DTE. If the DCE is optioned for constant carrier (without regard to RTS), then the RTS/CTS delay is usually set to zero.

Pin 6 and 20, Data Set Ready (DSR) and Data Terminal Ready (DTR)

Data Set Ready (Modem Ready) is used to indicate that the modem is powered on and is not in test mode. In dial-data applications, Data Terminal Ready is used to create the equivalent of an off-hook condition. When the modem is in auto-answer mode, DTR may be asserted in response to the ring indicator to tell the modem to answer the incoming call.

The Received Line Signal Detector (pin 8) is asserted when the modem receives a carrier of sufficient strength for reliable communications.

Pin 8, Received Line Signal Detector (DCD or CD)

This signal is usually called Data Carrier Detect (DCD) or Carrier Detect (CD); however, CD should be avoided as it can be confused with the EIA designation of CD for pin 20. The modem asserts DCD whenever it receives a signal on the telephone line that meets its internal criteria for amount of energy at the carrier frequency. Many DTEs require this signal before they will transmit or accept data; for this reason, in applications where no modem is present, pin 8 usually is tied to pin 20, which in most cases is asserted whenever the DTE is turned on.

When the DCE detects a ringing signal on pin 22, it tells the DTE that the phone is ringing. The DTE tells the modem to answer the phone by asserting pin 20.

Pin 22, Ring Indicator (RI)

The RI signal is the means by which the DCE tells the DTE that the phone is ringing. Virtually all modems that are designed to be directly connected (via an FCC-approved modular plug) to the telephone network are equipped for auto answer. This means that the modem is able to recognize standard ringing voltage, indicate the ringing to the DTE, and answer (take the line off-hook) when told to do so by the DTE. Pin 22 is asserted by the DCE in time with the cadence of the ringing signal on the line; that is, when the ringing voltage is present, RI is true (on); between rings, RI is not true (off). The DTE tells the modem to answer the phone by asserting pin 20, DTR.

The ten pins and signals described above are by far the most often used of those defined in RS-232C and V.24. It is very unlikely that equipment or cables for the small computer market will be equipped with any other proper RS-232C signals, although sometimes (unfortunately) other pins are used in non-standard ways which can cause problems when equipment is interconnected. We will now quickly cover the functions of the remaining pins in the specification.

Pins 15, 17, and 24 are used by synchronous modems to control bit timing. Pin 21 indicates that the quality of the received carrier is satisfactory.

Pins 15, 17, 21, and 24

Synchronous modems use the signals on these pins. Because the transmitting modem must send something (a 1 or 0) at each bit time, the modem controls the timing of the bits from the DTE. Similarly, the receiving modem must output a bit and associated timing whenever received. Pin 15 (Transmitter Signal Element Timing—DCE source), and pin 17 (Receiver Signal Element Timing—DCE source) are used for these purposes. In instances where the transmitter timing comes from a source other than the transmitting modem (such as another modem in a multiplexing situation), pin 24 (Transmitter Signal Element Timing—DTE source) is used. Pin 21 (Signal Quality Detector) indicates that the received carrier meets some predetermined criterion for quality.

Pin 23, Data Signal Rate Selector

This looks like two pins in the chart, but actually it is either Data Signaling Rate Selector (DTE source) *or* Data Signaling Rate Selector (DCE source). Some modems, called dual-rate or "gearshift" modems, allow switching between two transmission speeds. Sometimes the speed is selected automatically by the modem during the training (initializing) sequence, or it may be selected by the transmitting DTE. The signal on pin 23 controls whether the modem uses the high or low speed. Usually the modem at the calling end sets the speed for the connection and informs its DTE. The calling modem signals the speed to the answering modem, which informs the called DTE by asserting Data Signaling Rate Selector (DCE source).

In modems equipped with primary and secondary channels, the secondary channels carry signals in the reverse direction and at a much lower data rate.

Secondary Channels: Pins 12, 13, 14, 16, and 19

Some modems are equipped with both primary and secondary channels. The five secondary signals—Secondary Transmitted Data, Secondary Received Data, Secondary Request To Send, Secondary Clear To Send, and Secondary Received Line Signal Detector—allow control of the secondary channel in the same way as described for the primary channel. In these modems, the primary transmission channel usually has the higher data rate, and the secondary channel transmits in the reverse direction with a much lower data rate (for example, 75 bps).

SOME EXAMPLES OF RS-232C CONNECTIONS

Now that you understand the functions of the various signals, let's look at some examples of RS-232C interfaces. Let's begin with a simple example where the interface is used in the proper way; later we'll discuss how the interface is used in ways for which it was not intended.

Computer to Modem Interface

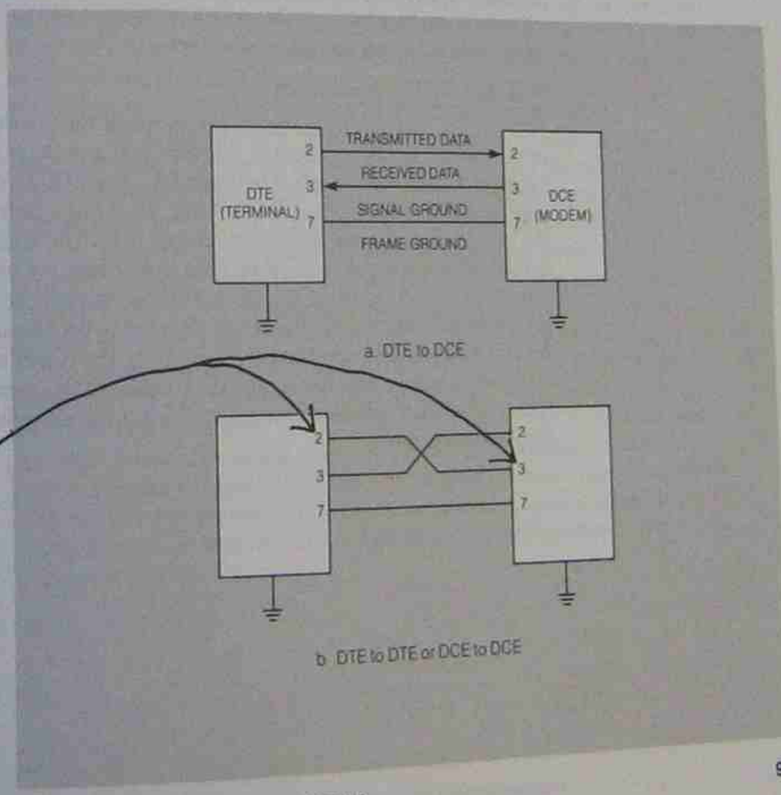
Not all of the RS-232C pins are used in every application, but the minimum DTE-DCE interface connections are: transmitted data, received data, signal ground and frame ground.

When you call from your home computer to another computer over telephone lines, you are using the Universal Seven-Part Data Circuit that we defined in Chapter 1 and illustrated in *Figure 1-9*. Your computer is connected to a modem through a RS-232C interface, and the modem is connected to the telephone network. At the far end, whether it is a local or long-distance call, the other computer is similarly connected to a modem through a RS-232C interface. *Figure 4-11a* shows an absolutely minimum RS-232C interface for the "normal" DTE-DCE connection.

Computer to Video Display Terminal Interface

The video display terminal (VDT) uses the simplest possible interface, as *Figure 4-13* shows. If both the computer and the VDT are set up as DTE, pins 2 and 3 must be interchanged at the VDT end of the cable, as *Figure 4-13b* shows. Otherwise, each would expect to transmit data on the same wire.

Figure 4-13. Minimum RS-232C Interface



Pins 2 and 3 must be interchanged if both computer and VDT are set up as DTE.

When a RS-232C cable is used to connect peripheral equipment to a computer, it may have to be modified.

This interchange usually is done at one of the cable connectors; however, an adapter with the wiring interchange done between its back-to-back connectors may be inserted between *one* end of a standard cable and the equipment. This adapter, often called a modem eliminator or null modem, permits the standard cable wiring to be left as is. These adapters are available in any combination of connector gender; that is, male-to-male, male-to-female, and female-to-female.

Computer to Serial Printer Interface

The label on the box for your new printer says, "Serial interface connects directly to the RS-232C port of your computer." When you open the box, you see a connector on the back of the printer that looks just like the RS-232C connector on your modem. (The manufacturer also charges you quite a bit extra for the absolutely essential printer cable, but that's another story.) So, it appears that the RS-232C interface on your computer does double duty; it can connect to a modem or to a printer. But the RS-232C was never intended for connecting peripheral devices such as printers to DTEs such as computers. It's not that it won't work; the problem comes in the confusion of identities. The computer is almost always (when talking about RS-232C, it *always* pays to say "almost always") internally wired as a DTE, and modems are always wired as DCEs (yes, always, because the role of the modem is defined precisely in the standard), but what about the poor printer? Is it DTE or DCE; fish or fowl? Does it use the same signals (and pins) that the modem uses, or are some different? Can you get by without buying that expensive special printer cable, or can you use your modem cable which, after all, looks just like the printer cable, even to having the same connectors.

Non-standard Use of a Standard

You probably have already guessed that the answer is bad news, or we wouldn't be talking about it. The RS-232C and other standards are intended to define the interface between DTEs and DCEs for communications over networks, usually telephone facilities. They don't address the idea of using the interface as a general serial input/output port. The wiring and control of a peripheral device in the same room, such as a printer (or another computer), is simply not considered in the standard. This means that the manufacturers are presented with a useful solution to the interfacing problem, but few guidelines (and *no* hard-and-fast rules) on how to apply it. The result is a certain amount of confusion in the marketplace, and the purpose of this section is to help the reader understand *what* happens across the interface so she or he can solve the problem if necessary.

Modified RS-232C cables often have pins that are not connected, or they are interchanged or jumpered to adapt the cable for the desired interface connections.

One source of the confusion is that printers and similar peripherals simply don't do the same things that modems do; therefore, they don't need the same control signals. The printer doesn't need to send and receive some of the signals that the modem needs, and needs some that the modem doesn't. One question is whether pin 2 or pin 3 of the RS-232C connector will be used for data transfer. In the sense of the standard, both the computer and printer are DTE, and when a DTE is connected directly to another DTE most of the signals defined in RS-232C are unnecessary. Although the manufacturer's special printer cable looks like an RS-232C cable, some of the pins aren't connected at all and some of the pin connections are interchanged or jumpered together to adapt the interface on both ends. So, the printer manufacturer's claim that its product is "RS-232C compatible" really means that the equipment accepts and generates only a small fraction of the RS-232C signals, and (hopefully) doesn't violate any other parts of the standard.

Flow Control

Let's imagine that we decide to write our own software (called a printer driver) to pass data through the RS-232C port to the printer. We bought the manufacturer's cable, and we've checked things like bit rate, parity, number of stop bits, and character code. Suppose the printer is set up for ASCII 7-bit characters, odd parity, and 1 stop bit. The printer manual says it can operate at 9600 bits per second, and since that's the fastest our computer can output serial data, we'll use that speed. Now we send a page of text to the printer, using our driver program. What happened? After the first line or two, there were more missing characters than were printed. Looks like they never made it to the printer! There doesn't seem to be any rhyme or reason, but every time we try to print using our program, data are lost. Why?

The RS-232C assumes that the flow control needed by modems would be furnished by the software at either end of the communications network.

Let's compare data rates. Asynchronous ASCII uses 7 bits for the character code, 1 for the parity bit, 1 for the stop bit, and 1 for the start bit, for a total of 10 bits per character. Since the computer is transmitting 9600 bits per second to the printer, the character rate must be 960 characters per second.

Digging out the printer manual one more time, we see that the printer's maximum rate is 80 characters per second. That means we are sending 12 times as much data as it can accept. No wonder the data were getting lost. The printer buffer would fill up, then data would be lost until buffer space became available again. What we need is a simple handshake mechanism to stop the computer from sending data when the buffer fills up, and turn it on again when the buffer empties. This is called flow control, and RS-232C is bound to have it, right? Wrong! RS-232C was designed for communicating between computers over telephone channels; the assumption is that any flow control mechanism needed is built into the software at either end. As we'll see next, various schemes have been used for flow control with varying degrees of success.

The RTS and CTS lines should not be used for flow control because it is very likely that data will be lost.

Methods Used For Flow Control

From the discussion about RTS and CTS (pins 4 and 5) above, it would seem that these signals could be used for flow control; although they often are, they shouldn't be, for many reasons. The problem is that the DCE is not allowed to drop Clear to Send until the DTE drops Request to Send. RTS and CTS are indeed handshake lines; the problem is that they indicate something besides that the DCE is ready to accept data. The RTS and CTS signals are intended to allow the terminal to request control of the communications link from the modem. The terminal assumes that it will keep the link as long as it needs it, and the DCE cannot arbitrarily drop Clear to Send whenever it desires to do so. That means that CTS cannot properly be used by the DCE as a flow control indicator; the DCE (or equipment such as a printer posing as DCE) can't raise and lower Clear to Send to tell the DTE to pause or restart transmission. Put another way, the DTE and the communications link are handshaking, not the DTE and the DCE; the handshake is for the purpose of establishing message transmission, not character-by-character or line-by-line acknowledgement for a printer.

Suppose that we decide to use RTS and CTS for flow control anyway (as some manufacturers who should know better have done). If the printer drops Clear to Send in the middle of a character transmission, what does that mean? If the transmitter stops immediately in the middle of the character, that character is sure to be garbled because timing is important in serial communications. If the transmitter waits until the character is transmitted and then stops, there may be no room in the printer's buffer for the last character. Because this possibility is not considered in the standard, the success of RTS-CTS as a flow control mechanism cannot be predicted without studying the manuals for both of the devices involved. Even doing that is no guarantee of success, but the odds are better.

A way to achieve flow control is for the peripheral equipment to send special characters to the computer, which uses software to recognize the characters.

Some manufacturers avoid the conflict by using the Data Terminal Ready or Data Set Ready (DTR or DSR) pins, depending on whether their printer thinks it is DTE or DCE. For example, if pin 6 (DSR) is used to provide flow control, the printer turns off pin 6 when it can accept no more data, and turns it on again when its buffer is nearly empty. Still, there is no assurance that the device on the other end of the cable will recognize the signal or interpret it correctly because none of these lines was intended to be used for flow control.

Another method for flow control is for the printer to send special characters for pause and return over the data line (pin 3, RD) to the computer. The software driver recognizes the special characters and controls the terminal output according to the character received. ASCII codes DC1 and DC3 (device control 1 and 3) are often used, but usually are called XON and XOFF. On the standard keyboard, these characters correspond to Control-Q (resume) and Control-S (pause). The printer logic transmits Control-S when its buffer is approximately 80% full and Control-Q when the buffer is below 20% full.

Some printer manufacturers use both the DSR line and the RD line (XON/XOFF). However, problems may occur with either method. In instances where the printer is connected to a modem (perhaps a short-haul modem on another floor of the building), there is no connection between the DSR circuits at the opposite ends (nor between RTS/CTS), so use of other pins is impossible. XON/XOFF should work fine in this instance, provided the software recognizes it, and provided that the printer logic anticipates buffer overflow in time. Problems also can occur if several terminals are multiplexed together because the reverse characters are stored for a short time in the multiplexer.

In summary, RS-232C is not very reliable for flow control because it was not designed for this application.

Of course, these problems occur because RS-232C was never intended for flow control or for DTE-to-DTE connections. We should not be surprised when individually arranged flow-control methods don't always work; they simply are not part of the standard.

LIMITATIONS OF RS-232C

It would seem that the lack of a flow control mechanism is the major drawback of RS-232C, but actually that isn't the case. Flow control is much better handled through software than through hardware, and that is the direction that new systems are moving. Small computer peripherals just haven't gotten there yet. However, RS-232C does have some limitations that restrict its use. After we discuss these, we'll look at some standards that don't have these limitations and offer other advantages as well.

Distance Limitations

The RS-232C standard of 2500 picofarads for stray capacitance may be restrictive in direct computer-to-terminal connections.

The principal problem with RS-232C is its distance limitation of 50 feet. With all but the highest data rates, much longer cable runs can be successful, but they never seem quite long enough and there's always a risk of losing data. The distance restriction is not a serious disadvantage when the modem is close to the computer or terminal, and the long-haul transmission to a remote computer takes place over telephone lines. In local applications, however, RS-232C cables often are used to connect terminals directly to computers because it is convenient to use the same terminal and computer interface whether or not a modem connection is used. (Also, the RS-232C connector is probably the only one provided.) For these connections, the 50-foot limit becomes restrictive.

An RS-232C transmitter generates a voltage between +5 and +25 volts for one of the two possible signal states (space), and a voltage between −5 and −25 volts for the opposite state (mark). Unfortunately, these voltage levels are not the same as those inside the computer and terminal, which use standard TTl and MOS logic. This means an additional power supply (usually ±12 volts) is required for the RS-232C levels. As *Figure 4-14* shows, an RS-232C receiver circuit recognizes voltage of above +3 V as spaces and voltages of below −3 V as marks. When a signal changes from one condition to the other, the specification limits the amount

of time in the undefined region to 4% of a bit period. This requirement determines the maximum amount of stray capacitance allowable in the cable because capacitance limits the rise time (or transition time) of the signal. RS-232 specifies that the capacitance must not exceed 2500 picofarads (pF). Because the cables generally used for RS-232C have a capacitance of 40 to 50 picofarads per foot, RS-232C limits cables to 50 feet (2500/50 = 50).

Improper grounding of control and data signal grounds can cause signals to be misinterpreted.

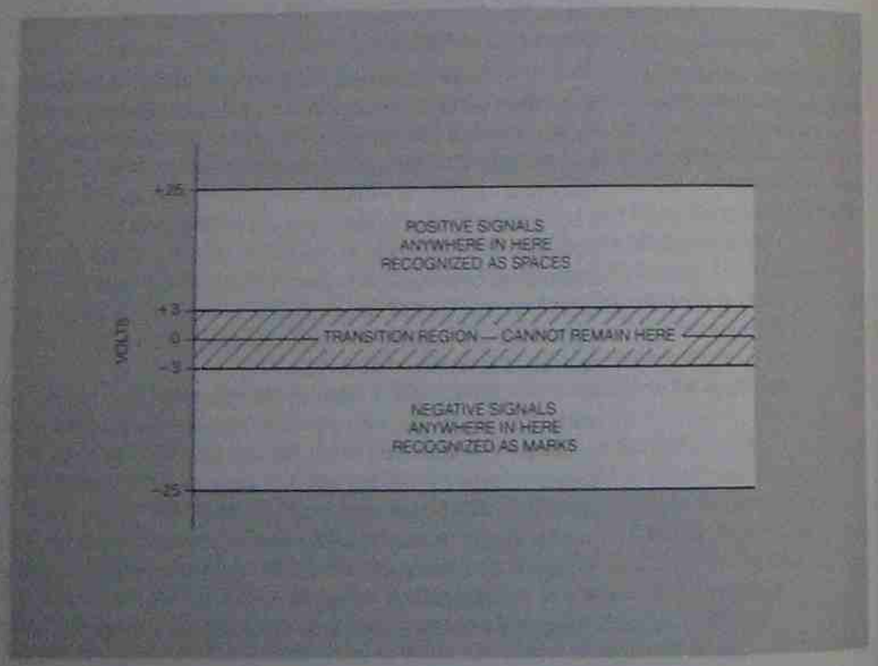
Speed Limitations

A second limitation is the maximum transmission speed of 20,000 bps, although this is not usually a disadvantage in applications between computers and terminals. The data rate between a computer and a terminal is usually 9600 bps at best, and it is difficult (and expensive, in terms of modem costs) to transmit data even at this rate over the switched telephone network.

Ground Limitations

The third disadvantage of RS-232C is, as we have discussed previously, its grounding method. The problem is not so much with the protective (chassis) ground, but with the control and data signals that all are referenced against the same signal ground wire (pin 7). This method, called unbalanced transmission, works satisfactorily most of the time.

Figure 4-14.
RS-232C Data Signal Levels at the Receiver



However, if there is a difference in ground potential between the two ends of the cable (quite likely for long runs), the transition region between a space and a mark is narrowed. When this occurs, the possibility exists that a signal will be misinterpreted, as *Figure 4-15* shows. Just as the various parts of the standard work together (speeds within the standard will usually work fine over cables that are also within the standard), so the various restrictions interact. If you try to drive cables that are too long at speeds that are too high, all these problems tend to bite you at once.

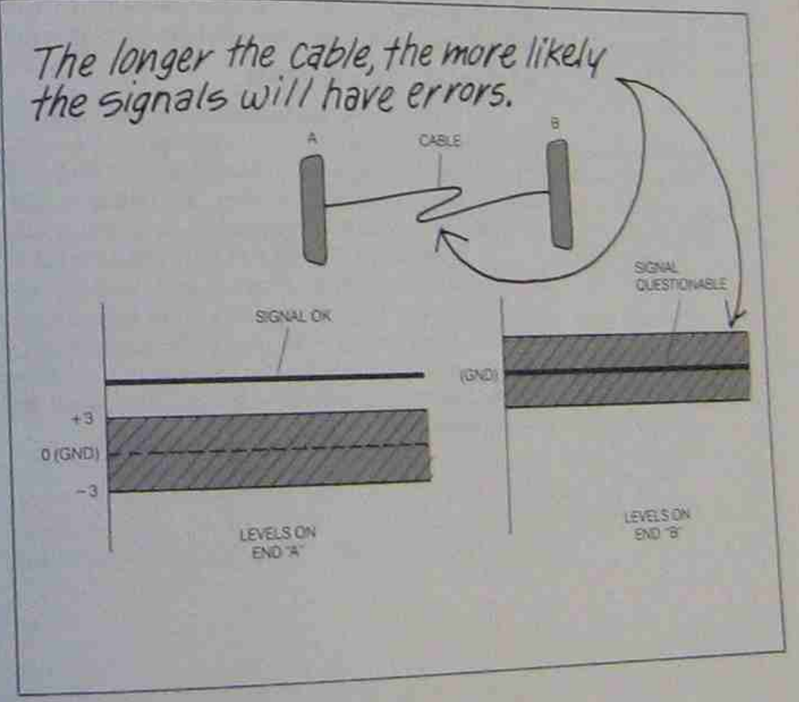
OTHER INTERFACES

An Old Standby: The Current Loop

When the serial data transmission format is defined by the presence or absence of current, the signal interface is called a current loop.

RS-232C, for all its faults, at least defines connector pins and signal levels. No such standards exist for teleprinters. The serial transmission format is defined, with start and stop bits, data bits, and parity bits, but the electrical interface varies with the manufacturer, and control signals are unknown. In teleprinters, the signal interface is called the current loop. Instead of positive and negative voltages representing logic 1s and 0s, the current loop uses presence or absence of current. Presence may be either 20 or 60 milliamperes, depending on the teleprinter model and manufacturer. No standard connector or standard pinouts exist for current-loop operation.

Figure 4-15.
Problems Because of Differing Ground Potentials in RS-232C



Data rates up to 9600 bps can be handled using current loops, even when cable lengths are longer than those allowed for RS-232C interfaces.

The idea of the current loop is to switch a current on and off. The active side of the transmission line generates the current, while the passive side switches or detects it. Since either the receiver or transmitter can be active, there are four possibilities: active receiver, active transmitter, passive receiver, and passive transmitter. Because of the convenience of locating the power supply at just one end of the link (usually at the computer), all four of these signals are found in a single computer-to-terminal connection.

Despite the lack of standards, designers of serial devices for computers like to include current-loop interface capability. It's a useful interface because teleprinters are still a cost-effective solution as a combination printer and terminal, and many devices other than teleprinters use it. Where RS-232C is limited to a 50-foot cable, a current loop can be up to 1500 feet long and handle data rates of up to 9600 bps without a special power supply. Unfortunately, the current-loop interface is completely incompatible with RS-232C and requires either hardware that has both interfaces built-in or interface boxes. The interface boxes can be built to adapt to any of the four possible interfaces. For example, if two microcomputers that are to be connected both contain the active interface, an active-to-active converter is required to directly connect them. Similarly, to connect a passive terminal directly to another passive terminal, a passive-to-passive converter is required. Fortunately, the converters can be easily constructed using an opto-isolator. The active-to-active converter does not even need a power supply, although the passive-to-passive must have one to generate the current.

RS-422A, RS-423A, and RS-449

Useful though the current loop may be in some cases, most of the world is *not* moving away from RS-232C back to teleprinters. Instead, the movement is toward two types of interfaces: one similar to RS-232C, but with some of the warts removed; the other quite different, with hardware functions replaced by software. We'll examine the one more like RS-232C first.

In the early 1970s, the EIA introduced the RS-422A, RS-423A, and RS-449 standards to overcome the defects of RS-232C while at the same time incorporating and improving upon the advantages of the current-loop interface. A major change was to unbundle the electrical, mechanical, and functional specifications of RS-232C into separate documents. RS-422A and RS-423A cover only electrical specifications, while RS-449 covers control functions.

By using balanced transmission based upon the RS-422A standard, greater speed in data transmission is obtained and ground potential problems are eliminated.

RS-422A

To allow transmission at high data rates, RS-422A uses two separate wires for each signal. This technique, called balanced transmission, doubles the number of wires in the cable, but permits very high data rates and minimizes the problem of varying ground potential. Since ground is not used as a voltage reference, RS-422A grounding requirements are much less critical than those of the RS-232C, and the use of the Signal Ground wire is optional with RS-422A.

Another major difference between RS-422A and RS-232C is the transition region between mark and space states. With the elimination of the ground potential problem, the transition region can be much narrower. In RS-422A, the difference between the voltages on the two wires determines whether a mark or a space is sent. This difference is only 0.4V in RS-422A, whereas it is 6V (+3V and –3V) in RS-232C. If the difference signal between the two wires is positive and more than 0.2V, the receiver reads a mark; if it is negative and more than –0.2V, the receiver reads a space. These voltage values allows suitable transmitters and receivers to be implemented with the ±5V power supply commonly available in computers.

The RS-423A standard defines an adapter for use between RS-232C and RS-422A which enables RS-232C equipment to interface with RS-422A equipment.

RS-423A

RS-423A transmits at lower speeds and uses one wire as a common return path for all signals in a given direction (unbalanced transmission like RS-232C, but with two return wires). The RS-423A standard operates in both RS-232C and RS-422A environments. It provides users of existing RS-232C interfaced equipment with a way to move to the new RS-422A regime, including a defined RS-232C to RS-422A adapter connector.

Because of the much smaller transition region, RS-422A transmitters will not drive RS-232C receivers correctly. RS-423A equipment, on the other hand, will interface with RS-232C signals. Each direction of transmission uses a common return path that is connected to ground only at the transmitter end. The receiver determines whether a mark or space is present by examining whether the signal wire is negative or positive with respect to the common return. Because this return path does not connect to ground in the receiver, the problem of ground currents does not arise.

In an RS-423A transmitter, the voltage difference between the signal line and the common return must be at least 4V, positive for a space and negative for a mark. This gives an 8V transition region, which is compatible with RS-232C receiver circuitry, but which presents the same power supply problem that occurs with RS-232C. Because the RS-423A receiver must properly respond to the same 0.4V transition region as a RS-422A receiver, RS-422A transmitters can be used with RS-423A receivers.

ST. GEORGE TECHNICAL COLLEGE LIBRARY

The RS-449 standard, which uses two connectors instead of one, is intended to replace the functional portion of RS-232C. Even so, it has no provision for automatic dialing.

RS-449

RS-449 is the intended successor to the functional portion of RS-232C, including the mechanical specification of the plugs and sockets. Apart from its improved speed and distance specifications, RS-449 offers some minor functional enhancements over RS-232C in automatic modem testing and provision for a standby channel, but it still does not incorporate outward dialing. The acceptance of RS-449 in the marketplace seems questionable at this time.

Figure 4-16 shows the RS-449 signals with the corresponding RS-232C and V.24 signals. Notice the similarity between the new and old standards. The major differences are in the grounding arrangements (Send Common and Receive Common) and testing facilities. Apart from these, only a few miscellaneous signals have been added. All signals use the RS-423A transmission standard except those that may optionally use the RS-422A for higher-speed links. (Two wires are specified for each of these.) The signals are divided between a 37-pin and a 9-pin connector, and the ground and common signals are handled separately for each cable. Many applications will not need the smaller cable because it contains only signals relevant to the secondary channel.

The RS-366 standard is different because it deals with automatic dialing under computer control.

RS-366

None of the RS- standards we've discussed provide for automatic dialing of calls by a computer. RS-232C and RS-449 provide specifications for answering calls, but not for dialing. A different standard, RS-366, covers automatic calling units. The reason is that, until recently, equipment for placing calls under computer control was expensive because the actions required in making a telephone call can be quite complicated. The dialing equipment must be able to determine whether the line is free, take the phone off hook (in effect), wait for and recognize dial tone, dial the number, and understand and respond to the various call progress tones (such as the busy signal) that the telephone system provides to indicate the status of a call.

The principal use of RS-366 automatic dialing equipment in data communications is for dial backup of private-line data circuits, and for automatic dialing of remote terminals to allow them to transmit data. For example, a company's central computer could be programmed to dial automatically the computer in each branch office each evening, and to cause (via software) the day's transactions to be transmitted to the main office.

Figure 4-16. Comparing RS-449 With RS-232C Signals *(Source: EIA Standard RS-449, Electronic Industries Association)*

RS-449		RS-232C		CCITT Recommendation V.24	
SG	Signal Ground	AB	Signal Ground	102	Signal Ground
SC	Send Common			102a	DTE Common
RC	Receive Common			102b	DCE Common
IS	Terminal In Service				
IC	Incoming Call	CE	Ring Indicator	125	Calling Indicator
TR	Terminal Ready	CD	Data Terminal Ready	108/2	Data Terminal Ready
DM	Data Mode	CC	Data Set Ready	107	Data Set Ready
SD	Send Data	BA	Transmitted Data	103	Transmitted Data
RD	Receive Data	BB	Received Data	104	Received Data
TT	Terminal Timing	DA	Transmitter Signal Element Timing (DTE Source)	113	Transmitter Signal Element Timing (DTE Source)
ST	Send Timing	DB	Transmitter Signal Element Timing (DCE Source)	114	Transmitter Signal Element Timing (DCE Source)
RT	Receive Timing	DD	Receive Signal Element Timing	115	Receiver Signal Element Timing (DCE Source)
RS	Request to Send	CA	Request to Send	105	Request to Send
CS	Clear to Send	CB	Clear to Send	106	Ready for Sending
RR	Receiver Ready	CF	Received Line Signal Detector	109	Data Channel Received Line Signal Detector
SQ	Signal Quality	CG	Signal Quality Detector	110	Data Signal Quality Detector
NS	New Signal			126	Select Transmit Frequency
SF	Select Frequency			111	Data Signaling Rate Selector (DTE Source)
SR	Signaling Rate Selector	CH	Data Signal Rate Selector (DTE Source)	112	Data Signaling Rate Selector (DCE Source)
SI	Signaling Rate Indicator	CI	Data Signal Rate Selector (DCE Source)		
SSD	Secondary Send Data	SBA	Secondary Transmitted Data	118	Transmitted Backward Channel Data
SRD	Secondary Receive Data	SBB	Secondary Received Data	119	Received Backward Channel Data
SRS	Secondary Request to Send	SCA	Secondary Request to Send	120	Transmit Backward Channel Line Signal
SCS	Secondary Clear to Send	SCB	Secondary Clear to Send	121	Backward Channel Ready
SRR	Secondary Receiver Ready	SCF	Secondary Received Line Signal Detector	122	Backward Channel Received Line Signal Detector
LL	Local Loopback			141	Local Loopback
RL	Remote Loopback			140	Remote Loopback
TM	Test Mode			142	Test Indicator
SS	Select Standby			116	Select Standby
SB	Standby Indicator			117	Standby Indicator

The CCITT standards are aimed toward direct digital data communications. CCITT X.21 covers both automatic dialing and answering, and uses full-duplex synchronous transmission.

CCITT X.21

The CCITT has charted a course different from the RS- standards, a course that was predicated on the eventual availability of direct digital connection to a digital telephone network. Then all data transmission will be synchronous, and the communication equipment will provide bit and byte timing signals. The CCITT X.21 recommendation, introduced in 1976, includes the protocol for placing and receiving calls and for sending and receiving data using full-duplex synchronous transmisson. Byte-timing signals are an option, but one which the vast majority of digital telephone exchanges will almost certainly provide. In sharp contrast to RS-449, X.21 uses only six signals. The electrical specifications are contained in recommendations X.26 (corresponding to EIA RS-422A) and X.27 (corresponding to EIA RS-423A).

The eventual minimum line speed under X.21 is likely to be 64,000 bps, the data rate now used to encode voice in digital form in the telephone network. In Europe, British Telecom offers X.21 customer data rate access to its Kilostream digital service at 2.4, 4.8, 9.6, 48 and 64 Kbps. Customer data rates up to 9.6 Kbps are converted to a line rate of 12.8 Kbps, whereas data rates of 48 and 64 Kbps both use a line rate of 64 Kbps for transmitting through the Kilostream network.

The circuits used in X.21 are given in *Figure 4-17*. The first two circuits listed provide the voltage reference and ground connections. The computer sends data to the modem on the Transmit line, and the modem returns data on the Receive line. The Control and Indication circuits provide control channels in two directions. The Signal Element Timing line carries the bit timing, and the Byte Timing line carries the byte timing. Although the Control and Indication lines are control lines, most of the controlling information uses the Transmit and Receive lines. The computer changes the state of Control when it wants to place a call, just as you lift the handset when you want to dial. To end the call, the computer changes

Figure 4-17.
CCITT X.21 Signals
(Source: CCITT Recommendation X21,CCITT, Geneva, C.H.)

Interchange Circuit	Name	To DCE	From DCE
G	Signal Ground or Common Return	(Note)	
Ga	DTE Common Return	X	
T	Transmit	X	
R	Receive		X
C	Control	X	
I	Indication		X
S	Signal Element Timing		X
B	Byte Timing		X

NOTE: See Recommendation X.24

Control back to the idle state. Similarly, the modem changes the state of Indication when the remote telephone is answered and changes it back if it shuts down. All the dialing information travels on the Transmit line in coded form, and all the information about tones comes back on the Receive line.

The major advantage of X.21 over RS-232C and RS-449 is that the X.21 signals are encoded in serial digital form. For example, when a dial tone is received, a continuous sequence of ASCII "+" characters is sent to the computer on the Receive wire. In effect, this is a digital dial tone. The computer dials the number by transmitting it as a series of ASCII characters on the Transmit line, one bit at a time. After dialing the call, the computer receives call progress signals from the modem on the Receive wire. These signals indicate such things as number busy, access disallowed, and network congestion.

X.21, with its signals encoded in serial digital form, permits a few signal lines to accomplish all functions.

By using serial digital coding instead of dedicated wires for special functions, X.21 establishes a basis for providing special services in computer communication. A short-code-dialing or repeat-last-call facility would be extrememly useful, for example, to reconnect a call everytime you complete a line of typing at the terminal. (Some European nations are trying this approach to networking right now.) If the line could be disconnected while you are typing, long-distance calls would be much cheaper. Of course, this would also depend on the tariff policies of the telecommunication carriers. Using X.21 could allow many of the advantages of telecommunications by the packet-switching technique without the associated complexity.

Although X.21 is defined as the lowest (or "physical") level of the international X.25 packet-switching protocol, it is far ahead of its time because direct digital connection to public telephone networks is hardly possible now. For this reason, CCITT offers the X.21 bis recommendation as an interim measure to connect existing computer equipment to packet communication services. X.21 bis is virtually identical to CCITT V.24 and RS-232C. (Packet switching and CCITT protocols are discussed more in a later chapter.)

TYPES OF MODEMS

Asynchronous (300 or 300/1200 Baud) Auto-Answer Modem

The asynchronous 300 or 300/1200 baud auto-answer modem is usually used with unattended personal computers that are connected to dial-up circuits.

These modems are usually used with personal computers. The signals needed are: Protective Ground, Signal Ground, Transmitted Data, Received Data, Request To Send, Clear To Send, Data Terminal Ready, Ring Indicator, Received Line Signal Detector, and possibly Data Set Ready. Pin 20 (DTR) indicates that the computer is ready to receive calls, while Pin 22 (RI) indicates that the modem is receiving the telephone ringing signal, going high and low as the ringing current goes off and on. (This is the way a modem can count how many rings to wait before answering.) If the computer leaves pin 20 (DTR) on all the time, the modem answers incoming calls without delay. If DTR is off, the computer will turn it on in response to RI to tell the modem to answer the call. At the completion of the call, the computer will turn off pin 20 (DTR) to cause the modem to disconnect the line (go on hook).

The following sequence of events is one possible set that can happen when a computer receives a call from another computer via a 300 baud asynchronous modem: When the modem detects ringing, it turns on pin 22 (RI), the DTE responds by asserting pin 20 (DTR), and the modem goes off hook to answer the phone. The answering DTE asserts pin 4 (RTS), which commands the modem to turn on its transmitter. After the RTS-CTS delay, the modem responds by asserting CTS. At the other end of the line, the computer operator hears the carrier signal and either pushes the data button or puts the telephone handset into the acoustic coupler. Now the calling modem's transmitter turns on, producing its own carrier tone. When the modem at the receiving computer hears this carrier, it turns on pin 8 (DCD). Upon receiving this signal, the receiving computer begins transmitting data to tell the sending computer to log in, etc. Some operating systems wait for the caller to transmit a special character to begin the log-on process.

At the end of the session, the caller logs off, causing the receiving computer to terminate the call. It turns off pin 4 (RTS), which causes the receiving modem to turn off its carrier signal. The receiving computer then drops pin 20 (DTR), causing the modem to go on hook to "hang up" the telephone line. At the calling end, pin 8 (DCD) goes off, the operator replaces the handset, and the action is complete.

Full-Duplex Asynchronous Private Line Modem

A private line modem does not have the capability for automatic answering because it's not needed, the modem is always on-line.

This modem could be used on switched lines as well; it is called a private line modem because it is not equipped for automatic answer. Signals used are: Protective Ground, Signal Ground, Transmitted Data, Received Data, Received Line Signal Detector (Pin 8), and Data Set Ready (Pin 6). Pin 8 (DCD) says in effect, "I hear something like a modem talking to me," in other words, that another modem is trying to make contact on the line. Pin 6 (DSR) usually is not used in asynchronous modems, but could indicate that the modem is ready and not in voice or test mode.

Half-Duplex Asynchronous Private Line Modem

The signals used in this modem are: Protective Ground, Signal Ground, Transmitted Data, Received Data, Request To Send, Clear To Send, Received Line Signal Detector, and possibly Data Set Ready. Pins 4 and 5, (RTS and CTS) control the transmission direction in the half-duplex operation. The computer asserts pin 4 (RTS) when it wants to transmit. The modem responds by turning on pin 5 (CTS) to indicate that it is ready to receive characters to be transmitted. There is a built-in delay (typically 200 milliseconds) between the Request To Send and the Clear To Send response because the modem must generate the carrier waveform and allow it to stabilize. When transmission is completed in that direction, the computer drops pin 4 (RTS), causing the modem to turn off the transmitter and to drop CTS. The process of determining who transmits when, and how to tell when the other end is finished, is the responsibility of the software protocol which is outside the modem and the RS-232C interface.

By combining a microprocessor with a modem, both data communications and automatic dialing can be realized in one unit—a "smart modem".

Smart Modems

The so-called "smart" or "intelligent" modem combines a standard modem (300, 300/1200 or 2400 baud) and a microprocessor to provide both data communications and automatic dialing in one unit. (One of the best known is the Smartmodem 1200 made by Hayes Microcomputer Products, Inc.) A smart modem accepts commands in ASCII form over the same RS-232C interface used for data transmission and can respond over the same interface with either status codes or sentences. For example, the computer operator would turn the modem on and load the terminal software, then type a code such as "AT TD1-800-772-5942". This tells the modem to take the phone line off hook, wait for a dial tone, then tone dial (using DTMF tones) the number 1-800-772-5942. The modem will attempt to do this. If the number is busy, the modem's response, "number busy," would appear on the computer's screen. Modems of this type usually can redial the last number entered, and many of these modems automatically keep trying a number at preset intervals if they encounters a busy signal. Some modems can store a list of frequently called numbers (such as the one above) that the user can select by entering only a two- or three-digit code.

The smart modem contradicts our previous statement that modems don't look at the data that passes through them, since it does examine some of the information coming over the RS-232C interface for its commands. Smart modems are being used increasingly by business and individual computer users, and a number of software products exist to support these modems. Virtually all of the software utilizes the Hayes interface protocol, and most (but not all) smart modems accept commands in the Hayes format, which has become a de facto standard. Thus, if you are considering the purchase of a smart modem, it would be wise to consider one that is compatible with the Hayes commands.

The Hayes command set consists of a basic set of commands and command extensions. The basic commands, such as placing the modem off hook and dialing a number, are applicable only to modems designed to transmit and receive data at that speed.

Commands in the Hayes command set are initiated by transmitting an attention code to the modem, followed by an appropriate command or set of commands. The attention code is the character sequence AT, which must be input by the user or sent to the modem by a software program in all uppercase or all lowercase letters. The requirement to prefix all command lines with the code AT has led many modem manufacturers to denote their modems as Hayes AT compatible or simply AT compatible.

A Hayes Smartmodem can hold up to 40 characters in its command buffer, permitting a sequence of commands to be transmitted to the modem on one command line. This 40-character limit does not include the attention code, nor does it include spaces placed in a command line to make it more readable. *Table 4-18* lists the major commands included in the basic Hayes Command Set. Other modems, such as the Telebit Trailblazer, employ a command set that is a superset of the Hayes command set. Although communications software that uses the Hayes command set operates with this modem, such software cannot use the

modem's full potential because the Hayes command set in 1987 supported a data rate selection only up to 2400 bps, whereas the Trailblazer modem could be set to data rates of 9600 and 19,200 bps by proprietary software commands.

Table 4-18. Major Commands of the Hayes Command Set

Command	Description
A	Answer call
A/	Repeat last command
C	Turn modem's carrier on or off
D	Dial a telephone number
E	Enable or inhibit echo of characters to the screen
H	Hang up telephone (on hook) or pick up telephone (off hook)
I	Request identification code or request check sum
M	Turn speaker off or on
O	Place modem on-line
P	Pulse dial
Q	Request modem to send or inhibit sending of result code
R	Change modem mode to "originate-only"
S	Set modem register values
T	Touch-Tone dial
V	Send result codes as digits or words
X	Use basic or extended result code set
Z	Reset the modem

The basic format required to transmit commands to a Hayes compatible modem is shown below.

AT Command[Parameter(s)]Command[Parameter(s)] ..Return

Each command line includes the prefix AT, followed by the appropriate command and the command's parameters. The command parameters are usually the digits 0 or 1, which serve to define a specific command state. As an example, C1 is the command that tells the modem to turn on its carrier signal, whereas C0 is the command that causes the modem to turn off its carrier signal. Up to 40 characters can be entered into a command line, and each command line must be terminated by a carriage return character.

To illustrate the use of the Hayes command set, suppose that you want to dial automatically information for New York City. The number is 212/555-1212. First, you tell the modem to go off hook, which is similar to picking up a telephone handset. Then, you tell the modem the type of telephone system that we are using—pulse or touch-tone—and the telephone

number to dial. If you have a terminal or personal computer connected to a Hayes compatible modem, the following commands would be sent to the modem:

AT H1
AT DT9,212-555-1212

In the first command, the 1 parameter used with the H command places the modem off hook. In the second command, DT tells the modem to dial (D) a telephone number using touch-tone (T) dialing. The digit 9 was included in the telephone number because we assumed that you had to dial 9 to dial long distance through your company's switchboard. The comma between the long distance access number (9) and the area code (212) causes the modem to pause for two seconds before dialing the area code. Usually, two seconds is long enough for you to hear the long distance dial tone before the modem dials the area code number.

Because a smart modem automatically goes off hook when dialing a number, the first command line is not required and is normally used for receiving calls. In the second command line, the type of dialing does not have to be specified if a previous call was made because the modem automatically uses the last type specified. Although users with only pulse dialing availability must specify P in the dialing command when using a Hayes Smartmodem, other vendors offer modems that can determine automatically the type of dialing facility that the modem is connected to, then use the appropriate dialing method without requiring the user to specify the type of calling.

WHAT HAVE WE LEARNED?

1. A modem (modulator-demodulator) changes data pulses to analog tones that the telephone channel will pass.
2. Frequency-shift keying is the type of modulation used by most low-speed modems. Phase modulation or a combination of phase and amplitude modulation is used in high-speed modems.
3. Data terminal equipment (DTE) is the equipment that originates or receives the data in digital form.
4. Data communications equipment (DCE) is the equipment that converts the data from the DTE to the form required for the transmission channel. A modem is DCE.
5. The signal levels, signal identification, and wiring between the DTE and the DCE are governed by standards.
6. The RS-232C and CCITT V.24 standards for the DTE-DCE interface are the ones used most in the present computer and communications world.
7. RS-422A, RS-423A, and RS-449 are relatively new standards that overcome the defects of RS-232C.

Quiz for Chapter 4

1. Flow control is:
   a. What people do who open and close floodgates on dams.
   b. The process whereby the modem matches the rate of the receiver.
   c. The process of starting and stopping the terminal output to avoid loss of characters by the receiving device.
2. Buffering is:
   a. The process of temporarily storing data to allow for small variations in device speeds.
   b. A method of reducing the severity of communications headaches.
   c. Storage of data within the transmitting modem until the receiver is ready to receive.
3. Direct machine-to-machine transmission over long distances without modems is not practical because:
   a. Copper wire does not transmit dc efficiently.
   b. A dc path that will handle data in pulse form does not exist.
   c. Data comes from the computer in the form of tones, not pulses.
4. Modulation is the:
   a. Varying of some parameter of a carrier, such as its amplitude, to transmit information.
   b. Utilization of a single transmission channel to carry multiple signals.
   c. Transmission of pulses in dc form over a copper wire.
5. RS-232C, RS-449, V.24, and X.21 are examples of:
   a. Standards for various types of transmission channels.
   b. Standards for interfaces between terminals and modems.
   c. Standards for interfaces between modems and transmission facilities.
   d. Standards for end-to-end performance of data communications systems.
6. A smart modem can:
   a. Detect transmission errors and correct them automatically.
   b. Correctly answer multiple-choice quizzes.
   c. Accept commands from the terminal via the RS-232C interface.
7. RTS/CTS:
   a. Is the way the modem indicates ringing, and the way the terminal indicates it is ready for the call to be answered.
   b. Is the way the DTE indicates that it is ready to transmit data, and the way the DCE indicates that it is ready to accept data.
   c. Are the pins that represent received transmissions and carrier transmissions.
8. Pin 7, the Signal Ground:
   a. Completes the circuit for control and data signals.
   b. Indicates a failure in the ground side of the transmission line.
   c. Completes the circuit for control signals but not data signals.
9. Pin 22, the Ring Indicator:
   a. Must be present on all modems connected to the switched network.
   b. Is used on acoustic-coupled modems but not on direct-connect types.
   c. Is asserted when ringing voltage is present on the line.
10. The reason that many cables have "RS-232" connectors with some wires crossed or connected to each other is:
    a. There are various RS-232 standards.
    b. Many computers and peripherals use RS-232 serial interfaces but not as DTE-to-DCE.
    c. Asynchronous modems reverse the direction of transmitted and received data from the standard.

Synchronous Modems and Digital Transmission

ABOUT THIS CHAPTER

In this chapter, you will be introduced to the signaling methods used by synchronous modems, the American and International standards which apply to synchronous data modems, and the concepts used in higher speed (above 15,000 bits per second) wideband devices. In addition, the techniques of direct transmission of binary signals without modems, called digital transmission, will be discussed.

SYNCHRONOUS SIGNALING AND STANDARDS

A Search for Higher Data Rates

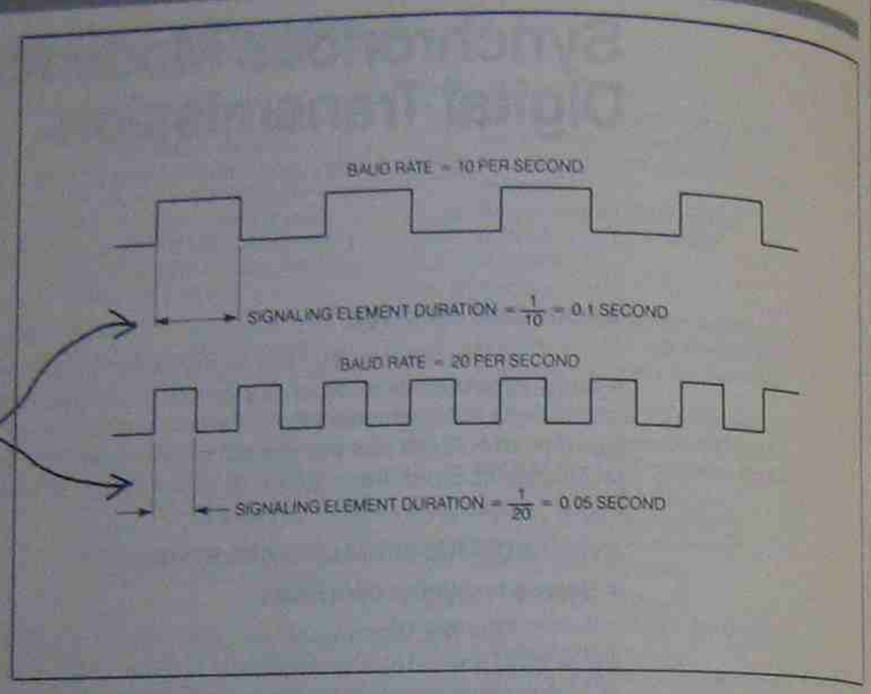
Like the Olympic athlete, whose motto is "alitus, cititus, fortitus," the aims of the data communications systems designer in the development of high-technology products are to pack more data per channel, deliver it at a faster speed, and deliver it at a lower cost. The transmission of a larger quantity of data through a channel of a given bandwidth with smaller and cheaper devices is a natural target. Since in many cases the channel available is a standard voice-grade line, and the cost of the channel dominates the cost of the system, a large amount of design effort has gone into the production of devices that will send the largest number of bits per unit time through standard (voice-grade) channels. One result has been synchronous signaling.

Sending the Clock With the Data

Binary signals have two basic parameters: amplitude and duration. As the number of information bits sent per unit of time increases, the bit duration decreases.

Binary signals sent over a channel represent the quantization of data in two dimensions: amplitude (voltage of the electrical wave), and time (the duration of each signaling element). As the number of signaling elements per unit time (the "baud rate") increases, the duration of each element must decrease as illustrated in *Figure 5-1*. If, as is the case with the asynchronous modems described in a previous chapter, the time base for the transmitter and receiver are independent, small differences between the two clocks become increasingly likely to cause errors due to the sampling of the data at the wrong time. This problem is overcome in synchronous modems by deriving the timing information for the received data.

Figure 5-1.
Baud Rate Versus Element Duration



Duration decreases as baud rate increases.

The data rate is limited by the bandwidth of the voice-band channel to about 2400 bauds. One way to increase the data rate is to pack two or more bits in a baud.

More Bits per Baud

Recall that in a previous chapter it was stated that for a channel of given bandwidth, the maximum signaling rate is fixed by the channel bandwidth and the signal-to-noise ratio (Shannon's formula). Since voice-grade channels have a fixed bandwidth (about 3000 Hz), they have a fixed maximum baud rate (about 2400 baud). The solution to this dilemma is to pack more than one bit into one signaling element; that is, two or more bits per baud.

TYPICAL SYNCHRONOUS COMPONENTS

The greater complexity and cost of synchronous modems over asynchronous units is due to the circuitry necessary to derive the timing from the incoming data and to pack more than one bit into one baud. Synchronous modems typically consist of four sections as shown in *Figure 5-2:* transmitter, receiver, terminal control, and power supply. We will look at the first three of these in some detail.

Transmitter

As shown in *Figure 5-3*, the transmitter section of a synchronous modem typically consists of timing (clock), scrambler, modulator, digital-to-analog converter and equalizer circuits.

Figure 5-2.
Synchronous Modem Block Diagram

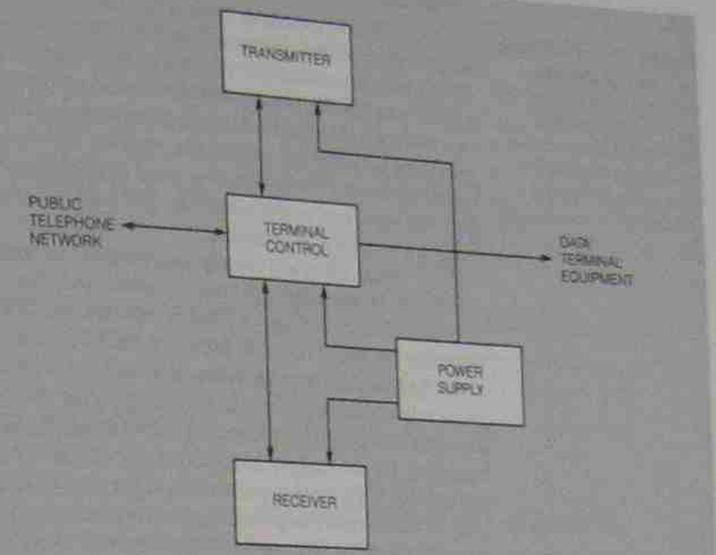
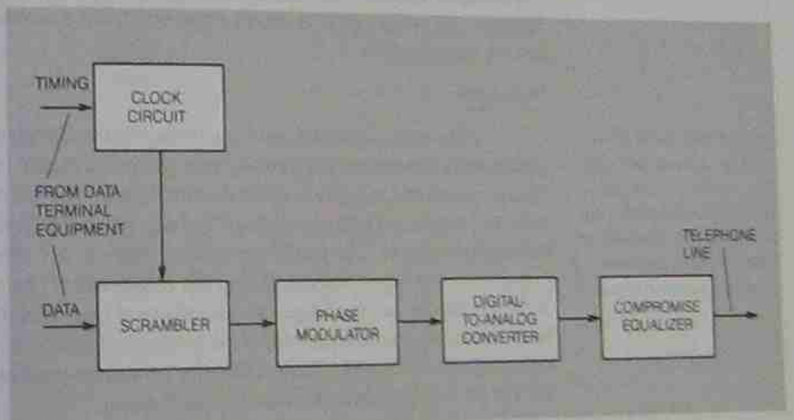


Figure 5-3.
Modem Transmitter Block Diagram



The internal clock of the transmitting modem is phase locked with the external clock source when the DTE or another modem provides the timing.

Timing Circuit

The timing circuit provides the basic clocking information for both the modem and the data terminal equipment (DTE) that is providing the data to be transmitted. Certain data circuit arrangements require that the clocking for the transmitted data be supplied by the DTE (which may be another modem). In these cases, an option is usually provided in the modem to phase lock the internal clock to an external clock source input through the DTE interface. The internal timing is usually controlled by a crystal oscillator to within about 0.05% of the nominal value.

Enough 0 to 1 to 0 changes must exist in the data stream to keep the receiving modem's clock synchronized. The scrambler ensures that the changes occur by modifying the data stream in a controlled way.

Scrambler

Since the receiver clock is derived from the received data, those data must contain enough changes from 0 to 1 (and vice versa) to assure that the timing recovery circuit will stay in synchronization. In principle, the data stream provided by the associated terminal or business machine can consist of any arbitrary bit pattern. If the pattern contains long strings of the same value, the data will not provide the receiver with enough transitions for synchronization. The transmitter must prevent this condition by changing the input bit stream in a controlled way. The part of the transmitter circuitry that does this is called the scrambler.

Scramblers are usually implemented as feedback shift registers, which may be cascaded (connected in series) as shown in *Figure 5-4*. Input bit #1 is modified by adding the state of selected bits of the 7-bit register to the incoming bit. Input bits 2, 3, etc are modified by the preceding bits as indicated in the figure. The scramblers are designed to ensure that each possible value of phase angle is equally likely to occur to provide the receiver demodulator with enough phase shifts to recover the clocking signal. Although scrambling is necessary for the reasons cited above, it increases the error rate, since an error in one bit is likely (and in some cases is certain) to cause an error in subsequent bits. To counteract this problem, some modems encode the input to the scrambler into the Gray code so that the most likely error in demodulation (picking an adjacent phase state) will cause only a 1-bit error when decoded at the receiver. As shown in *Table 5-1*, the Gray code has the property that, between any two successive binary numbers (tribit numbers in this case), only one bit changes state.

Modulator

Specific modulator characteristics such as the baud rate, number of bits per baud and modulation techniques vary for modems of different rates and from different manufacturers.

The modulator section of the transmitter converts the bit patterns produced by the scrambling process into an analog signal representing the desired phase and amplitude of the carrier signal. The carrier frequency, baud rate, and number of bits represented by each baud is different for modems of different data rates. The modulator collects the correct number of bits and translates them into a number giving the amplitude of the electrical signal that is correct for the carrier frequency and phase of the carrier at that instant in time. Modulation techniques differ for modems of different speed and from different manufacturers. Some of the more common ones are discussed in the section on standard modems later in this chapter.

Digital-to-Analog Converter

The digital-to-analog converter changes the binary encoded signal from the modulator into the required analog voltage.

The binary encoded signal from the modulator is fed to a digital-to-analog converter which produces the actual analog voltage required. This in turn goes through a low-pass filter to remove frequencies outside of the voice band, then through a circuit called an equalizer which compensates for transmission impairments on the line.

Figure 5-4. Tribit Scrambler

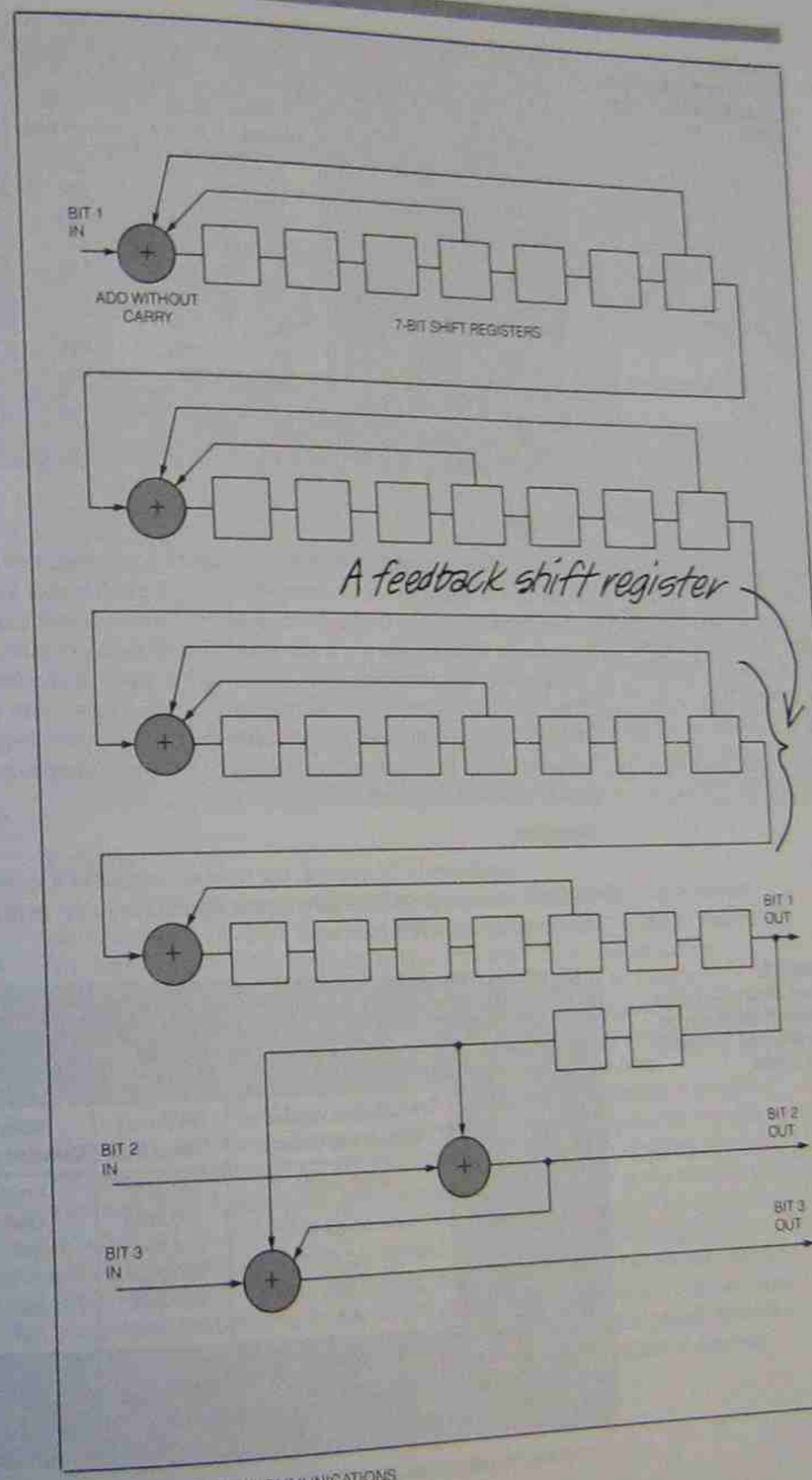


Table 5-1. Binary and Gray Code Equivalence for 3-bit Code

Decimal	Binary	Gray Code
0	000	000
1	001	001
2	010	011
3	011	010
4	100	110
5	101	111
6	110	101
7	111	100

Equalizer

The transmitter equalizer corrects amplitude distortion and group delay based upon the average characteristics of the transmission medium.

The transmitter equalizer is called a compromise or statistical equalizer because it is set to compensate for the nominal or average characteristics of the transmission medium. The equalizer compensates for amplitude distortion in the medium and for the problem called group delay. Group delay measures the amount by which a signal of one frequency travels faster in the transmission medium than a signal of a different frequency. Group delay usually is expressed in microseconds (us) at a given frequency. Nominal values of delay variation versus frequency for some commonly available voice grade channels are given in *Table 5-2*.

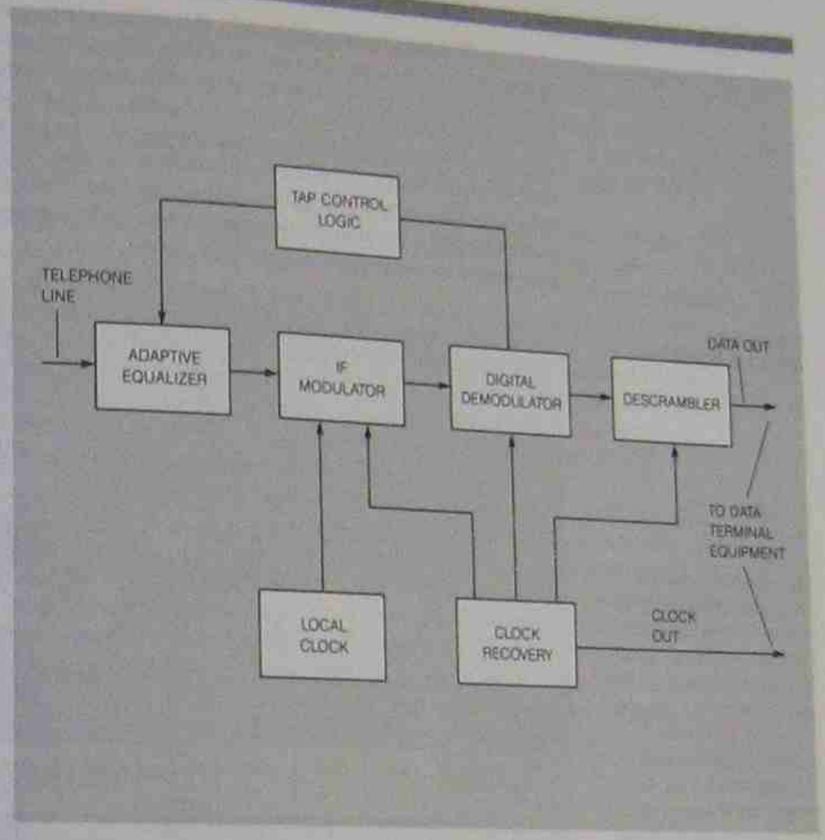
Receiver

As shown in *Figure 5-5*, the receiver section of a synchronous modem typically consists of an equalizer, timing (clock) recovery, demodulator, descrambler, and DTE interface.

Table 5-2. Group Delay Variation Versus Frequency for Conditioned Transmission Lines

Conditioning Applied to Voice Grade Channel	Frequency Range (Hz)	Delay Variation (μs)
Basic (none)	800-2600	1750
C1	800-2600	1750
C2	600-2600	1500
C3	600-2600	260-300
C4	800-2800	500
C5	600-2600	300

Figure 5-5. Modem Receiver Block Diagram



Equalizer

The receiving equalizer must deal with actual errors in the received signal. The errors are measured and corrected by adjusting specific circuit parameters.

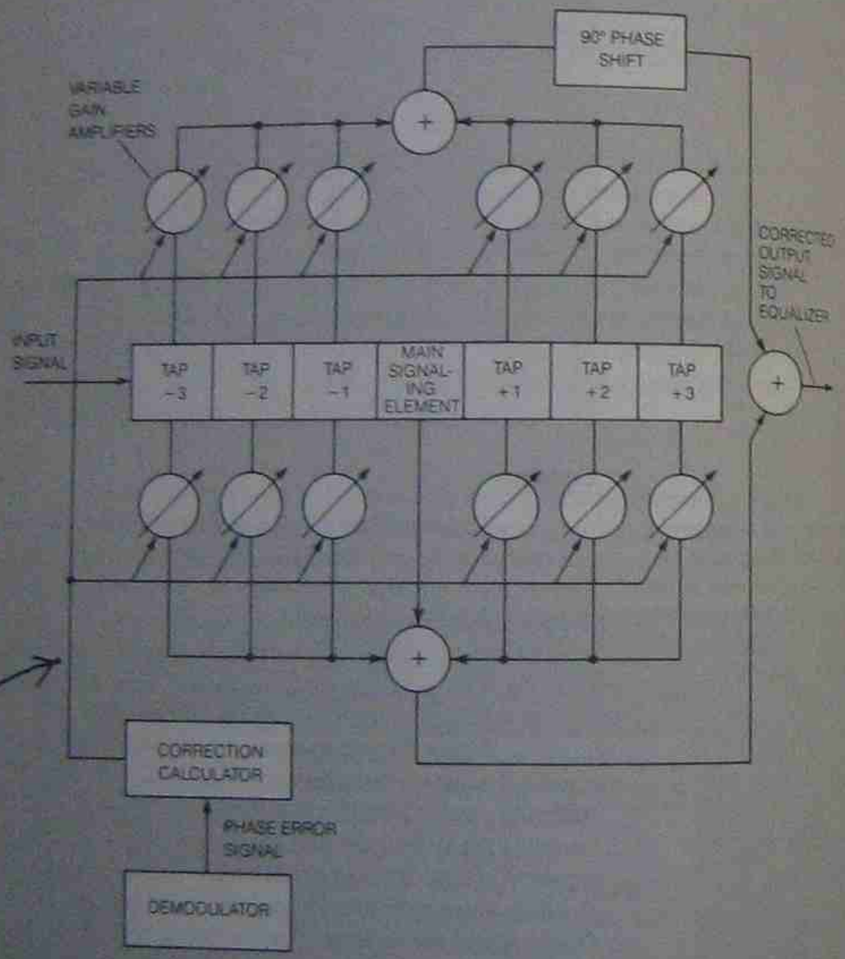
The equalizer section of the transmitter is relatively simple since it can compensate only for the average of expected errors on the output channel. The receiver equalizer, however, must compensate for the actual errors introduced in the transmission path. This is done by using an adaptive equalizer which measures errors observed in the received signal and adjusts some parameter of the circuit (usually the receiver clock frequency) to track slowly varying changes in the condition of the transmission line.

Delay distortion has the greatest effect on transmission of an analog signal. As mentioned above, analog signals of different frequencies travel at different rates through a transmission medium. Since each signaling element contains many frequencies, each signaling element arrives at the receiver over a period of time instead of all at once. The frequencies which travel faster (leading frequencies) arrive earlier and those traveling slower (lagging frequencies) arrive later. The leading and lagging frequencies not only fail to make their proper contribution to the proper signaling element, but also cause interference with signaling elements behind and ahead of the proper element. The equalizer must get the parts of each element back together and cancel their effects on other elements.

The adaptive equalizer does this by use of a tapped delay line which stores the analog signal for a period of time. This time period includes the main signaling element to be corrected at the center, and several element times before and after the center as shown in *Figure 5-6*. The taps in the delay line allow the analog voltage representing the signal to be picked off at specified time intervals before and after the element of interest. The time interval between each tap is the reciprocal of the baud rate. The voltage from each tap is amplified by a variable gain amplifier whose gain is controlled by an amount determined by the correction calculator circuit.

Figure 5-6. Adaptive Equalizer

The adaptive equalizer compensates for delay distortion by temporarily storing the analog signal in a tapped delay line. The signals from each tap, amplified by a different amount as determined by the amount of error detected, are summed to form the corrected signal.



Controls the amplifier gain.

In response to a phase error signal developed by the demodulator circuit, the correction calculator determines the amplitude and polarity of the signal needed from each tap amplifier. What makes the equalizer adaptive is that the error signal is continuously derived from the difference between the phase of the received signal and a nominal phase value determined by the demodulator.

Start-up bits (training sequences) are sent at the beginning of a transmission to allow the adaptive equalizer to prepare itself for the incoming message.

The circuit adapts itself fairly slowly to changing line conditions because of all of the delays. This can be a disadvantage for half-duplex operation, or at start-up on a one-way link, so a start-up or training sequence of bits usually is sent first to allow the receiver to adapt more rapidly than it would on random data. This is called training. The amount of time required by the modem to retrain after changing direction of transmission must be less than the turnaround time or request-to-send, clear-to-send (RTS-CTS) delay. This delay effectively reduces the data transmission rate on the circuit. Contemporary 4800 bps modems have RTS-CTS delays in the 10 to 20 millisecond (ms) range (about 50 to 100 bit times).

The equalizer section of some modern modems attempts to minimize the amount of line time lost in turnaround by a dual speed or "gearshift" technique. This technique causes the modem pair to start at a lower data rate (say 2400 bps) at which training is very fast, and to transmit data at that rate while the adaptive equalizer goes through its training cycle. The modems then "shift gears" to run at the full data rate as shown in *Figure 5-7*. Since for short messages the RTS-CTS delay is the controlling factor in throughput, and for long messages the transmission speed is dominant, the speed changing technique minimizes the first while still providing the second.

Timing Recovery

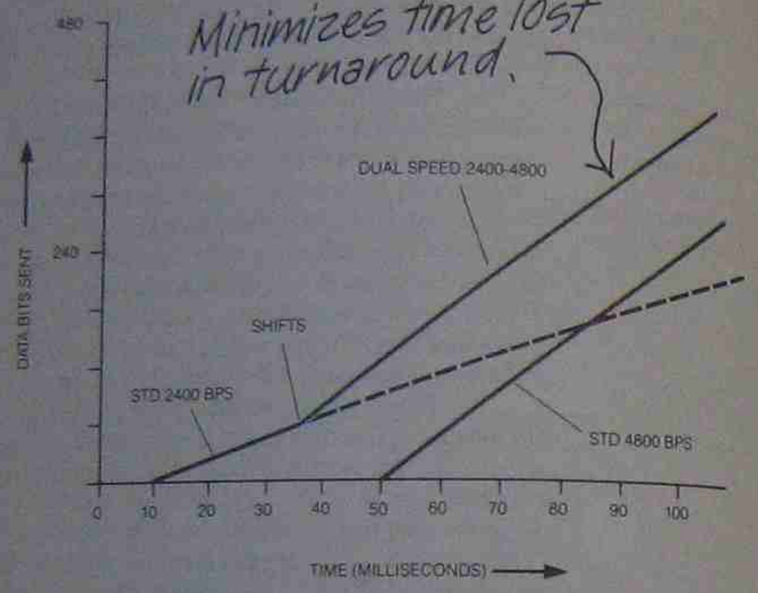
The received signal is modulated with an internal clock to produce an intermediate frequency. It is from this intermediate frequency that the final clock signal is produced.

At the receiver, the incoming signal from the line is modulated or frequency translated using an internal clock. The resulting intermediate frequency is processed to produce a clock signal at the rate at which the data are actually being received. This signal is applied as the reference to a phase locked loop oscillator. The output of this oscillator is a stable signal locked to the incoming line frequency in both phase and frequency.

Descrambler

The descrambler section of the receiver performs an operation that is the inverse of the scrambler described in the section above on the transmitter. If the data have been Gray coded by the transmitter prior to scrambling, they are converted back to straight binary, then applied to the DCE interface circuit.

Figure 5-7. Dual-Speed Modem Performance



Terminal Control Section

Each of the terminal control interface points requires different control and response signals.

The control section of synchronous modems must deal with two external interfaces: the telephone line at one end and the business machine at the other. If it is to be used on the public dial telephone network, the modem must: 1) sense the ringing signal, 2) provide line supervision (connecting and disconnecting the modem with the telephone line), and 3) in some cases it must be able to provide a busy indication to the incoming line without tying up telephone central office equipment.

At the other end, it must connect to the associated business machine or data terminal equipment (DTE). The DTE interface for most modems conforms to one of two standards: 1) the Electronic Industries Association (EIA) RS-232C, or 2) the International Telephone and Telegraph Consultative Committee (CCITT) Recommendation V.24. These standards were covered in a previous chapter.

Interface Circuits

Most synchronous modems support the set of interface circuits shown in *Table 5-3*. The backward channel circuits (CCITT circuit numbers 118 through 122) are required only in modems that provide a reverse or backward channel for control information in the opposite direction to normal data flow. The Data Terminal Ready circuit (CCITT circuit no. 108.2) is required only for modems used on the dial-up connections.

Table 5-3. Synchronous Modem DTE Interface Circuits

Circuit Number		
EIA	CCITT	Designation
AA	101	Equipment Ground
AB	102	Signal Ground
BA	103	Transmitted Data
BB	104	Received Data
CA	105	Request to Send
CB	106	Clear to Send
CC	107	Modem Ready
CD	108.2	Data Terminal Ready
CE	125	Ring Indicator
CF	109	Received Line Signal Detector (Carrier)
CG	110	Signal Quality Detector
CH	111	Data Signal Rate Detector (DTE Source)
CI	112	Data Signal Rate Detector (DCE Source)
DA	113	Transmitted Signal Element Timing (DTE Source)
DB	114	Transmitted Signal Element Timing (DCE Source)
DD	115	Received Signal Element Timing (DCE Source)
SBA	118	Secondary (Backward Channel) Transmitted Data
SBB	119	Secondary Received Data
SCA	120	Secondary Request to Send
SCB	121	Secondary Clear to Send
SCF	122	Secondary Received Line Signal Detected (Carrier)

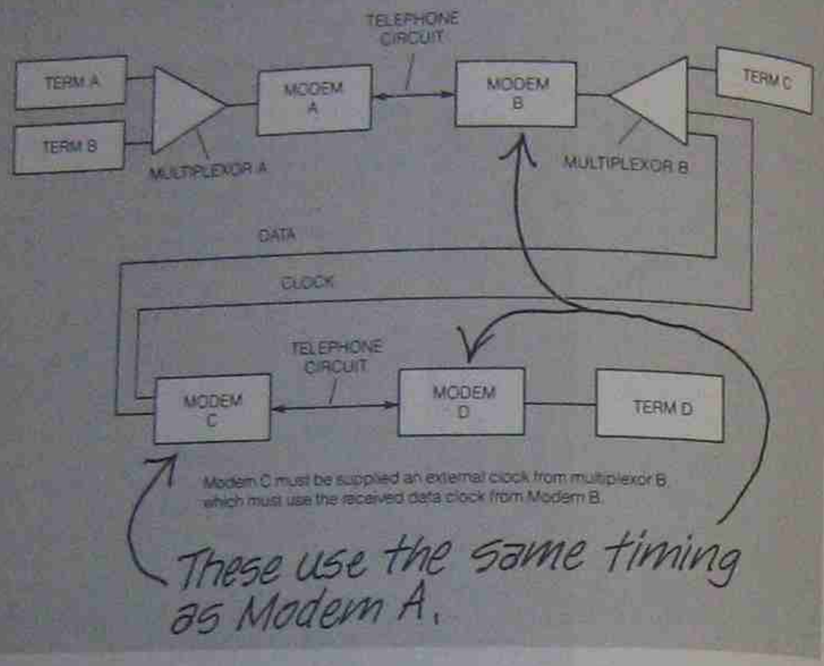
The Clear-to-Send signal is returned to the DTE in response to assertion of the Request-to-Send signal, but is delayed by the modem by the amount of time necessary for the modem to 'turn the line around'; that is, to send the required training pattern to the distant end.

External Timing

Systems which have terminals remote from the terminating modem on the main channel require that the remote modem be synchronized to the originating modem.

Standard 9600 bps modems usually contain internal multiplexors which allow more than one terminal to operate simultaneously over a single channel. In many cases, one or more channels are carried to terminals remote from the first modem terminating the circuit as shown in *Figure 5-8*. In such cases, which apply only to full-duplex operation, the clock signal for all of the modems in the path must be synchronized to only one of the modems. The external clock input (EIA circuit DA) for the extension modem circuit is fed from the receive clock from the primary modem (EIA circuit DD), causing the clocks in all of the modems to be slaved to the single modem at the originating end.

Figure 5-8.
Multiple Modem Circuit Requiring External Clocking



STANDARD MODEMS

Let's now examine the implementations of some currently available modems, and the CCITT recommendations to which synchronous modems used in Europe are designed. Four types of modems are considered:

1. 2400 bps half-duplex and full-duplex
2. 4800 bps half-duplex and full-duplex
3. 9600 bps half-duplex and full-duplex
4. 1200 bps full-duplex modems that operate over 2-wire dial-up circuits.

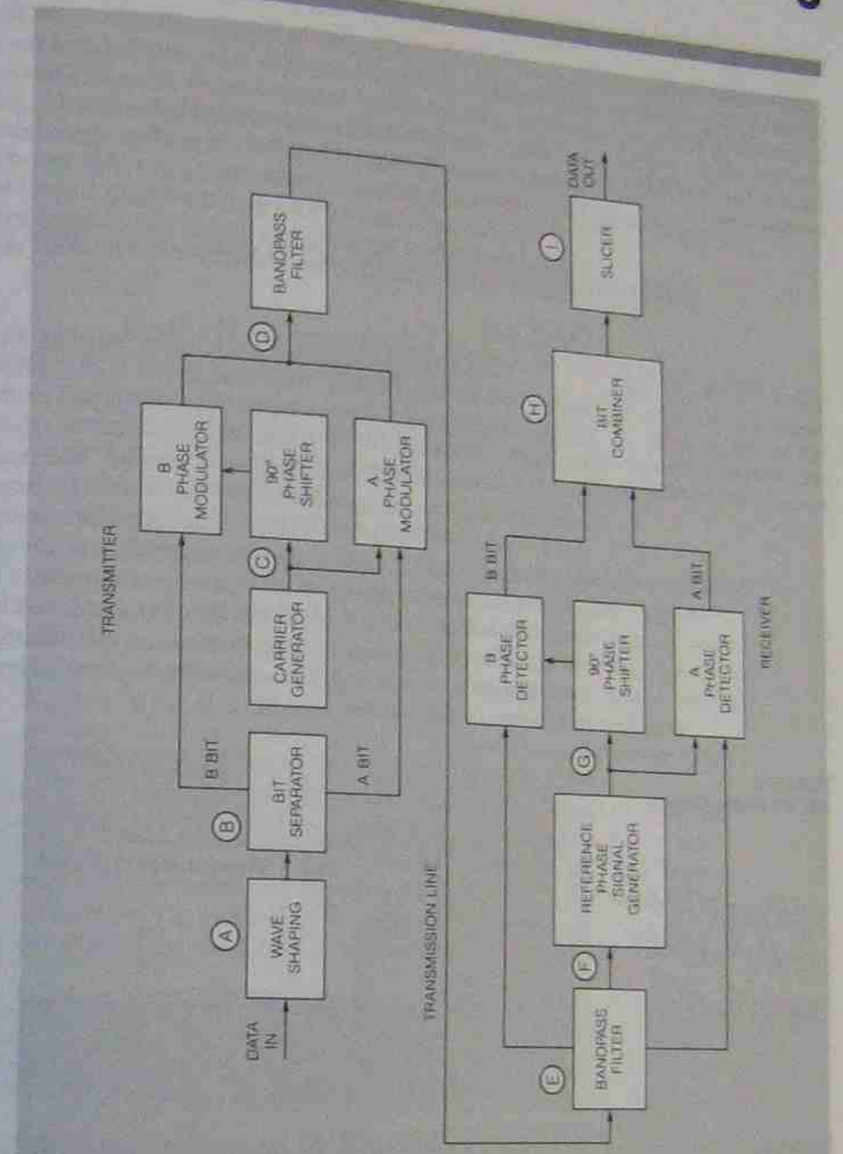
2400 bps Half-Duplex 2-Wire or Full-Duplex 4-Wire (DPSK)

Western Electric 201

The Western Electric 201 is representative of standard 2400 bps synchronous modems. This modem encodes two bits per baud using differential phase shift keying (DPSK).

A good example of the basic synchronous modem is the Western Electric (WE) 201 diagrammed in *Figure 5-9*. It operates at 1200 bauds with a carrier frequency of 1800 Hz. It uses a kind of phase shift modulation called the differential phase shift keying (DPSK) technique, and encodes two bits (called a dibit) into one signaling element (2 bits per baud). Phase shift keying encodes the binary values as changes of the phase of the carrier signal as discussed in a previous chapter. Differential phase shift means that the reference point from which the phase angle is measured at any signaling interval is the phase angle of the immediately previous interval. The WE 201 shifts the phase by multiples of 45 degrees according to *Table 5-4*.

Figure 5-9.
WE 201 Modem Block Diagram



A At the transmitter input, pulses are generated according to the data input, then filtered to shape them for modulation.
B Alternate bits in the stream become A bits and the others become B bits.
C The two bit streams phase modulate a sine wave carrier. The carrier for the B bit stream is phase shifted by 90 degrees with respect to the A carrier.
D The outputs of the modulators are combined and the bandpass filter eliminates unwanted signals before transmission.
E At the receiver, another bandpass filter removes noise and other unwanted signals.
F The input signal is used to synchronize a sine wave reference phase signal.
G The reference signal is applied directly to the A phase detector, but is phase shifted 90 degrees before application to the B phase detector.
H The A and B bit streams from the detectors are combined into a single bit stream.
I The slicer output consists of clean rectangular pulses that convey the same data as the transmitter input.

DPSK requires no absolute phase information to be transmitted. It has its own clock which is phase locked to the phase of the received signal.

With this technique, the signal is self-referenced; that is, no separate absolute phase information needs to be transmitted, and the most likely error (picking an adjacent phase value) will cause only a 1-bit error. The WE 201 contains a local clock which is phase locked to the detected phase of the incoming signal. Differences between the nominal phase values given in *Table 5-4* and the detected phase of the incoming signal referenced to the phase of the local clock are used to measure the quality of the incoming line signal. Since the phase of the received signal changes for each signaling interval, there is always sufficient energy at the 1200 baud rate to recover the baud clock.

CCITT V.26

The CCITT has published a recommendation for a standard 2400 bps modem with the designation, "Recommendation V.26 - 2400 bits per second modem standardized for use on leased telephone-type circuits." It specifies a full-duplex, 1200 baud DPSK modem with a modulation rate of up to 75 baud in each direction on an optional "backward channel." The recommendations are generally similar to the specifications for the WE 201 (in large part because the WE 201 had by 1968 become a de facto standard). One interesting difference is that the assignment of dibits to phase change values is allowed alternative sets of values. Alternative A is given in *Table 5-5* and alternative B is the same as that given in *Table 5-4* for the WE 201. Use of alternative A leads to the problem of no phase change at all for long strings of 00 dibits; thus, the energy content in the transmitted stream at the 1200 baud rate is greatly reduced, making baud clock recovery uncertain.

The CCITT V.26 2400 bps modem is similar to the WE 201, but one significant difference is the alternate set of phase codes.

Table 5-4.
WE 201 Phase Coding

Dibit	Phase Change in Degrees
00	45
10	135
11	225
01	315

Alternative B is the same as WE 201.

Table 5-5.
V.26 Alternative A Phase Coding

Dibit	Phase Change in Degrees
00	0
01	90
11	180
10	270

Recommendation V.26 specifies the form of the synchronizing signal between two modems to be continuous transmission of dibit 11 (giving continuous 180 degree phase shifts) during the RTS-CTS sequence. It also specifies incoming signal levels of -26 dBm or higher for the modem to assert the Carrier Detect (Circuit 109) and -31 dBm or lower for it to negate Carrier Detect. It is recommended that the modem be designed so that the operator cannot control the send level or receive sensitivity.

4800 bps Half-Duplex 2-Wire or Full-Duplex 4-Wire Modems (DPSK)

CCITT V.27

The CCITT Recommendation V.27 specifies a 4800 bps modem for use on leased telephone-type circuits. The modem:

1. Can operate in either half or full-duplex.
2. Uses differential phase shift keying with eight phases.
3. Provides an optional 75 bps backward channel in both directions for supervisory signaling.
4. Has a manually adjustable equalizer.

Although the manual equalizers mentioned in CCITT V.27 are no longer used, most of the standard is still used as the basis for many modems today, particularly in Europe.

This recommendation was promulgated in 1972, and the state of the modem art has progressed far beyond the manually adjustable equalizer. However, the remainder of the recommendation is the basis for many current modems, particularly for the European market. The carrier frequency is 1800 Hz, the modulation rate is 1600 baud, and the basic phase shift interval is 45 degrees. To signal at 4800 bps with 1600 baud requires that each baud represent three bits (tribits). The tribit to phase-change amount for V.27 is given in *Table 5-6*.

The synchronization pattern for the V.27 modem is comprised of continuous ones in tribits for at least 9 ms to train the demodulator, plus continuous ones at the input to the transmit scrambler until the CTS lead (circuit 106) is asserted by the transmitting modem to synchronize the scramblers. The general operation of a scrambler has been described earlier in the chapter, but the details of the V.27 scrambler follow.

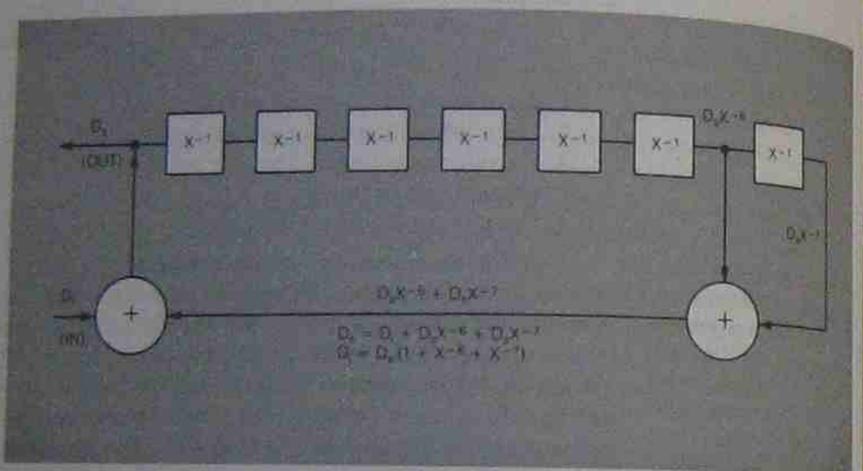
Table 5-6.
V.27 Phase Coding

Tribit Value	Phase Change in Degrees
001	0
000	45
010	90
011	135
111	180
110	225
100	270
101	315

The scrambler implements dividing the message polynominal by a generating polynominal by means of a feedback shift register.

The transmitter scrambler divides the message polynomial, of which the input data bits are the coefficients in descending order, by a generating polynomial, which for V.27 is equal to $1 + X^{-6} + X^{-7}$. This division is continuously performed as each bit enters the scrambler and the result taken as the transmitted data pattern. The transmitted pattern is then searched continuously over 45 bits for repeating patterns of 1, 2, 3, 4, 6, 9 and 12 bits, which are eliminated. The result is sent as the scrambled data pattern. *Figure 5-10* shows a feedback shift register which implements the polynomial division process.

Figure 5-10.
CCITT V.27 Scrambler



Western Electric 208

The WE 208 modem is another 4800 bps modem which is widely used in the U.S. It has been supplied by AT&T and the Bell Operating Companies when they were part of the Bell System. It is a DPSK unit, using eight phase angles to represent eight different groups of three bits. The coding pattern (after Gray coding the binary input) for the WE 208 is given in *Table 5-7*.

Table 5-7.
WE 208 Phase Coding

Tribit	Relative Phrase in Degrees
001	22.5
000	67.5
010	112.5
011	157.5
111	202.5
110	247.5
100	292.5
101	337.5

The WE 208 uses DPSK to encode tribits to transmit 4800 bps at 1600 bauds. Its design and operation are similar to the typical standard synchronous modem.

The carrier frequency for the WE 208 is 1800 Hz and the baud rate is 1600 baud. The WE 208 is not equipped with an internal multiplexor because the 1600 baud rate of the modem is not a standard data rate for any other modem or business machine. The design of the WE 208 is as covered in the section above on typical synchronous modem components, including the scrambler implementation as four 7-bit feedback shift registers to randomize the first bit of each tribit.

The demodulator in the WE 208 multiplies the carrier frequency by eight, then measures the interval between the 1600 baud clock transition and the first positive value of the current carrier cycle. This measurement is done digitally, and successive values of the count are compared to determine the phase shift which occured during the previous signaling interval. The measured phase shift, with any errors that may be present due to line noise, demodulator errors, baud clock errors, etc is rounded to the nearest permitted value (22.5 degrees, 67.5 degrees, etc) and the transmitted tribit sequence is extracted and decoded from Gray code to binary. An error feedback signal is developed as the difference between the actual phase of the signal, as demodulated, and the nominal value. This error signal is monitored and used to cause the modem to retrain if it becomes too large.

The WE 208 modem normally operates with a basic retraining time of 50 ms, but it can be changed with internal strapping (jumper wires) to 150 ms for operation on long circuits (over 2000 miles). Even this delay probably is still too short for operation on satellite links, due to the very long (about 650 ms) round trip propagation delay.

9600 bps Half-Duplex 2-Wire or Half/Full-Duplex 4-Wire

Modulation Types

The data rate of the standard 9600 bps modem is increased by increasing its baud rate to 2400 bauds and increasing the number of bits per baud to four.

In order to increase the data rate of a synchronous modem over 4800 bps, it is necessary to either increase the baud rate, or increase the number of bits per baud. Standard 9600 bps modems do both. The baud rate is increased to 2400 bauds, and the input data are taken four bits at a time (quadbits) for controlling the output of the modulator. Two modulation schemes are commonly used to reach these extreme data rates: vestigial sideband modulation (VSB), and quaternary amplitude modulation (QAM).

The choice between modems using one or the other of the two modulation schemes is less a matter of performance than of economics. Some VSB modems require little or no line conditioning, whereas a few QAM unit manufacturers generally recommend at least C2 conditioned circuits. The VSB units are less complex internally, which usually allows a price advantage over QAM units.

The VSB transmitter modulator takes two bits at a time and constructs a four-level signal which modulates the carrier to transmit 9600 bps.

Vestigial Sideband Modulation

Since it is possible to transmit a double-sideband frequency-modulated (FM) signal (with a WE 202 type modem) at 1800 bps, one would think that it should be possible to send at 3600 bps using only one sideband since the entire bandwidth would be available for the one sideband. However, this is not possible with an FM signal (the signal can't be recovered), but it is possible to use an amplitude-modulated (AM) single-sideband signal to transmit at 4800 bauds. Four different carrier amplitudes are used to represent the four different values of bits taken two at a time, thus, each baud represents two bits which gives an effective rate of 9600 bps.

A VSB transmitter has a scrambler and Gray coder similar to those of 4800 bps modems. The modulator, however, takes the bits two at a time to form a four-level signal which modulates a 2853 Hz carrier. The carrier and upper sideband are suppressed, and a 2853 Hz pilot signal is added back to what remains, but 90 degrees out of phase. A signal that is 90 degrees out of phase is called a quadrature signal because 'in quadrature' means at right angles to the original carrier. The quadrature signal does not interfere with the modulated single-sideband signal, but it is necessary for the receiver to recover the exact frequency and phase of the original carrier so it can recover the data.

The receiver's local 2853 Hz oscillator is locked to the phase and frequency of the pilot signal recovered from the received signal. This adjusted clock is used to demodulate the received single-sideband signal into the four-level 4800-baud baseband signal. The signal is processed by a 60-tap adaptive equalizer to remove intersymbol interference, then put through a decision circuit to recover the dibits, which are then unscrambled.

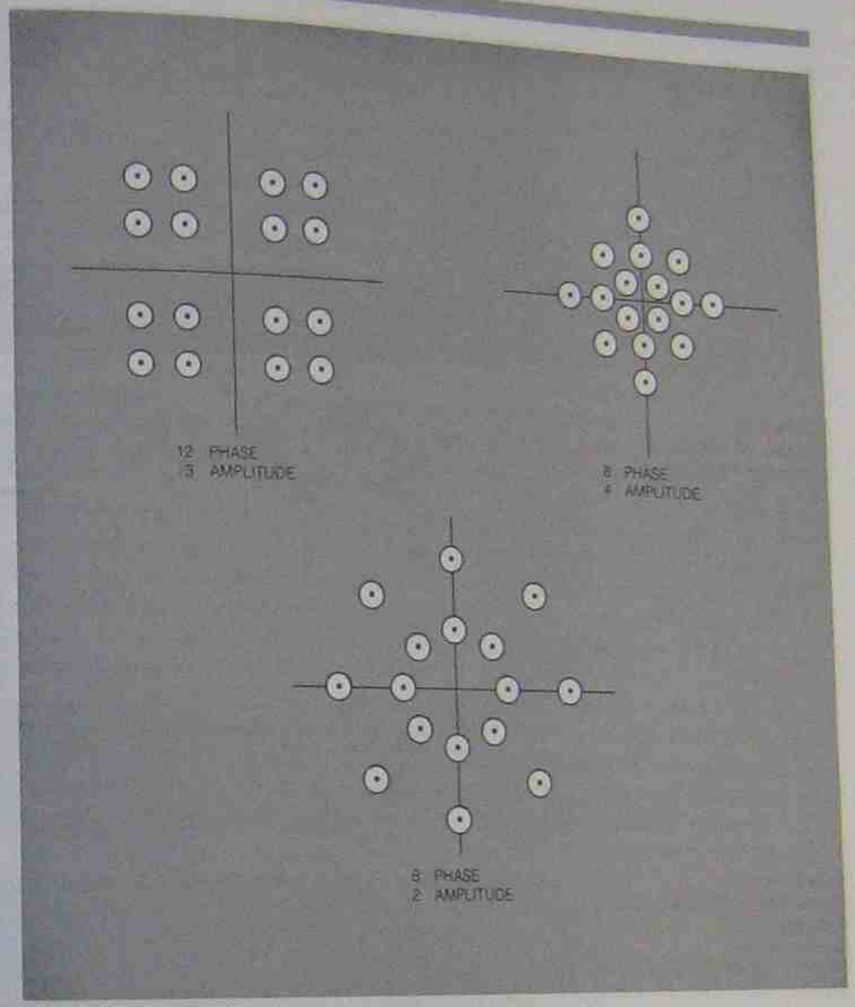
Training a 60-tap equalizer is a lengthly process which reduces the cost effectiveness of using 9600 bps modems on half-duplex circuits. A four-segment training sequence must be sent to: 1) set the automatic gain control circuits, 2) acquire the phase of the carrier tone, 3) set the equalizer tap coefficients, and 4) synchronize the scrambler. In the CCITT Recommendation V.29 for a 9600 bps modem, the total four-segment sequence takes about 250 ms.

Quadrature Amplitude Modulation

Modems using quadrature amplitude modulation achieve a data rate of 9600 bps by encoding four bits per baud, and by using a baud rate of 2400.

QAM modems use some combination of differential phase shifts and amplitudes totaling sixteen states to encode four bits per baud. This combined with a baud rate of 2400 gives the desired 9600 bps. The CCITT Recommendation V.29 specifies eight phases with two possible signal amplitudes at each phase change. The WE 209 modem uses twelve phases and three amplitudes, but some combinations of phase and amplitude are not allowed. Constellation patterns are shown in *Figure 5-11*. The choice of the number of phases and level spacing represents the modem manufacturer's best guess at the type and severity of various transmission line errors the modem is likely to experience.

Figure 5-11. QAM Modem Constellation Patterns



CCITT Recommendation V.29 9600 bps Modem

The CCITT V.29 modem's maximum data rate is 9600 bps, but has automatic fallback to 7200 and 4800 bps if the error rate is too high.

The carrier frequency for the CCITT V.29 modem is 1700 Hz and it operates at 2400 bauds. This modem is specified to operate at 9600 bps with fallback rates of 7200 and 4800 bps. It has an automatic adaptive equalizer, and may optionally contain a multiplexor for carrying up to four separate data streams whose aggregate (combined) rate does not exceed 9600 bps. The combined data stream to be transmitted is scrambled, then divided into sets of four bits (quadbits). The first bit (in time) of each quadbit determines the amplitude of the transmitted signal as illustrated in *Figure 5-12*, and the remaining three bits determine the phase shift to be applied as shown in *Figure 5-13*. The resultant "constellation pattern" resembles *Figure 5-14*.

At the fallback rate of 7200 bps, only three bits per baud are encoded and there is no amplitude change. At the 4800 bps fallback rate, two bits per baud are encoded and they specify only four phase shifts.

Figure 5-12.
Bit Q1 Determines Amplitude
(Source: CCITT Recommendation V.29, CCITT, Geneva, CH.)

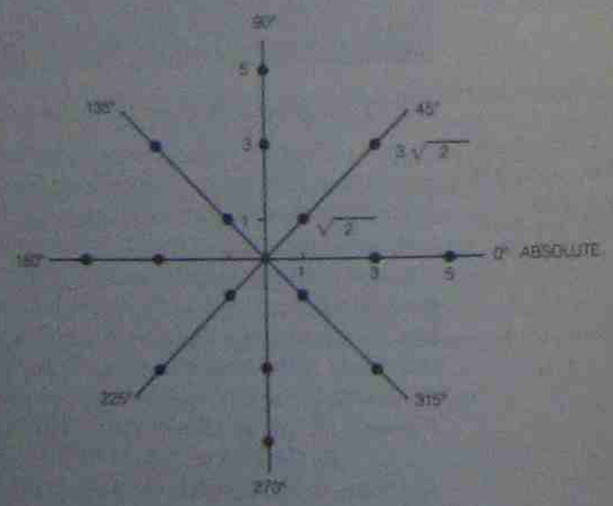
Absolute Phase	Q1	Relative Signal Element Amplitude
0°, 90°, 180°, 270°	0	3
	1	5
45°, 135°, 225°, 315°	0	$\sqrt{2}$
	1	$3\sqrt{2}$

Figure 5-13.
Bits Q2, Q3 and Q4 Determine Phase Shift
(Source: CCITT Recommendation V.29, CCITT, Geneva, CH.)

Q2	Q3	Q4	Phase Change*
0	0	1	0°
0	0	0	45°
0	1	0	90°
0	1	1	135°
1	1	1	180°
1	1	0	225°
1	0	0	270°
1	0	1	315°

*Note — The phase change is the actual on-line phase shift in the transition region from the center of one signaling element to the center of the following signaling element.

Figure 5-14.
Signal Space Diagram at 9600 bps
(Source: CCITT Recommendation V.29, CCITT, Geneva, CH.)



Except for the WE 209 type modem, most 9.6 Kbps devices manufactured throughout the world adhere to the V.29 standard. This makes a V.29 modem popular for use on international leased lines because the selection of the modem to be placed on either end of the line may be restricted to one or a few models due to some countries' government restrictions. By using V.29, modems' compatibility among different vendor products is assured, which enables data transmission compatibility on international leased lines.

CCITT V.32

A recent CCITT recommendation that warrants attention is the V.32 standard. This standard is based on a modified quadrature amplitude modulation technique and is designed to permit full-duplex 9.6 Kbps transmission over the switched telephone network. The key to the operation of a V.32 modem is a built-in echo canceling technique that enables transmitted and received signals to occupy the same bandwidth.

When the V.32 modem connects to another V.32 modem, two high speed channels in the opposite direction of one another are established as illustrated in *Figure 5-15*. Each of these channels occupies roughly the same bandwidth as the other channel. Intelligence built into the receiver in each modem permits the modem to cancel the effects of its transmitted signal, which enables each device to distinguish its sending signal from the signal being received.

Early V.32 modems required sophisticated circuitry to carry out echo canceling. The circuitry made the cost of V.32 modems two to three times the cost of V.29 modems, which provide the same data transmission rate on leased lines. Because it is reasonable to expect firms manufacturing semiconductors to develop chip sets to implement echo canceling, it also is reasonable to expect the cost of V.32 modems to decrease. When the price drops, V.32 modems are expected to increase in popularity.

Higher Speed QAM Modems

To increase the data rate of modems beyond 9.6 Kbps, several vendors developed modulations schemes whereby data was collected 2,400 times per second into a 6-bit symbol. This technique resulted in the use of a 64-point signal constellation to obtain a 14.4 Kbps data transmission rate. The resulting constellation pattern was four times as dense as the V.29 constellation pattern *(Figure 5-14)*, so the modem had high susceptibility to phase jitter and noise.

When phase jitter or noise occur, the signal points in the modem's constellation pattern are shifted or rotated during transmission from their proper location. The receiving modem, which is designed to select the nearest valid signal point, misinterprets a shifted or rotated signal point, which results in a transmission error. Because of the susceptibility of conventional 14.4 Kbps QAM modems to transmission impairments, their throughput in many instances is lower than that obtainable with 9.6 Kbps devices.

Figure 5-16 shows the proportion of noise to signal level increasing on a line when 9.6 Kbps and 14.4 Kbps QAM modems are used.

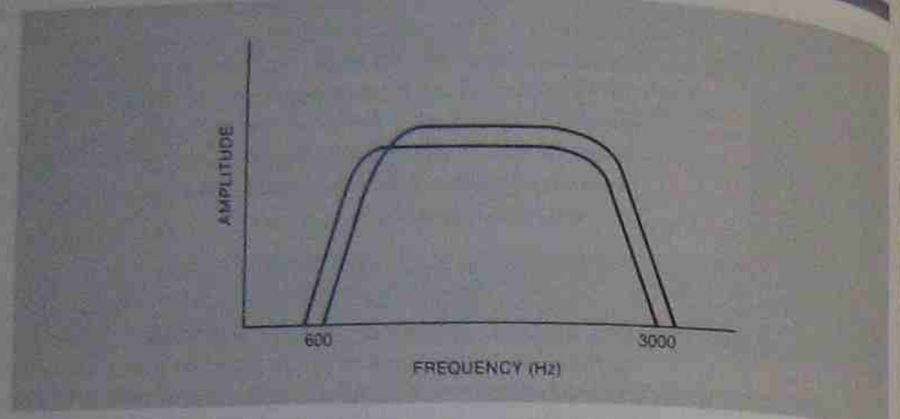


Figure 5-15.
CCITT V.32 Channel Derivation

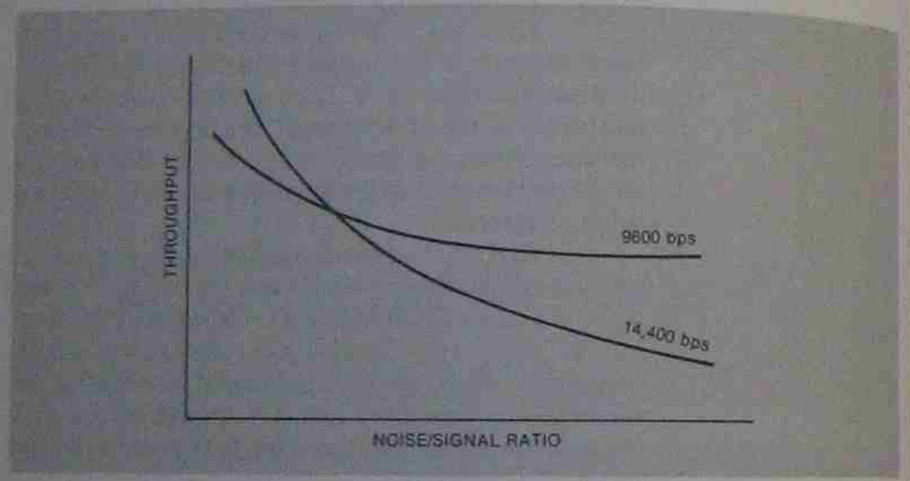


Figure 5-16.
Modem Throughput Variance

Because the constellation pattern of a 14.4 Kbps modem is four times the density of a 9.6 Kbps modem, 14.4 Kbps modems are more susceptible to noise, which results in their throughput at a certain noise to signal ratio decreasing under that obtainable with a 9.6 Kbps device. The reason the throughput declines is that the erroneous received signal points cause the transmission protocol of devices using the modems to attempt to correct the error by retransmission. The resulting retransmissions lower the modem's effective data throughput rate.

CCITT V.33

Conventional QAM modems' sensitivity to transmission impairments makes them unsuitable for use at data rates above 9.6 Kbps on noisy circuits. Many long distance circuits have a fair degree of noise, so a mechanism was required to provide to the modulation method some immunity from the noise on the line. The application of Trellis Coding to QAM provided a significant degree of immunity, which resulted in the development of a series of modems that based operations on Trellis Coded Modulation (TCM).

In the TCM process, an encoder is employed to add a redundant bit to each group of bits used to generate a signal change. In a 14.4 Kbps TCM modem, the resulting signal point is selected from a 128-point signal constellation pattern. The extra bit introduced by the encoder results in only certain sequences of signal points being valid. Thus, if a transmission impairment occurs that causes a signal point to be shifted, the receiving modem knows that certain points are invalid and compares only the observed point to valid points by selecting the valid signal point closest to the observed signal. By using TCM, the modem becomes only half as susceptible to noise as a conventional QAM modem, which permits higher data rates without decreasing the throughput of the modem that can occur because of the standard magnitude of noise commonly encountered on leased lines.

The V.33 modem is a 14.4 Kbps device that uses TCM, is designed for operation on leased lines, and has a fallback rate of 9.6 Kbps. Although the CCITT has yet to promulgate recommendations covering higher speed modems, several vendors have introduced 19.2 Kbps TCM-based devices. In 1986, Codex Corporation became the first to introduce an "Eight-State Trellis Coded Modulation" modem, but Paradyne, NEC, and Fujitsu followed shortly with their own versions. Although the initial cost of 19.2 Kbps modems exceeded $10,000 per unit, by mid-1988, several vendors were offering them for under $2,000, a price that encourages their widespread use.

DIGITAL TRANSMISSION

It is advantageous to transmit data in binary form, rather than converting it to analog form. However, digital transmission must be restricted to channels with greater bandwidth than the voice band.

In transmission systems which are not restricted by the bandwidth limitations of the voice-grade channel, it is advantageous to transmit the binary information from the source as binary data without conversion to analog form. The reasons for this were explained in Chapter 3. For end-user applications, the devices that connect between the business machine (DTE) and the channel usually are called limited distance adapters, wireline modems or channel interface units.

The most obvious difference between voice-band modems and direct digital transmission units is the increased bandwidth required by the latter. Recall from Chapter 3 that this increase is about a factor of eight. Thus, digital transmission is restricted to channels with bandwidths greater than the 3 kHz voice band.

Line Coding

Unipolar Code

The NRZ code works well for the shielded and short travel paths within a machine, but is not suited for long distance use because residual dc shifts the "zero" level.

The waveform of binary signals normally used in computers and terminals is called unipolar; that is, the voltage representing the bits varies between 0V and +5V as *Figure 5-17a* shows. (The code shown is called NRX, Non-Return to Zero, because the voltage does not return to zero between adjacent 1 bits.) This representation works well inside machines where the transmission paths are short and well shielded, but it is unsuitable for long paths because of the presence of residual DC levels and the potential absence of enough signal transitions to allow reliable recovery

of a clocking signal. Signal conditioning devices (the aforementioned channel interface units) are used to convert the unipolar waveform to one of several different patterns that meet the goals of no long-term DC residual in the signal and strong timing content. Several of these signal patterns are shown in *Figure 5-17b* through *5-17e*.

The bipolar (alternate mark inversion) coding scheme eliminates residual dc and has zero power in the spectrum at zero frequency. The polarity of every other 1 bit is inverted.

Polar NRZ Code

The simplest pattern that eliminates some of the residual DC problem is called a polar NRZ line code. It is shown in *Figure 5-17b*. This coding merely shifts the signal reference level to the midpoint of the signal amplitude. This has the effect of reducing the power required to transmit the signal by one-half compared with unipolar but has the disadvantage of having most of the energy in the signal concentrated around zero frequency, as illustrated in *Figure 5-18*.

A more satisfactory coding scheme is that used in the so-called T1 digital transmission system. This scheme, called bipolar or alternate mark inversion, is illustrated in *Figure 5-17c*. This format has no residual DC component and has zero power in the spectrum at zero frequency, as *Figure 5-18* shows. It achieves these goals by transmitting pulses with a 50% duty cycle (only half as wide as the pulse interval allows) and by inverting the polarity of alternate 1 bits that are transmitted. Note that the bipolar format is really a three-state signal (+V, 0V, −V). This increase in the code space of the signal adds redundancy without increasing the bandwidth of the signal and makes performance monitoring easier. Assuring that long strings of zero bits do not occur (with an accompanying lack of timing information) is left to the transmitting terminal or channel interface.

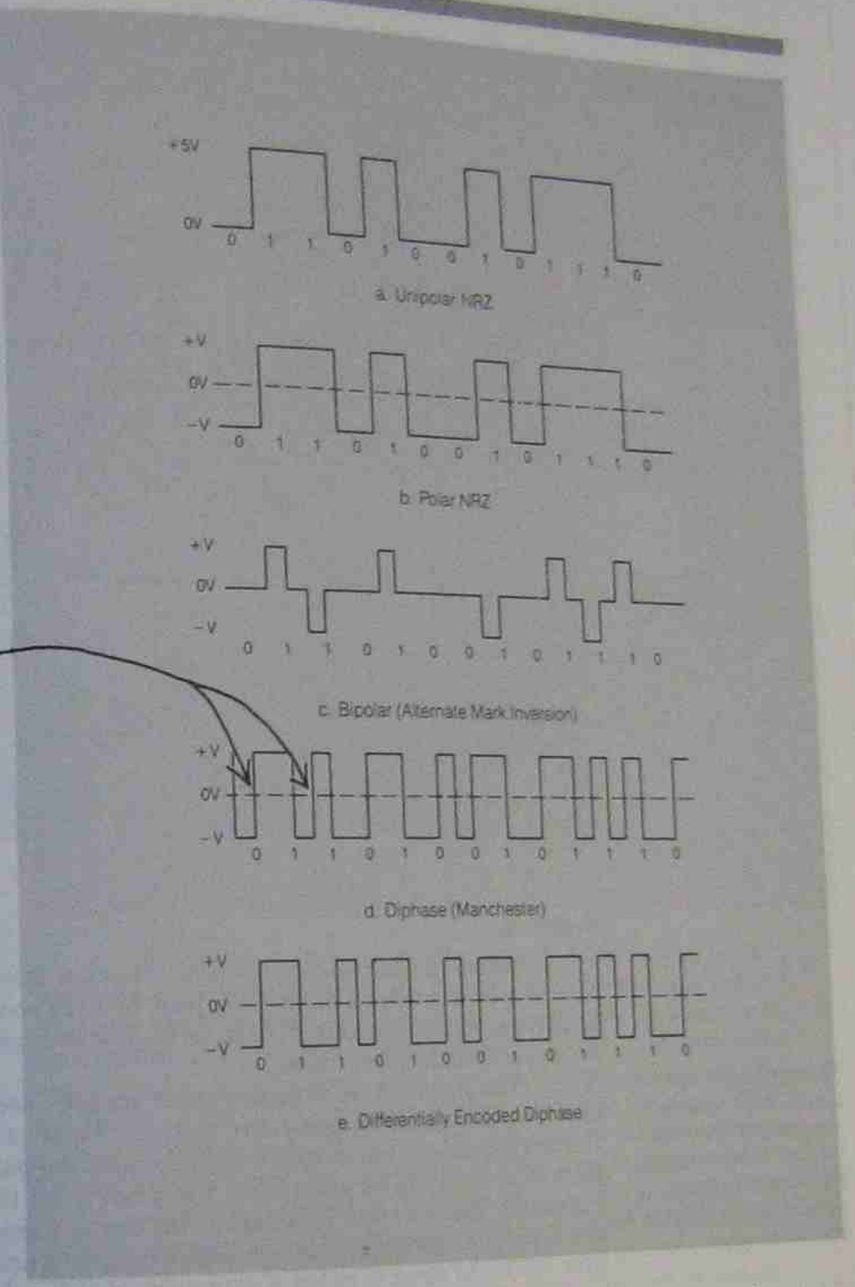
In diphase coding, strong timing information is provided by having a transition for every bit, and residual dc is eliminated by having both polarities for every bit.

Diphase Code

Another technique that has received considerable acceptance is called digital biphase, diphase, or Manchester coding. It is illustrated in *Figure 5-17d*. The diphase code provides strong timing information by supplying a transition for every bit, whether it be a one or a zero. If the diphase signal is differentially encoded as *Figure 5-17e* shows, it is even possible to determine the absolute phase of the signal because only a zero has a transition at the beginning of an interval. It also eliminates the residual DC problem by providing both a positive and negative polarity for every bit. However, all these desirable characteristics are gained at the expense of requiring twice the bandwidth of a bipolar signal. Also, for differentially encoded diphase, the error rate is doubled. The diphase scheme is used on the Ethernet[2] local area network developed by the Xerox Corporation, Digital Equipment Corporation, and Intel Corporation.

[2]Ethernet is a trademark of Xerox Corporation.

Figure 5-17. Line Coding of Binary Pulses



The block diagram of a circuit to receive the bipolar signal and decode it into straight binary is shown in *Figure 5-18*. Note that the alternate mark inversion (AMI) pulse format is rectified into polar pulses inside the receiver where suppressing the DC component of the signal is no longer critical.

ST. GEORGE TECHNICAL COLLEGE LIBRARY

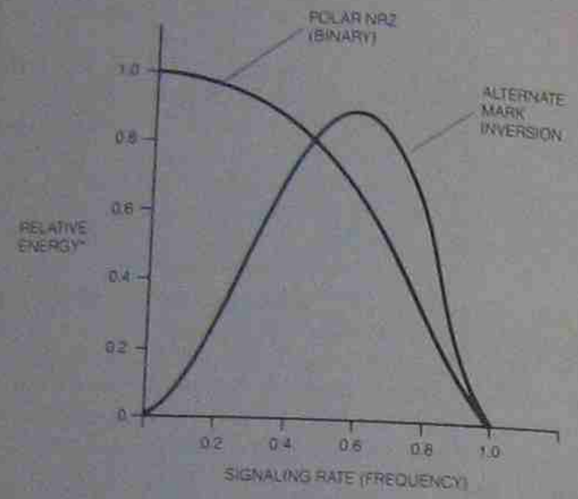


Figure 5-18.
Energy Distribution of Encoded Pulses

Repeaters

Because it is not possible to separate the noise from an analog signal once the two are mixed on a transmission path, both the noise and the desired signal are amplified in repeaters. The signal to noise ratio gets progressively worse as the path length increases.

The regenerative repeater used in digital transmission doesn't just amplify the input signal, it actually regenerates the binary signal so it is just like the original signal.

This is not true when digital signals are used because a different kind of repeater, called a regenerative repeater *(Figure 5-20)*, can be used. This repeater doesn't just amplify but regenerates the pulses to restore the shape of the binary signal exactly as it was when it left the originating transmitter. This is a major advantage of the digital transmission technique as it is possible in principle, and to a large extent in practice, to reduce the error rate of the signal to as low a value as desired merely by putting the repeaters closer together.

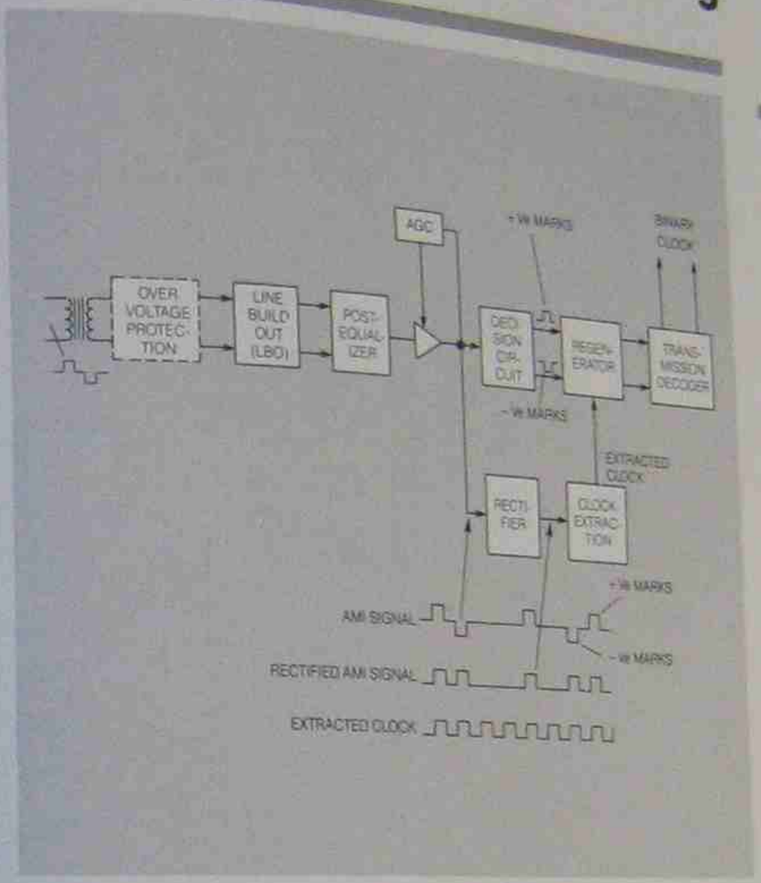


Figure 5-19.
Bipolar Decoder
(Source: Frank F. E. Owen, PCM and Digital Transmission Systems, © 1982 McGraw-Hill, Inc., used by permission from McGraw-Hill, Inc.)

The regenerative repeater has phase and frequency correction circuits to compensate for non-linearities in the transmission path.

The regenerative repeater, like the sending line terminal circuit, has some of the same functions as a modem. For example, it does certain electrical transformations on the input waveform to make it suitable for the transmission path. Both the sending and receiving ends of these operations can be studied by referring to *Figure 5-20*. On the input end, the line build-out (LBO) circuit matches the post-equalizer circuit with the length of the transmission cable on the input end. The post-equalizer circuit compensates for the frequency and phase nonlinearities of a standard length of input circuit, which is represented by the output of the LBO. The equalized input signal is amplified by an amplifier controlled by an automatic gain control (AGC) circuit, then the stabilized and normalized input goes to a clock extraction circuit and a decision circuit/regenerator. The clock extraction circuit, which is driven by the received data transitions, clocks the regenerator so that the original signal is regenerated exactly.

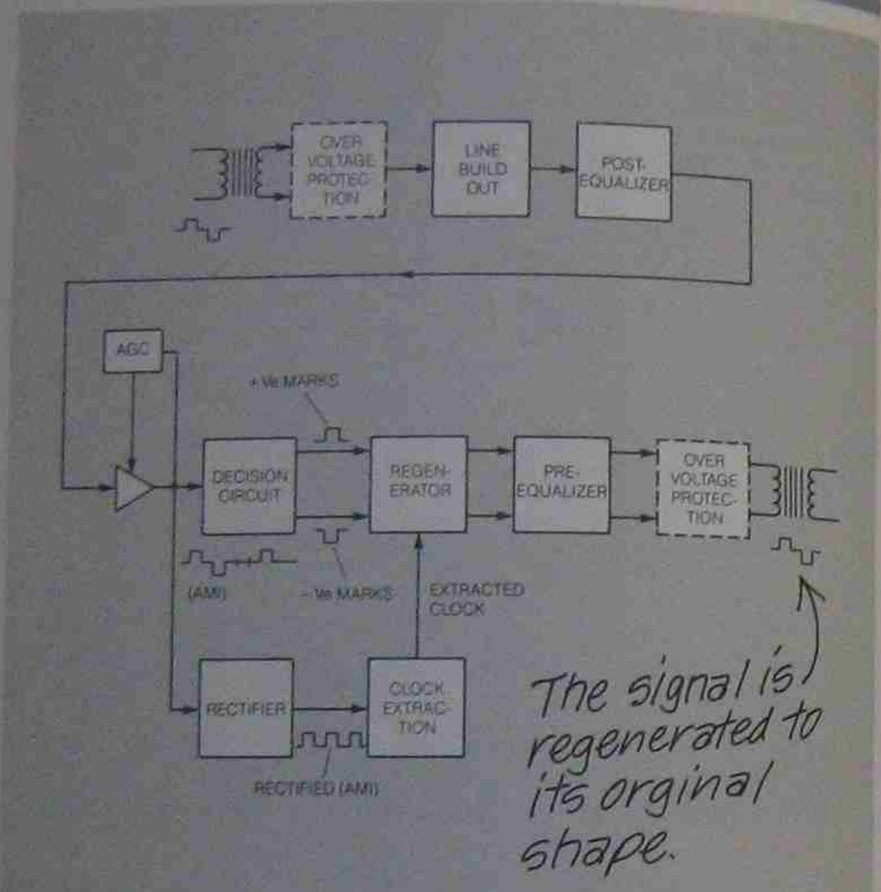


Figure 5-20. Regenerative Repeater *(Source: Frank E. Owen, PCM and Digital Transmission Systems, © 1982, McGraw-Hill, Inc., used by permission from McGraw-Hill, Inc.*

WHAT HAVE WE LEARNED?

1. The bandwidth of a channel sets a limit on the modulation or baud rate of the channel.
2. The data rate through a channel can be increased by sending more than one bit per baud.
3. Synchronous modems modulate a carrier wave in phase or amplitude or both to send more than one bit per baud.
4. Synchronous modems send the clock signal along with the data, and are more expensive than asynchronous modems because of the additional circuitry necessary to recover the clock from the data.
5. Usually only phase modulation is used in 2400 bps modems, but both phase and amplitude modulation is used in 4800 and 9600 bps modems.
6. The most important transmission impairment that synchronous modems must deal with is differential delay distortion. Such modems have adaptive equalizers to correct for this distortion.

7. Accurate recovery of the signal in both synchronous modems and digital transmission of binary signals depends on having a sufficient number of one bits transmitted.
8. Digital transmission uses regenerative repeaters instead of amplifiers to correct the line signal.

Quiz for Chapter 5

1. What is one principal difference between synchronous and asynchronous transmission?
   a. The bandwidth required is different.
   b. The pulse heights are different.
   c. The clocking is mixed with the data in asynchronous transmission.
   d. The clocking is derived from the data in synchronous transmission.
2. Synchronous modems cost more than asynchronous modems because:
   a. They are larger.
   b. They must contain clock recovery circuits.
   c. The production volume is larger.
   d. They must operate on a larger bandwidth.
3. The scrambler in a synchronous modem is in the:
   a. Control section.
   b. Receiver section.
   c. Transmitter section.
   d. Terminal section.
4. Binary codes are sometimes transformed in modems into:
   a. Hexadecimal.
   b. Huffman codes.
   c. Gray code.
   d. Complementary codes.
5. The digital-to-analog converter in a synchronous modem sends signals to the:
   a. Modulator.
   b. Transmission line.
   c. Terminal.
   d. Equalizer.
6. The receive equalizer in a synchronous modem is called:
   a. A compromise equalizer.
   b. A statistical equalizer.
   c. An adaptive equalizer.
   d. An impairment equalizer.
7. The receive equalizer reduces delay distortion using a:
   a. Tapped delay line.
   b. Difference engine.
   c. Descrambler.
   d. "Gearshift".
8. A Western Electric 201 modem operates with a carrier frequency of:
   a. 1000 Hz.
   b. 1200 Hz.
   c. 1800 Hz.
   d. 600 bauds.
9. The CCITT V.26 modem has a modulation rate of:
   a. 1200 Hz.
   b. 1200 bauds.
   c. 1560 cps.
   d. None of the above.
10. The transmission signal coding method for Tl carrier is called:
    a. Binary.
    b. NRZ.
    c. Bipolar.
    d. Manchester.

Fiber Optic and Satellite Communications

ABOUT THIS CHAPTER

This chapter is about transmission of data using light as a carrier. We will learn some of the history of the development of this form of communications, and why the technique has generated so much excitement in the communications world. We will learn why fiber optic systems work, and some of the terms used in describing their operation. Some current terrestrial and undersea fiber optic systems will be described. We will learn about the application of fiber optics to local area data networks. Finally, we will look at some of the techniques and considerations in transmitting data via geostationary satellites.

INTRODUCTION AND HISTORICAL PERSPECTIVE

Alexander Graham Bell was a very curious and inventive man. In 1880, four years after the invention of the telephone, he patented an "Apparatus for signaling and communicating, called Photophone." This device, illustrated in *Figure 6-1*, transmitted a voice signal over a distance of 200 meters using a beam of sunlight as the carrier. As the speech into the speaking trumpet vibrated the mirror, the light energy reflected onto the photovoltaic cell was varied. The electric current produced by the cell varied in accordance with the varying light energy.

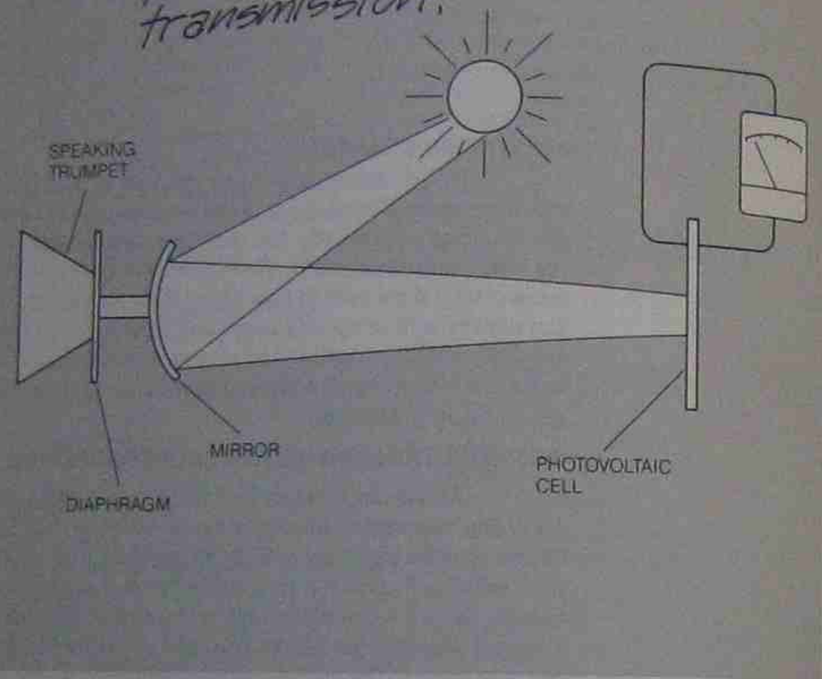
The device demonstrated the basic principle of optical communications as it is practiced today. The two requirements for commercial success, however, were almost a hundred years away. These were a powerful and reliable light source, and a reliable and low-cost medium for transmission.

The use of light as a commercially feasible communications system had to wait until the development of the laser as a powerful, reliable light source, and the clad glass fiber as an inexpensive, reliable medium.

In 1960 the laser was recognized as the long-sought light source, and systems were tried using both the atmosphere and beam waveguides as the transmission medium. The application of a glass fiber with a cladding was proposed in 1966, and by 1970 fibers with losses of only 20 decibels per kilometer (dB/km) were demonstrated. Since then, progress in the invention and application of fiber optics has been startling. Fibers with losses of less than 0.2 dB/km have been demonstrated in the laboratory (in 1979), as have systems which can transmit at data rates in excess of 400 million bits per second (Mbps) for distances in excess of 100 km without repeaters or amplifiers. Advances in fiber optics threaten to obsolete satellite systems for some kinds of communications (point-to-point where large bandwidths are required, such as transoceanic telephone systems) only ten years after the satellite systems were commercially employed as the communications systems of the future.

Figure 6-1.
Alexander Bell's Photophone

Demonstrates basic principles of optical transmission.



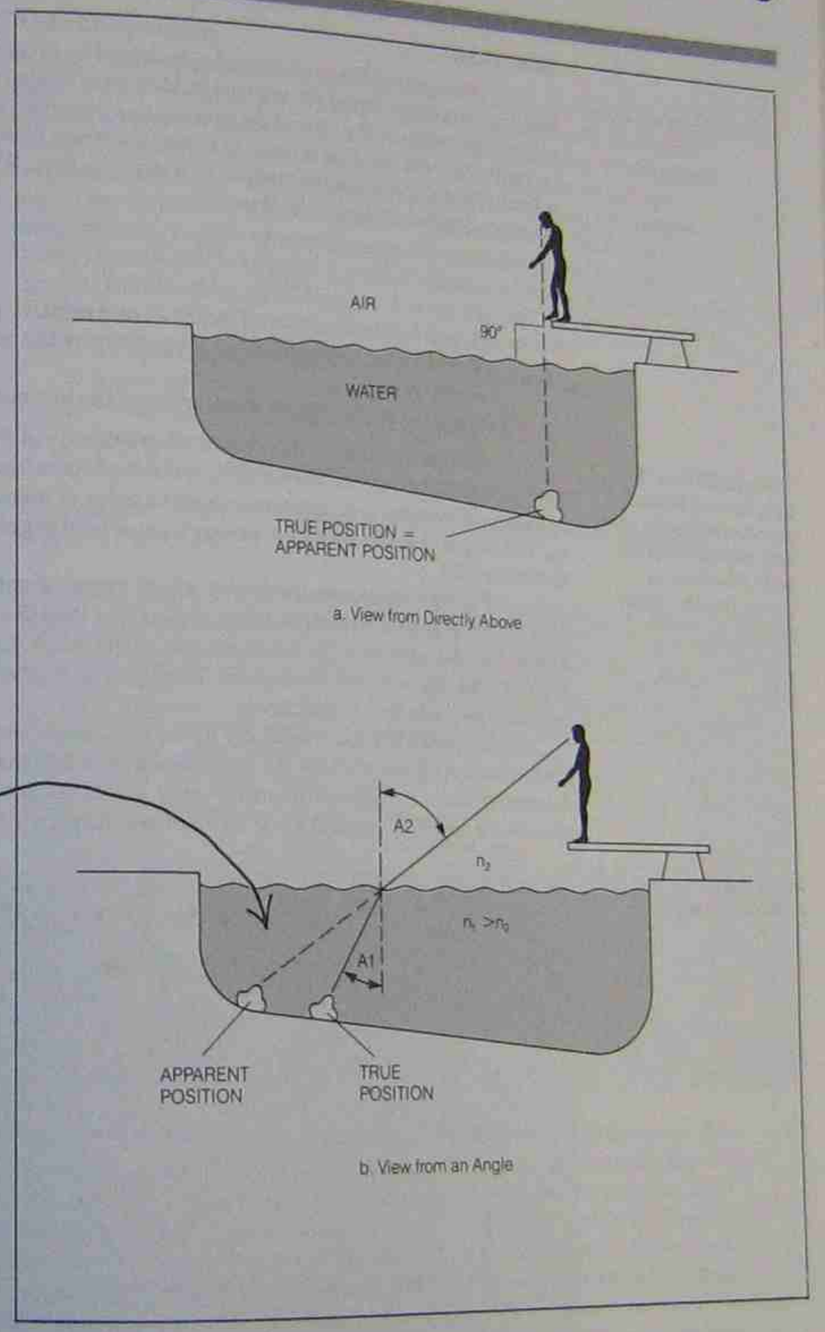
Optical fiber transmission has come of age as a major innovation in telecommunications. Such systems offer extremely high bandwidth, freedom from external interference, immunity from interception by external means, and cheap raw materials (silicon, the most abundant material on earth).

FUNDAMENTALS OF FIBER OPTIC SYSTEMS

Light rays are guided within the optical fiber by the phenomenon of light refraction. Refraction is the change in direction of light rays caused by the change in speed of propagation when they pass from one medium to another.

Optical fibers guide light rays within the fiber material. They can do this because light rays bend or change direction when they pass from one medium to another. They bend because the speed of propagation of light in each medium is different. This phenomenon is called refraction. One common example of refraction occurs when you stand at the edge of a pool and look at an object at the bottom of the pool. Unless you are directly over the object as shown in *Figure 6-2a*, it will appear to be farther away than it really is as indicated in *Figure 6-2b*. This is because the speed of the light rays from the object increases as they pass from the water to the air. This causes them to bend, changing the angle at which you perceive the object.

Figure 6-2.
Bending of Light Rays



Speed of light rays slower in water than air.

Snell's Law

How optical fibers work can be explained by Snell's Law, which states that the ratio of the sine of the angle of incidence to the sine of the angle of refraction is equal to the ratio of the propagation velocities of the wave in the two respective media. This is equal to a constant which is the ratio of the refractive index of the second medium to that of the first. Written as an equation, Snell's Law looks like this:

$$\frac{\sin A_1}{\sin A_2} = \frac{V_1}{V_2} = K = \frac{n_2}{n_1}$$

where: A_1 and A_2 are the angles of incidence and refraction, respectively;
V_1 and V_2 are the velocities of propagation of the wave in the two media; and
n_1 and n_2 are the indices of refraction of the two media.

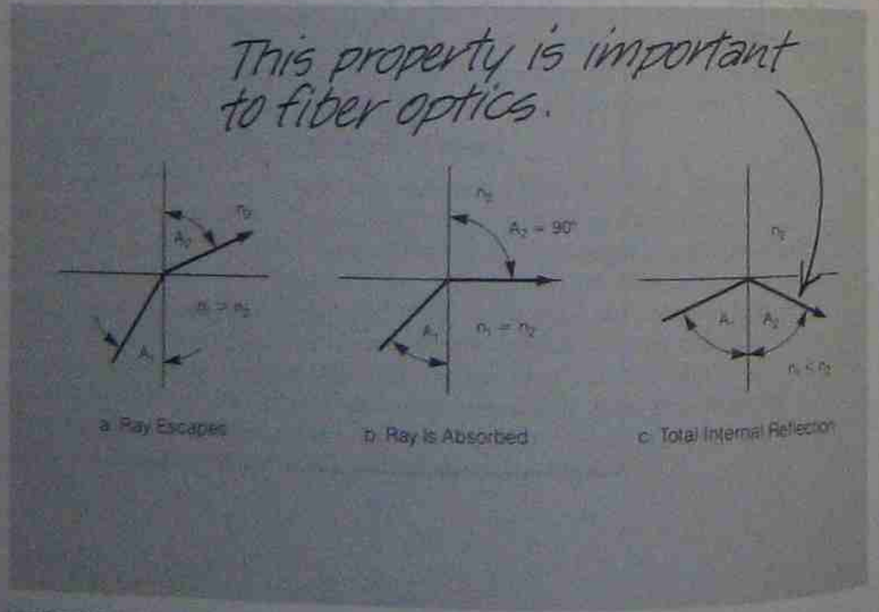
Using Snell's Law, the light rays will be totally contained within the fiber optic material when the angle of incidence is greater than the critical angle.

The parameters are demonstrated graphically in *Figure 6-3*. In each case, A_1 is the angle of incidence and A_2 is the angle of refraction. The index of refraction of material 1, n_1, is greater than the index of refraction of material 2, n_2. This means that the velocity of propagation of light is greater in material 2 than material 1.

Figure 6-3a demonstrates how a light ray passing from material 1 to material 2 is refracted in material 2 when A_1 is less than the critical angle. *Figure 6-3b* demonstrates the condition that exists when A_1 is at the critical angle and the angle A_2 is at 90 degrees. The light ray is directed along the boundary between the two materials.

As shown in *Figure 6-3c*, any light rays that are incident at angles greater than A_1 of *Figure 6-3b* will be reflected back into material 1 with angle A_2 equal to angle A_1. The condition in *Figure 6-3c* is the one of particular interest for optical fibers and it will be discussed further in following sections.

Figure 6-3. Index of Refraction



Fiber Composition

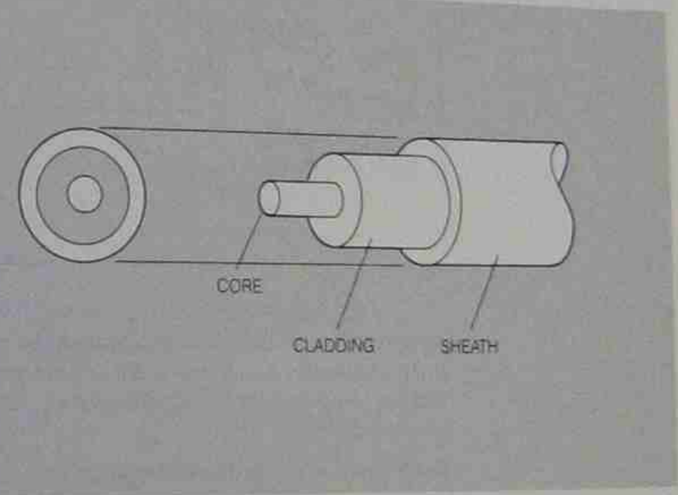
Optical fibers are constructed with three separate regions: the fiber core, the cladding around the core, and the protective outer sheath.

An optical fiber is a dielectric (nonconductor of electricity) waveguide made of glass or plastic. As shown in *Figure 6-4*, it consists of three distinct regions: a core, the cladding, and a sheath or jacket. The index of refraction of the assembly varies across the radius of the cable, with the core having a constant or smoothly varying index of refraction called n_c, and the cladding region having another constant index of refraction called n. For a fiber designed to carry light in several modes of propagation (called a multimode fiber), the diameter of the core is several times the wavelength of the light to be carried, and the cladding thickness will be greater than the radius of the core. Some typical values for a multimode fiber might be:

1. An operating light wavelength of 0.8 micrometers (μm).
2. A core index of refraction n_c of 1.5.
3. A cladding index of refraction n of 1.485 (= 0.99 × n_c).
4. A core diameter of 50 μm.
5. A cladding thickness of 37.5 μm.

The clad fiber would have a diameter of 125 μm and light would propagate as shown in *Figure 6-5*.

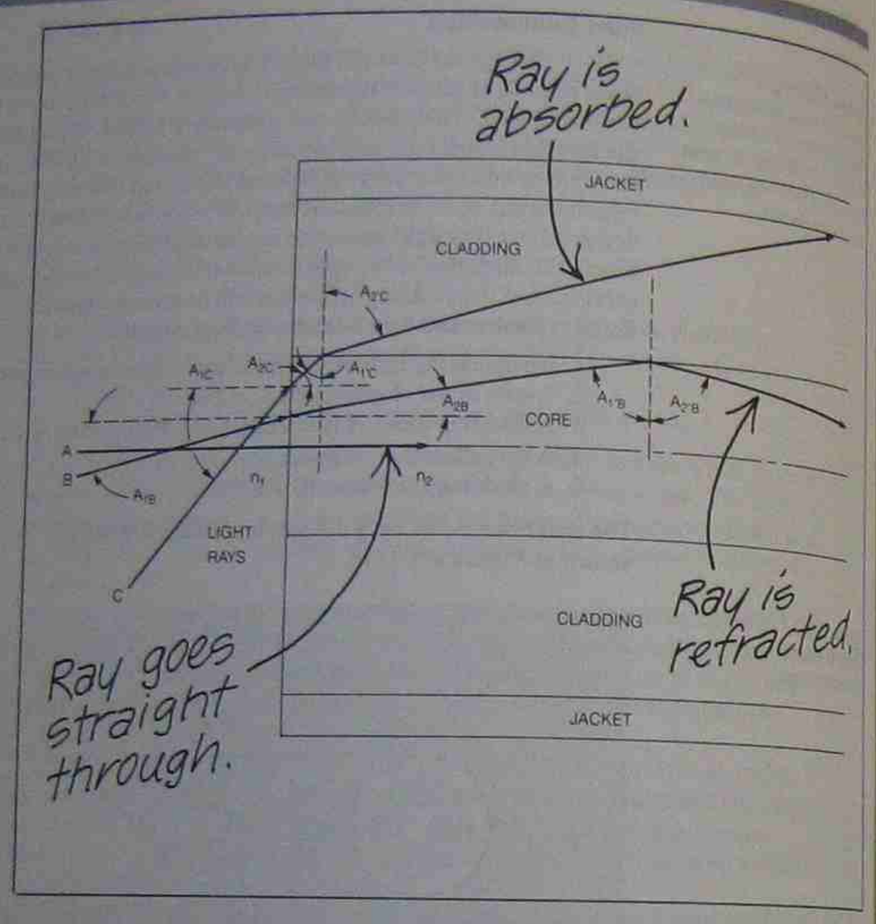
Figure 6-4. Optical Fiber Construction



The angle at which light rays enter an optical fiber determines how the rays will propagate in the fiber — either parallel to the fiber axis, in a zig-zag path, or not at all.

A light source emits light at many angles relative to the center of the fiber. In *Figure 6-5*, light ray A enters the fiber perpendicular to the face of the core and parallel to the axis. Its angle of incidence A_1 is zero, therefore, it is not refracted and travels parallel to the axis. Light ray B enters the fiber core from air (n_1 = 1) at an angle of incidence of A_{1B} and is refracted at an angle A_{2B} because n_2 is greater than n_1. When light ray B strikes the boundary between the core and the cladding, its angle of incidence, $A_{1'B}$, is greater than the critical angle. Therefore, the angle of refraction, $A_{2'B}$, is equal to $A_{1'B}$ and the light ray is refracted back into the core. The ray propagates in this zig-zag fashion down the core until it reaches the other end.

Figure 6-5.
Light Ray Paths in Multimode Fiber



If the angle of incidence, A_{1C}, is too large, as it is for light ray C, the light ray will strike the boundary between the core and the cladding with an angle of incidence, $A_{1'C}$, less than the critical angle. The ray will enter into the cladding and propagate in, or be absorbed in, the cladding and jacket (which is opaque to light).

Multimode and Single-Mode Propagation

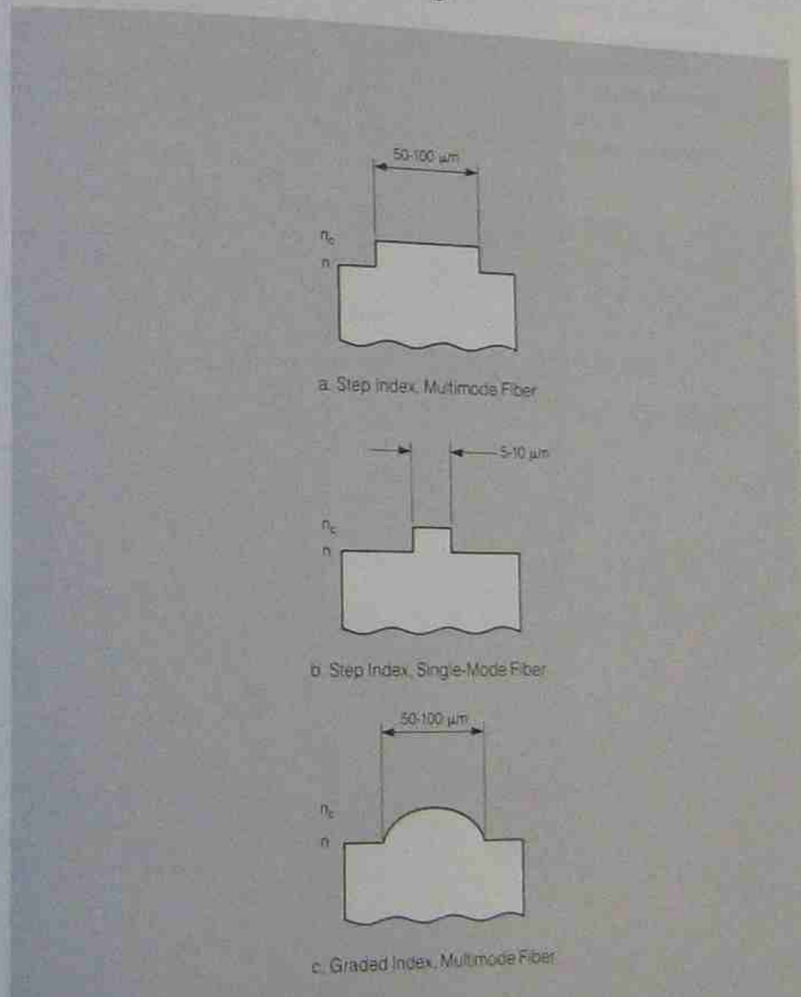
The angle at which light enters an optical fiber is called its mode of propagation. When a fiber carries more than one mode, it is a multimode fiber.

For optical fibers in which the diameter of the core is many times the wavelength of the light transmitted, the light beam travels along the fiber by bouncing back and forth at the interface between the core and the cladding. Rays entering the fiber at differing angles are refracted varying numbers of times as they move from one end to the other, and consequently do not arrive at the distant end with the same phase relationship as when they started. The differing angles of entry are called modes of propagation (or just modes), and a fiber carrying several modes is called a multimode fiber. Multimode propagation causes the rays leaving the fiber to interfere both constructively and destructively as they leave the end of the fiber. This effect is called modal delay spreading.

Much of the power in a single-mode fiber, which does not have a core diameter large enough to transmit more than one mode, is propagated in the cladding.

If, on the other hand, the diameter of the fiber core is only a few times the wavelength of the transmitted light (say a factor of 3), only one ray or mode will be propagated, and no destructive interference between rays will occur. These fibers are called single-mode fibers, and are the media that will likely be used in most future transmission systems. *Figures 6-6a* and *6-6b* show the distribution of the index of refraction across, and typical diameters of, multimode and single-mode fibers. One of the principal differences between single-mode and multimode fibers is that most of the power in the multimode fiber travels in the core, while in single-mode fibers a large fraction of the power is propagated in the cladding near the core. At the point where the light wavelength becomes long enough to cause single-mode propagation, about 20% of the power is carried in the cladding, but if the light wavelength is doubled, over 50% of the power travels in the cladding.

Figure 6-6.
Refractive Index Profiles



Graded-index multimode fibers have fairly equal propagation delays due to a reduction in inter-modal dispersion.

Figure 6-6c shows the distribution of another kind of multimode fiber, called a graded-index fiber. The index of refraction varies smoothly across the diameter of the core, but remains constant in the cladding. This treatment reduces the inter-modal dispersion by the fiber, as rays traveling along a graded-index fiber have nearly equal delays. Other refractive-index profiles have been devised to solve various problems, such as reduction of chromatic dispersion. Some of these profiles are shown in *Figure 6-7*; the step and graded profiles are repeated for comparison.

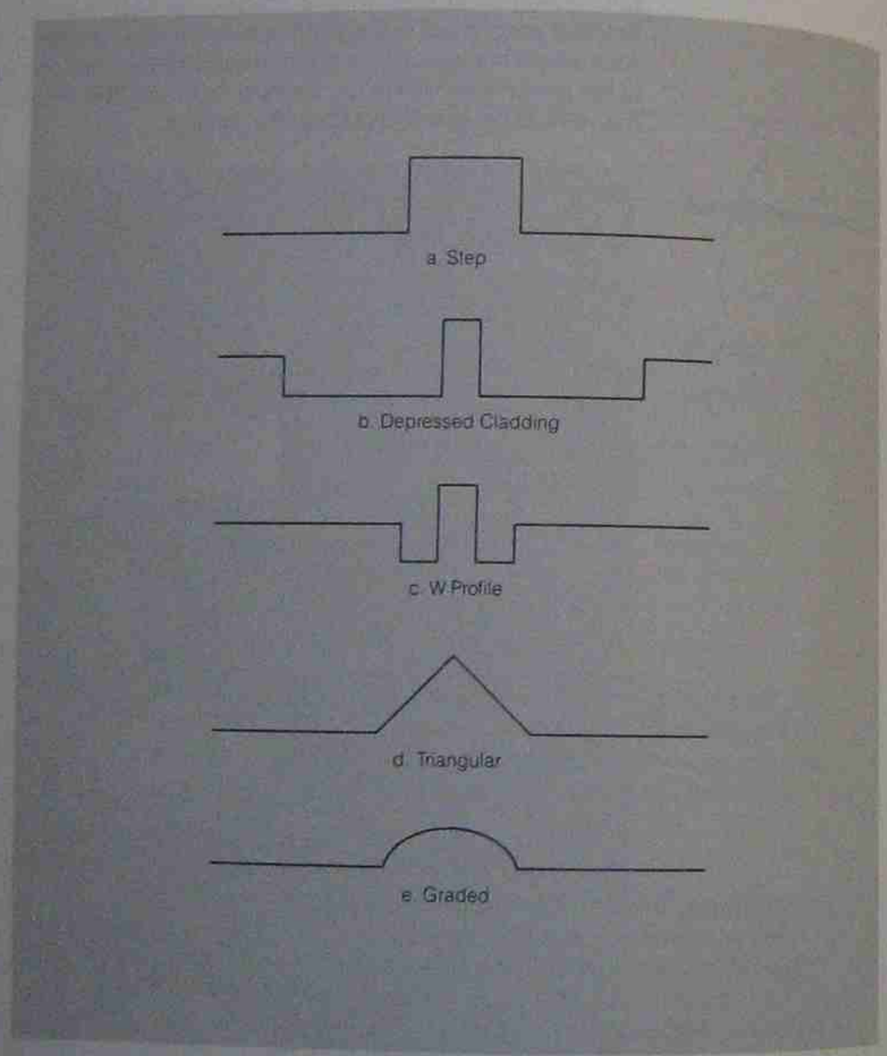


Figure 6-7.
Different Refractive Index Profiles for Optical Fibers

Material dispersion is the spreading of pulses that occurs when the light source sends light through the optical fiber at more than one wave length.

Bandwidth

The limitations on bandwidth in fiber optic systems arise from two main sources: modal delay spreading and material dispersion. Modal delay spreading was described above, and is evident primarily in multimode fibers. Material dispersion arises from the variation in the velocity of light through the fiber with the wavelength of the light.

If the light source, such as a light-emitting diode (LED), emits pulses of light at more than one wavelength, the different wavelengths will travel at different velocities through the fiber. This causes spreading of the pulses. At a typical LED wavelength of 0.8 µm, the delay variation is about 100 picoseconds (ps) per nanometer (nm) per km. If the width of the spectrum emitted by the LED is 50 nm, pulses from the source will be spread by 5 nanoseconds (ns) per km. This will limit the modulation-bandwidth product to about 50 to 100 MHz/km. Fortunately, at certain wavelengths (near 1.3 and 1.5 µm for some types of fibers), there is a null in the material dispersion curve, giving much better modulation-bandwidth performance. *Figure 6-8* shows the relationship of loss in doped silicon glass fibers versus light wavelength. Most current development work is aimed at making fibers, light sources and detectors that work well at the loss nulls at 1.3 and 1.5 µm.

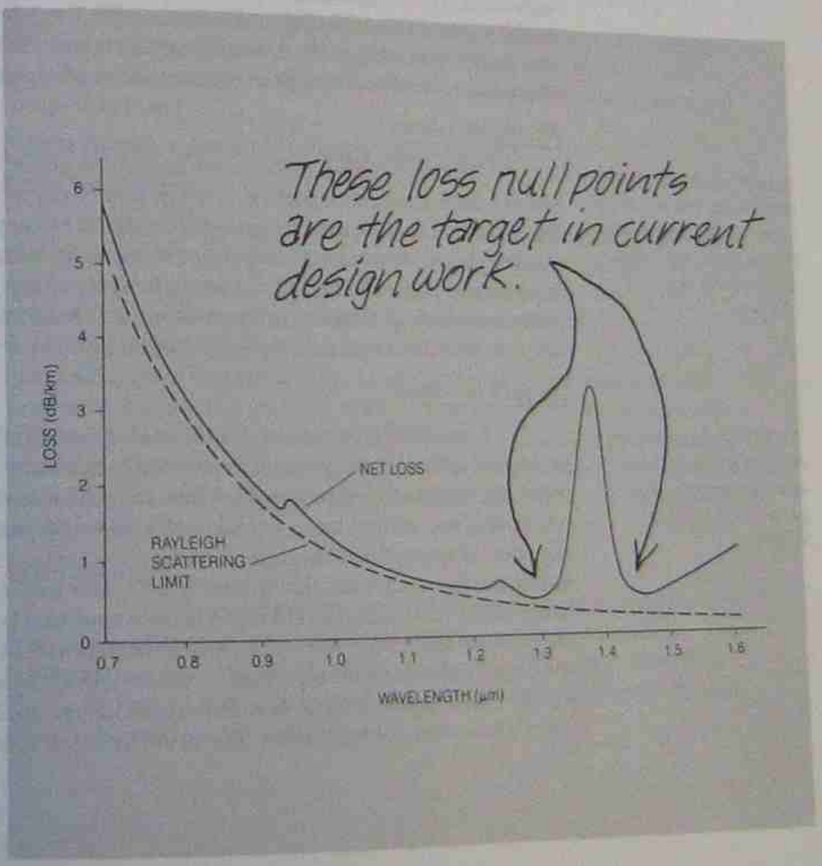


Figure 6-8.
Net Spectral Loss Curve for a Glass Core
(Source: James Martin, Telecommunications and the Computer, Prentice-Hall, Inc. 1969)

Attenuation

Signal power loss (attenuation) is a function of four factors: scattering, absorption, connections, and fiber bending.

The loss in signal power as the light travels down the fiber is called attenuation. Attenuation in the fibers is controlled mainly by four factors: radiation of the propagated light, called scattering; conversion of the light energy to heat, called absorption; connection losses at splices and joints in the fiber; and losses at bends in the fiber.

Scattering Losses

Scattering occurs because of microscopic imperfections in the fiber.

Scattering arises due to microscopic imperfections in the fiber, such as the inclusion of water in the glass. The effect of impurities in the transmission medium is evident when we look up at the sky and see a blue color. In fact, deep space has no color (appears as black), but due to the scattering of sunlight by the dust in the atmosphere, the sky appears as a bright blue.

There is a limit below which scattering cannot be reduced, no matter how perfectly the glass fiber is made, because of irregularities in the molecular structure of glass. This limit, called the Rayleigh scattering limit, is shown graphically in *Figure 6-8*. For light with a wavelength of 0.8 µm, it is about 2.9 dB/km. At a wavelength of 1.3 µm, the value is about 0.3 dB/km, and at 1.55 µm wavelength, the limit is about 0.15 dB/km. Commercially available glass fibers exhibit losses of about 3.5 dB/km at 0.8 µm, and 0.7 to 1.5 dB/km at 1.3 and 1.5 µm. There is less attenuation through 6 meters (about 20 feet) of good quality optical fiber glass than through an ordinary clean windowpane.

Absorption Losses

Opaque imperfections also cause signal power loss because, when a light beam strikes them, some of the light's energy is converted into heat.

Absorption refers to the conversion of the power in the light beam to heat in some material or imperfection which is partially or completely opaque. This property is useful, as in the jacket of the fiber, to keep the light from escaping the cable, but is a problem when it occurs as inclusions or imperfections in the fiber. Current fiber optic systems are designed to minimize intrinsic absorption by transmitting at 0.8, 1.3, and 1.5 µm, where there are reductions in the absorption curve for light.

Connection Losses

A relatively large amount of signal power is lost at every connection point, particularly at repair splices.

Connection losses are inevitable and represent a large source of loss in commercial fiber optic systems. In addition to the installatiion connections, repair connections will be required because a typical line will be broken accidentally two or three times per kilometer over a 30-year period. The alignment of optical fibers required at each connection is a considerable mechanical feat, somewhat akin to kissing with your eyes closed and your hands behind your back. The full effect of the connection is not obtained unless the parts are aligned correctly. The ends of the fibers must be parallel to within one degree or less, and the core must be concentric with the cladding to within 0.5 µm. Production techniques have been developed to splice single-mode fibers whose total diameter is less than 10 µm by using a mounting fixture and

small electric heater. Mechanical connectors have been developed which allow random mating of fibers with average connection losses of less than 0.3 dB.

Bending Losses

Bending the fibers also can cause signal loss because the light rays on the outside of a sharp bend cannot travel fast enough to keep up with the other rays and they are lost.

Bending an optical fiber is akin to playing crack-the-whip with the light rays. As light travels around the bend, the light on the outside of the bend must travel faster to maintain a constant phase across the wave. As the radius of the bend is decreased, a point is reached where part of the wave would have to travel faster than the local speed of light — an obvious impossibility. At that point the light is lost from the waveguide. For commercial single-mode fiber optic cables operating at 1.3 and 1.5 µm, the bending occuring in fabrication (the cables are made with the fibers wound spirally around a center) and installation does not cause a noticeable increase in attenuation.

Numerical Aperture and Acceptance Angle

The numerical aperture is the measure of a fiber's ability to gather light. The acceptance angle is the largest angle at which a light ray can enter and still propagate down the fiber.

The numerical aperture of the optical fiber is a measure of its light-gathering ability (much like the maximum *f*-stop of a camera lens). The optical power accepted by the fiber varies as the square of the numerical aperture, but unlike the camera lens *f*-stop, the numerical aperture does not depend on any physical dimension of the optical fiber.

The acceptance angle is the maximum angle that an entering light ray may have relative to the axis of the fiber and still propagate down the fiber. A large acceptance angle makes the end alignment less critical when splicing and connecting fibers.

FIBER OPTIC SUBSYSTEMS AND COMPONENTS

Fiber Production

Optical fibers are fabricated in several ways, depending on the vendor and purpose of the system. The core and cladding regions of the fiber are doped to alter their refractive indices. This doping is carried out by heating vapors of various substances such as germanium, phosphorus, and flourine, and depositing the particles of resulting oxidized vapor or "soot" on high-quality fused-silica glass mandrels, called preforms. The preforms are a large-scale version of the core and cladding which is then heated to a taffy-like consistency and drawn down into the actual fiber. The core and cladding dimensions have essentially the same relationship in the final fiber as in the preform. Deposition of the dopants is done in one of three standard ways: outside, inside, and axial vapor deposition.

Light Sources

Light sources for fiber optic systems must convert electrical energy from the computer or terminal circuits feeding them to optical energy (photons) in a way that allows the light to be coupled effectively to the fiber. Two such sources currently in production are the surface light-emitting diode (LED) and the injection laser diode (ILD).

To convert the electrical energy of a computer or terminal to light energy for use by an optical fiber system, LEDs or ILDs are used.

Light-Emitting Diodes

A cross section of a surface LED is shown in *Figure 6-9*. It emits light over a relatively broad spectrum, and in addition disperses the emitted light over a rather large angle. This causes the LED to couple much less power into a fiber with a given acceptance angle than does the ILD. Currently, LEDs are able to couple about 100 microwatts (uW) of power into a fiber with a numerical aperture of 0.2 or more with a coupling efficiency of about 2%. The principal advantages of LEDs are low cost and high reliability.

Injection Laser Diodes

A cross section of a typical ILD is shown in *Figure 6-10*. Because of its narrow spectrum of emission and its ability to couple output efficiently into the fiber lightguide, the ILD supplies power levels of 5 to 7 milliwatts (mW). At present, ILDs are considerably more expensive than LEDs and their service life is generally less by a factor of about 10. Other disadvantages of laser diodes are that they must be supplied with automatic level control circuits, the laser power output must be controlled, and the device must be protected from power supply transients.

Light Detectors

Light detectors are used at the receiving end of the optical fiber to detect, amplify, and convert the light signal back into its original electrical form.

At the receiving end of the optical communications system, the receiver must have very high sensitivity and low noise. To meet these requirements, there is a choice of two types of devices to detect the light beam, amplify it, and convert it back into an electrical signal: the integrated p-i-n field-effect transistor (FET) assembly and the avalanche photodiode (APD).

In the p-i-n FET device, a photodiode with unity gain (the p-i-n device) is coupled with a high-impedance front-end amplifier. This device combines operation at low voltage with low sensitivity to operating temperature, high reliability, and ease of manufacture.

The avalanche photodiode produces a gain of 100 or more; however, it also produces noise which may limit the receiver sensitivity. The APD devices require high voltage bias which varies with temperature. Receivers using APDs are so sensitive that they require as few as 200 photons to be detected at the receiver per bit transmitted at data rates of 200 to 400 Mbps.

WAVELENGTH MULTIPLEXING

A combination of single-mode fiber (low dispersion by the transmission medium), narrow output spectrum (power concentration at a single frequency), and narrow dispersion angle (good power coupling) from ILDs makes possible the extreme bandwidth-distance characteristics given for systems at the beginning of this chapter. The narrow ILD emission spectrum also makes it possible to send several signals from different sources down the same fiber by a technique called wavelength multiplexing. The ability to multiplex several analog signals in the frequency domain (FDM) has been

Figure 6-9. Construction of Light-Emitting Diode for 1.3μm Operation

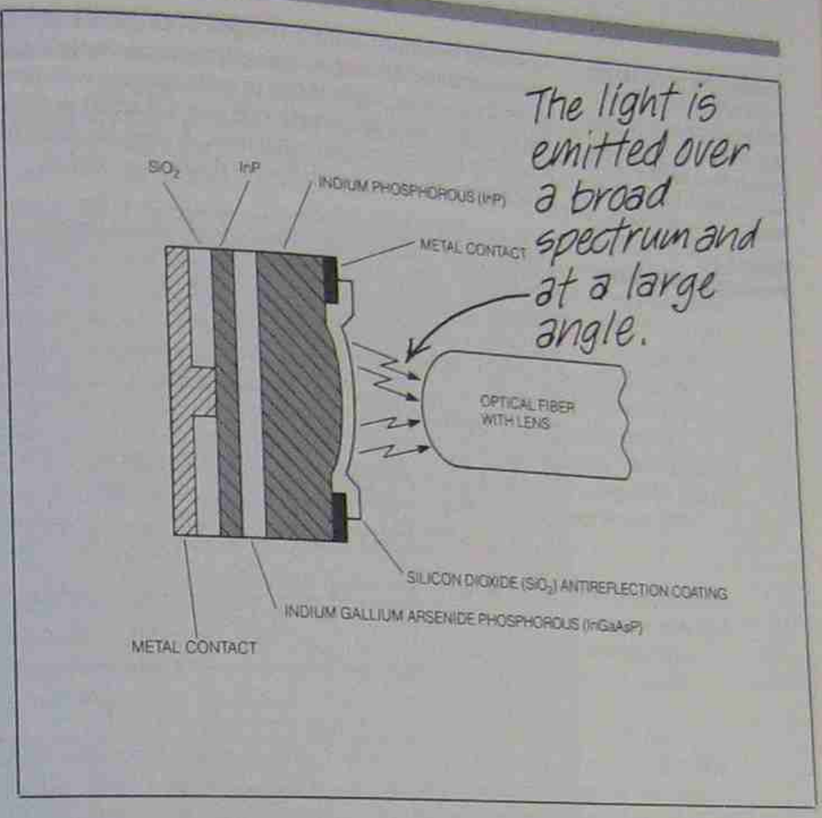
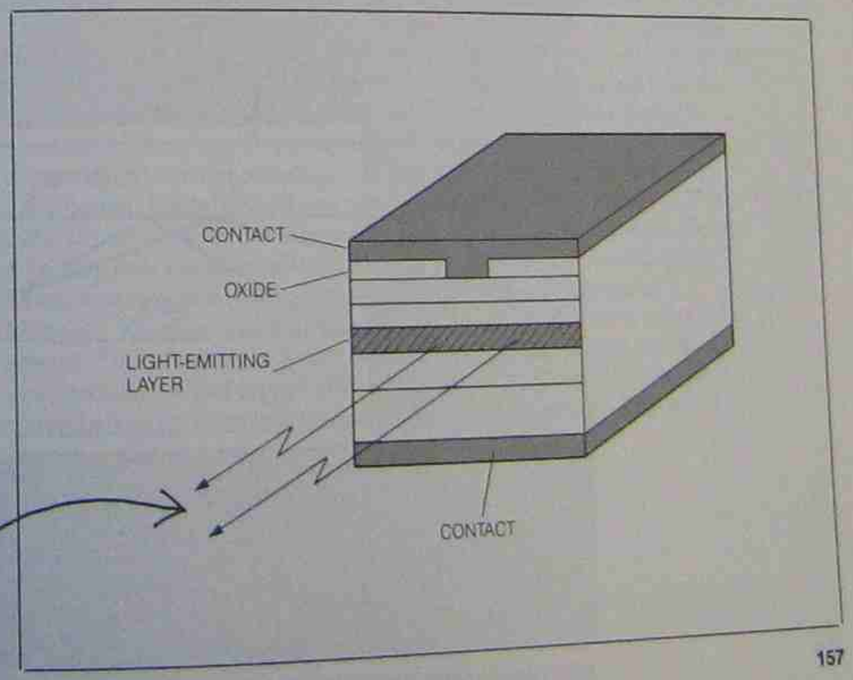


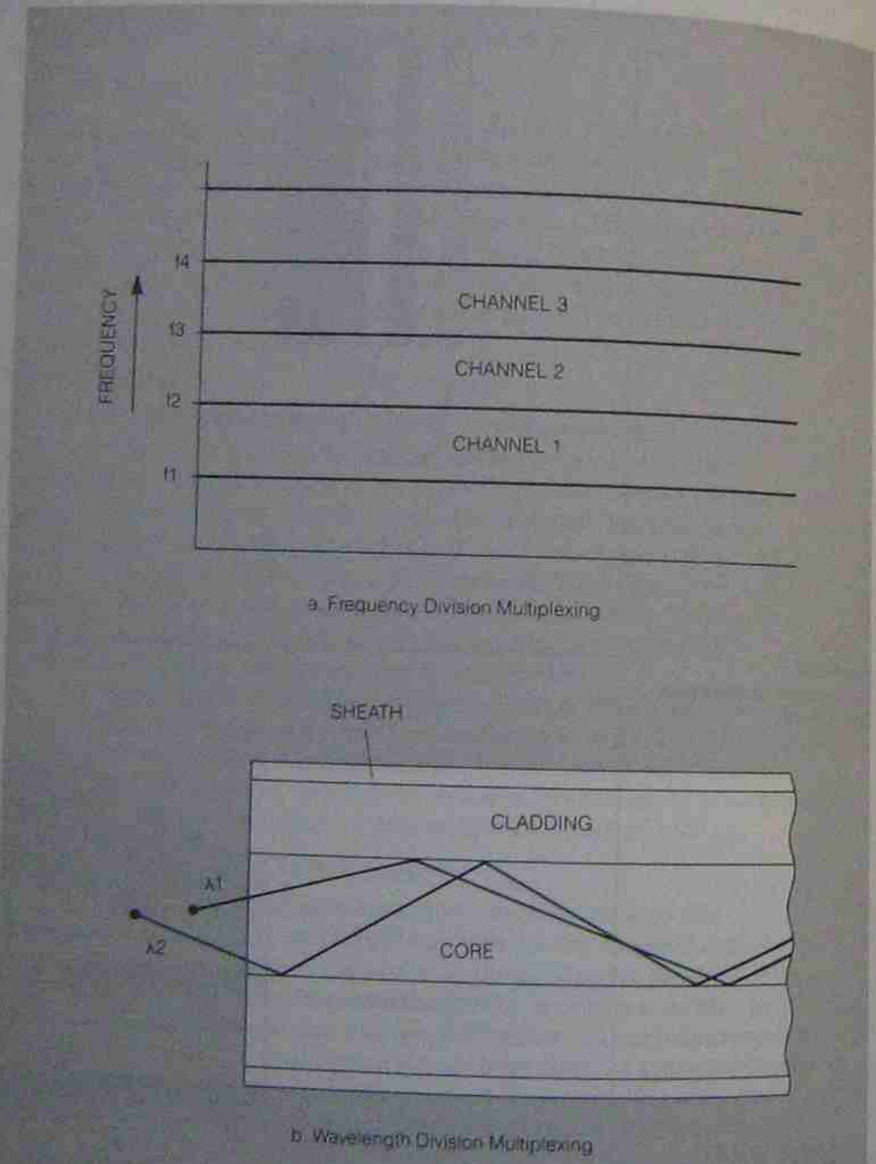
Figure 6-10. Construction of Injection Laser Diode



Wavelength multiplexing, a technique similar to frequency division multiplexing, is used to send light signals from several sources down the same fiber.

described in detail in previous chapters. As illustrated in *Figure 6-11*, wavelength multiplexing at optical frequencies is the equivalent of FDM at lower frequencies. Light at two or more discrete wavelengths is coupled into the fiber with each wavelength carrying a channel at whatever modulation rate is used by the transmission equipment driving the light source. Thus, the information capacity of each fiber is doubled or tripled.

Figure 6-11. Wavelength Multiplexing



TRANSMISSION SYSTEMS

Local and Intercity Systems — The FT3C System

Studies by AT&T of a digital lightwave transmission system indicated it would cost less than traditional systems because of differences in terminal multiplexing equipment requirements and cost.

An AT&T study in 1978 noted a striking advantage of installing a digital transmission system over comparable analog systems: $900.00 per circuit termination when interconnecting a number of digital switching machines. The savings come from the difference in requirement for terminal multiplexing equipment. The AT&T FT3C lightwave system was devised to provide the most economical digital transmission system possible with the current state of the fiber optic art. It uses wavelength multiplexing techniques to send three 90 Mbps signals over the same fiber, giving over 240,000 digital channels at 64,000 bps in a cable containing 144 optical fibers. First application of the system has been in the Northeast Corridor project by AT&T, between Boston and Washington, and in the North/South Lightwave Project on the West coast of the U.S. by Pacific Telesis, between San Francisco and San Diego. *Figure 6-12a* is a map of the Northeast Corridor system, which contains 78,000 fiber-kilometers of lightwave circuits. *Figure 6-12b* is a map of the North/South Lightwave Project. The two systems were placed in service in 1983.

International Systems — The SL Underwater Cable

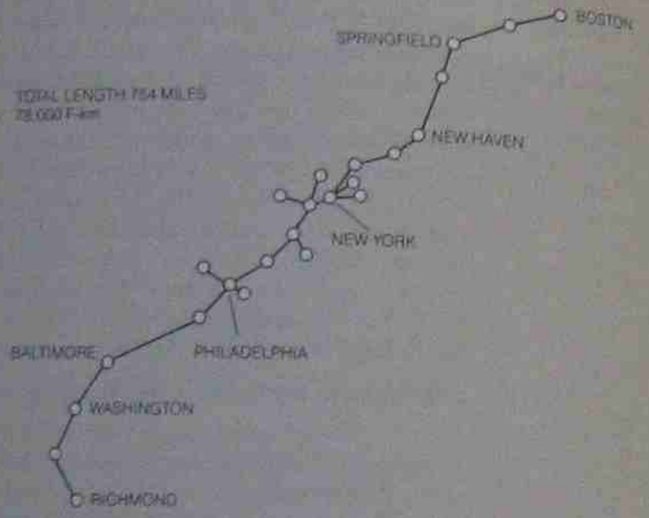
Undersea optical fiber cables can provide a data communications channel with a carrying capacity equal to satellite systems, but with greater security, less interference, less noise, and lower cost.

The advent of optical fiber technology for undersea cables promises point-to-point channel capacity equal to satellite systems at reduced cost and without long transmission delays, unstable environmental interference, induced noise, and broadcast of potentially sensitive information to half the world.

Undersea cable systems have some understandably difficult environmental requirements. The environment includes pressures of 10,000 pounds per square inch (psi) at depths of 7300 meters, salt water, and the possibility of mechanical damage from anchors and earth movement in shallow waters. An important requirement for these systems is that the regenerator spacing be as wide as possible to cut down on the system failure probability and the power requirements, since power must be fed from the ends of the cable.

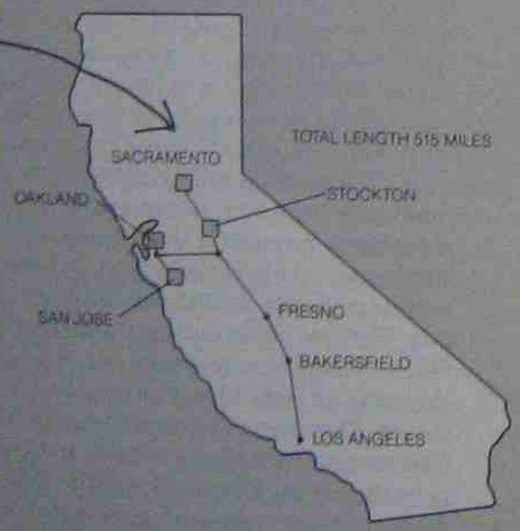
A schematic diagram of the SL Undersea Lightguide System is shown as *Figure 6-13*. It is comprised of a high-voltage power supply, a supervisory terminal, a multiplexor with inputs for several types of information, the cable light source, the cable itself, and the repeaters. The cable is made of a central core and a surrounding support as shown in *Figure 6-14*. The core has an outside diameter of 2.6 mm and consists of 12 optical fibers wound helically around a central copper-clad steel wire called a kingwire, all embedded in an elastomeric substance and covered with a nylon sheath. This assembly is in turn covered with a number of steel strands, a continuously welded copper tube, and finally an insulation of low-density polyethelene for electrical insulation and abrasion resistance. The outside diameter of the completed cable is 21 mm (about 0.8 in.).

Figure 6-12.
Fiber Optic Networks
(Source, J.R. Sauffer, IEEE Journal on Selected Areas in Communications, Vol. SAC-1, No. 3, April 1983, © 1983 IEEE)



a. Fiber Optic Network for Northeast Corridor

Digital transmission systems that use state-of-the art fiber optics.



b. Pacific Telephone North/South Lightwave Project

Figure 6-13.
SL Undersea Lightguide System
(Source, J.R. Sauffer, IEEE Journal on Selected Areas in Communications, Vol. SAC-1, No. 3, April 1983, © 1983 IEEE)

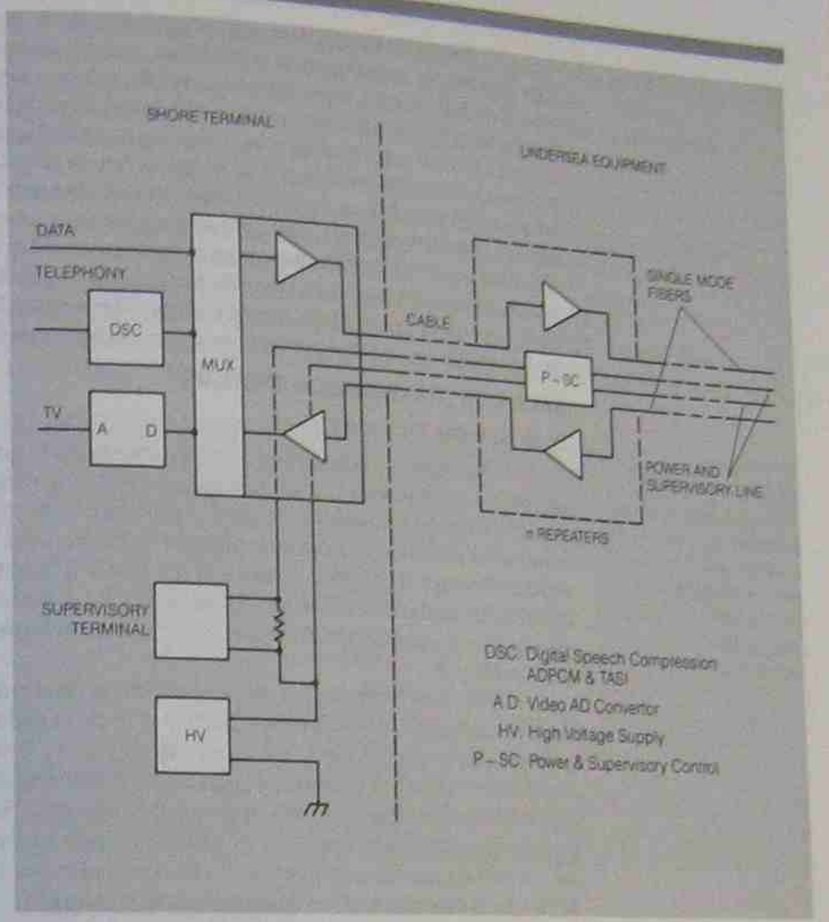
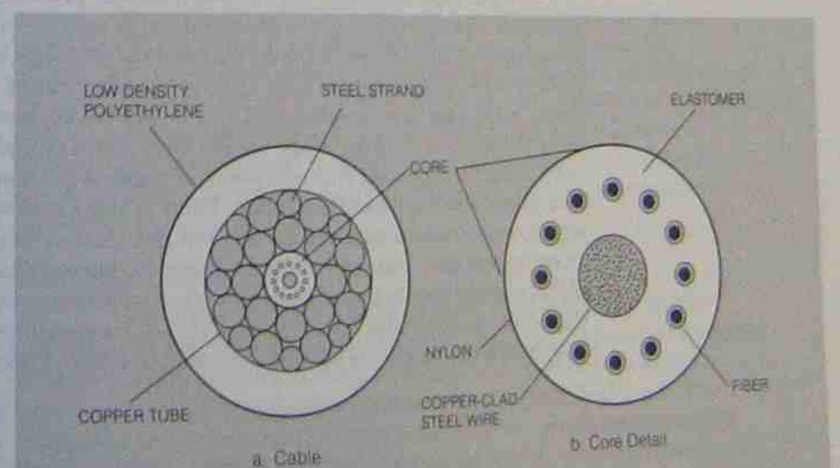


Figure 6-14.
Embedded Fiber Core Cable
(Source, J.R. Sauffer, IEEE Journal on Selected Areas in Communications, Vol. SAC-1, No. 3, April 1983, © 1983 IEEE)



The fibers are single-mode optical lightguides operating at 1.312 µm. The data rate on each fiber pair is 280 Mbps, and repeaters are spaced every 35 km. The total capacity of the system is over 35,000 two-way voice channels. Inputs from binary data sources are multiplexed directly into the stream. Analog channels are first converted to binary by the adaptive delta modulation technique, then processed by a type of digital circuit (Digital TASI) that interleaves inputs from a number of input speech channels onto a smaller number of output channels. The light sources for the system are ILDs operating near 1.3 µm with an average output power of 1 mW (0 dBm). Light detectors are indium-gallium-arsenide (InGaAs) p-i-n diode receivers followed by silicon bipolar transimpedance amplifiers. Three spare laser diodes, which can be switched in remotely if a failure occurs, are provided per circuit.

SATELLITE TRANSMISSION SYSTEMS

Basic Satellite Technology

Broad focus of the downlink signal and the propagation delay caused by the great distance between earth and satellite can present problems in two-way satellite communications.

Satellite communications systems are basically line-of-sight microwave systems with a single repeater. As stated in Chapter 3, the satellite is said to be in geostationary orbit, and is called a geostationary satellite, because the speed of the satellite is matched to the rotation of the earth at the equator. Because of the great distance of the satellite from the earth (about 22,300 miles) and antenna size limitations which limit focusing ability, the cone of coverage for a single satellite transmitter may be as large as the entire continental U.S.

For those transmission services which originate at a single point and flow to many points in one direction, such as television and radio signals, the large area of coverage is ideal. The relatively long delay between the instant a signal is sent and when it returns to earth (about 240 milliseconds (ms)) has no undesirable effect when the signal is going only one way. However, for signals such as data communications sessions and telephone conversations, which go in both directions and are only intended to be received at one other point, the large area of coverage and the delay can cause problems.

Data and telephone conversations usually proceed as a series of messages in one direction which are answered or acknowledged by messages in the other direction. When the delay between the message being sent and the reply or acknowledgement is long, the transmission rate of information slows down. In the case of voice messages, long delays between utterance and reply make the speaker think he or she has not been heard or understood. This leads to requests for repeating (the equivalent of negative acknowledgement in the data world) and increased frustration. There is also a serious privacy issue with communications intended for only one destination, but broadcast so that an entire continent can receive them. One of the factors causing the rush to all-digital transmission is the ease of encrypting information in digital form, so that when the inevitable interception of broadcast signals occurs, the information intercepted is at least somewhat difficult to decipher.

The main components of a satellite system are the earth sending and receiving stations and a geosychronous satellite equipped with transponders.

A satellite transmission system consists of one or more earth stations and a geostationary satellite which can be seen by the stations. Separate frequencies are assigned for sending to the satellite (the uplink) and receiving from the satellite (the downlink). The current frequency assignments for satellite systems are shown in *Table 6-1*.

Satellites are equipped with multiple repeater units called transponders. Current systems have 10 or 12 transponders, but the next series of international satellites, called INTELSAT VI, will have 46 transponders. Transponders are assigned different uses, but in the case of those used for voice or voice-equivalent data communications channels (nominal 4 kHz bandwidth), the transponder capacity may be as large as 3000 channels. The INTELSAT VI satellite will provide a total bandwidth, using frequency reuse techniques, of 3460 MHz.

Table 6-1. Frequency Assignments for Satellite Systems

Uplink Frequencies	Downlink Frequencies
5.925-6.425 GHZ	3.700-4.200 GHz
7.900-8.400 GHz	7.250-7.750 GHz
14.00-14.50 GHz	11.70-12.20 GHz
27.50-30.00 GHz	17.70-20.20 GHz

Multiple Access Systems

Telephone switching systems and data multiplexors are designed based on the fact that not every telephone or terminal that can send information wishes to do so at the same time; or alternatively, the telephone or terminal user does not need or wish to pay for the entire capacity of a channel. These conditions are also true for the users of satellite systems. Several methods have been devised to allow sharing of the satellite and earth station resources among several users so that it appears that the transmission channel is dedicated to each user.

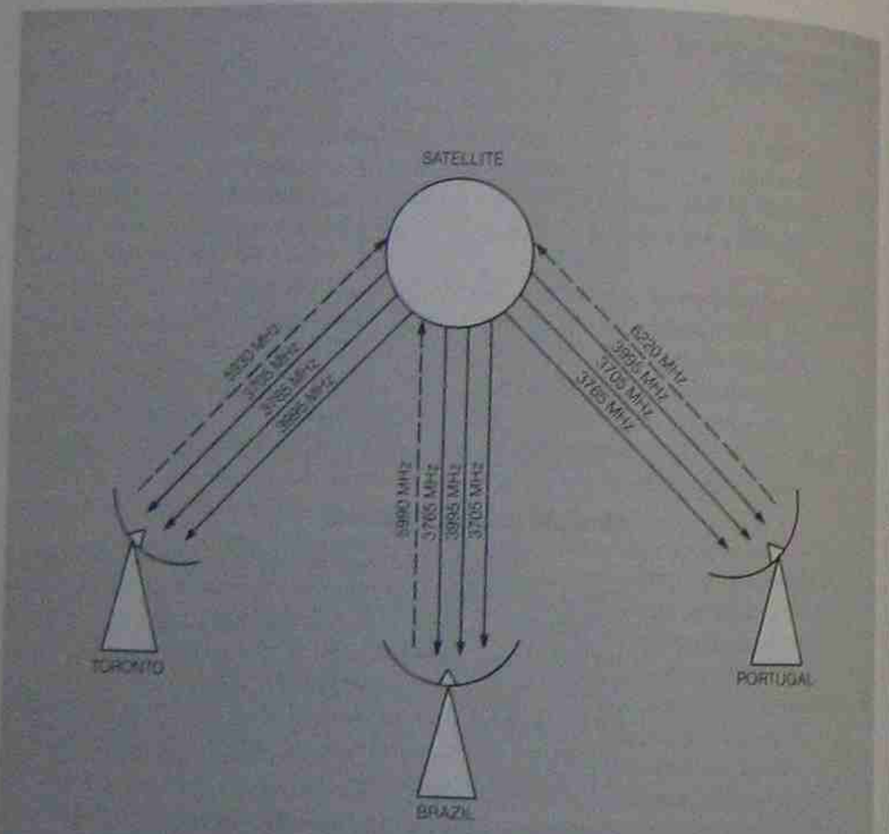
Frequency Division Multiple Access

Satellites using FDMA have the available bandwidth of a transponder divided into smaller segments.

Frequency Division Multiple Access (FDMA) is simply another example of the familiar data and voice transmission technique called frequency division multiplexing (FDM) used to allocate small portions of a large bandwidth (500 MHz for satellite transponders) to individual users. For instance, a telecommunications common carrier in a particular country, say Brazil, may want 132 voice-grade channels for sending voice and analog-coded data to various other countries. The bandwidth required on the current

international satellite systems for this many channels is 10 MHz. Since one transponder has a bandwidth of 500 MHz, it could accommodate 50 users each requiring 132 channels. The Brazilian user might be allocated the frequency band between 5990 and 6000 MHz for the uplink to the satellite, and the corresponding downlink frequencies would be 3765 to 3775 MHz. Other users might be assigned similar portions of the bandwidth in the same transponder. For example, a Portugese user might be allocated the 6220 to 6230 MHz uplink band and the 3995 to 4005 MHz downlink band. A Canadian user might operate on the 5930 to 5940 MHz uplink band and the 3705 to 3715 MHz downlink band. *Figure 6-15* illustrates how the three users each have one uplink, all into a single transponder, but all users can receive all three downlinks. This arrangement makes possible simultaneous two-way transmissions between any of the three sites using only a part of one satellite transponder.

Figure 6-15. Multiuser Satellite System



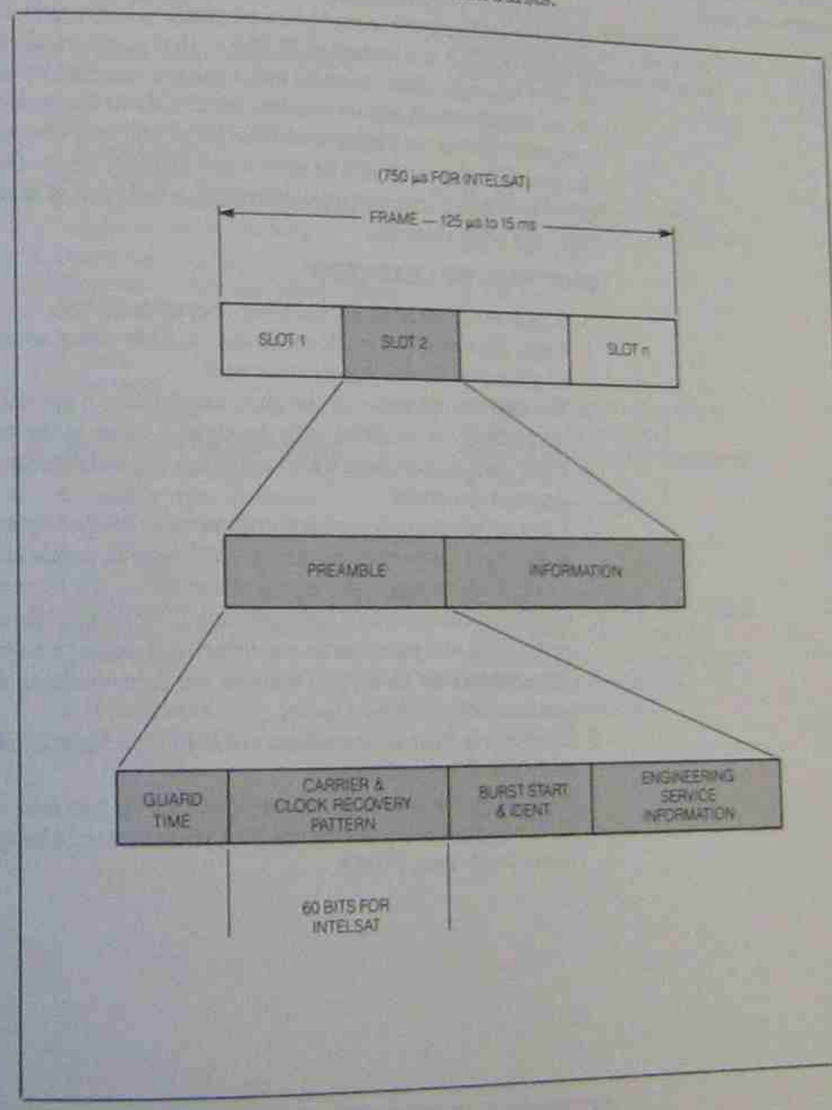
Each user has a permanently assigned channel.

Time Division Multiple Access

In TDMA, access to a satellite transponder is divided into discrete time frames.

Time Division Multiple Access (TDMA) is the equivalent of FDMA, but in the time domain rather than the frequency domain. TDMA works just like time division multiplexing (TDM) for land-based data and voice transmissions, with each satellite transponder having a data rate capacity of between 10 and 100 Mbps. During the time a station is sending on the uplink, it has available the entire data rate capacity of the transponder, but then it must stop sending for a short time to allow another station access to the transponder. The information flows to the satellite in frames, each frame containing one burst of information from each earth station allowed access to the single transponder. A typical format for a frame is shown in *Figure 6-16*. The sum of the durations of the individual bursts do not quite equal the frame time in order to give some guard time between bursts.

Figure 6-16. TDMA Transmission Frames



Timing for keeping the station bursts apart is a major problem, and is complicated by two facts: 1) The satellites are not perfectly stationary in orbit (each one appears to travel in a small figure-8 pattern), and 2) the time of travel of the signal between different earth stations and the same satellite is different because of different slant range distances. TDMA techniques allow the satellite transmitter to be operated at higher power levels than FDMA. This is permissible because only one carrier at a time occupies the transponder, reducing the amount of intermodulation distortion generated.

Demand Assignment Multiple Access

DAMA is similar to FDMA except that channels are not assigned on a permanent basis.

Using the Demand Assignment Multiple Access (DAMA) technique, each satellite transponder is used much like a telephone switch; that is, a subchannel is assigned only when traffic is available to be carried on it. This is in contrast to the FDMA and TDMA techniques described above, where channels are assigned to users on a permanent basis, even if there is no traffic demand. DAMA is a variant of FDMA in that a part or all of a transponder is divided into individual channels which can be accessed by all DAMA terminals on the ground which are serviced by the DAMA transponder. A central computer system, or a system of distributed and cooperating computers, on the ground keeps track of who gets which channel when. DAMA is particularly useful for loading transponders efficiently when each of several ground sites need only a few channels.

WHAT HAVE WE LEARNED?

1. Light travelling in an optical fiber obeys Snell's law.
2. Light can travel in a multi-mode optical fiber along several paths, or in a single-mode fiber along only one path.
3. The index of refraction of the glass making up an optical fiber varies across the diameter of the fiber, with the highest index in the middle of the fiber.
4. Fiber optic system bandwidth is limited by modal delay spreading and material dispersion.
5. Fiber optic system transmission distance is limited by scattering, absorption, losses at connections, and losses at bends in the cable.
6. Fibers with varying index of refraction across the fiber are made by heating a preform which has been doped with metal salts such as germanium and phosphorus and drawing it down to a small diameter.
7. Transmitters for fiber optic systems are light-emitting diodes or semiconductor injection lasers.
8. Receivers in fiber optic systems are avalanche photodiodes or integrated p-i-n field-effect transistors.
9. Satellite transmissions have four uplink bands and four downlink bands.
10. TDMA techniques allow the satellite transmitter to be operated at higher power levels than FDMA.

Quiz for Chapter 6

1. The requirements for a successful transmission system using light are:
   a. Powerful, reliable light source.
   b. Strong glass.
   c. Reliable, low-cost transmission medium.
   d. Powerful amplifiers.
2. The core of an optical fiber has:
   a. A lower index of refraction than air.
   b. A lower index of refraction than the cladding.
   c. A higher index of refraction than the cladding.
   d. None of the above.
3. For single-mode fibers, the core diameter is about:
   a. 10 times the fiber radius.
   b. 3 times the wavelength of the light carried in the fiber.
   c. 15 micrometers.
   d. 10 times the wavelength of the light carried in the fiber.
4. Over a period of 30 years, a kilometer of fiber optic cable is likely to be broken:
   a. Not at all.
   b. Once.
   c. 10 times.
   d. 2 or 3 times.
5. Deposition of dopants on fiber preforms is done by:
   a. Outside vapor deposition.
   b. Axial vapor deposition.
   c. Inside vapor deposition.
   d. All of the above.
6. A light-emitting diode is able to couple how much power into an optical fiber?
   a. 10 watts.
   b. 10 milliwatts.
   c. 100 microwatts.
   d. 1 picowatt.
7. Avalanche photodiode receivers can detect bits of transmitted data by receiving:
   a. 1 photon.
   b. 10 photons.
   c. 100 photons.
   d. 200 photons.
8. One unsolved problem with satellite systems is:
   a. Coverage.
   b. Privacy.
   c. Bandwidth.
   d. Access.
9. The AT&T FT3C fiber optic transmission system is designed to use how many light wavelengths?
   a. One.
   b. Two.
   c. Three.
   d. Ten.
10. The SL Undersea Lightguide system can carry how many two-way voice channels?
    a. 10,000
    b. 100,000
    c. 35,000
    d. 760
11. When the index of refraction is greater in Material 1 than it is in Material 2, the velocity of propagation in Material 1 compared to Material 2 is:
    a. Equal to or greater.
    b. Greater.
    c. Lesser.
    d. Equal.
12. The different angles of entry of light into an optical fiber where the diameter of the core is many times the wavelength of the light transmitted are called:
    a. Emitters.
    b. Modes.
    c. Sensors.
    d. Refractors.

13. In single-mode fibers, a large fraction of the power is propagated in the:
    a. Sheath.
    b. Core.
    c. Cladding.
    d. Air.
14. The loss in signal power as light travels down a fiber is called:
    a. Propagation.
    b. Scattering.
    c. Absorption.
    d. Attenuation.
15. The coupling efficiency of a LED light source to an optical fiber with a numerical aperture of 0.2 or more is:
    a. 60%.
    b. 10%.
    c. 2%.
    d. 0.1%.
16. The FT3C lightwave system contains the following number of fibers:
    a. 12.
    b. 144.
    c. 128.
    d. 64.
17. A geostationary satellite used for communications systems is:
    a. Stationary in the sky.
    b. Rotates with the earth.
    c. Positioned over the equator.
    d. Remains stationary relative to the earth's surface.
    e. a and c.
    f. b, c and d.
18. Multiple repeaters in communications satellites are known as:
    a. Transponders.
    b. Detectors.
    c. Modulators.
    d. Stations.
19. In the current frequency assignments, how many frequency bands are there for the up-link frequencies?
    a. 16.
    b. 8.
    c. 4.
    d. 2.
20. The bandwidth required to send 132 voice-grade channels by FDM on an international satellite system is:
    a. 500 MHz
    b. 10 MHz
    c. 1320 MHz
    d. 50 MHz

Protocols and Error Control

ABOUT THIS CHAPTER

This chapter is about the rules that data communications systems use when communicating with each other, how they discover that an error has been made in transmission, and some methods that are used in correcting those errors. Since there are many data transmission schemes in use (for example, asynchronous and synchronous, half- and full-duplex), many sets of rules, called protocols, and many error detection and correction schemes have been devised. The most common of these techniques are explained in this chapter.

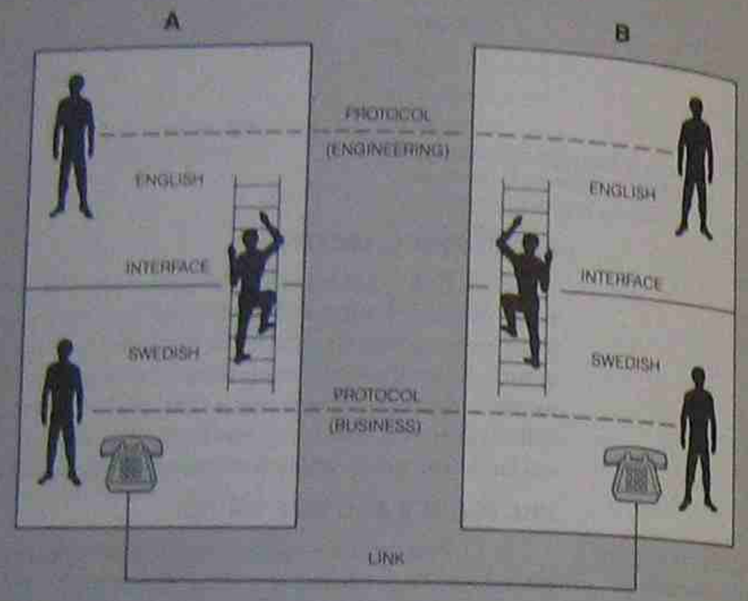
PROTOCOLS AND INTERFACES

A protocol is a set of rules defining the interactions between two machines or processes that are alike or that have similar functions. An interface is a set of rules controlling the interaction of two *different* machines or processes.

Protocols are agreements between people or processes (usually nowadays the processes are computer programs) about which may do what to whom, and when. A protocol is different from an interface. An interface is a set of rules, often embodied in pieces of hardware, controlling the interaction of two different machines or processes, such as a computer and a modem. A protocol, on the other hand, is a set of rules defining the interactions between two machines or processes that are alike or that have similar functions. The difference is illustrated in *Figure 7-1*. House A and House B look alike, but are occupied by a curious set of people. The British engineer on the top floor of house A needs to communicate with another engineer, also British, on the top floor of B. The only telephone, however, is guarded by a Swedish businessman on the bottom floor of each house. The engineer speaks no Swedish, and the Swede no English.

The two engineers, like most people who talk on the telephone, have a routine, worked out over long years, for communicating. The one answering says "Hullo." The one calling says "Hullo, this is Reggie." Then the first says "Oh, hullo, Reggie. How are you, old boy?" This is called a preamble. They then begin discussing Wimbledon, the weather, and even some business. At the end of most sentences, the one being questioned makes a reply, even if it is only a sound rather than a word or sentence. If one party does not hear some sort of reply from the other at fairly short intervals, he may say, "I say, are you still there?" If he still receives no reply, he may terminate the conversation and begin the whole process over. If one party does not understand the other, he will say, "What?", and the second party will repeat the last sentence. When the conversation is over, the person terminating the call will say, "Well, goodbye for now." The called party will reply, "Goodbye.", and both will hang up. The

Figure 7-1.
Protocols and Interfaces



Swedish businessman has a similar technique for conversing with his counterpart. All of these conventional conversational actions make up a protocol. As we shall see in the remainder of this chapter, protocols developed for communicating between processes in computers have all of the same mechanisms as those used between humans, for exactly the same reasons.

Protocols function for machine communication much like the formats and rules for human conversations and are used for the same reasons.

Even though our two engineers are able to communicate reliably with each other because they have a communications protocol, they still have the problem of no direct access to the communications channel, which is jealously guarded by the large Swedish businessmen. The solution is an interface. The engineers engage a person who stands midway up the stairs and who speaks both engineering English and business Swedish, and whose task it is to relay the conversation from the engineer (speaking in English) to the businessman (listening in Swedish), who relays it over the telephone to his Swedish counterpart in house B, who relays it in Swedish to another person halfway up the stairs, who finally conveys the messages in English to the called party. This seems like a clumsy and inefficient procedure and in many respects it is, but it is the same procedure used in millions of data communications systems to allow an applications program, communicating in one set of symbols (perhaps ASCII) and at one rate of symbol production, to communicate through a teleprocessing monitor using another set of symbols (perhaps EBCDIC) at a vastly different rate, to a line interface program dealing with yet a third set of symbols (bits). The story of standard interfaces is told in chapter four. The story in this chapter is about protocols.

Elements of a Protocol

There are three main elements to any protocol: 1) a character set, 2) a set of rules for message timing and sequence, and 3) error determination and correction procedures.

The basic elements of a communications protocol are a set of symbols, called a character set, a set of rules for the sequence and timing of messages constructed from the character set, and procedures for determining when an error has occurred in the transmission and how to correct the error. The character set will consist of a subset which is meaningful to people (usually called printing characters), and another subset which conveys control information (usually called control characters). There is a correspondence between each character and a group of symbols on the transmission channel. For instance the printing character A might correspond to the binary code 1000001. Several standard codes with equivalent sets of ones and zeros (bits), such as the ASCII discussed in a previous chapter, have been defined over the years. The set of rules to be followed by the sender and receiver gives the meaning, permissible sequences, and time relationships of the control characters and messages formed from the symbols. The error detection and correction procedure allows for the detection of and orderly recovery from errors caused by factors outside the control of the terminal at either end.

TELETYPEWRITER AND XMODEM PROTOCOLS

TTY protocols are the simplest ones in general use. The 58 characters in the character set of one TTY protocol were encoded by a 5-bit binary code.

The simplest protocols in general use are the ones associated with the transmission of start-stop or asynchronous data between teleprinter machines. For a long period after the development of the teleprinter, the machines were electromechanically controlled as we discussed in a previous chapter, and protocols had to be appropriate to a very simple mechanical controller in each machine. One of the first teletypewriter (TTY) protocols used a character set containing 58 symbols: 50 printing characters, a space, and seven control characters as shown in *Figure 7-2*. These symbols were encoded into a 5-bit binary code as shown (the octal representation is shown only for readers familiar with octal). The printing characters are the upper case letters, the numbers, and 14 punctuation marks. Three control characters allow for carriage return (CR), line feed (LF), and ringing the terminal bell (BELL) to announce a message. Also provided are a NULL (or blank) character (which is not the same as a space) and a WRU character. WRU stands for "Who Are You", and causes a mechanical device in the teleprinter to send a character sequence identifying that particular terminal. The other two control characters are the FIGS character, which causes the machine to print numbers and punctuation, and the LTRS character, which causes the machine to print the upper case alphabet. As mentioned in a previous chapter, this character set and associated binary code is commonly referred to as the Baudot code, after Emile Baudot,who invented the first code with symbols of a fixed length.

Figure 7-2.
CCITT Alphabet #2

BINARY	OCTAL	LTRS	FIGS
0 0 0 0 0	00	BLANK	BLANK
0 0 0 0 1	01	E	3
0 0 0 1 0	02	LF	LF
0 0 0 1 1	03	A	-
0 0 1 0 0	04	SPACE	SPACE
0 0 1 0 1	05	S	'
0 0 1 1 0	06	I	8
0 0 1 1 1	07	U	7
0 1 0 0 0	10	CR	CR
0 1 0 0 1	11	D	WRU
0 1 0 1 0	12	R	4
0 1 0 1 1	13	J	BELL
0 1 1 0 0	14	N	,
0 1 1 0 1	15	F	!
0 1 1 1 0	16	C	:
0 1 1 1 1	17	K	(
1 0 0 0 0	20	T	5
1 0 0 0 1	21	Z	+
1 0 0 1 0	22	L	)
1 0 0 1 1	23	W	2
1 0 1 0 0	24	H	£
1 0 1 0 1	25	Y	6
1 0 1 1 0	26	P	0
1 0 1 1 1	27	Q	1
1 1 0 0 0	30	O	9
1 1 0 0 1	31	B	?
1 1 0 1 0	32	G	&
1 1 0 1 1	33	FIGS	FIGS
1 1 1 0 0	34	M	.
1 1 1 0 1	35	X	/
1 1 1 1 0	36	V	=
1 1 1 1 1	37	LTRS	LTRS

The protocol for operation of a message system using the Baudot code goes something like this (the procedure varies somewhat according to the operator of the system):

1. A communications channel is connected between the two machines. This may be a temporary, dialed connection, or a permanent private line.
2. The sending machine sends a WRU, to verify that the receiving machine is the proper one for the message to be sent.
3. The sending machine sends a "Here is ...", which is normally the same sequence of characters which is sent when it receives a WRU, to identify the sender.
4. The sending machine sends a preamble or message header which identifies the name and address of the intended message recipient, the date and time of entry of the message into the system, the date and time of transmission, and the message sequence number assigned by the sender. In some systems, the sender will use two sequence numbers; one to count the number of messages sent by the sender that day, and the second to count the number of messages sent to the receiver's terminal that day.
5. At the end of the message, the sending machine sends another "Here is ..." and another WRU. The second WRU is to determine if the receiving machine is still connected to the line. The operator of the sending machine will, if the message is being sent manually, sometimes send several BELL characters, to alert a person at the receiving end that a message is ready for delivery.

This protocol has developed over many years of sending messages manually between individual terminals operated by people. You will note that it has all of the elements described in the opening paragraphs: a previously defined set of symbols and corresponding codes suitable for electronic transmission, a preamble, a message, and even a rudimentary error detection procedure (the invocation of the WRU at the beginning and end of the session). The same sort of protocol is used for teleprinter message systems even when one end of the system is a computer, as are all of the current public teleprinter networks.

Because the simple TTY protocol does not check for errors in the message, even low error rates will cause the received message to be gibberish and the sender will not know it.

The simple teletypewriter protocol described above does not work very well in the presence of even fairly low error rates on the path between the machines. Individual characters can be mutilated in such a way as to cause the receiving machine to begin printing gibberish, and the transmitter will never know. The WRU check at the end of a message assures only that the receiving machine received the WRU correctly and that the circuit is still good in the backwards direction, but says nothing about whether the message got through. To provide better assurance of correct transmission, two more techniques are sometimes used for simple protocols: character parity and echoplex.

Parity

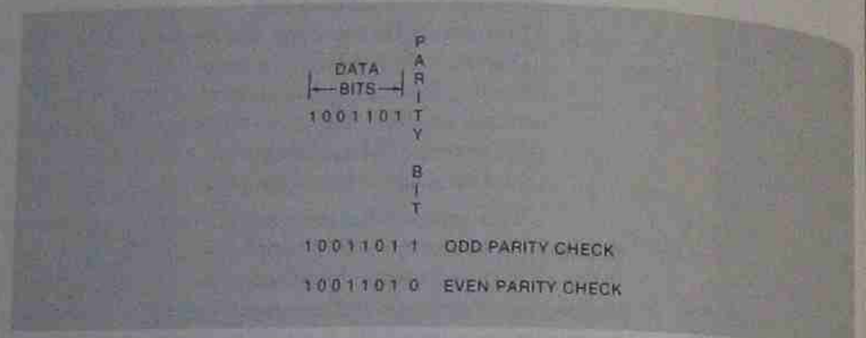
Parity is the technique where a bit is added to every symbol for the purpose of error detection. It can either be even or odd parity.

The Communicator's Creed: "Now abideth faith, hope, and parity, but the greatest of these is parity."

The addition by the transmitter of another bit to the bits that encode a symbol for the purpose of error detection is called parity. The bit is always transmitted, and is usually set to the value that will cause the coded symbol to have an even number of one bits; thus, the scheme is called even parity. The parity bit is recomputed by the receiver. If the newly computed value gives the correct parity, all is well. If not, some indication is given to the receiving terminal, usually by substituting a special error character for the one received.

For an example of parity checking, consider the ASCII character M, which has a bit composition of 1 0 0 1 1 0 1 (see *Figure 1-5*). Because the character has four 1 bits, a 1 bit would be added if odd parity checking is used, whereas a 0 bit is used if even parity checking is employed. *Figure 7-3* illustrates the formation of the parity bit for odd and even parity checking for the ASCII character M.

Figure 7-3.
Parity Bit Formation for ASCII Character M



The parity scheme will detect single-bit errors in the transmitted symbols, but not multiple-bit errors. Error correction can occur by the receiver sending a message back which requests retransmission.

Echoplex

Echoplex is an error detection technique in which each character of a message is retransmitted by the receiver back to the originator as it is received.

Echoplex is perhaps not properly classed as an element of protocols, but we will discuss it here anyway. It is the technique of sending back (or echoing) each character by the receiver as the characters are received. The sender can then see by the copy printed locally whether the characters are making the round trip without being mutilated. When an echoed character is received in error by the original sender, it is not possible to tell whether the data were received correctly at the destination and scrambled on the return path, or were erroneous when received at the original destination. But at least some error indication is immediately available to the sender.

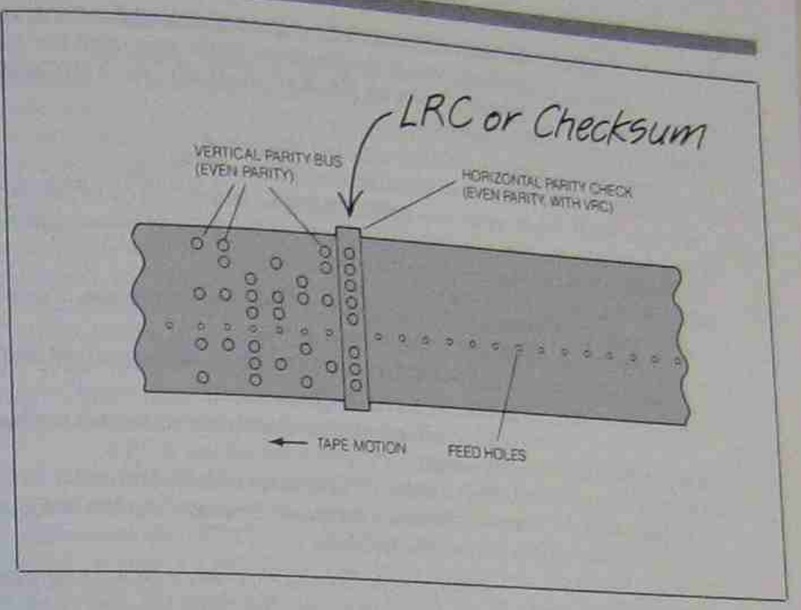
Checksums

In addition to character (or vertical) parity where a parity bit is added to each character, a block parity character may be added at the end of each message block.

In the discussion on parity above, the bit required to make the number of 1-bits in an individual character was added onto the end of the character and sent along with it. This scheme is sometimes called vertical redundancy checking (VRC), or vertical parity, because if one holds a punched paper tape with its length in the horizontal, each character will appear as a vertical column of holes across the tape. It is possible, and is common in some systems, to include a horizontal check character, which performs the parity function for each row of holes in the tape.

Figure 7-4 shows a set of characters represented as holes punched in a tape, with the vertical or character parity bit at the top of each character, and the horizontal or block parity character at the right. The block parity character is usually called a checksum, because it is formed by performing a binary addition without carry of each successive character. It also is sometimes referred to as a longitudinal redundancy check (LRC)

Figure 7-4.
Vertical and Horizontal Parity



character. It is sent as an extra character at the end of each message block. A system that uses both vertical parity and a checksum usually can detect all single-bit errors in a single character and some multiple-bit errors within a single character.

XMODEM Protocol

XMODEM protocol has an error checking technique that can be used between microcomputers. Messages are sent in blocks of 128 characters surrounded by control characters.

The XMODEM protocol was devised by Ward Christensen as a simple error checking scheme suitable for use between microcomputers. It requires that one terminal or computer be set up by an operator or computer program as the sender and the other be set up as the receiver. Once the protocol is started, the transmitter waits for the receiver to send a Negative Acknowledge (NAK) character. The receiver meanwhile is set to send NAKs every 10 seconds. When the transmitter detects the first NAK, it begins sending messages as blocks of 128 data characters, surrounded by some protocol control characters. The beginning of each block is signaled by a Start Of Header (SOH) character. This is followed by a block number character in ASCII, followed by the same block number with each bit inverted. The bit inversion is known as the 1's complement and results in the block number being followed by the 1's complement of

the block number. A 128-character piece of the file is sent, followed by a checksum that is the remainder of the sum of all the 128 bytes in the message divided by 255. Mathematically, the XMODEM checksum can be represented as

$$\text{CHECKSUM} = R\left[\frac{\Sigma \text{ ASCII Value of Character}_1^{128}}{255}\right]$$

where R is the remainder of the division process. *Figure 7-5* shows the XMODEM protocol block format.

The receiver checks each part of the received block as follows:

1. Was the first character an SOH?
2. Was the block number exactly one more than the last previous block received?
3. Were exactly 128 characters of data received?
4. Was the locally computed checksum identical with the last character received in the block?

If the receiver is satisfied, it sends an Acknowledge (ACK) back to the transmitter, and the transmitter sends the next block. If not, a NAK is sent, and the transmitter resends the block found in error. This process is continued, block by block, until the entire message or file is sent and verified. At the end of the data, the transmitter sends an End Of Text character. The receiver replies with an ACK and the session is terminated.

Even though the XMODEM protocol is very simple and easy to operate, it requires a computer at each end, and every file to be transferred must be set up manually.

There are several points to be made about the XMODEM protocol. First, it is very simple and easy to implement with a small computer, but it does require a computer at each end. Second, it requires manual setup for each file to be transferred. Third, the error detection technique (ordinary sum of the data characters) is unsophisticated and unable to detect reliably the most common type of transmission error, which is a noise burst which may last on the order of 10 milliseconds (the duration of about 12 bits at 1200 bps). Fourth, it is a half-duplex protocol; that is, information is sent, then the sender waits for a reply before sending the next message. Since operation of the XMODEM protocol generally assumes a full-duplex line, it is inefficient in use of the transmission facility.

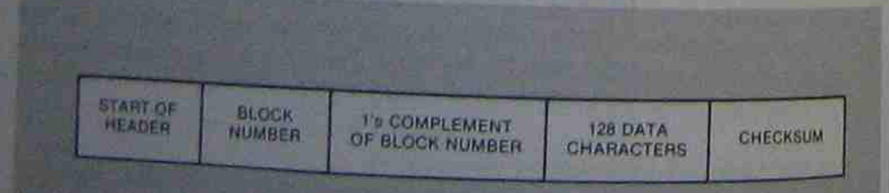


Figure 7-5. XMODEM Protocol Block Format

CONVOLUTIONAL CODING — CYCLIC REDUNDANCY CHECKS

Convolutional coding attaches a BCC to the end of each message block. The receiver recomputes BCC and compares it to the one transmitted to determine if the message was correctly received.

Several schemes have been devised to detect errors in binary communications systems using feedback or convolutional coding. These methods all append information computed at the transmitter to the end of each message transmitted to enable the receiver to determine if a transmission error has occurred. The added information is mathematically related to the messages and is therefore redundant. The receiver recomputes the value and compares the recomputed number with the number received. If the two are the same, all is well. If not, the receiver notifies the transmitter that an error has occurred, and the message is re-sent. These methods go by the name of Cyclic Redundancy Checking (CRC) and the values appended to the messages are called CRCs or Block Check Characters (BCCs). A CRC is calculated by dividing the entire numeric binary value of the block of data by a constant, called a generator polynomial. The quotient is discarded, and the remainder is appended to the block and transmitted along with the data.

The check character is determined by dividing the total binary value of the entire block by a constant called a generator polynominal.

CRC's usually are computed using multiple section feedback shift registers with eXclusive-OR (XOR) logic elements between each section and at the end. A typical arrangement is shown in *Figure 7-6*. This register implements the CCITT/ISO High-level Data Link Control (HDLC) CRC, which is called CCITT-CRC. The circles with a + in the middle represent XOR logic elements. For B = 0 or B = 1, the two rules for XOR are

1. B XOR V = 0
2. B XOR 0 = B

The shift register is initialized to all ones at the beginning of CRC calculation for a message. As each bit of the transmitted characters is applied to the transmission facility, it is also applied at point A of *Figure 7-6*, then the entire register is shifted right one bit. As the bits are transmitted and shifted, each one-bit that appears at A also affects the state of the other XOR elements, and that effect is propagated throughout the register for several bit times after the bit initially appears. Thus, any bit continues to have an effect on the transmitted data for a considerable time after that bit is sent. When the last data bit has been sent, the bits in the CRC shift register are complemented and transmitted.

At the receiver, the identical process is performed, and when the end of the message including the CRC is detected, the CRC is tested for the unique value 0001110100001111. If this value is found, all is well and the CRC register is reset to all ones for the next message. If the special value is not found, the program is notified that a transmission error has occurred, and a negative acknowledgement is returned to the sender.

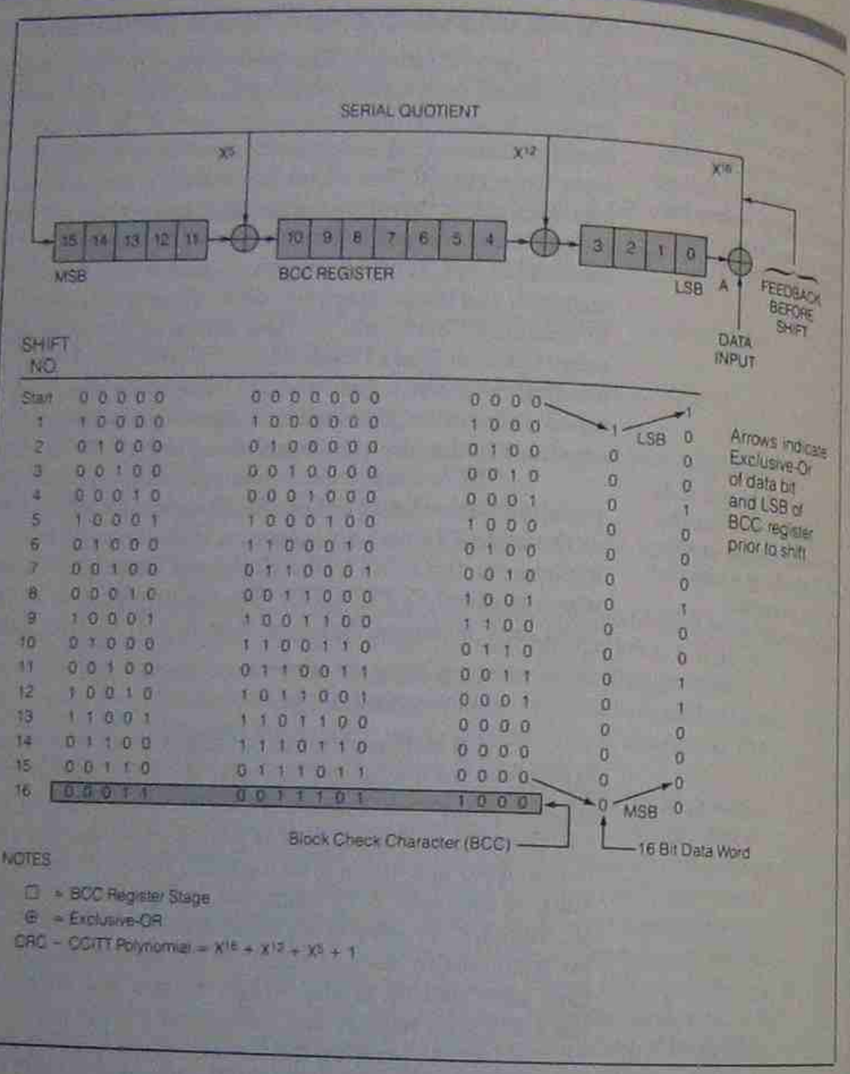


Figure 7-6. Shift Register Implementation of CCITT-CRC *(Source: John E. MacNamara, Technical Aspects of Data Communication, © 1977, Digital Press, Maynard, Mass.)*

The CRC error detection process has a very high rate of success in accurately detecting error bursts of up to 16 bits because of the history held in the shift registers.

The CRC process has the advantage that the current state of the shift register is the result of considerable past history. It is therefore unlikely that a burst of errors, such as normally occurs in serial data transmission, will produce a calculated CRC at the receiver that is the same as that which was originally sent. In fact, the 16-bit CRC calculation procedures, CRC-16 and CRC-CCITT, will detect all error bursts of 16 bits or less in length. They will also detect over 99.9% of error bursts of greater than 16 bits. Another advantage of CRC is that it does not require the addition of another bit per character sent as do the VRC and LRC schemes. It does, however, require the sending of one or two (two for CRC-16 and CRC-CCITT) extra characters at the end of each transmitted block.

CRC algorithms are usually implemented in hardware, and integrated circuits have been developed to do the entire process for more than one method (e.g., CRC-12, CRC-16, and CRC-CCITT). It is also possible to do the calculations in software, with table lookup techniques providing reasonable performance.

HALF-DUPLEX PROTOCOLS

Links

Data links, either point-to-point or multipoint, are the data communications equipment and circuits that allow two or more terminals to communicate.

The basic notion in link protocols is that of the data link. A data link is an arrangement of modems or other interface equipment and communications circuits connecting two or more terminals that wish to communicate. Probably the most widely used link protocol is comprised of the Binary Synchronous Communications procedures defined by IBM[1] (usually abbreviated as BiSync or BSC). These procedures allow for operation on a data link in one direction at one time. BiSync can be operated on full-duplex circuits, but information still flows in only one direction at a time, so that in many cases the advantage of the full-duplex circuit is minimized.

For the purpose of discussion of protocols, the physical form of the link is important in that the procedures for connecting and disconnecting the link may be different depending on whether the link is permanently connected or dial-up, and also depending on the delay between when a data bit is sent and when it is received. This latter factor is of particular concern on satellite links due to the large difference in round-trip propagation delay between satellite links and terrestrial links. Regardless of the physical form of the link, however, the data are sent over it as a serial stream of symbols encoded as bits, and the control procedures between the ends of the link are effected using the transmission and recognition of line-control characters or control codes.

Point-to-Point Links

A point-to-point link is one that connects only two stations at one time as *Figure 7-7a* shows. Point-to-point links may be established on dedicated or dial-up circuits, and they may be half-duplex or full-duplex.

Multipoint Links

Multipoint links connect more than two stations at one time as *Figure 7-7b* shows. Obviously, some control procedure must be in place to designate which stations may use the link at any one time. For this purpose, one station on the multidrop lines is designated as the control station, and all other stations are designated as tributaries. The control station is the traffic director, designating which stations are to use the link by a polling and selection process. At any instant in time, transmission on a multipoint link will be between only two stations, with all other stations on the link in a passive receive mode.

[1]IBM is a trademark of International Business Machines Corporation.

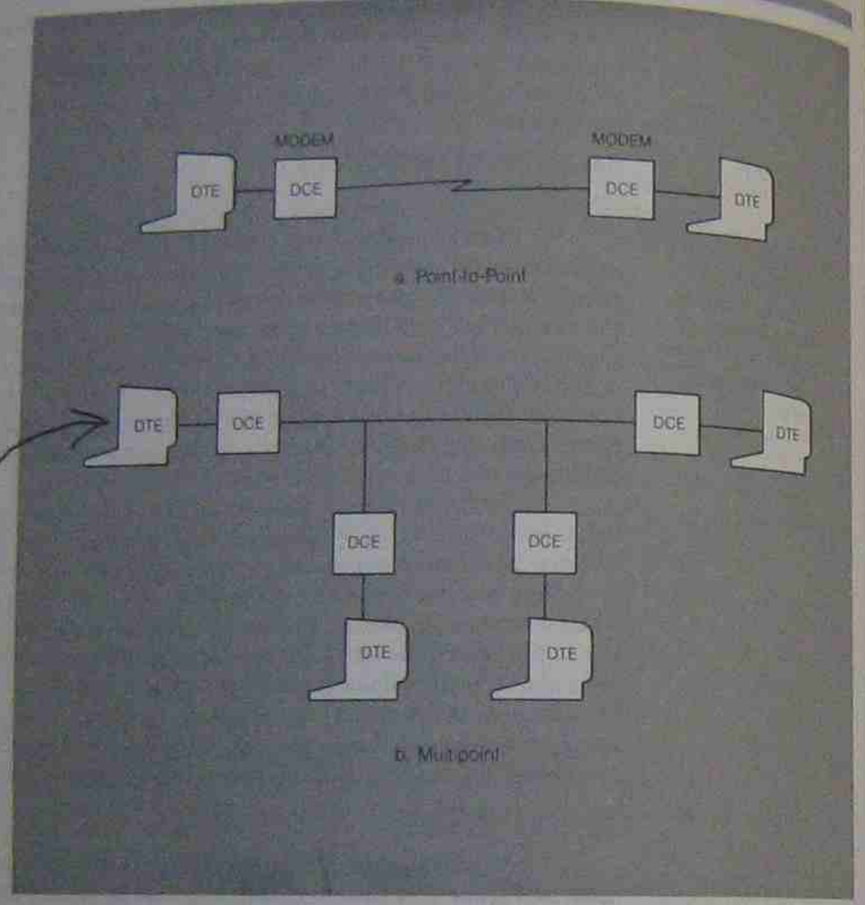


Figure 7-7.
Point-to-Point and Multipoint Links

One station is the control station; all others are tributaries.

Transmission Codes — Character Sets

IBM's BiSync protocol is one of the most widely used link protocols. It can be used with three character sets: Six-Bit Transcode, Extended Binary Coded Decimal Interchange Code, and American Standard Code for Information Interchange.

BiSync is defined by IBM to accommodate three character sets and their associated binary codes. Each set consists of a set of graphics (letters, numbers, and punctuation), a set of terminal control and format control codes (BELL, Form Feed, WRU, etc), and a set of data link control codes (Start of Text, End of Transmission, etc). The three sets are the: 1) Six-Bit Transcode (SBT), 2) Extended Binary Coded Decimal Interchange Code (EBCDIC), and 3) United States of America Standard Code for Information Interchange (USASCII, or more commonly, just ASCII).

The codes differ in number of bits per symbol encoded (6 for SBT, 7 for ASCII, and 8 for EBCDIC), and different numbers of characters in the character set (64 for SBT, 128 for ASCII, and 144 for EBCDIC). There are significant differences between the sets in such properties as the sorting order between letters and numbers.

Various link control codes are used in BiSync to gain control of the data link and ensure that the proper actions occur.

Link Control Codes

Link control is effected by the proper recognition of control characters, and the taking of appropriate actions. The control codes used in BiSync are defined as follows:

Synchronous Idle (SYN) — Establishes and maintains synchronization between DCEs, and is used as a filler between data blocks.

Start of Header (SOH) — Identifies the beginning of a block of heading information. Headers are control information (such as addresses, priority, and message numbers) used by the system to process the text part of the message.

Start of Text (STX) — Identifies the end of a header and the beginning of a block of text. Text is the part of the message from an applications program which is destined for another applications program and which must pass through the communications system unchanged.

End of Transmission Block (ETB) — Identifies the end of a block started with an SOH or STX. A BCC is sent immediately following an ETB. Receipt of an ETB requires a status reply (ACK0, ACK1, NAK, WACK or RVI).

End of Text (ETX) — Terminates a block of data started by an STX or SOH which was transmitted as an entity. A BCC is sent immediately following an ETX. ETX also requires a status reply.

End of Transmission (EOT) — Indicates the end of message transmission by this station; the message may have contained one or more blocks. Receipt of an EOT causes all receiving stations to reset. The EOT is also used as a response to a poll when the polled station has nothing to send, and as an abort signal when the sender cannot continue transmission.

Enquiry (ENQ) — Requests a repeat transmission of a response to a message block if the original response is garbled or is not received. ENQ may also indicate the end of a polling or selection sequence, and is used to bid for the line when the line is a point-to-point connection.

Affirmative Acknowledgement (ACK0 or ACK1) — Indicates correct receipt of the last previous block, and that the receiver is ready to accept the next block. ACK0 and ACK1 are used alternately, and receipt of the wrong one is an indication of an error in the protocol. ACK0 is the correct response to a station selection (on a multipoint circuit) or a line bid (on a point-to-point circuit).

Wait-Before-Transmit Affirmative Acknowledgement (WACK) — Indicates a temporary not-ready-to-receive condition to the sending station and affirmative acknowledgement of the last previously received data block. The normal response to WACK by the sending station is ENQ, but EOT and DLE EOT are also valid. If ENQ is received after sending WACK, the sending station continues to send WACK until it is ready to continue sending data.

Negative Acknowledgement (NAK) — Indicates that the last previously received block was received in error, and that the receiver is ready to receive another block. NAK is also used as a station-not-ready reply to a station selection or line bid.

Data Link Escape (DLE) — Indicates to the receiver that the character following the DLE is to be interpreted as a control character.

Reverse Interrupt (RVI) — Indicates a request by a receiving station that the current transmission be terminated so that a high-priority message can be sent (such as a shutdown notice). RVI is also an affirmative acknowledgement to the last previous block received. In a multipoint circuit, RVI sent from the control station means the control station wishes to select a different station. When a sending station receives an RVI, it responds by sending all data that would prevent it from becoming a receiving station.

Temporary Text Delay (TTD) — Indicates that the sending station is not ready to send immediately, but wishes to keep the line. The receiving terminal reply to TTD is NAK. TTD normally is sent after two seconds if the sender cannot transmit the next block; this holds off the normal three-second abort timer at the receiving terminal. The response to TTD is NAK, and TTD may be repeated several times. TTD also is used as a forward abort, by sending EOT in response to the NAK reply.

Switched Line Disconnect (DLE EOT) — Indicates to a receiver that the transmitter is going to hang up on a switched line connection.

Code Sequences

Some control characters require a two-character sequence of standard characters.

Reference to the ASCII and EBCDIC code charts will reveal no code assigned to several of the control characters mentioned above. Such control codes are represented by two-character sequences of the characters that are defined in the charts. *Table 7-1* gives the correspondence between these control characters and sequences of standard characters.

Table 7-1. Character Sequences for BiSync Control Characters

BiSync Data Link Character	Character Sequences for Various Code Sets		
	ASCII	EBCDIC	SBT
ACK0	DLE 0	DLE 70	DLE -
ACK1	DLE 1	DLE /	DLE T
WACK	DLE :	DLE ,	DLE W
RVI	DLE	DLE @	DLE 2
TTD	STX ENQ	STX ENQ	STX ENQ

Polling and Selection

In a BiSync link, the control station determines which tributary station is active on the link, as well as the direction of transmission.

The active participants on a BiSync link are managed by a control station which issues either a Poll or a Select message addressed to the desired tributary. The Poll is an invitation-to-send from the control to the tributary. This allows the tributary to send any messages desired to the control station. The Select is a request-to-receive notice from the control station, telling the tributary that the control station has something to send to it. The control station thus controls which station has the link, and the direction of transmission.

Stations on the data link are assigned unique addresses which may consist of from one to seven characters. The first character defines the station, and succeeding characters define some part of the station, such as a printer, as required. Some BiSync implementations repeat the station address for reliability purposes.

Message Blocks

Messages consist of one or more blocks of text, called the body of the message, surrounded by synchronization, header and error control characters. The beginning of each block is identified by the STX control character, and all blocks except the last in a message are ended either by an ETB or an ITB character. The last block of the message is ended by an ETX character.

Error Checking

Three types of error detection modes are used by BiSync: An odd parity check of each character including VRC/LRC, a BCC comparison using CRC-12, and a BCC comparison using CRC-16.

BiSync uses three types of error detection: VRC/LRC, CRC-12, and CRC-16. *Table 7-2* shows under what conditions each is used. Transparency mode is described later in this section. The VRC is an odd-parity check performed on each character transmitted, including the LRC character at the end of the block. Each bit in the LRC character provides odd parity for the corresponding bit position of all characters sent in the block. *Figure 7-8* illustrates the formation of the VRC and LRC for an eight character data block. Note that the LRC is formed in the same manner as the VRC; however, the LRC covers all characters in the data block, whereas the VRC covers individual characters.

In BiSync, the LRC character is called the Block Check Character (BCC) and is sent as the next character following an ETB, ITB, or ETX character. The BCC sent with the data is compared at the receiver with one accumulated by the receiver, and if the two are equal, all is well. The BCC calculation is restarted by the first STX or SOH character received after the direction of transmission is reversed (called a line turnaround). All characters except SYN characters received from that point until the next line turnaround are included in the calculation. If the message is sent in blocks with no line turnaround, each block will be followed by an ITB, then the BCC. The BCC calculation is then restarted with the next STX or SOH character received.

Figure 7-8.
VRC/LRC Parity Check

CHARACTER	DATA BITS								CHARACTER PARITY BIT (VRC)
1	1	0	1	0	1	1	0	1	0
2	0	1	1	0	1	0	1	0	1
3	0	1	1	1	0	1	0	1	0
4	1	0	1	0	0	1	0	1	1
5	1	0	0	1	0	0	1	0	0
6	0	1	0	0	1	0	0	1	0
7	1	0	0	1	0	0	1	0	0
8	0	1	1	0	1	0	0	1	1
BLOCK PARITY CHARACTER (LRC)	1	1	0	0	1	0	0	1	0

Table 7-2.
Type of Redundancy Check

Code Set	No Transparency	Transparency Installed and Operating	Transparency Installed but Not Operating
EBCDIC	CRC-16	CRC-16	CRC-16
ASCII	VRC/LRC	CRC-16	VRC/CRC-16
SBT	CRC-12	CRC-12	CRC-12

The Cyclic Redundancy Check codes are used for error checking in the same fashion as the LRC code described above. If the transmission code set is SBT, the CRC-12 method is used, since each transmitted character is only six bits. For EBCDIC, the CRC-16 scheme is always used. For the ASCII code set, IBM has specified that the VRC/LRC scheme be used in the non-transparency (standard) mode, and CRC-16 be used in transparency mode. Transparency mode is described in detail below.

Message Formats

Information is carried in BiSync as messages, and each message may have several parts: a synchronization sequence, a header, some text, and a block check sequence. Each part is identified by one or more control characters. In the case of messages used only for control, some parts such as the header, the text, or the BCC may be missing.

Synchronization

Establishing the synchronization required between sender and receiver requires a three step sequence: Bit synchronization, character synchronization, and message synchronization.

Unless the transmitter and receiver agree on the exact (to the bit) starting point of a message, they cannot communicate. Achieving this synchronization requires three steps:

1. The modems or other data communications devices at both ends of the circuit must acquire bit synchronization. The methods for doing this are discussed in Chapter 5.
2. The link interface equipment in the DTE must acquire character synchronization. This is done by searching the bit stream for a specific pattern of bits called a Synchronization Character (SYN). In ASCII, the bit pattern for SYN is 0010110. To help insure against recognition and a false SYN, most systems including BiSync require transmission and detection of two successive SYNs before the next step in the synchronization process is taken. The two SYN characters taken as a control set are called a Sync Sequence, and are shown in most diagrams as a 0. In order to insure that the first and last characters of a transmission are correctly sent, some BiSync stations add a PAD character before the first SYN and after every BCC, EOT, and ACK. This strategem was devised primarily to overcome hardware limitations in some early hard-wired BiSync terminals.
3. The program operating the protocol must acquire message synchronization. In other words, the program must be able to find the beginning of each message or control character sequence. This is achieved by a program search of the characters received after the sync sequence for a control character which is defined to begin a block of data or control sequence. Such characters are SQH, STX, ACK0/1, NAK, WACK, RVI, and EOT.

Heading

The character SOH begins the heading used in identifying the message type, numbering, priority and routing.

A heading is a sequence of characters beginning with the character SOH that is used for message type identification, message numbering, priority specification, and routing. Receipt of the SOH initiates accumulation of the BCC (but the SOH is not included in the BCC). The heading may be of fixed or variable length, and is ended by an STX.

Text

The text part of the message contains the information to be sent between application programs. The text begins with an STX and ends with an ETX. Text may be broken up into blocks for better error control. Each block begins with an STX and ends with an ETB, except for the last block which ends with an ETX. The normal end of a transmission is signaled by sending an EOT following the ETX. Control characters are not allowed within the body of a text block. If a control character is detected, the receiving station terminates reception and looks for a BCC as the next two characters.

Timeouts

Four BiSync time-outs functions prevent indefinite delays due to data errors or missing line turnaround signals.

Timeouts must be provided by the communications program or terminal to prevent indefinite delays caused by data errors or missing line turnaround signals. Four functions are specified in BiSync for timeouts: transmit, receive, disconnect, and continue.

Transmit timeout is normally set for one second, and defines the rate of insertion of synchronous idle (SYN SYN or DLE SYN) sequences in the data to help maintain bit and character synchronization.

Receive timeout is normally set for three seconds, and does several things:

1. It limits the time a transmitting station will wait for a reply.
2. It signals a receiving station to check the line for synchronous idle characters, which indicate that transmission is continuing. The receive timeout is reset each time a sync sequence is received.
3. It sets a limit on the time a tributary station in a multipoint circuit may remain in control mode. The timer is reset each time an end signal such as an ENQ or ACK is received as long as the station stays in control mode.

Disconnect timeout causes a station on a switched network to disconnect from the circuit after 20 seconds of inactivity.

Continue timeout causes a transmitting station sending a TTD to send another if it is still unable to send text. A receiving station must transmit a WACK if it becomes temporarily unable to receive within a two-second interval.

Transparent-Text Mode

Data communications often requires the sending of information that is in binary instead of in a character code. A transparent-text mode is used for this purpose.

It is often necessary in communicating between machines to send data that do not represent characters but instead represent some purely arbitrary quantity or object. An example is the binary representation of a computer program. In such a case, it is likely that some of the data will have the same bit pattern as a BiSync control character. Transparent-text mode, sometimes called transparency, allows such data to be sent without being misinterpreted by the communications program. The basic technique in transparency is to precede each true control character with the DLE character as *Figure 7-9a* shows. If a DLE bit pattern appears in the text of the message, it also is preceded by a DLE. Thus, a bit pattern is interpreted as a control character only if preceded by a DLE. The resulting message after processing by the link interface program to remove the DLE is shown in *Figure 7-9b*. Note that any DLE DLE sequence in the text has the first DLE suppressed and the second sent along as part of the data.

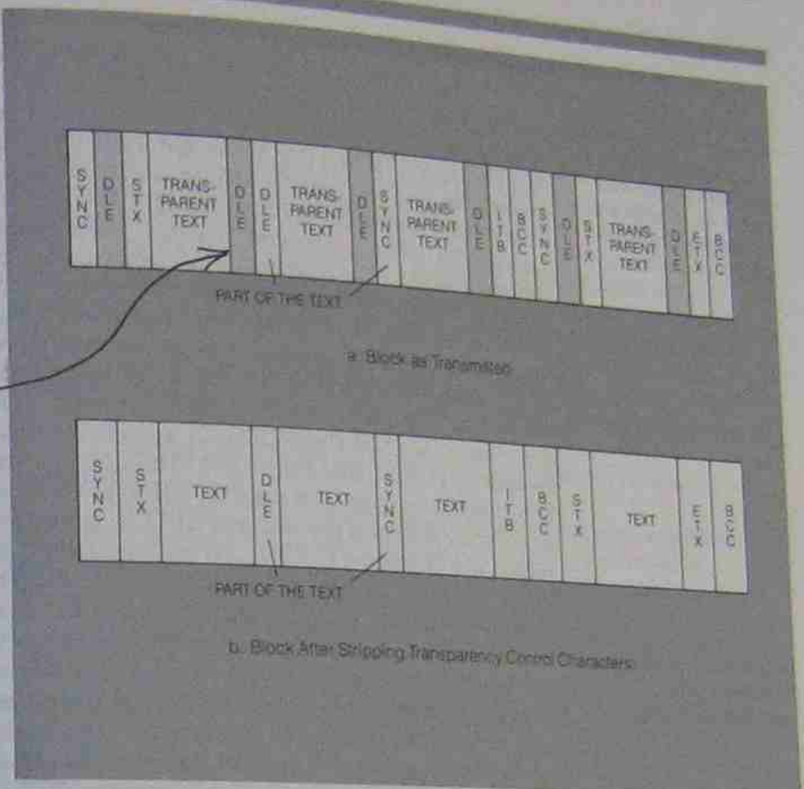
FULL-DUPLEX PROTOCOLS

On-line applications using CRT terminals required a protocol for full-duplex operations that contained a powerful error detection and correction system that prevented aliasing.

The Binary Synchronous Communications procedures were devised at a time when most data communications circuits were operated at 2400 bits per second, and were two-wire half-duplex circuits connecting remote job entry card reader/printer terminals to mainframes. As on-line applications using remote CRT terminals became more cost-effective, and four-wire private line circuits became available, the demand for a communications protocol that could handle full-duplex operations arose. The requirements for such a protocol were:

1. Messages must flow in both directions simultaneously.
2. It must be possible to have more than one message in the channel at one time (BiSync allows only one).
3. Transparency must be designed in, not tacked on as an afterthought.

Figure 7-9. BiSync Transparency Mode



Added if DLE appears in text.

4. The protocol must allow switched network, half-duplex, and multipoint operation, as well as full-duplex point-to-point.
5. The error detection and correction scheme must be powerful, and must prohibit the problem of aliasing.

The last condition is a particularly difficult one. Aliasing occurs when a message fragment caused by a transmission error is interpreted as a good message at the receiver; that is, when a bad or broken message "looks like" a good message. This can be a serious problem, especially in critical applications such as funds transfer and military command and control. Much design effort in the international standards organizations and in private companies was devoted to development of protocols that prevented the aliasing problem. This led to the development of three widely used schemes: two of them were alike in principle and operation and in requiring special hardware, but promulgated by different organizations. The third, developed at Digital Equipment Corporation, uses a different technique, but can operate with standard character-oriented line interface equipment.

High-level Data Link Control Procedures

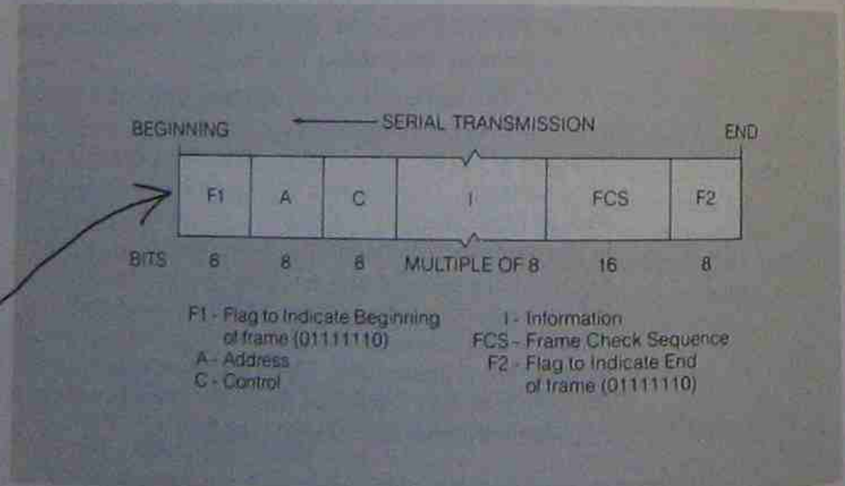
Protocols to prevent aliasing must determine where a true message block begins and ends and what part of the message is to be included in the CRC. HDLC is one of these.

The critical issue in preventing aliasing is the determination of where a legal message block begins and ends, and exactly what part of the information taken in by the receiver is to be included in the CRC. As a result of work done at IBM by J. Ray Kersey and others in the early 1970s, a protocol was proposed and later adopted as an international standard by the International Standards Organization in 1979. It is called the High-level Data Link Control (HDLC) procedures. In HDLC, the message synchronization indicator (called a Flag) is generated by a hardware circuit, and other hardware circuits prevent any data being transmitted from having the same pattern as the Flag. The Flag then becomes a kind of out-of-band framing signal, much like the break signal in the teletypewriter protocol. Since the data being transmitted are examined on a bit-by-bit basis to screen out possible aliases of the Flag; HDLC and other similar protocols, such as IBM's Synchronous Data Link Control (SDLC), are referred to as bit-oriented protocols, or BOPs. Unlike BiSync and DDCMP (which will be described later) the text part of messages sent using the HDLC protocol may, in principle, be any arbitrary number of bits long, and HDLC is defined to allow this. In practice, like BiSync and DDCMP, real implementations of BOPs restrict the text (and in fact all of the message including the Flag) to be an integral multiple of the number of bits in a character (almost always eight).

Data is examined on a bit-by-bit basis in HDLC, thus, it is called a bit-oriented protocol.

In HDLC, all information is carried by frames that may be of three types: information frames (I-frames), supervisory control sequences (S-frames), or unnumbered command/responses (U-frames). *Figure 7-10* shows one information frame as a rectangular block divided into its six fields: a beginning Flag (F1) field, an Address field (A), a Control field (C), an Information field (I), a Frame Check Sequence (FCS) field, and a final Flag (F2) field. S-frames and U-frames have the same fields except that the I field is left out.

Figure 7-10.
HDLC Format *(Source: J.L. Fike and G.E. Friend, Understanding Telephone Electronics, © 1983, Texas Instruments Incorporated)*



Each information frame is divided into six fields.

I-frames perform information transfer, and independently carry message acknowledgements, and Poll or Final bits. S-frames perform link supervisory control such as message acknowledgements, retransmit requests, and requests for temporary holds on I-frame transmissions (like a WACK in BiSync). U-frames provide a flexible format for additional link control data by omitting the frame sequence numbers and thus providing a place for an additional 32 command and 32 response functions. The fields in the HDLC frame are used as follows:

Flag field: Every frame begins and ends with a Flag, which is the bit pattern 01111110. The same Flag may end one frame and begin the next. Every station connected to a link must continuously search the received data for the Flag.

Address field: In command frames, the address identifies the destination station for the command. In response frames, the address identifies the station sending the response.

Control field: The control field carries commands and responses, according to *Table 7-3*.

In HDLC, there are six fields in an information frame, but only five fields in other frames, which are supervisory control sequences or unnumbered command/responses.

Information field: The information field may contain any sequence of bits, which in principle need not be related to a particular character set or data structure. In practice, the information field is almost always an integral multiple of one character in length, which is usually eight bits.

Frame Check Sequence (FCS) field: The FCS (or CRC) for HDLC is the CCITT-CRC as discussed above.

In order to ensure that the Flag is unique, the transmitting hardware must monitor the bit stream continually between the beginning and ending

Table 7-3.
Control Field Contents

HDLC Frame Format	Bits in Control Field								
	1	2	3	4	5	6	7	8
I-frame (Information transfer commands/responses)	0	N(S)[1]			P/F[2]	N(R)[3]		
S-frame (Supervisory commands/ responses)	1	0	S[4]		P/F	N(R)		
U-frame (Unnumbered commands/ responses)	1	1	M[5]		P/F	M		

Notes:
[1] N(S) is the transmitting station send sequence number. Bit 2 is the low-order bit.
[2] P/F is the Poll bit for primary station transmissions, and the final bit for secondary station transmissions.
[3] N(R) is the transmitting station receive sequence number. Bit 6 is the low-order bit.
[4] S are supervisory function bits.
[5] M are modifier function bits.

Flags for the presence of strings of five one-bits in a row. If such a string occurs, a zero-bit is inserted (called bit stuffing) by the hardware after the fifth one-bit so the string will not "look like" a Flag. The added zero-bit is removed at the receiver. This procedure makes any appearance of strings of one-bits greater than five in number either a Flag, an error on the transmission line, or a deliberately sent fill pattern between frames of all one-bits. One method of aborting transmission of a frame is to begin transmitting continuous ones.

Specification in the protocol of a unique, hardware-generated Flag pattern and the length of the Address, Control, and FCS fields provides complete transparency for the Information field. The hardware prevents any bit pattern sent by the applications program from having more than five continuous ones, and the hardware-added zeros are stripped from the data stream as it passes through the receiver. Also, the position of the FCS is defined by receipt of the ending Flag, so that the residual pattern in the CRC register can be immediately compared with the fixed value (0001110100001111).

The order of bit transmission is defined for addresses, commands, responses, and sequence numbers to be low-order bit first, but the order of bit transmission for the information field is not specified. The FCS is transmitted beginning with the coefficient of the highest order term (x^{15}). An invalid frame is defined as one that is not properly bounded by Flags, or is shorter than 32 bits between Flags. Invalid frames are ignored (in contrast with frames that have a bad FCS, which require a NAK).

HDLC Semantics

A normal message sequence involves sending frames from a transmitting station (data source) to a receiving station (data sink), and having the receiving station acknowledge receipt of the transmission by sending a frame back to the sender.

The above discussion dealt with what is called the syntax of HDLC. The syntax is the definition of the bit patterns and order of sending bits that will make correctly formed messages; that is, those that are legal in the protocol. In order to understand what happens when HDLC is used, however, it is necessary to define the semantic content of messages; that is, the meanings that should be assigned to correctly formed messages.

The normal sequence of messages in HDLC consists of the transfer of one or more frames containing I-fields from a data source (the transmitting station) to a data sink (the receiving station). The receipt is acknowledged by the sink by sending a frame in the backwards direction. The source must retain all transmitted messages until they are explicitly acknowledged by the sink. The value of N(R) indicates that the station transmitting the N(R) has correctly received all I-frames numbered up to N(R)-1. I-frames and S-frames (see *Table 7-3* above) are numbered, the number going from 0 to 7 (for unextended control fields). An independent numbering sequence is carried for each data source/data sink combination. The response from a sink may acknowledge several (but not more than 7) received messages at one time, and may be contained in an I-frame being sent from the sink to the source.

A data link consists of two or more communicating stations, so for control purposes it is necessary to designate one station as the primary station, with responsibility for managing data flow and link error recovery procedures. Primary stations send command frames. Other stations on the link are called secondary stations, and they communicate using response frames. Primary stations can send to secondaries using the Select bit in the control field of an I-frame, or a primary may allow a secondary to send by sending a Poll bit.

Secondary stations may operate in one of two modes; a Normal Response Mode (NRM) or an Asynchronous Response Mode (ARM). In NRM, the secondary may send only in response to a specific request or permission by the primary station. The secondary explicitly indicates the last frame to be sent by setting the final bit (F) in the control field. In ARM, the secondary may independently initiate transmission without receiving an explicit permission or Poll from the primary.

Space does not permit covering HDLC in greater detail. The interested reader is referred to International Standards ISO 3309-1979 (E), ISO 4335-1979 (E), and ISO 6256-1981 (E) for the complete definition of the HDLC protocol. There is a data link control procedure standard promulgated by the American National Standards Institute (ANSI) in the United States called Advanced Data Communications Control Procedures (ADCCP), which is functionally equivalent to HDLC.

Synchronous Data Link Control

SDLC developed by IBM has the same frame format and is similar in function to HDLC.

The standard full-duplex synchronous data link control protocol used by the IBM Corporation in products conforming to their System Network Architecture is called Synchronous Data Link Control (SDLC). It is functionally equivalent to HDLC, but with these exceptions:

1. SDLC information fields must be an integral multiple of 8 bits long.
2. Current IBM products support only the Unbalanced Normal operation with Normal Disconnected mode class or procedures as defined in ISO DIS 6159 and DIS 4335/DAD1.
3. SDLC contains additional commands and responses not defined in the ISO Elements of Procedure, for example, a TEST command and response. *Figure 7-11* shows the frame structure of SDLC (which is the same as HDLC) along with the common modes, commands, and responses. A close examination of this figure will help you understand the SDLC format and the differences between HDLC and SDLC.

Digital Data Communications Message Protocol

DDCMP is character oriented rather than bit oriented.

At about the same time SDLC was being developed at IBM, George Friend, Steven Russell, and Stuart Wecker were given the task of developing a synchronous protocol at Digital Equipment Corporation. The requirements were similar to those given above for full-duplex protocols, but with one important additional item: the protocol must run on existing data communications hardware, and preferably on asynchronous as well as synchronous links. The result was the Digital Data Communications Message Protocol, or DDCMP.

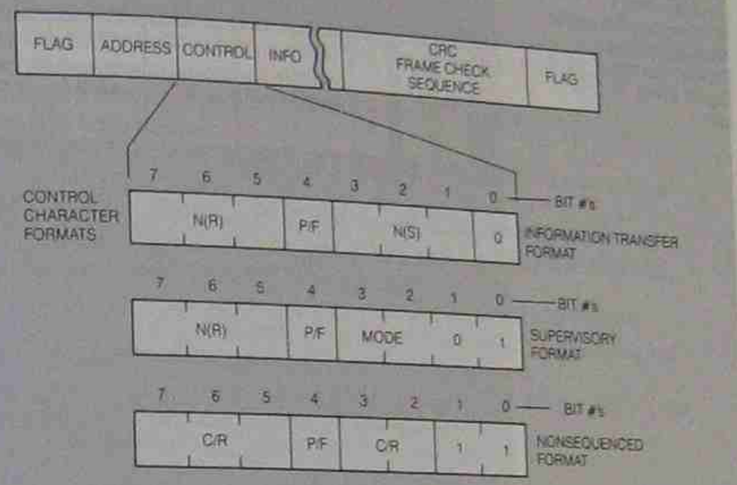
DDCMP is a character oriented protocol rather than a bit oriented protocol as is SDLC. Therefore, DDCMP requires no special bit stuffing and destuffing hardware and can be used with various types of line interfaces, including asynchronous units. The method of specifying the length of a message in order to look for the BCC at the right time is inclusion of a 14-bit count field in the header. This is a count of the number of characters in the information field of the message. Because error-free

transmission depends on the count field being detected correctly, the header of DDCMP messages is a constant length and has its own BCC, which is checked before setting up to receive the information part of the message. The format of DDCMP messages is shown in *Figure 7-12*. A detailed study of this figure reveals the bit pattern of the various fields in the frame for different types of messages.

The DDCMP does not require any special hardware to perform successfully on synchronous, asynchronous or parallel data channels.

DDCMP has two principal advantages and some disadvantages. One advantage is that it is usable without special hardware on asynchronous, synchronous, or even parallel data channels. Another advantage is that the sequence number field allows for up to 255 messages to be outstanding on the channel at one time, a requirement when operating full-duplex channels over satellite links. On the other hand, it does not support a go-back-N or selective reject mode of operation, and it is, in principle, more vulnerable to the aliasing problem than bit oriented protocols. It also has been mistakenly criticized for inefficiency because of the inclusion of a BCC after the header on the equivalent of I-frames. In fact, for messages of average length, DDCMP is somewhat more efficient than bit-stuffing protocols because of the extra bits added by the hardware in the bit oriented schemes. In practice, aliasing has not caused noticeable problems with DDCMP.

Figure 7-11. SDLC Format *(Source: Kenneth Sherman, Data Communications: A User's Guide, © Reston Publishing Co., Reston, VA.)*



Flag is always 01111110 (If a sequence of 5 "1" bits occurs in the middle of a transmission, the transmitter "stuffs" a "0" bit and the receiver removes it.)
Address is the Station Address on a particular line
N(S) is Sending Sequence Number = 000 through 111 and then repeats
P/F is Primary Station Poll when = 1 from Primary Station
Secondary Station Final when = 1 from Station on the line
N(R) is Receiving Sequence Number = 000 through 111 and then repeats
Mode is Receive Ready (RR) when 00
Receive Not Ready (RNR) when 10
Reject (REJ) when 01
C/R is Command from Primary Station or Response from Secondary Station

Commands:
- Nonsequenced Information (NSI) = 000-00
- Set Normal Response Mode (SNRM) = 001-00
- Disconnect (DISC) = 010-00
- Optional Response Poll (ORP) = 100-00
- Set Initialization Mode (SIM) = 000-10

Responses:
- Nonsequenced Information (UI) = 000-00
- Nonsequence Acknowledgment (UA) = 110-00
- Request for Initialization (RIM) = 000-10
- Command Reject (FRMR) = 001-10
- Request Online (DM) = 000-11

Info is Information = Variable Length for Information Transfer
Prohibited for Supervisory Format
Variable Length for Nonsequenced Format with NSI
Fixed Format with CMDR
CRC = Cyclic Redundancy Check Remainder

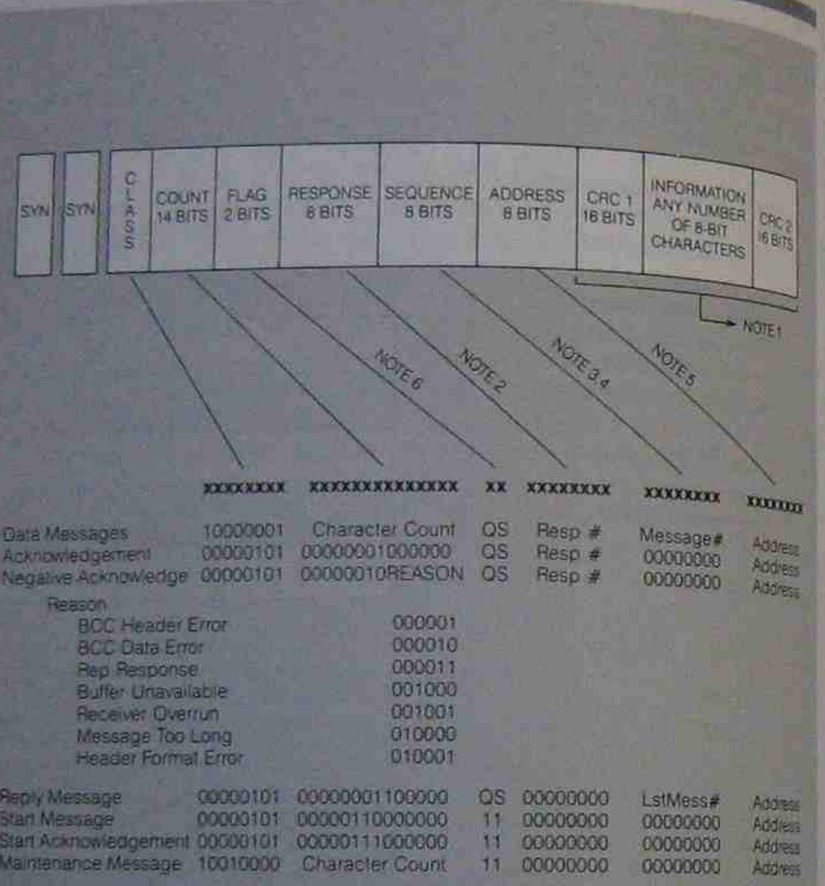


Figure 7-12. DDCMP Message Format in Detail *(Source: John E. MacNamara, Technical Aspects of Data Communication, © 1977, Digital Press, Maynard, MA.)*

	XXXXXXXX	XXXXXXXXXXXXXX	XX	XXXXXXXX	XXXXXXXX	XXXXXXX
Data Messages	10000001	Character Count	QS	Resp #	Message#	Address
Acknowledgement	00000101	00000001000000	QS	Resp #	00000000	Address
Negative Acknowledge	00000101	00000010REASON	QS	Resp #	00000000	Address
Reason						
BCC Header Error		000001				
BCC Data Error		000010				
Rep Response		000011				
Buffer Unavailable		001000				
Receiver Overrun		001001				
Message Too Long		010000				
Header Format Error		010001				
Reply Message	00000101	00000001100000	QS	00000000	LstMess#	Address
Start Message	00000101	00000110000000	11	00000000	00000000	Address
Start Acknowledgement	00000101	00000111000000	11	00000000	00000000	Address
Maintenance Message	10010000	Character Count	11	00000000	00000000	Address

Notes

1. Only the Data Message and the Maintenance Message have character counts, so only these messages have the information and CRC2 fields shown in the message format diagram above.
2. "RESP #" refers to Response Number. This is the number of the last message received correctly. When used in a negative acknowledge message, it is assumed that the next higher numbered message was not received, was received with errors, or was unaccepted for some other reason. See "Reasons."
3. "Message#" is the sequentially assigned number of this message. Numbers are assigned by the transmitting station modulo 256, i.e., message 000 follows 255.
4. "LstMess#" is the number of the last message transmitted by the station.
5. "Address" is the address of the tributary station in multipoint systems and is used in messages both to and from the tributary. In point-to-point operation, a station sends the address "1" but ignores the address field on reception.
6. "Q" and "S" refer to the quick sync flag bit and the select bit.

WHAT HAVE WE LEARNED?

1. Protocols are rules for communications between processes that are alike. Interfaces are rules for communicating between processes that are different.
2. Data communications protocols are made up of symbols to be communicated, a code set translating those symbols to a binary code, and rules for the correct sequencing of the symbols.
3. Parity is a method of adding redundancy to coded symbols to help detect errors in transmission.
4. XMODEM is a simple protocol used in asynchronous transmission between microcomputers.
5. A Cyclic Redundancy Check is a number calculated from the data transmitted by the sender and recalculated by the receiver that allows the data to be checked for errors in transmission.
6. The most widely used code sets in the U.S.A. are ASCII and EBCDIC.
7. Protocols must be suitable for use on dial and private line connections, and for point-to-point and multipoint network arrangements.
8. The most widely used protocols in the U.S.A. are TTY, BiSync, SDLC, and DDCMP.

Quiz for Chapter 7

1. Communications protocols always have a:
   a. Set of symbols.
   b. Start of header.
   c. Special Flag symbol.
   d. BCC.
2. The Baudot code uses how many bits per symbol?
   a. 9
   b. 7
   c. 5
   d. 8
3. When transmitting odd-parity coded symbols, the number of bits that are zeros in each symbol is:
   a. Odd.
   b. Even.
   c. Unknown.
   d. None of the above.
4. In the XMODEM protocol, the sender waits for what character from the receiver before beginning transmission?
   a. WACK
   b. ACK
   c. RVI
   d. NAK
5. Which of the following is not a valid rule for XOR:
   a. 0 XOR 0 = 0
   b. 1 XOR 1 = 1
   c. 1 XOR 0 = 1
   d. B XOR B = 0
6. Which of the following BiSync control codes are not defined in the EBCDIC character set?
   a. STX
   b. ACK0
   c. ENQ
   d. TTD
7. How many messages may be outstanding (unacknowledged) on a BiSync link?
   a. 1
   b. 2
   c. 4
   d. 8
8. Which code set is used in BiSync when using VRC/LRC but not operating in transparency mode?
   a. EBCDIC
   b. ASCII
   c. SBT
   d. Fieldata
9. The escape character that identifies control characters in BiSync Transparency mode is:
   a. ESC
   b. SYN
   c. DLE
   d. RVI
10. One primary difference between DDCMP and SDLC is:
    a. DDCMP does not have a transparent mode.
    b. SDLC does not use a CRC.
    c. DDCMP has a message header.
    d. DDCMP does not require special hardware to find the beginning of a message.

Alternatives in Local Area Networks

ABOUT THIS CHAPTER

This chapter begins by describing the characteristics of an ideal local area network (LAN) and the major obstacles to achieving the ideal. The most common types of existing LAN architectures; contention access based coaxial cable systems, polled access based coaxial cable systems, and frequency division broadband systems, are discussed through examples of existing LANs. The chapter continues with a discussion of the application of digital telephone switching systems as LANs. These digital systems have been developed during the same time as the LANs and have the capability to provide many of the same services as an LAN.

WHAT IS A LAN?

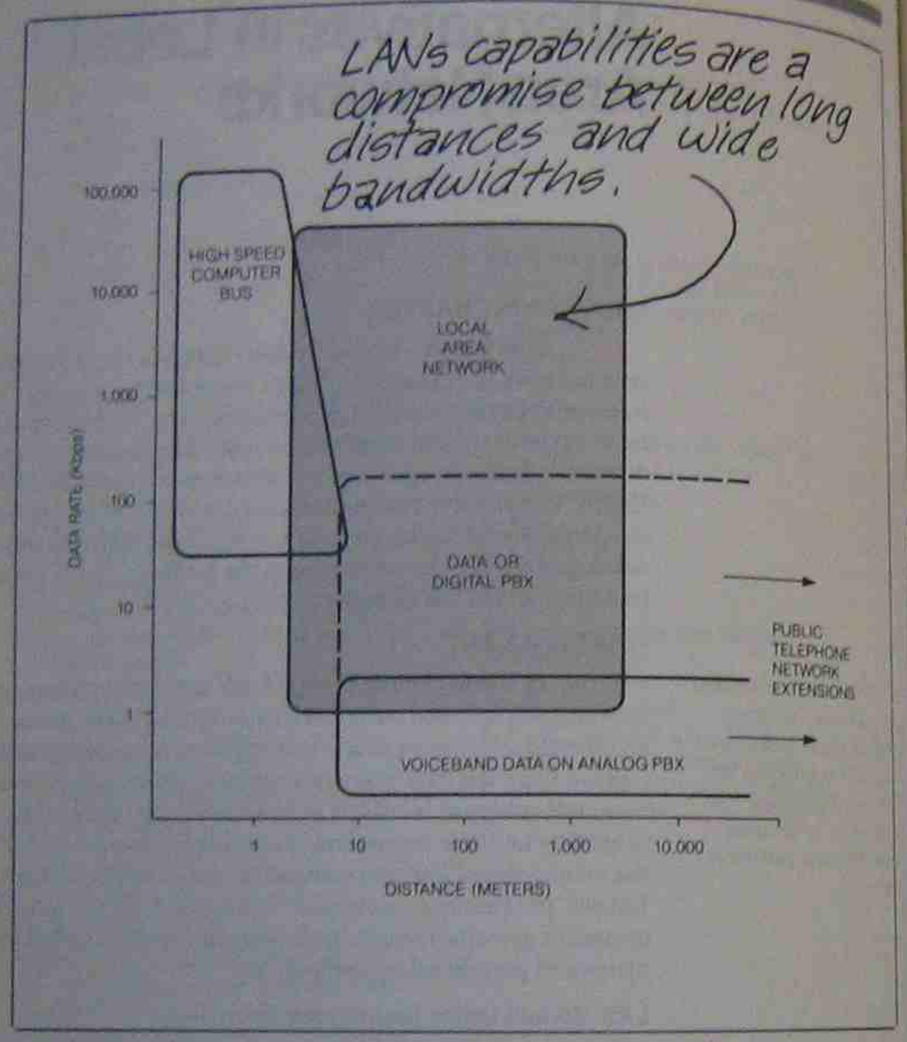
A LAN is a data communications facility located within a single building or campus that provides all high-speed switched connections to its terminal processors and peripheral devices.

In its most basic form, a LAN is a data communication facility providing high-speed switched connections between processors, peripherals, and terminals within a single building or campus. Historically, LANs have evolved from the data processing or office processing industry where economics suggest relatively expensive storage devices and printers should be shared by multiple computers. Thus, LANs have evolved partly in response to the emergence of low-cost computing and its need for high-cost peripherals. Indeed, the first generation LANs were actually one form of distributed computer systems using proprietary high-speed data links between processing nodes and peripheral equipment.

LAN Versus Other Techniques

A perspective of a local area network as a digital communications facility is shown in *Figure 8-1*, which depicts the bandwidth versus distance capabilities of three data communications technologies: computer buses, voiceband data links, and LANs. The computer buses are internal to a computer mainframe and achieve very high data transfers by using high clock rates (for example, a few MHz) and parallel transmission (for example, 32 bits per clock cycle). Such systems are economical and practical only at distances up to a few meters. Voiceband data transmission uses readily available telephone wire pairs with modems. These systems are virtually unlimited in distance, but the standard analog telephone channel usually restricts voiceband data rates to 9600 bps for leased lines or 4800 bps for switched (dial-up) connections. From a cost effectiveness point of view, 300 and 1200 bps is very popular for dial-up data communications.

Figure 8-1. Distance Versus Data Rate Comparison



Each of the three data communications technologies have their own particular strengths and weaknesses when considering bandwidth vs. distance trade-offs.

As can be seen in *Figure 8-1*, the capabilities of a typical LAN represent a compromise between long distances and wide bandwidths. Some computer architectures have moved in on the region occupied by LANs by moving away from high-speed parallel buses to more loosely coupled distributed systems. Thus, the LAN region is overlapped on the left edge by distributed computer systems. Similarly, newly developing digital telephone systems are overlapping the LAN region on the bottom by providing moderate data rates (up to a few hundred kilobits per second) as a low-cost addition to basic telephone service.

The distance versus data rate aspect of a local area network depicted in *Figure 8-1* actually represents only one level of definition of an LAN. As the market and technology have grown, so has the recognition that an LAN should solve or alleviate more problems pertaining to data communications than merely the data transmission problem. The data communications manager or

service organization responsible for a particular building or campus is typically concerned with a variety of applications and associated types of equipment. If these different applications can be supported by one transmission facility, one type of hardware and software interface, and one set of maintenance procedures, total systems costs are reduced and overall administration is greatly simplified. For example, *Figure 8-2* illustrates a local area network serving several independent applications.

The next section presents the attributes of an ideal local area network — one that is *all* supportive of *all* applications. In the succeeding sections, present day LANs are described and compared to the ideal as a means of contrasting one type of LAN to another.

THE IDEAL LAN

The ideal LAN should provide the same ease of access and equipment utilization that is found in a building's ac circuit.

The ideal LAN would be an information distribution system that is as easy to use as the conventional ac power distribution system in a building. Thus, adding a data terminal, processor, or peripheral to a local area network should require nothing more than plugging it into a conveniently located access port. Once plugged in, it should communicate intelligently with any other device on the network. This ideal system is summarized by the features that make the ac power system so easy to use:

1. One-time installation.
2. Widespread access.
3. Application independence.
4. Excess capacity.
5. Easy maintenance and administration.

The features identified above have been developed over the years to minimize life cycle costs for supplying ac power within a building. Although the costs of a typical installation can be reduced by wiring only to locations with immediate electrical needs, the costs of only small amounts of rewiring to supply additional locations at a later date would more than offset the initial savings. In order to achieve a one-time installation, it is necessary to provide widespread access; that is, an electrical outlet at virtually every location where one may be needed. Furthermore, the capacity at every outlet is generally greater than the needs of the electrical appliances that connect to it. Adding wiring or rewiring becomes necessary only for particularly high-power consuming devices (for example, those needing 240 volts or dedicated, high-current circuits).

If an information distribution system were available with all the desirable properties listed above, it would mean that telephones, data terminals, printers, and storage devices could be moved as easily as unplugging and plugging in a lamp. Moreover, the equipment could be supplied by a variety of vendors. Although such an ideal system does not now exist, local area networks of several forms represent some of the first steps in the development of such a system.

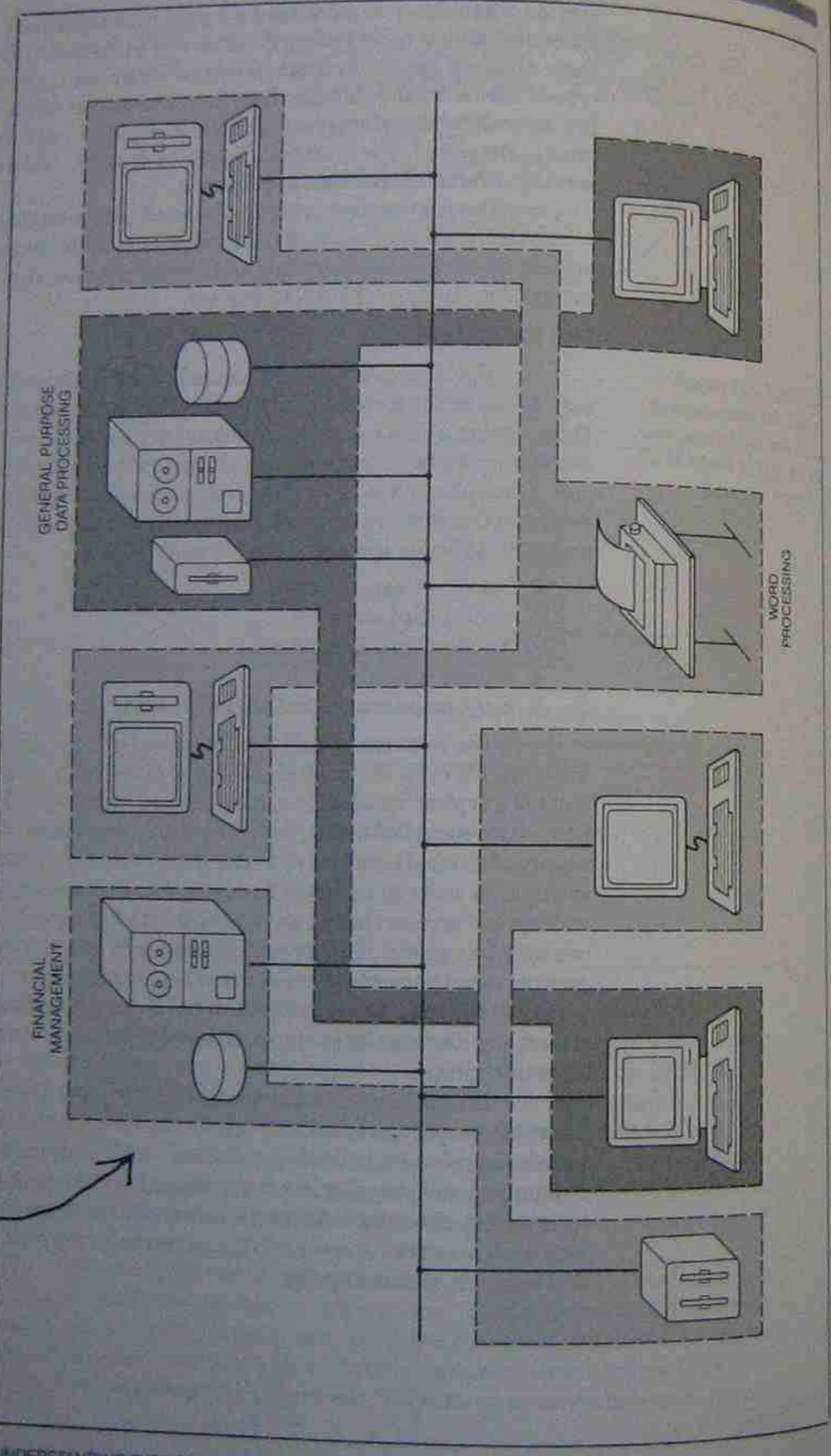


Figure 8-2. Local Area Network Supporting Multiple Independent Applications

Three different applications are served by common transmission facility.

There are four major obstacles that must be overcome in the development of the ideal LAN.

Major Obstacles to the Ideal LAN

The major obstacles to overcome in the development of the ideal system are summarized below.

No Single Standard

Due to the continually changing status of LANs and competitive nature of the vendors, a variety of local area network standards exist — both official and de facto. The situation is improving, however, because even the dominant suppliers who have been protecting their proprietary interfaces are being pressured by a maturing market to release interface specifications.

Diverse Requirements

The communications needs of a modern office building include voice, video, high-speed data, low-speed data, energy management, fire alarm, security, electronic mail, etc. These systems present transmission requirements that vary greatly in terms of data rates, acceptable delivery delays, reliabilty requirements, and error rate tolerance.

Costly Transmission Media

Being able to deliver tens of megabits per second to one device and only a few bits per second to another implies that the lower rate devices are burdened with a costly transmission media. The best economic solution must involve a hierarchical network design (one with stepped levels of capacity) that allows twisted pair connections for low and medium data rate devices (a low step) feeding into a backbone high bandwidth transmission system (a higher step) such as coaxial cable or optical fibers.

Sophisticated Functional Requirements

Providing a network with the desired data rates and distances is only one item that must be considered in the data communications problem. Before one data device can communicate intelligently with another, numerous higher level communications functions must be compatible. These include codes, formats, error control, addressing, routing, flow control, access control, configuration management, and cost allocations.

The ISO Model

A formal heirarchical identification of all data communications network functions has been established by the International Standards Organization (ISO) and referred to as the *ISO Model for Open Systems Interconnection* (OSI). This model, shown in *Figure 8-3*, identifies seven distinct levels of functional requirements pertaining to data communications networks. The lower three levels of this model were originally proposed by CCITT to encourage vendor commonality in interfacing to public packet switching networks. This standard, designated as CCITT recommendation X.25, is described in detail in Chapter 9, along with the ISO model.

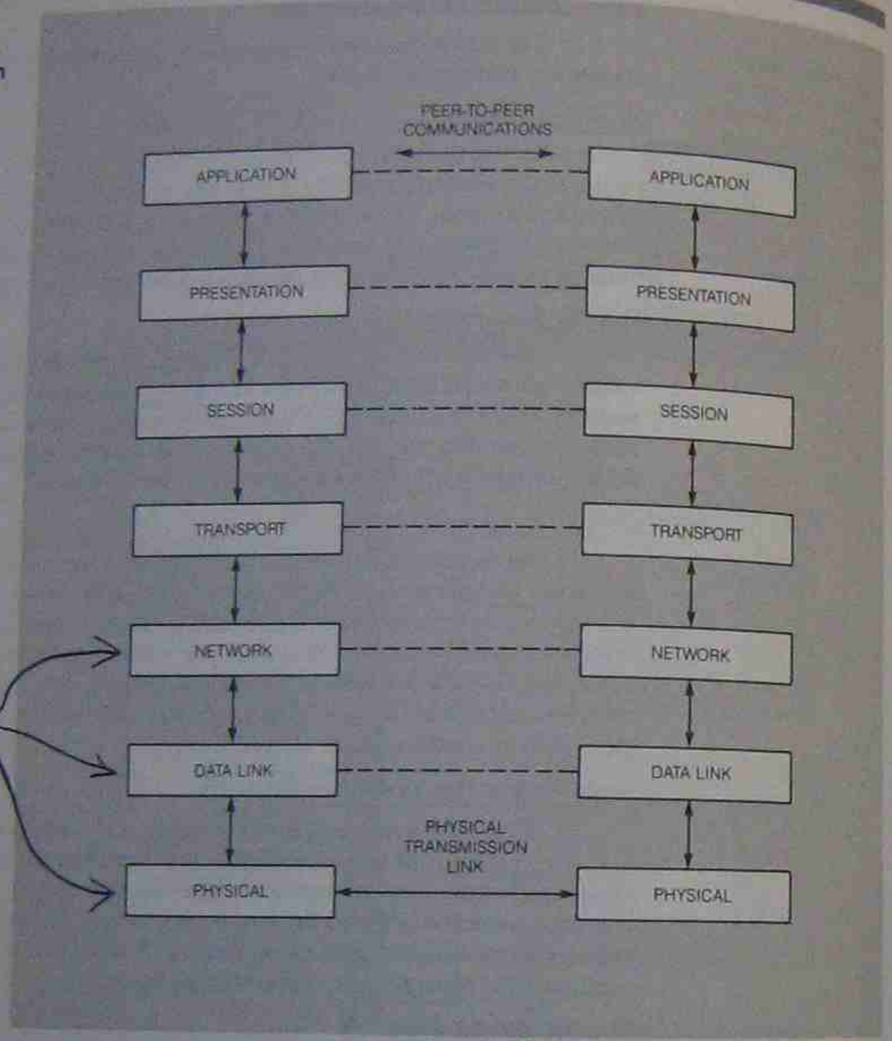


Figure 8-3.
ISO Model for Open Systems Interconnection

A formal data communications hierarchy model containing seven levels of functional requirements has been established by the Internal Standards Organization.

Realization of the ideal LAN would require all levels of functions included in the OSI standard; however, not all levels of the OSI standard need to be implemented to provide effective communications in an LAN. If only the lower levels of the standard exist, a LAN can usefully support the multiple applications shown in *Figure 8-2*. In essence, the transmission media and lower level interfaces are common so that data can be exchanged within virtual subnetworks; for example, the financial management group indicated in *Figure 8-2*. However, a device in one subnetwork cannot communicate intelligently with a device in another subnetwork because each application is using unique higher level implementations. This LAN would allow messages to be exchanged between these dissimilar terminals, but the message would not be understood. Nevertheless, the backbone LAN permits flexible location and relocation of the various attached equipment.

ETHERNET[1] (CSMA/CD)

Ethernet, which uses baseband coaxial cable as its carrier medium, was a joint effort by Xerox, Digital Equipment and Intel and was the first commercially available LAN.

Although the Ethernet was not the first data communications network to technically qualify as a local area network, it is the most important because it represents the first major product offering with non-proprietary communications interfaces and protocols. Using an experimental design developed in a Xerox Corporation research laboratory; Xerox, Digital Equipment, and Intel teamed to define some commercial products based upon jointly published communications standards. This action has invited other manufacturers to develop compatible products. The marketing strategy has been successful enough to establish the Ethernet interface as a standard even on systems utilizing transmission media different from the baseband coaxial cable used in the Ethernet.

The Ethernet architecture is based in concept on the Aloha satellite communications network developed at the University of Hawaii. The Aloha system allows multiple distributed devices to communicate with each other over a single radio channel using a satellite as a transponder. One station communicates with another by waiting until the radio channel is idle (determined by carrier sensing) and then sending a packet of data with a destination address, source address, and redundant check bits to detect transmission errors. All idle stations continuously monitor incoming data and accept those packets with their address and valid check sums. Whenever a station receives a new packet, the receiving station returns an acknowledgement to the source. If an originating station receives no acknowledgement within a specified time interval, it retransmits the packet under the assumption that the previous packet was interfered with by noise or by a transmission from another station at the same time. (The latter situation is referred to as a collision.) The Ethernet employs the same basic system concept using coaxial cable distribution throughout a building or campus.

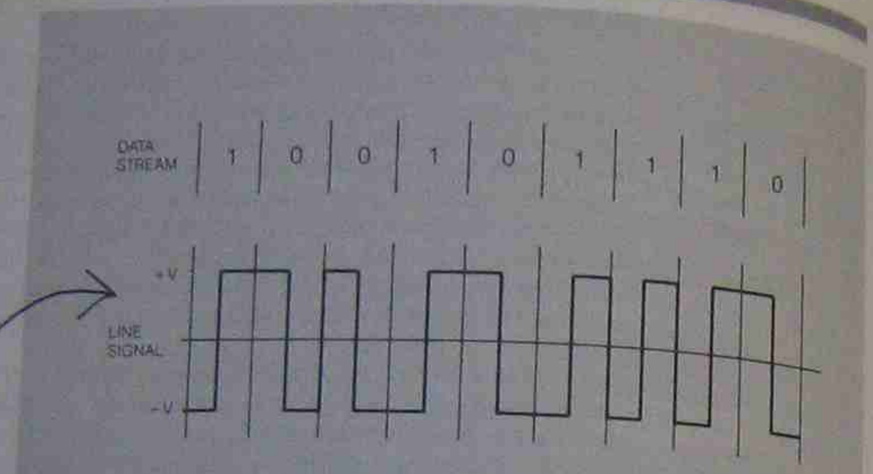
Physical Layer

The coaxial cable used in Ethernet can handle transmissions at 10 Mbps over one channel encoded in a Manchester line code system.

The transmission media of the Ethernet is a coaxial cable using baseband transmission at 10 Mbps. Baseband transmission implies data is transmitted without the use of a carrier and with only one channel defined in the system. When a station is transmitting, it uses the entire 10 Mbps capacity of the system. Data are encoded using a Manchester code as shown in *Figure 8-4*. This line code provides a strong timing component for clock recovery because a timing transition always occurs in the middle of every bit. The Manchester line code has the additional property of always maintaining equal amounts of positive and negative voltages. This prevents the build-up of a dc component which simpifies the implementation of decision thresholds in the data detectors.

[1]Ethernet is a trademark of Xerox Corporation

Figure 8-4.
Manchester Line Code



Continuous transitions provide the equivalent of a carrier.

The Manchester Code provides a form of carrier in its continuous transition. Multiple stations can enter or leave the line without causing traffic interruptions by using passive taps.

Although the data are not transmitted with a carrier per se, the continous transitions of the Manchester code provide the equivalent of a carrier so the channel is easily monitored for activity (for example, by a carrier sense technique). Multiple access to the coaxial cable is provided by passive taps, thereby allowing station connections (called drops) to be added or removed without disrupting traffic in the system.

Another requirement of the transmission link and its associated access electonics is that, while transmitting, a transceiver must be capable of detecting the existence of another active transmittter. This is referred to as collision detection. Thus, the three basic steps for accessing an Ethernet are denoted CSMA/CD for: Carrier Sense, Multiple Access with Collision Detection.

The ability to detect collisions allows colliding stations to release the channel after using it for only a short period of time. A conventional Aloha system (one without collision detection), on the other hand, transmits entire messages without knowing if a collision has occurred. Because an Ethernet station checks for carrier presence before transmitting, collisions occur only if two stations begin transmitting within a time interval equal to the propagation delay between the stations. By restricting the maximum distance between transceivers to 2500 meters, the collision window is limited to 23 microseconds (us) including amplifier delays.

Once a station begins transmitting, it can be sure that no collision will occur if none is detected within a round-trip propagation time of 46.4 us. Since one bit time is 0.1 us at 10 Mbps, the collision detection decision is made within 464 bit times. The maximum length of a frame is 12,144 bits, thus the collision detection feature will detect a collision very early in the frame to save significant transmission time that would be wasted if the entire frame were transmitted before detection.

A transceiver must be able to detect any other active transmitters on the line while it is transmitting. This is collision detection.

Whenever a collision occurs, all colliding stations detect the condition and wait individually random amounts of time before retrying the transmission. Thus, the stations do not wait for acknowledgement time outs to trigger a retransmission. (Acknowledgements are unnecessary at this level of the protocol, but may be used at a higher level to ensure that the message was received at the final destination.) By waiting random amounts of time before transmitting, the probability of repeated collisions is reduced. In heavy traffic conditions, the average delay before retransmission begins to increase after 10 unsuccessful attempts. After 16 collisions, no further attempts are made to transmit that message and the station is notified of the error by the transceiver.

In preparation for transmitting a data frame, the physical layer must insert a 64-bit preamble so that all receivers on the network can synchronize to the data stream before the desired data frame begins. The preamble consists of alternating 1s and 0s, ending in two 1s to signify the start of a frame.

Physical Layer Interface

Ethernet transceiver cable connectors are 15-pin D-shell connectors (MIL-C-24308 or equivalent). The transceiver has a male connector while the station apparatus has a female connector, thus, the interconnecting cable must have one of each type. The pin assignments are:

1. Shield
2. Collision Presence +
3. Transmit +
4. Reserved
5. Receive +
6. Power return
7. Reserved
8. Reserved
9. Collision Presence –
10. Transmit –
11. Reserved
12. Receive –
13. Power
14. Reserved
15. Reserved

Data Link Layer

The data link is concerned with message frame format, message addressing and error checking; and as such is largely independent of the medium-dependent physical channel.

The data link layer is primarily concerned with message packaging and link management. It is largely independent of the medium-dependent physical channel. The message packaging function includes:

1. Framing: Identifying the beginning and end of a message.
2. Addressing: Specified fields for source and destination addresses.
3. Error checking: Redundant codes for detecting channel errors.

The format of a frame is shown in *Figure 8-5*. The preamble is not shown because its generation and removal are functions of the physical layer. Similarly, the end of a frame is provided by the removal of the carrier sense signal detected by the absence of a bit transition following the last bit of the cyclic redundancy code (CRC). Notice that frame sizes must be an integral number of bytes, ranging from 64 to 1518 bytes. (At 8 bits per byte, that's 512 to 12,144 bits). The TYPE field is reserved to indicate which of several possible higher level protocols might be in use.

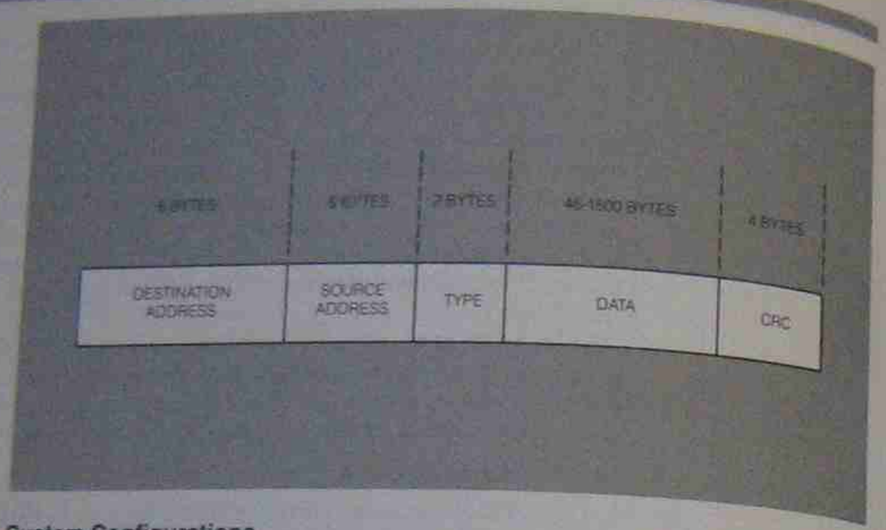


Figure 8-5.
Ethernet Data Link Frame Format

By using point-to-point cables and repeaters, several multidrop segments can be connected to add flexibility to Ethernet systems.

System Configurations

Besides the single cable system indicated in *Figure 8-3*, Ethernet configurations are possible where several multidrop segments (possibly one in each of several buildings) are interconnected with point-to-point cables and repeaters. *Figure 8-6* shows a typical configuration of a multisegment system. The following specifications apply:

1. Maximum Lengths:
   - 1500 meters of multidrop cable.
   - 1000 additional meters of point-to-point cable between cable segments.
   - 300 additional meters in six transceiver cables.
   - 50 meters in a single transceiver cable.
2. Maximum Number of Stations:
   - 1024

Other CSMA/CD Systems

Besides the Ethernet, numerous other LANs have been developed using CSMA/CD for network access. Most of these systems have focused on supporting communications between personal computers and utilize both coaxial cables and twisted pair for transmission media. Various broadband LANs (discussed in a later section) also support CSMA/CD access on one or more rf channels.

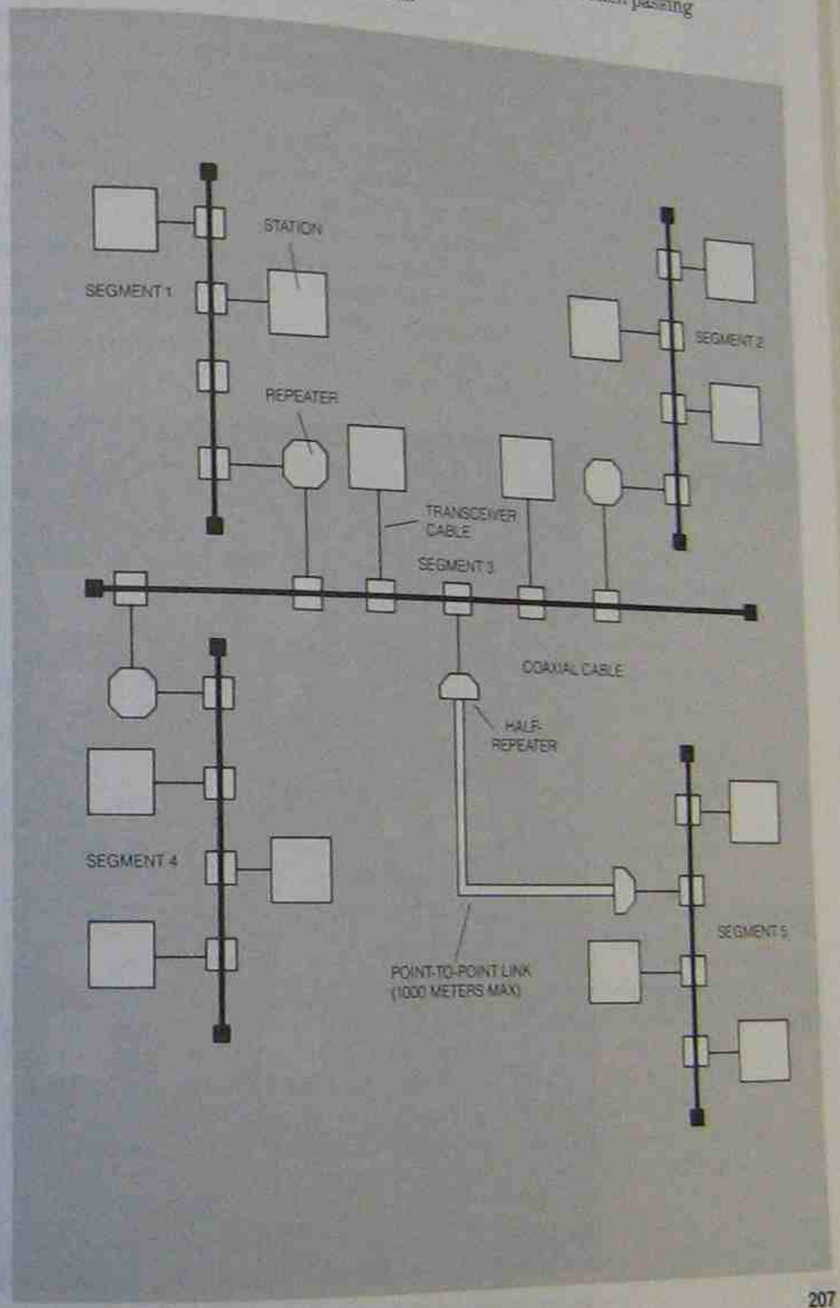
Token passing networks provide access to the network by only one station at a time; the one with the token that is ready to send a message.

TOKEN PASSING NETWORKS

Because access to a CSMA network involves a certain amount of contention (competition) between stations trying to send a message at the same time, the behavior of the network must be analyzed and controlled in a statistical manner. Token passing networks, on the other hand, provide a

different access procedure. Access is determined by which station has the token; that is, only one station at a time, the one with the token, is given the opportunity to seize the channel. The token is passed from one idle station to another until a station with a pending message receives it. After the message is sent, the token is passed to the next station. In essence, a token passing network is a distributed polling network.

Figure 8-6.
A Typical Large-Scale Ethernet Configuration
(Source: Ethernet Specification Version 2.0)



Token topologies can be either a token passing ring or a token passing bus; each has its own peculiar strengths and weaknesses.

Two basic topologies (configurations or arrangements) exist for token passing networks: Token Passing Rings and Token Passing Buses. In a token passing ring, shown in *Figure 8-7*, the closed loop topology defines the logical topology (that is, the order in which the token is circulated). A token passing bus, shown in *Figure 8-8*, has more operational flexibilty because the token passing order is defined by tables in each station. If a station (for example, a printer) never originates communications, it will be a terminate-only station and need not be in the polling sequence. If a station needs a high priority, it can appear more than once in the polling sequence.

Besides the operational differences between bus and ring topologies, other major contrasts are:

1. A ring requires an active interface module in series with the transmission link. A bus uses a passive tap like a CSMA system.
2. The point-to-point nature of a ring's transmission links implies the signal quality is easily controlled. Multiple taps in a linear bus system, on the other hand, each contribute to signal distortions.
3. A ring has no inherent (built-in) distance limitation (as does the bus) because there is no delay dependence in its operation, and the access taps at each station serve as regenerative repeaters to the digital signals.

Figure 8-7.
Token Passing Ring

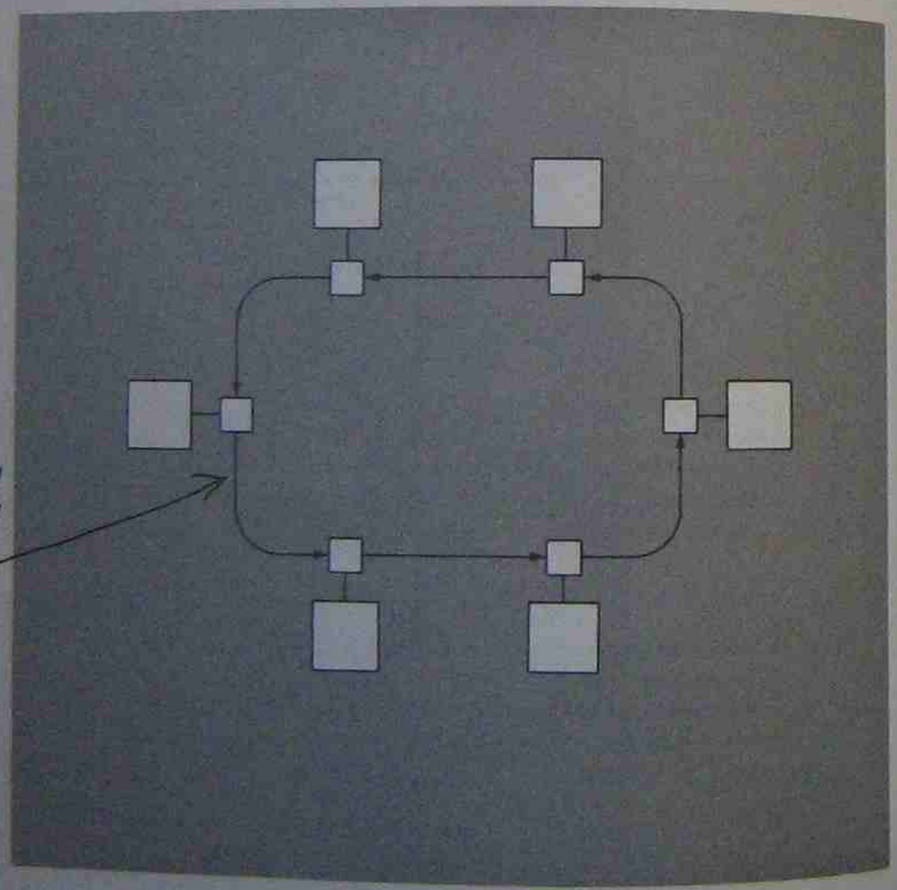
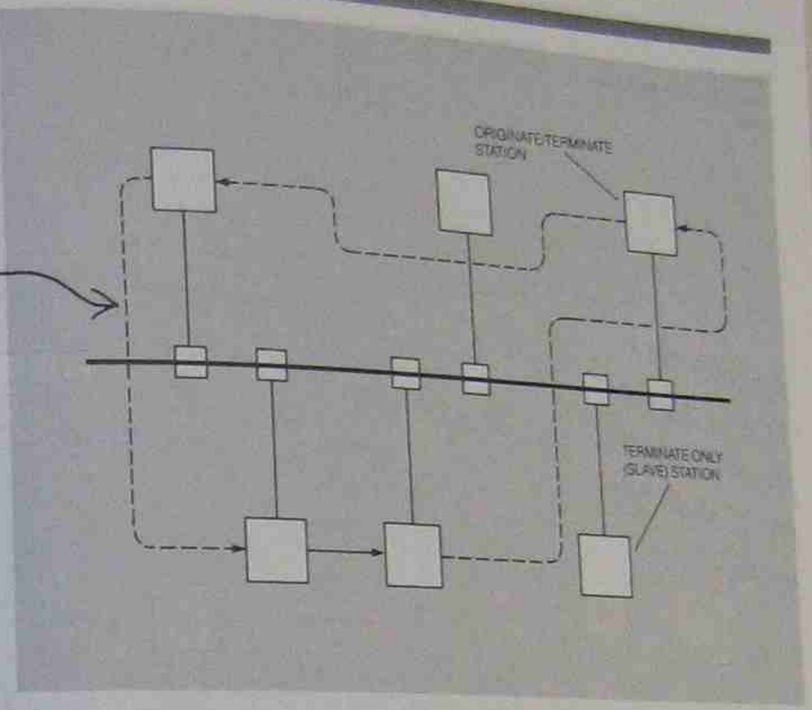


Figure 8-8.
Representative Access Sequence in Token Passing Bus



4. A bus allocates the entire channel to an active device. The channel of a ring can be subdivided into time division multiplexed subchannels for time continuous lower data rate applications. *Figure 8-9* shows a ring with time division multiplexed subchannels. Such a system is also referred to as a loop, and typically requires a unique control node to define framing for the subchannels. Access to the TDM channels is distributed by passing tokens.
5. The ring topology is ideally suited to the particular capabilities of fiber optic transmission. As discussed in a previous chapter, optical fibers provide very wide bandwidth and complete elimination of electrical interference, but they are presently practical only as point-to-point transmission links.

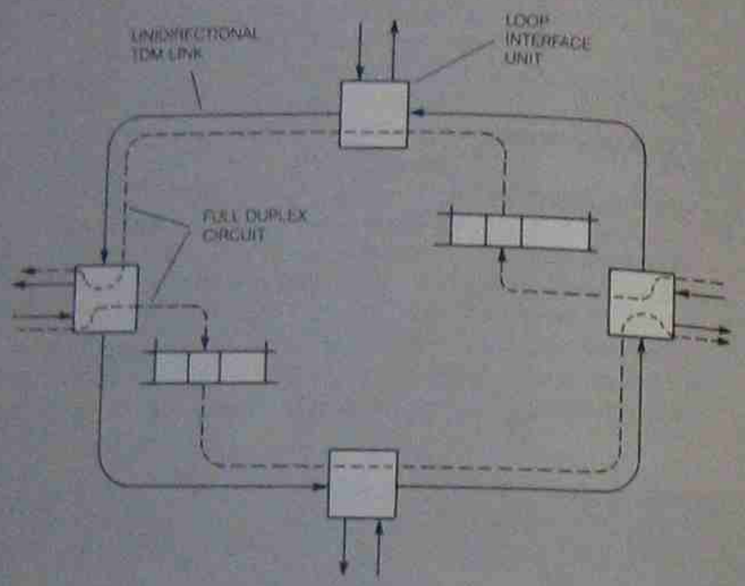
ARCNET[2]

ARCNet is a type of token passing bus. Its topology is a hybrid bus/star.

The forerunner of token passing networks in the U.S. is the Attached Resource Computer Network, ARCNet, developed by Datapoint Corporation. Initially, the network and protocol were kept proprietary, but the data link protocol, interface specs, and even integrated circuits were made publicly available in 1982. Functionally, the ARCNet is a token passing bus, but the physical topology, shown in *Figure 8-10*, is a hybrid bus/star. Rather than distribute taps along a linear bus as suggested in *Figure 8-8*, the ARCNet uses hubs with individual ports to connect Resource Interface Modules (RIMs) to the transmission media.

[2]ARCNet is a trademark of Datapoint Corporation.

Figure 8-9.
Time Division Multiplex Loop



Access to TDM channels is distributed by passing tokens.

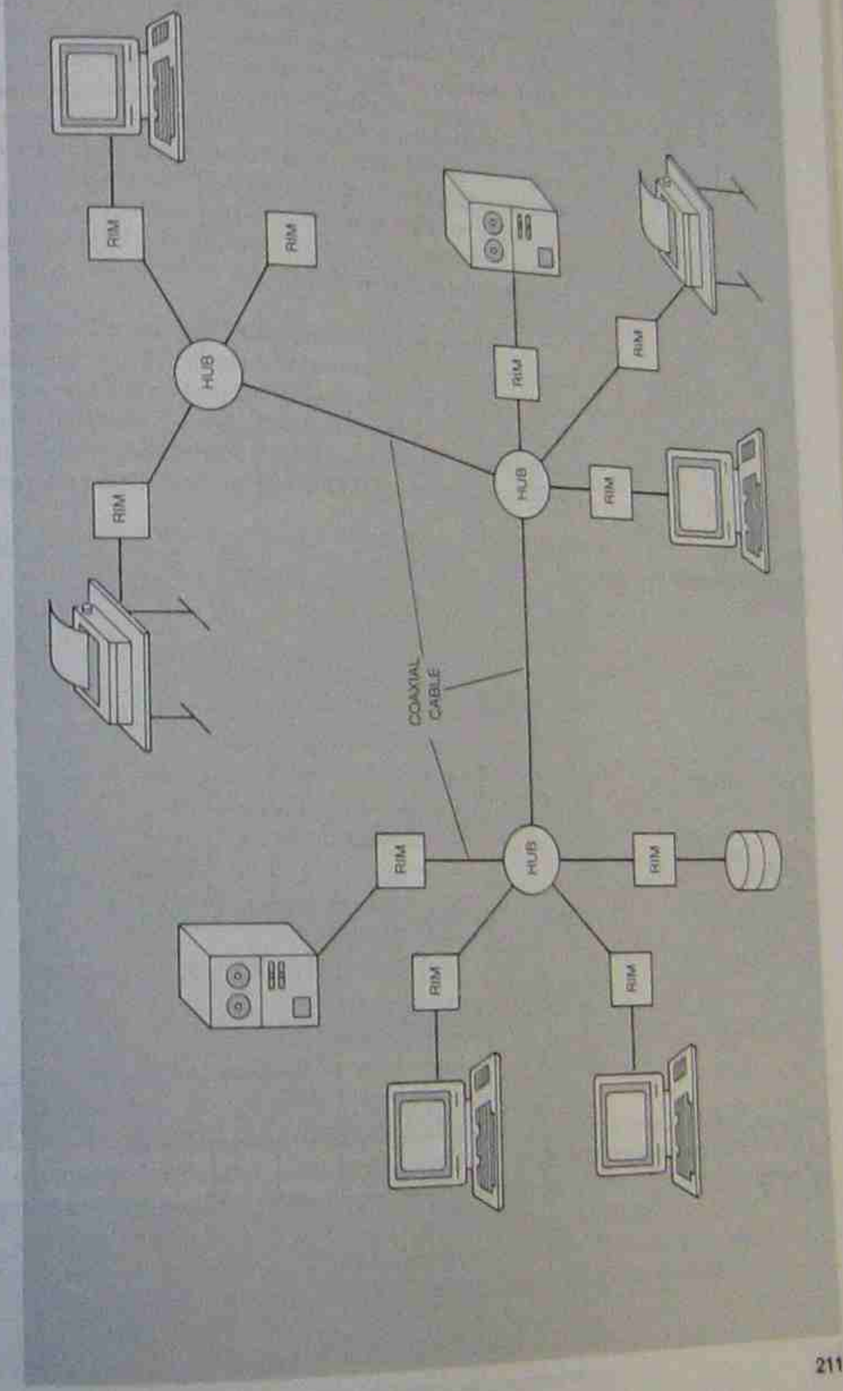
The hub based architecture is an effective means of controlling the signal quality because the hub isolates each RIM port from the main coaxial cable. Unidirectional (one-way) amplifiers in the hubs provide zero insertion loss and suppress reflections because only one direction of transmission is enabled at a time. Amplifier switching is possible because a token passing network only transmits in one directon at a time.

Physical Layer

RG62 coaxial cables with a baseband transmission rate of 2.5 Mbps are used to connect the hubs and RIMs in a ARCNet.

The ARCNet interconnects the hubs and RIMs with RG62 coaxial cable using baseband transmission at 2.5 Mbps. Although 2.5 Mbps is a relatively low data rate, ARCNet uses inexpensive coax and can be configured (laid out) with as much as four miles between stations. The cable length between a hub and a RIM is limited to 2000 feet, but a four-mile span can have up to a maximum of ten hubs in a series path.

Figure 8-10.
Representative ARCNet Configuration



There are five basic message formats used in link protocol: four control message and one data carrier. Eight bits are used in all address fields.

Link Protocol

The ARCNet employs five basic message formats as shown in *Figure 8-11*. The first four formats are used for control messages while the fifth carries data between stations. All address fields consist of 8 bits which restricts the number of stations to 255 (address 0 is reserved as a broadcast address to all stations). As indicated in *Figure 8-11*, the destination address (DID) is duplicated with every message for error protection.

A logical flow diagram of the token passing procedure implemented in every RIM is shown in *Figure 8-12*. Upon receiving a token (message type 1) with the proper address, a RIM chooses one of two paths, depending on whether it has transmit data pending. If data are pending, the RIM sends a free-buffer request (message type 2) to the desired destination RIM inquiring whether it is ready to receive. The destination RIM responds with an ACK (message type 3) if it has buffer space available, or a NAK (message type 4) if it does not.

After a RIM has transmitted its data or determined that it cannot, it passes the token to the station with the next higher address. After sending the token, the RIM monitors the channel to see if it is accepted. The occurrence of a message type 1, 2 or 5 indicates the token is accepted. No response on the channel within 74 us implies the intended station is off-line and the token should be passed to the station with the next higher address.

Figure 8-11. Message Formats in ARCNet

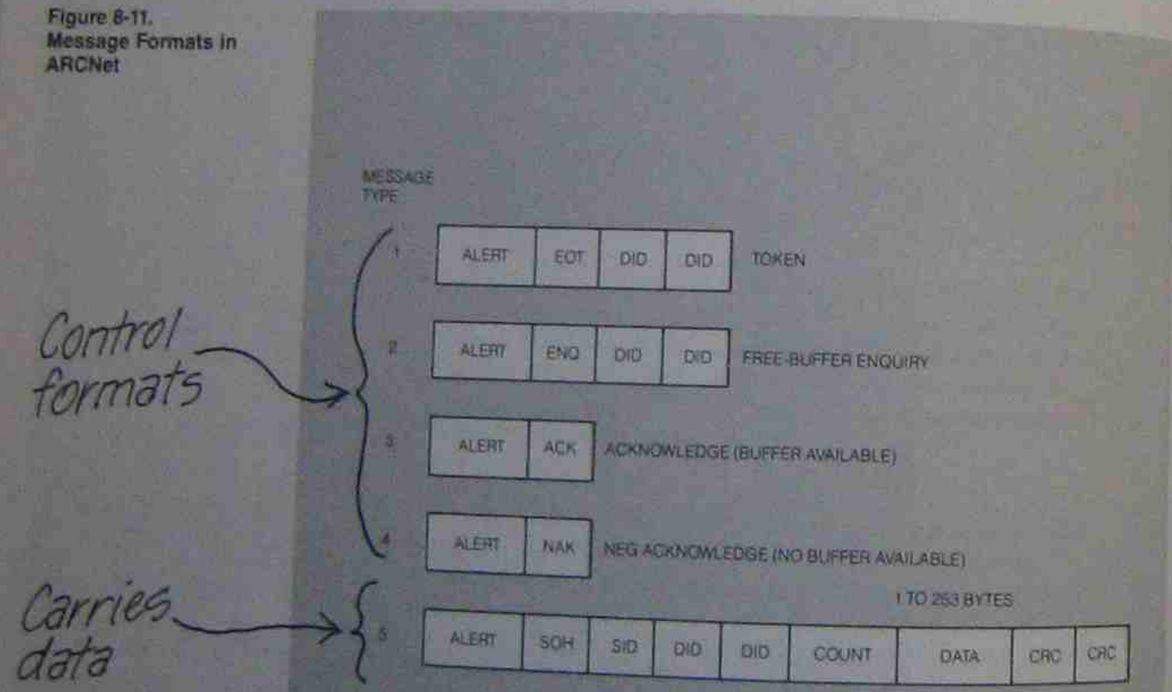
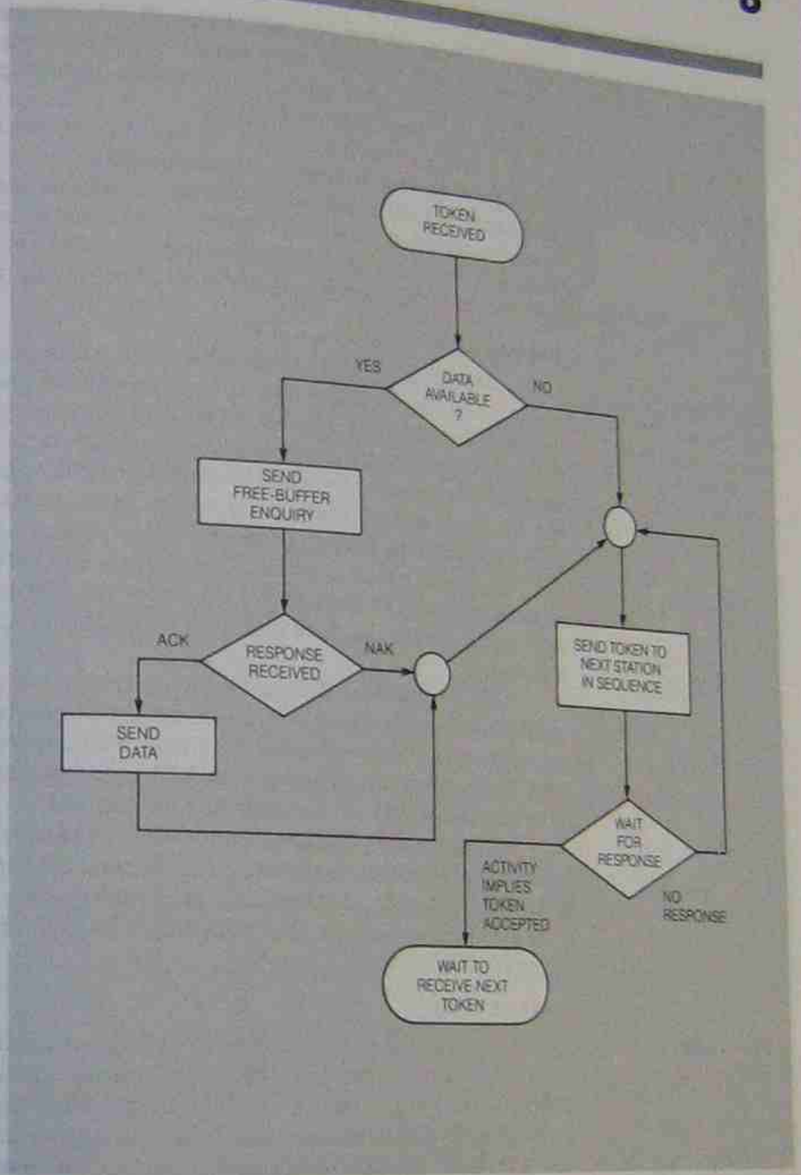


Figure 8-12. Flow Diagram of Token Passing Procedure in ARCNet



Frequency division multiplexing over wide bandwidth coaxial cables is used by broadband local area networks to transmit on multiple channels. This technique was developed by the cable TV industry.

BROADBAND NETWORKS

Broadband local area networks use wide bandwidth coaxial cables with frequency divison multiplexing (FDM) to establish multiple channels within a system. The FDM channels are typcially 6 MHz wide following the standards and technology developed by the CATV (cable TV) industry. Since the usable bandwidth of a CATV cable is 300 to 400 MHz, there can be as many as 50 to 60 separate channels. Often separate FDM channels are used for separate applications.

The power and attraction of a broadband network is evident in *Figure 8-13*. Notice that the transmission system supports not only a data network, but also closed-circuit TV distribution, point-to-point voice telephone circuits, and point-to-point data circuits.

Use of a broadband system for data transmission requires modems in the physical interface equipment. If the modems are assigned to a fixed frequency (channel), the associated station can communicate only with other stations assigned to the same frequency. If frequency agile modems (modems that have a tuner to allow changing channels) are used, the stations can switch between the FDM channels to gain access to much larger networks.

Physical Layer

Single-cable systems use a dual-frequency architecture; one frequency for sending, one for receiving on the same cable.

Broadband LAN systems utilize either one or two cables routed to every station. Single-cable systems are patterned after customary CATV systems where the FDM channels are separated into two groups as shown in *Figure 8-14*. One group of frequencies carries signals from the transmitter of all station interfaces to a frequency translation module referred to as the head end. The head end amplifies the received signals and shifts them to the second group of frequencies, whereupon they travel back along the cable to the receivers of every station interface. Thus, a single channel in a functional sense requires a pair of FDM channels in a physical sense. The dual-frequency architecture permits the use of unidirectional amplifiers as shown in *Figure 8-14*.

An alternative to the dual-frequency approach is the dual-cable architecture shown in *Figure 8-15*. Each station interface transmits on a particular channel in one cable and receives on the same channel in the second cable. The head end no longer provides FDM group translation; in fact, it may be nothing more than a passive cable-to-cable connection.

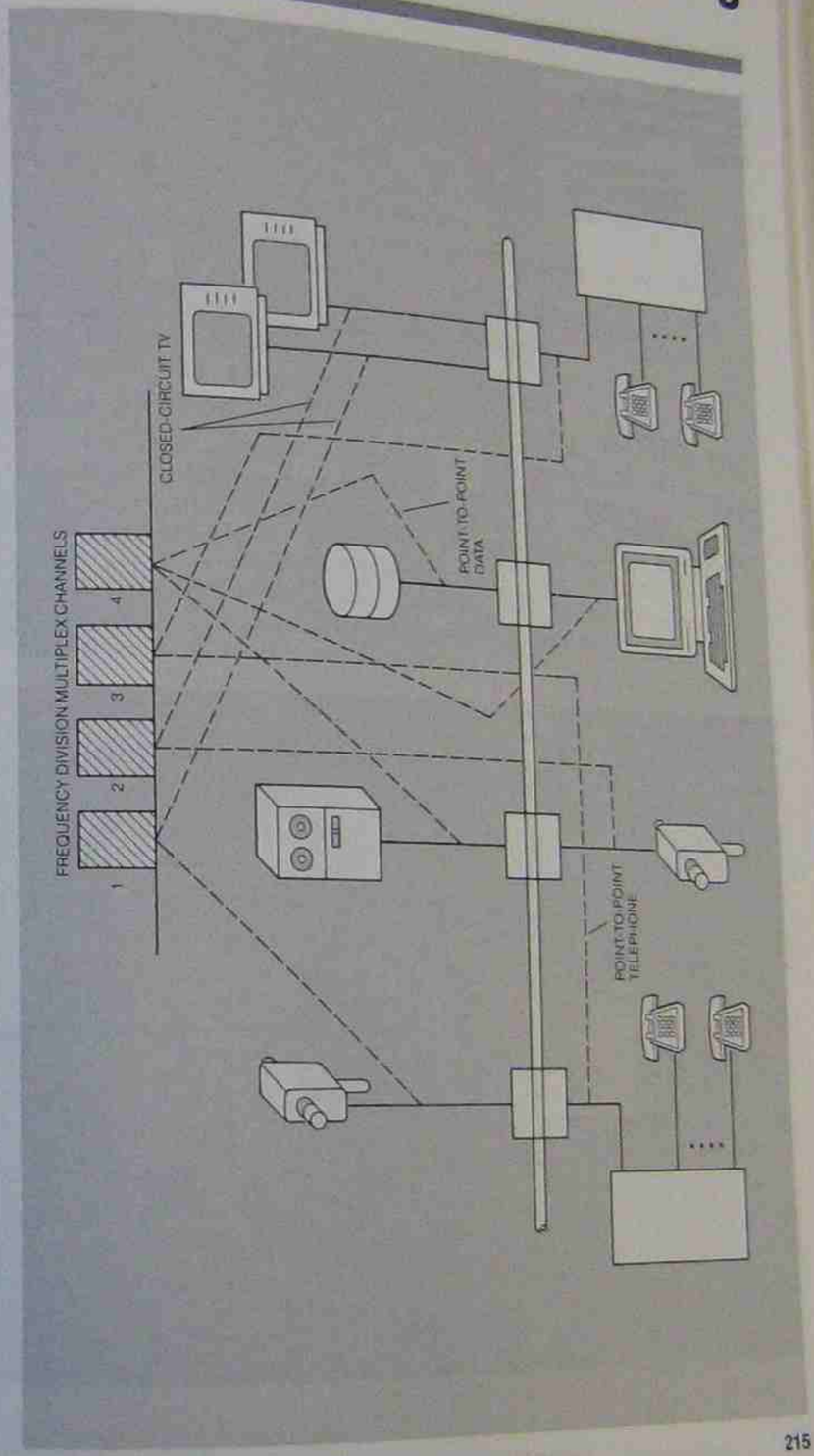
Whichever physical architecture is used, the advantage of unidirectional amplification appears in the maximum distance specifications of broadband networks. Distance limits of 5 to 30 miles are typical except in cases where access protocols like CSMA/CD restrict the allowable propagation times.

Data Link Layer

There is no single standard for broadband networks. Only the FDM channels have been given specific frequency assignments to allow interface with CATV systems.

Due to the flexibility and multiple services supported by broadband networks, no one standard configuration or agreement on operation has come about or is likely to come about. The only thing that appears to be a standard is the assignment of FDM channels which must be standardized to interface with CATV systems. Suppliers of data communications services on broadband cables generally have decided to use the data link protocols of the baseband systems. In particular, both Ethernet and ARCNet access protocols are available for rf channels on a broadband network. Thus, the OSI goal of separating the physical layer from the higher layers has proven to be useful by allowing data communications suppliers these options.

Figure 8-13. Broadband LAN Supporting Multiple Services



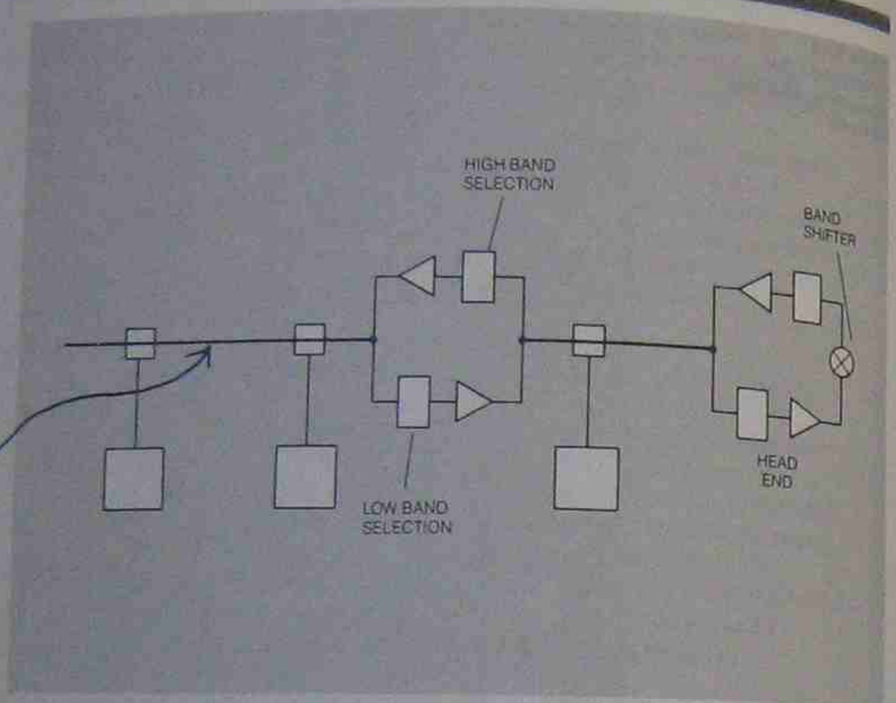


Figure 8-14. Dual-Frequency/Single-Cable Broadband Network

One cable, two frequencies

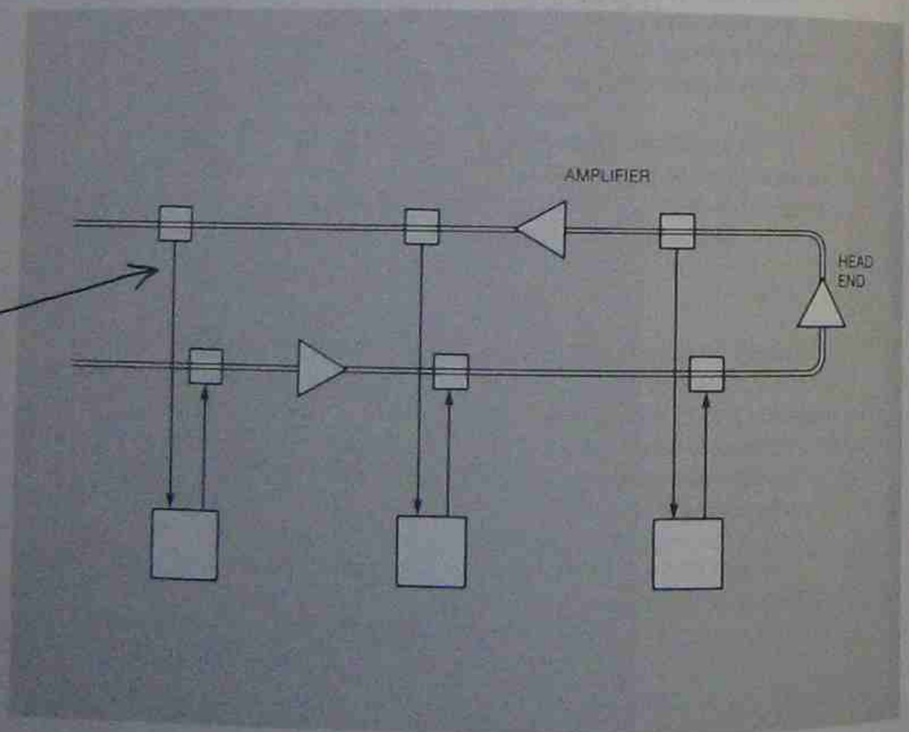


Figure 8-15. Dual-Cable/Single-Frequency Broadband Network

Two cables, one frequency

LAN STANDARDS

Local area network standards (as with other communication standards) get established in two ways: by dominant manufacturers who attract plug compatible competitors, and by official standards organizations. The leading official standards organization for LANs in the U.S. is the IEEE 802 Standards Committee. This committee has several working groups reponsible for establishing these LAN standards:

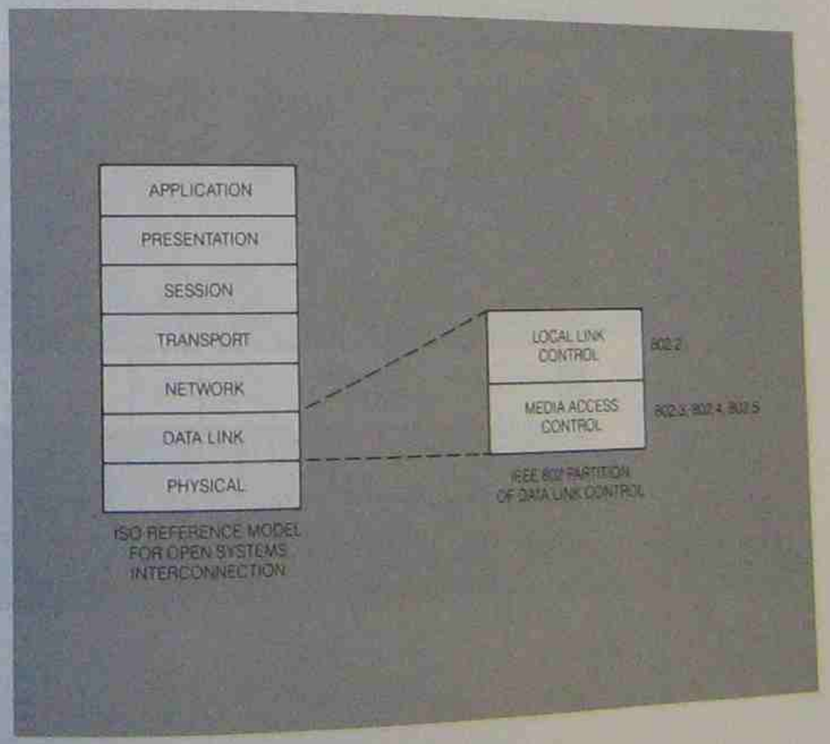
1. 802.1 — Coordinating the interface between OSI Levels 1 & 2 with the five higher level layers.
2. 802.2 — Logical data link standard similar to HDLC and ADCCP.
3. 802.3 — CSMA/CD standard similar to Ethernet.
4. 802.4 — Token Bus standard.
5. 802.5 — Token Ring standard.

Logical Link Control — IEEE 802.2

The IEEE 802 standards committee separated the OSI Data Link Layer into LLC and MAC layers.

In order to accommodate multiple LAN access methods, the IEEE 802 standards committee separated the OSI Data Link Layer into two sublayers: a logical link control (LLC) sublayer and a media access control (MAC) sublayer as shown in *Figure 8-16*. Under working group 802.2, the LLC control procedures have been defined to be basically the same as the CCITT X.25 Link Access Procedure HDLC in a balanced mode (LAPB). The balanced mode is used for peer structured networks wherein any station can originate communication directly with any other station.

Figure 8-16. Relationship of IEEE 802 Partition to ISO Reference Model



ST. GEORGE TECHNICAL COLLEGE LIBRARY

LLC control procedures are similar to CCITT's X.25 Link Access Procedure HDLC in a balanced mode.

Two basic types of services are provided within 802.2 LLC. The Type 1 service involves unacknowledged "connectionless" operation wherein the source station sends a message to another station (or stations) without having established a logical connection for sequencing and acknowledging messages. This mode of operation is intended for transmission of messages that are not essential and for systems wherein higher levels provide error recovery and sequencing functions (as in Ethernet).

The Type 2 service is the more conventional balanced data communications service which establishes logical connections between two LLCs. Each LLC can send and receive both messages and responses. Each LLC also has the responsibility of ensuring complete and accurate deliveries of its outgoing messages.

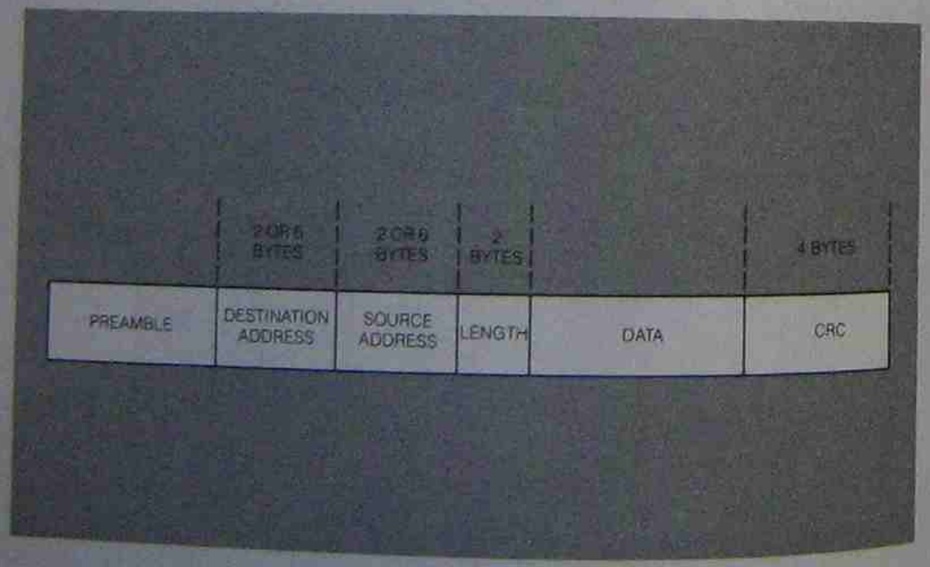
CSMA/CD — IEEE 802.3

The IEEE 802.3 standard is similar to the Ethernet specification but has 16-bit address fields, instead of 48.

The IEEE 802.3 standard defines the MAC sublayer for CSMA/CD and a corresponding physical layer for connection to a baseband coaxial cable. The standard is basically patterned after the Ethernet specification with certain options in the frame format shown in *Figure 8-17*. The address fields can be 16 bits long instead of Ethernet's 48 bits, but all addresses in a particular network must be of the same length. The 2-byte type field of Ethernet has been replaced by a 2-byte length field which designates the number of data bytes in the following message data field.

Variations in the physical layer allow signaling rates of 1, 5, 10, and 20 Mbps. A multidrop coaxial cable segment is limited to 500 meters (with 100 stations) instead of 1500 meters in Ethernet, but there can be as many as five coaxial cable segments in a IEEE 802.3 system.

Figure 8-17.
IEEE 802.3 Format for CSMA/CD

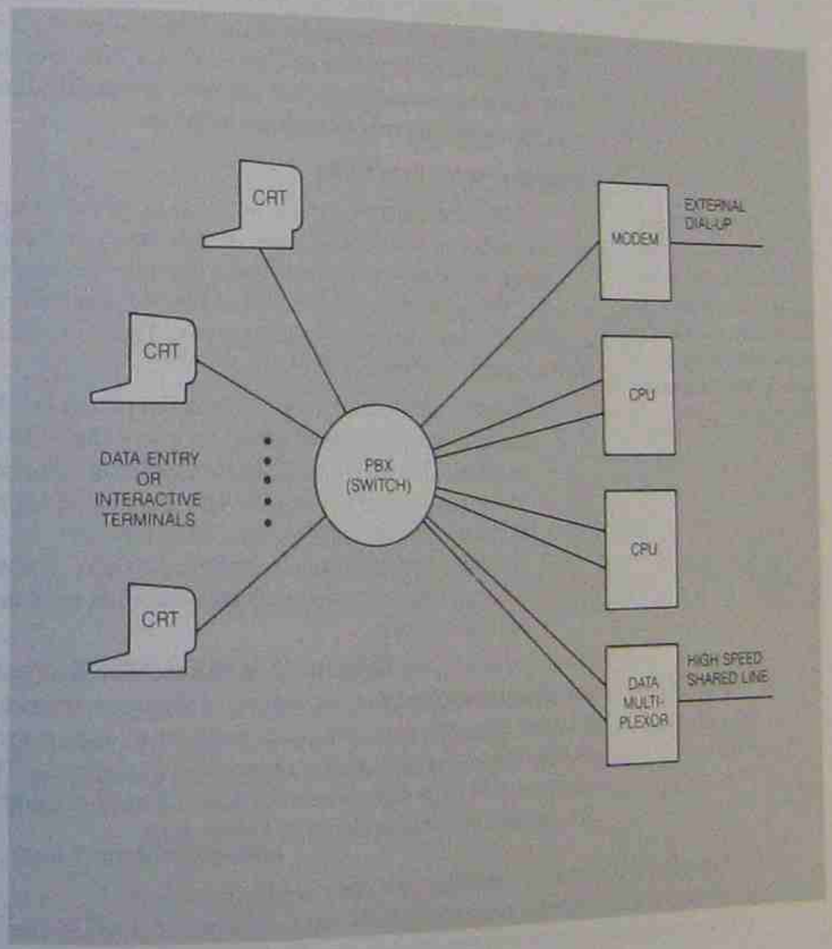


As LAN technology has been developing, PBXs have been evolving and using more digital electronics in voice switched services.

VOICE AND/OR DATA PBX

The local area networks described in the preceding sections have evolved primarily from the data processing or office equipment industry; therefore, they are strongly influenced by data communications applications. At the same time that LAN technology has been developing, voice telephone systems called private branch exchanges (PBXs) have been incorporating more and more digital electronics as a means of providing switched voice service. As shown in *Figure 8-18*, the conventional PBX architecture utilizes dedicated twisted pair wires from every station to the central node (switch). Connection requests are signaled by an off-hook signal and dialed digits or "hot-line" addresses stored in the memory of the switch controller. Once a connection is established, the associated transmission channels are dedicated to that connection for its duration.

Figure 8-18.
Representative Application of a Data PBX



PBXs are not capable of effectively handling high speed processor-to-processor communications because of 50 kbps to 1 Mbps transmission rate range.

The data PBX depicted in *Figure 8-18* provides an economical solution to several data communications problems.

1. A terminal can be switched to different computers (CPUs) for different applications, to modems for external dial-up connections, or to data multiplexers for efficient use of external leased lines.
2. The distance between the computer ports and the data terminals is extended to as much as several thousand feet. The standard EIA specification for the RS-232C interface is limited to 50 feet (although RS-232C interfaces are often used over longer distances).
3. The switch acts as a port concentrator; that is, many external device ports can be handled by a few computer ports. The switch accomplishes this by connecting only active terminals to the computer ports. This greatly reduces the number of ports required in each computer.

Although twisted pair wires are relatively inexpensive and easy to install, they restrict the capacity of the data channels to a range of 50 kbps to 1 Mbps, depending on distance. Thus, the data PBX is a viable solution for data entry and medium rate interactive terminal communications, but not for high-speed processor-to-processor communications.

Integrated Voice/Data PBXs

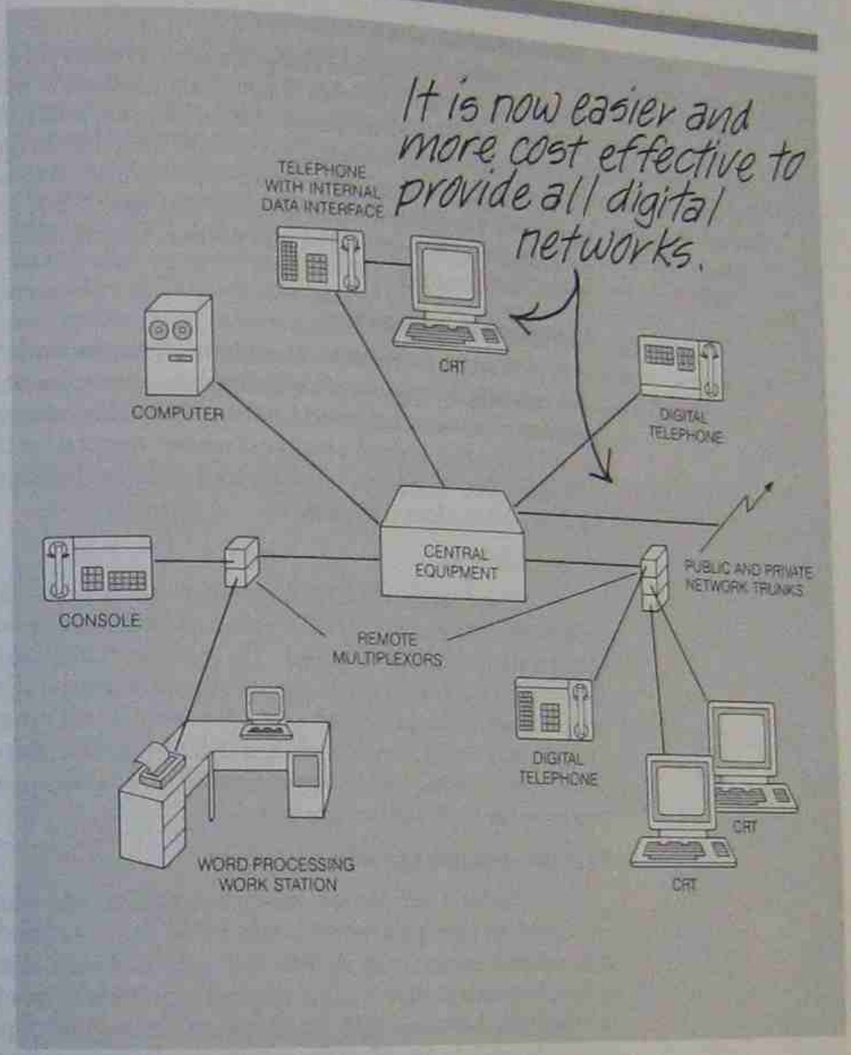
Digitization of PBX Systems will enable control information to be multiplexed onto the voice signal wires allowing only one or two wires to handle multiple lines instead of the normal 25 pairs.

Beginning in the mid-1970s, PBXs began utilizing digital technology in their switching matrices because, despite the cost of voice digitization at line interfaces, the total switch cost is lower due to economics in the matrix. More recently, the point of digitization has begun to move from within the central equipment to the voice stations themselves. The main motivations for digitizing in the instruments are:

1. To easily multiplex control information (which is inherently digital) onto the same pair or pairs of wires that carry the voice signals (typically 64 kbps). This allows sophisticated multiline telephones to operate over one or two pairs of wires. Conventional analog telephone systems use expensive 25-pair cables for multiline key sets.
2. To multiplex data channels on the same pairs of wires used for voice so one installation can simultaneously provide voice and data communications.

Initially, the digital instruments could be cost justified only in sophisticated applications. As the cost of digital instruments decreased, however, it became possible to cost justify an all-digital system, thus providing 64 kbps digital channels at every location of a telephone. The widespread presence of 64 kbps digital channels makes it easier to add other digital transmisson services as shown in *Figure 8-19*.

Figure 8-19. Integrated Voice/Data PBX with Digital Instruments



LAN COMPARISONS AND TRENDS

Each of the LAN system architectures presented in the previous sections have unique technical and operational advantages and disadvantages. The following paragraphs contrast the systems in terms of the properties of an ideal LAN.

One-Time Installations

The overall goal of a one-time installation is not satisfied in practice by any of the LAN architectures. Even a telephone system, which is a necessity in every building, typically requires significant rewiring as people move, if only from one side of an office to another.

Widespread Access

Before widespread access becomes a reality for an information distribution network, the cost of significant numbers of unused ports must be reduced. Even though telephone twisted pairs are relatively inexpensive, unused telephone ports typically represent a significant hidden cost in terms of the electronics interfaces that the wires are connected to in the central equipment. Distributed access architectures like CSMA or token passing do not produce hidden costs for unused ports as long as the maximum number of ports is not exceeded.

Application Independence

A major prerequisite for an information network to support multiple applications is the support of higher level data communications functions. The suppliers of coaxial cable based LANs are generally most advanced in this regard although it is more a matter of market orientation than a technology difference. The CSMA and token passing systems also provide a high data rate for those applications that need it.

Excess Capacity

In one sense, the CSMA or token passing architectures are ideal because the entire capacity of the system is available at every port. A port can use as much bandwidth as it needs. On the other hand, this feature limits the system capacity. If the total bandwidth requirements of the stations exceed the system bandwidth, the needed expansion is either very costly or impossible.

In contrast, circuit switched (PBX) architectures can be expanded almost without limit. The bandwidth delivered to every port remains unchanged as more stations are added.

Easy Maintenance and Administration

The most difficult systems to maintain are the broadband systems because of the tight gain requirements of the amplifiers and equalizers. From a data communications point of view, the integrated voice/data PBX is the easiest to maintain since well established procedures already exist for the voice services. Furthermore, PBX suppliers are accustomed to providing turn-key system support for the life of the installation. However, due to product background, they typically do not support higher level communication requirements.

No single LAN system architecture has sufficient capabilities to provide all possible data communication transmission needs. Each one has its own special qualities and technical advantages and disadvantages.

No presently available single LAN system architecture can economically satisfy the needs of all communications within a building or campus. Nor is it likely that one system will ever evolve to economically fulfill these needs. Thus, there will always be a need for either separate systems tailored to specific applications or possibly hybrid systems employing the best features of selected individual architectures. One such hybrid architecture that is beginning to evolve is shown in *Figure 8-20*. This figure depicts an integrated voice/data PBX that collects terminal traffic and switches it onto high-speed multiplexed computer to PBX interfaces (CPIs). Communications between processors (CPU) and their shared high-speed peripherals is best accomplished with a high-speed coaxial cable or fiber optic LAN. The advantage of the configuration in *Figure 8-20* is that low-speed, short messages to and from the terminals are not contending for the high-speed facility optimized for file transfers.

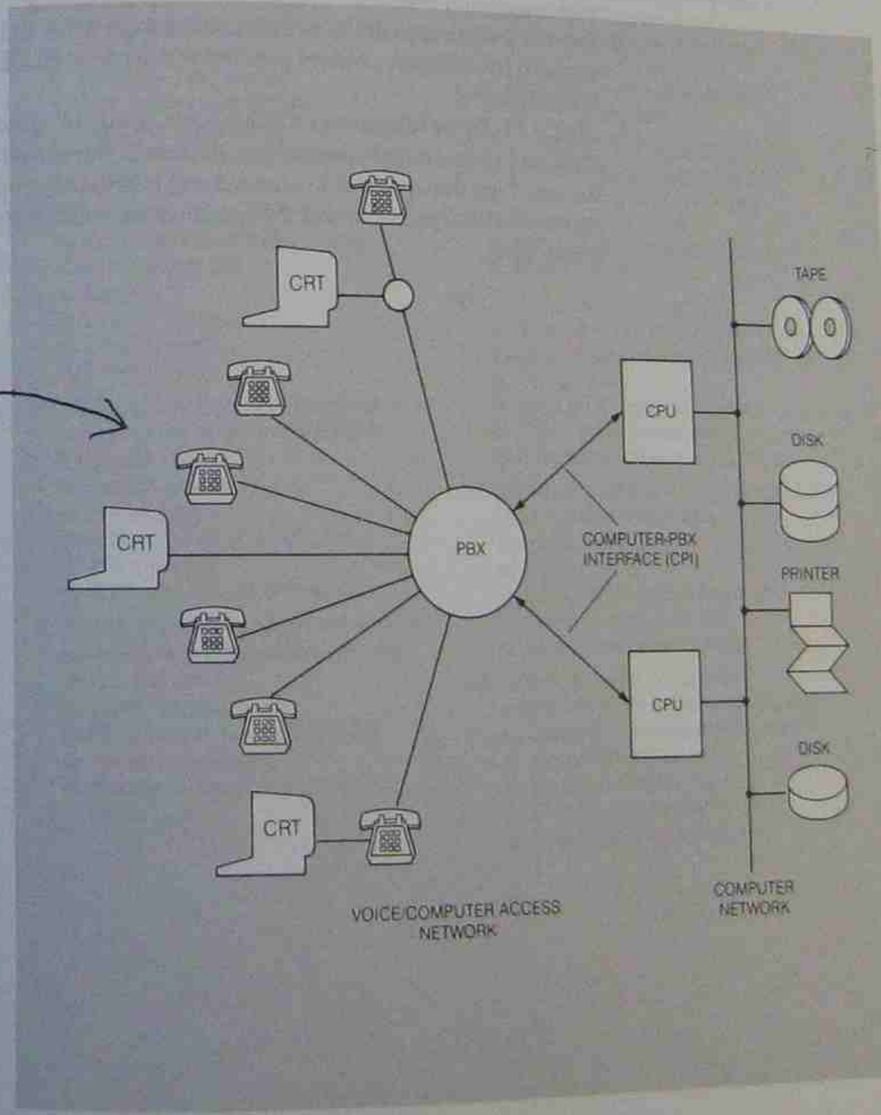


Figure 8-20. Hybrid PBX/LAN Network

A hybrid system allows tailoring to specific needs.

WHAT HAVE WE LEARNED?

1. LANs provide high-speed switched connections between such data equipment as computers, storage devices, printers, word processors, and display terminals at distances up to a few thousand meters.
2. The usefulness of a LAN is maximized if it provides higher level data communications support in addition to basic transmission and switching functions.
3. Communications functions within a network should be partitioned in a hierarchial manner to isolate the implementation of major functional components from each other.
4. A myriad of LAN architectures are possible, but the most prevalent are: 1) contention access based coaxial cable systems (for example, Ethernet); 2) polled access based coaxial cable systems (for example, ARCNet); and 3) frequency division broadband systems based on cable television technology and practice.
5. Some of the incompatibilities that traditionally arise in LAN equipment supplied by different vendors is being resolved by IEEE standards committee 802.
6. Digital PBXs for telephone switching are being designed to carry data in addition to basic telephone service. Although these systems do not provide the very high data rates of a conventional LAN, they represent a very economical alternative to an LAN for moderate data rate switched connections.

Quiz for Chapter 8

1. Which of the following transmission systems provides the highest data rate to an individual device?
   a. Voiceband modem.
   b. Local area network.
   c. Computer bus.
   d. Digital PBX.
2. Which of the following systems provides the longest digital transmission distances?
   a. Voiceband modem.
   b. Local area network.
   c. Computer bus.
   d. Digital PBX.
3. Which of the following is a characteristic of a LAN?
   a. Parallel transmission.
   b. Unlimited expansion.
   c. Low cost access for low bandwidth channels.
   d. Application independent interfaces.
4. Which of the following transmission media is not readily suitable to CSMA operation?
   a. Radio.
   b. Optical fibers.
   c. Coaxial cable.
   d. Twisted pair.
5. Which of the following functions is not provided as part of the basic Ethernet design?
   a. Access control.
   b. Addressing.
   c. Automatic retransmission of a message.
   d. Multiple virtual networks.
6. Which of the following is not a useful property of a Manchester line code for an Ethernet?
   a. Continuous energy.
   b. Continuous clock transitions.
   c. No dc component.
   d. No signal change at a 1 to 0 transition.
7. Which of the following data communications functions is generally provided for in a LAN?
   a. Data link control.
   b. Applications processing.
   c. Flow control.
   d. Routing.
8. The purpose of the preamble in an Ethernet is:
   a. Clock synchronization.
   b. Error checking.
   c. Collision avoidance.
   d. Broadcast.
9. Which of the following is possible in a token passing bus network?
   a. Unlimited number of stations.
   b. Unlimited distances.
   c. Multiple time division channels.
   d. In-service expansion.
10. Which of the following is not possible in a token passing loop network?
    a. Unlimited number of stations.
    b. Unlimited distances.
    c. Multiple time division channels.
    d. In-service expansion.
11. Which of the following LAN architectures can be expanded to the greatest total system bandwidth?
    a. Digital PBX.
    b. CSMA/CD baseband system.
    c. Token passing network.
    d. Broadband cable system.
12. Which of the following systems is the most capable of servicing a wide range of applications?
    a. Digital PBX.
    b. CSMA/CD baseband system.
    c. Token passing network.
    d. Broadband cable system.
13. Which of the following is a characteristic of a token passing ring as opposed to a token passing bus?
    a. Signal quality control.
    b. Passive interface (tap).
    c. Flexible polling sequences.
    d. Priority access capabilities.

14. Which of the following is not a characteristic of the hub architecture of ARCNet?
    a. Directionalized transmission.
    b. RIM port isolation.
    c. Zero insertion loss amplifiers.
    d. Alternate routing.
15. Which of the following can not be provided in a broadband LAN?
    a. Frequency agile modems.
    b. Closed-circuit TV.
    c. Voice circuits.
    d. Fiber optic transmission.
16. Which of the following is not possible in a digital (data) PBX using twisted pair transmission?
    a. Computer port concentration.
    b. 64 kbps data circuits.
    c. High speed file transfers.
    d. Transmission up to several thousand feet.
17. Which of the following is not a motivation for digitizing a voice signal in the telephones of a digital PBX?
    a. Simplified control signaling.
    b. Lower cost telephones.
    c. Fewer wire pairs.
    d. Multiplexed voice and data channels.

Architectures and Packet Networks

ABOUT THIS CHAPTER

This chapter deals with several data networking architecture concepts; that is, the way data networks are arranged or structured. These include the concept of architectural levels or layers. Without emphasizing architectural structure or terminology, previous chapters have actually described data networking and switching on two different levels: the physical level and the link level. Packet networks, which involve a third architectural level, are discussed in this chapter. The CCITT X.25 standard packet switching architecture is described; then the Open Systems Interconnection (OSI) Reference Model is explained along with its relationship to X.25 and other systems.

PROTOCOL LAYERING

A protocol is a set of rules governing a time sequence of events that take place between peer entities; that is, between equipment or layers on the same level.

Physical Layer

Level one protocols deal with the physical layer of data communication; that is the passing of the bits through the wires to and from DTEs. Link protocols cover a higher level of architecture; they are concerned with all aspects of maintaining order within the link.

In a previous chapter, we discussed the mechanical, physical, electrical, logical, and functional relationships between the various wires and signals in a serial interface and how the bits of data are passed through them between a Data Terminal Equipment (DTE) and Data Circuit-terminating Equipment (DCE). We showed, for example, that bits passing through the transmit data wire from a source DTE should eventually pass through a receive data wire to a sink DTE. These are called physical layer or level one protocols.

Link Layer

In a previous chapter, we discussed means by which various fields, such as address, text, and error checking, may be present within the same bit stream. A number of specific link protocols were discussed, showing techniques which are used to define (without any possibility of confusion) the boundaries of the fields within a bit stream, send information to a specific terminal on a multipoint link, check for and correct transmission errors, and generally maintain order within the link.

Architecturally speaking, link protocols are at a higher level (a more "intelligent" level) than physical protocols; yet all of the information used in the link protocol is actually contained in the bit stream transmitted through the serial interface. Conceptually, we speak of the link protocol fields as embedded (contained within) or layered within the physical protocol bits.

Network Layer

The highest level of protocol architecture is the network layer. This layer is used to route data link-to-link through a network.

At the third architectural level is the network layer. Protocols at this layer may be used to route data from link to link through a network containing intelligent nodes called packet switches. The method by which this routing and associated administration takes place is called the network layer protocol. The network layer protocol information, called the packet header, is embedded in the information field of a link-level frame. The packet switch uses part of this information to route the data to the next link, and so on. The packet header, together with the user data, is called a packet.

PACKET NETWORKS

Packet networks function with the help of special intelligent switching nodes. A network node joining three or more links is called a DSE.

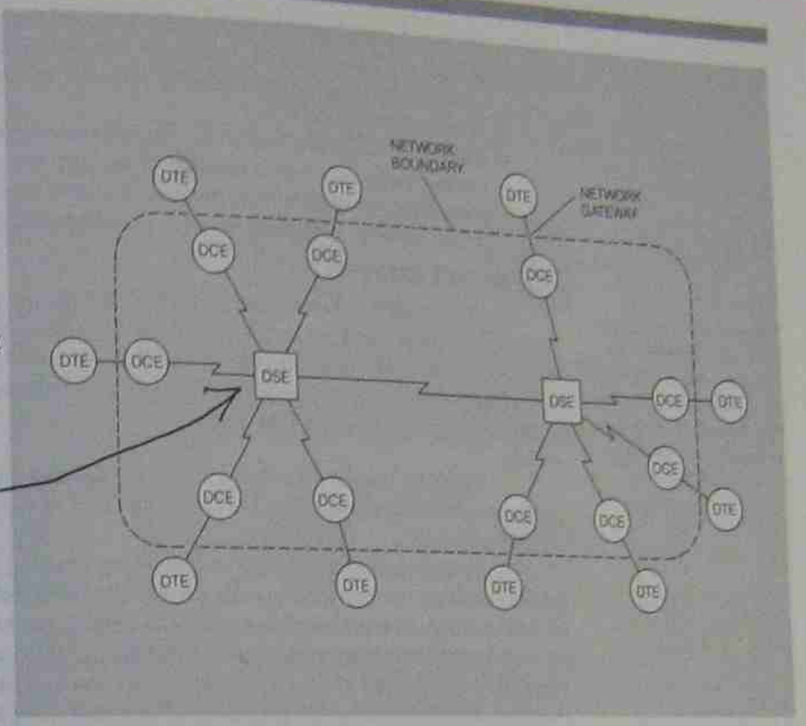
A packet network is a special kind of data network containing intelligent switching nodes. Packet networks have the following general characteristics:

1. Prior to transmission, each data message is segmented into short blocks of specified maximum length, and each block is provided with a header containing addressing and sequencing information. Each packet becomes the information field of a transmission at the link protocol level which usually contains error control capabilities.
2. The packets are passed very quickly from node to node, arriving in a fraction of a second at their final destination.
3. The node computers do not archive (store) the data. Messages are "forgotten" by the sending node as soon as the next node checks for errors (if required), and acknowledges receipt.

The terms DTE and DCE have been discussed previously. In packet networks, an additional term is introduced — the Data Switching Exchange (DSE). A DSE is a network node joining three or more links. *Figure 9-1* shows the relationship between DTEs, DCEs, and DSEs in a packet network. The point where the serial interface cable connects to the DCE is sometimes referred to as the network gateway.

Unlike level one switching, in which a specific link is dedicated to a particular message or group of messages, data packets are forwarded from DSE to DSE in such a way that packets from many sources and to many sinks may pass through the same internode link at different moments during the same short period of time.

Figure 9-1. Two-Node Packet Network



Advantages of Packet Switching

Packet switching can offer significant advantages over other data communication techniques.

Depending on the situation, packet switching can offer several possible advantages over other data communications techniques:

1. For data applications in which the amount of traffic between terminals cannot justify a dedicated circuit, packet switching may be more economical than transmission over private lines.
2. For applications in which data communication sessions are shorter than a minimum chargeable time unit for a telephone call, packet switching may be more economical than dialed data.
3. Because destination address information is inherently a part of the packet, a large number of messages may be sent to many different destinations as fast as the source DTE can turn them out. Depending on the type of packet service being used, there might not be any connection time delay before transmitting packets containing actual data.
4. Because of the intelligence built into the network (that is, computers at each node), dynamic routing of data is possible. Each packet travels over the route established by the network as the best available path for that packet at that time. This characteristic can be used to maximize efficiency and minimize congestion.

5. Built-in intelligence also facilitates a "graceful degradation" property of the packet network because, whenever there is a failure of a link or node, packets may be automatically rerouted around the defective portion of the network.
6. Because of the intelligence within the network, a rich array of basic communication services is possible. Examples are: error detection and correction, message delivery verification, group addressing, reverse billing, message sequence checking, and diagnostics.

X.25 PACKET SYSTEMS

The primary emphasis of the X.25 standard is to describe what a terminal and its associated network must be able to do; not describe the terminal and network themselves.

The CCITT X.25 standard for packet switching systems is one of the most significant networking architectures affecting data communications for the present and the foreseeable future. The X.25 standard is actually only a part of a much larger collection of CCITT recommendations on public data networks. To fully understand the subject, one must study the entire "X" series.

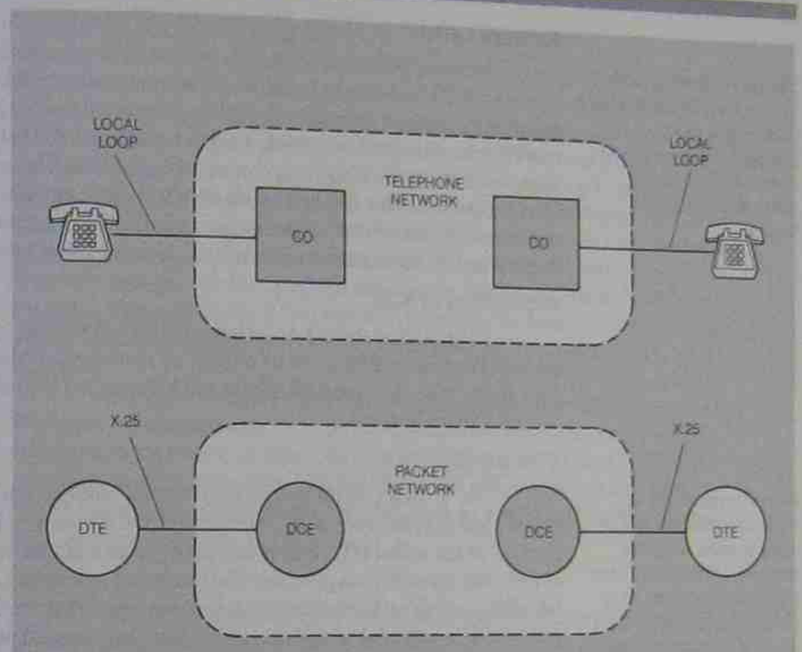
The X.25 standard itself describes the physical, link, and network protocols *in the interface between the DTE and DCE at the gateway* to a packet switching network.

At first glance, the concept of describing a packet network in terms of gateway parameters may seem peculiar, but from the users' point of view, it's all that's needed. In voice telephone network terms, if we were to collect all of the standards relating to the physical, electrical, logical, and functional properties of local loops and their signals into one document, we would have the telephone network equivalent of X.25. As shown in *Figure 9-2*, X.25 relates a data terminal to the gateway of a packet network in precisely the same way that local loop signaling standards relate a telephone to the central office of a telephone network.

If we had never seen a telephone or placed a telephone call, our knowing all there is to know about a local loop and its signaling wouldn't teach us much about the design of the telephone or of the network; but we would know how to use them. Similarly, the X.25 standard describes neither the data terminal nor the packet data network, but X.25 reveals a great deal about what the terminal and the network must be able to do. It is for this reason that the X.25 has taken on so much importance.

As already mentioned, the X.25 standard specifies three separate protocol layers at the serial interface gateway. The physical-level characteristics of X.21 (or X.21 bis) are specified as the physical layer protocol. ("bis" is the Swiss French equivalent of "alternate form.") X.21 uses a 15-pin synchronous interface which has not enjoyed enthusiastic acceptance in the U.S. because of the popularity of RS-232 and V.35 which existed before X.21.

Figure 9-2. Analogy Between X.25 and the Total Specifications for a Local Loop



In recognition of this fact, the CCITT has endorsed X.21 bis as a suitable alternate. X.21 bis specifies V.24/V.28 (essentially the equivalent of RS-232C), X.26/IS-4902 (essentially the equivalent of RS-449), and V.35 at their appropriate respective bit rates. Since U.S. vendors generally provide RS-232C or V.35, no change is required.

Link Level

The protocol at the link level, LAPB, provides for two-way simultaneous communications between the DTE and DCE at the network gateway.

At the link level, the protocol specified is a subset of HDLC, referred to in X.25 terminology as LAPB (Link Access Procedure Balanced). (An older version of X.25 used a link protocol called LAP which is being phased out.) LAPB provides for two-way simultaneous communication between the DTE and DCE at the network gateway. LAPB frame structure is identical with that of SDLC which was explained in a previous chapter. The delimiter (flag), abort, idle channel, transparency (zero insertion), frame sequencing, flow control, and error control mechanisms are identical with those of SDLC. The differences in the LAPB procedures and the SDLC procedures are mainly in the areas of line control and addressing; these will be explained later.

The packet level of X.25 is the network level and it is transmitted as an information field in the command frame. At least three octets are used in the header.

Network Level

The network level of X.25 is referred to in the standard as the packet level. All X.25 packets are transmitted as information fields in LAPB information command frames. A packet contains at least a header of three or more octets. (An octet is a group of eight bits.) Most packets also contain user data, but some packets are only for control, status indication, or diagnostics. (The preceding terms will be explained in later text and illustrations.) The maximum amount of user data that can be included in a data packet is determined by the network vendor, but is usually 128 octets.

Capabilities of X.25

The X.25 standard, together with its references, specifies two essential services which must be offered by carriers in order to be in full compliance with the standard. These are Virtual Call Sevice and Permanent Virtual Circuit Service.

Virtual connections are made and disconnected by using special, non-data carrying packets that have a unique bit stream. In a virtual connection no fixed physical path actually exists in the network.

Virtual Call Service

In Virtual Call Service (VC), a virtual "connection" normally must be established first between a logical channel from the calling DTE and a logical channel to the called DTE before any data packets can be sent. Establishment of a virtual connection is precisely the functional equivalent of placing a telephone call prior to beginning a telephone conversation. The virtual connection is established and disestablished using special packets having unique bit streams, but usually containing no data. Once the connection has been established, the two DTEs may carry on a two-way dialog until a "clear request packet" (disconnect) is sent.

The VC connection is referred to as "virtual" because a fixed physical path does not really exist through the network. The intelligence in the network simply relates a specified logical channel number at one DTE to that at the other DTE.

A given DTE may have many logical channel numbers active in the same X.25 interface at the same time. (That statement was important. Read it again.) Once a virtual connection has been established, say, between logical channel #319 on one DTE and, say, #14 on another DTE, then the actual data packet headers need refer only to logical channel numbers. The intelligent network keeps up with their conversational relationship; that is, a data packet from logical channel #319 at the one DTE will appear on logical channel #14 at the other, with the network performing the conversion. Further details on this subject will be given later.

After a PVC has been established only data packets are sent between the terminals on the carrier-assigned logical channels.

Permanent Virtual Circuit Service

Permanent Virtual Circuit (PVC) service is the functional equivalent of a private line service. As in VC, no end-to-end physical pathway really exists; the intelligent network relates the respective logical channels of the two DTEs involved. No special packets are sent by the DTEs to establish or disestablish connections. A PVC is established by requesting it in writing from the carrier providing the packet network service and it remains in effect until disestablished by written request. Terminals between which a PVC has been established need only send data packets on the carrier-assigned logical channels as required. Since a DTE may have many logical channels active at the same time within the X.25 interface, some may be assigned as PVCs while others are used for setting up VCs as required — all in the same X.25 interface!

For both VC and PVC, the network is obligated to deliver packets in the order submitted, even if the physical path changes due to circuit loading, failure, or whatever. Costwise, the relationship between VC and PVC is similar to that between long distance telephone and private line services. Other standard packet network features, such as error control, diagnostics, etc., are equivalent for VC and PVC.

Facilities

In CCITT terminology, a facility is an optional feature offered with a service, sometimes at extra cost. X.25 specifies a rich array of facilities for VC and PVC services. In any case, the final choice of services and facilities is up to the customer.

Facilities are optional features available to customers to customize VC service.

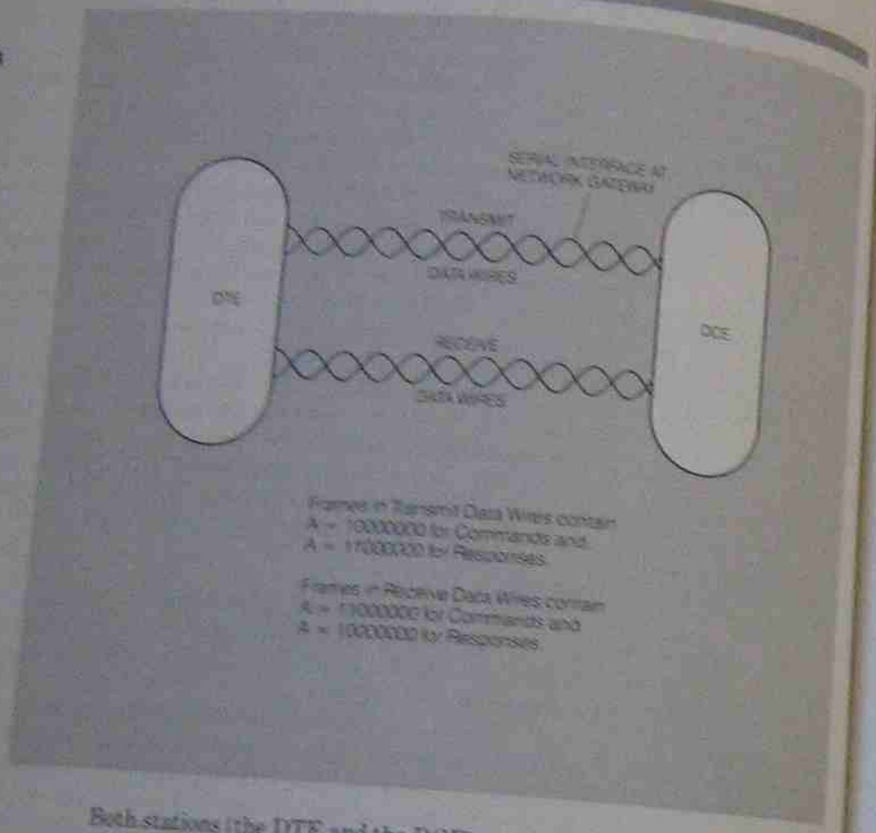
Examples of facilities for VC service are fast select, incoming calls barred, outgoing calls barred, closed user group, and reverse billing. Fast select gives the user the capability to send and receive data to and from a number of different remote stations very quickly and efficiently without actually establishing regular virtual connections. Such a capability has many potential uses for distribution of electronic mail, polling of inventories, etc.

LAPB Procedures

Space does not allow a complete discussion of the link layer protocol of X.25; however, the following information provides general knowledge of the LAPB procedures. A thorough treatment is included in the X.25 standard, and examples of operation are included in ISO 4335 Addendum 2.

LAPB provides for two-way simultaneous transmissions on a point-to-point link between a DTE and a DCE at the packet network gateway (*Figure 9-1*). Since the link is point-to-point, only the address of the DTE or the address of the DCE may appear in the A (address) field of a LAPB frame. These addresses are shown in *Figure 9-3*. The A field refers to a link address, not to a network address. The network address of the destination terminal is embedded in the packet header (PH), which is part of the I field.

Figure 9-3.
Commands Versus Responses in LAPB



Commands and responses may be made by both the DTE and the DCE. The function of a frame depends on the direction it is moving and the value of A.

Both stations (the DTE and the DCE) may issue commands and responses to each other as shown in *Figure 9-3*. Whether a frame is a command or a response depends on a combination of two factors:

1. Which direction is it moving: that is, is it on the transmit data wires from the DTE or the receive data wires toward the DTE?
2. What is the value of A?

Because of the addressing scheme, there can be no uncertainty even if frames are moving in opposite directions at the same time between DTE and DCE.

Tables 9-1 and *9-2* show the legitimate commands and responses, respectively, in the LAPB frame, along with their respective control octet values. Explanations of the abbreviations and terms follow the tables.

Table 9-1.
LAPB Commands

Command Name	Content of Control Octet				
	Bit Number				
	8 7 6	5	4 3 2	1	
I (Information)	N(R)	P	N(S)	0	
RR (Receiver Ready)	N(R)	P	0 0 0	1	
RNR (Receiver Not Ready)	N(R)	P	0 1 0	1	
REJ (Reject)	N(R)	P	1 0 0	1	
SABM (Set Asynchronous Balanced Mode)	0 0 1	P	1 1 1	1	
DISC (Disconnect)	0 1 0	P	0 0 1	1	

Table 9-2.
LAPB Responses

Response Name	Content of Control Octet				
	Bit Number				
	8 7 6	5	4 3 2	1	
RR (Receiver Ready)	N(R)	F	0 0 0	1	
RNR (Receiver Not Ready)	N(R)	F	0 1 0	1	
REJ (Reject)	N(R)	F	1 0 0	1	
UA (Unnumbered Acknowledgement)	0 1 1	F	0 0 1	1	
DM (Disconnect Mode)	0 0 0	F	1 1 1	1	
FRMR (Frame Rejected)	1 0 0	F	0 1 1	1	

Most frames in LAPB operations are commands. Response frames are issued only when a command frame with P = 1 is received.

Explanation of LAPB Commands and Responses

During LAPB operation, most frames are commands. A response frame is compelled when a command frame is received containing P = 1; such a response contains F = 1. All other frames contain P = 0 or F = 0.

SABM/UA is a command response pair used to initialize all counters and timers at the beginning of a session. Similarly, DISC/DM is a command response pair used at the end of a session. FRMR is a response to any illegal command for which there is no indication of transmission errors according to the frame check sequence (FCS) field.

I commands are used to transmit packets; that is, in the I field. Packets are never sent as responses. N(S) is a 3-bit packet counter capable of counting from 0 through 7 (000 through 111 in binary). After seven packets have been sent, the counter simply rolls over to 000 again for the next value of N(S).

While N(S) counts packets sent from one end of the link, the value of N(R) indicates the next value of N(S) expected to be returned from the other end of the link. By updating the value of N(R), a station acknowledges correctly received packets in the same way as SDLC frames are acknowledged.

For example, if a DTE sends an I command (a packet) for which N(S) = 5 and P = 1, and the DCE returns an RR response with N(R) = 6 and F = 1, the DTE knows that the packet was received correctly. If N(R) = 5 and F = 1, the packet was received incorrectly. No more than seven unacknowledged packets can be outstanding, as an ambiguity would result.

In checkpointing, a type of LAPB error detection, RR is the response by a station that has no packets to send but must answer a compelled response.

RR is what a station sends when it needs to send something, but has no packets to send. For example, a compelled response to an I command might be an RR with F = 1. This procedure is called checkpointing. Checkpointing is one of the techniques which LAPB makes available for error detection.

REJ is another way of requesting retransmission of frames. RNR is used for flow control to indicate a busy condition. It prevents further transmissions until cleared by an RR.

Packet-Level Procedures

The most interesting feature of X.25 is the network (packet) layer protocol. Remember that the packets are contained in the I command frames at the link (frame) level, and that each packet has a header of at least three octets. Most, but not all, packets contain user data; other packets are used for control, status indication, or diagnostics. *Figure 9-4* illustrates the general layout of X.25 packet headers.

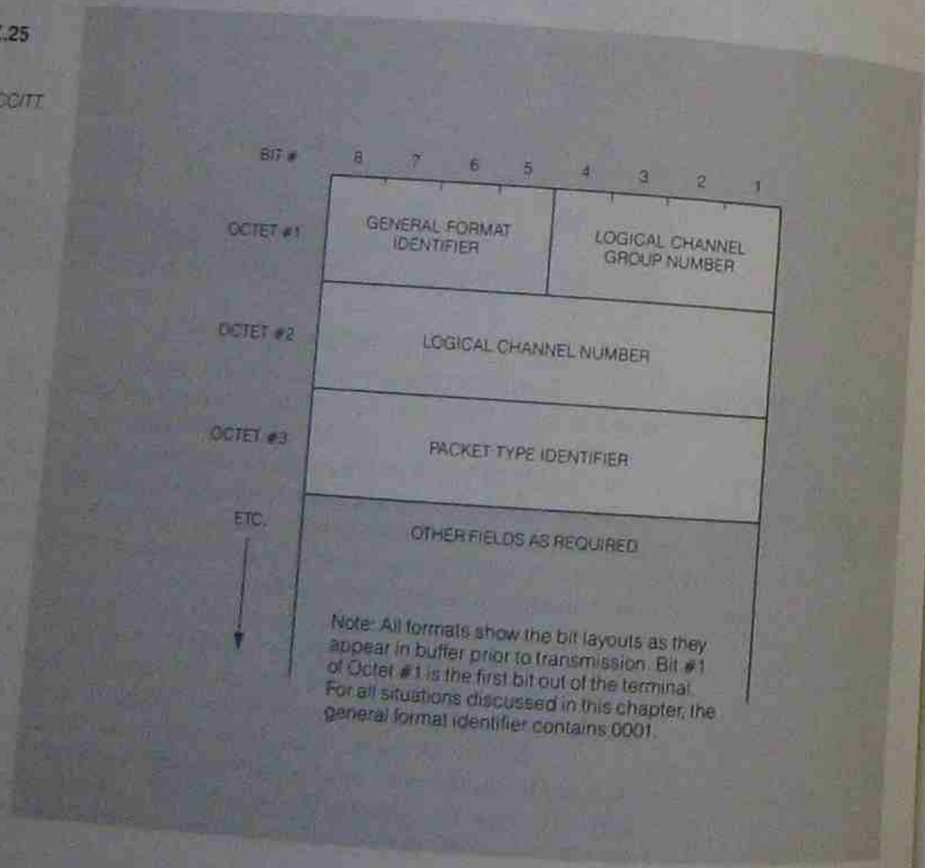


Figure 9-4.
General Layout of X.25 Packet Headers
(Source: CCITT, Recommendation X.25, CCITT, Geneva, C.H.)

Data Packets are packets where Bit #1 in Octet #3 of the header has a value of 0. A data packet can have up to 128 octets of user data.

Data Packets

A packet which has 0 as the value of Bit #1 in Octet #3 of the header is a Data Packet. Data packet headers normally contain three octets. A standard X.25 data packet contains up to 128 octets of user data following the header. *Figure 9-5* illustrates the layout of a typical data packet. (This chapter is not intended to cover all possible ramifications and extensions of X.25. There is, for example, the provision for modulo 128 P(S) and P(R) counters, each of which requires 7 bits. A data packet header in such a system would require a minimum of four octets.)

It was pointed out earlier that X.25 specifies two essential services; namely, Virtual Call (VC) and Permanent Virtual Circuit (PVC) services. Once a VC has been set up, the operation of data packets under VC service is identical with that under PVC. In the paragraphs that follow, operation under PVC will be described first; then the additional concerns of VC, such as call setup, will be described.

Theoretically, every X.25 gateway interface may support up to 16 logical channel groups, each containing up to 256 logical channels, for a grand total of 4096 simultaneous logical channels per gateway. It is up to the company that establishes the network to say how many it will actually support in each type of service; that is, PVC and VC.

There are three parts to Octet #3 of the data packet header: two 3-bit fields, P(S) and P(R) and a 1-bit field M. When M has a value of 1, it indicates that more data packets are to follow as a single unit.

Notice in *Figure 9-5* that Octet #3 of the data packet header contains two 3-bit fields, P(S) and P(R), and a 1-bit field M. M is the "more data" mark; when it is 1, additional data packets will follow which are to be considered as a unit; when it is 0, no more packets will follow in this unit. P(S) and P(R) are data packet counters, each of which may vary from 000 through 111 (0-7). They roll over to 000 again after passing 111. (These are not the same as N(S) and N(R) discussed earlier.) Only data packets have a P(S) counter; there will be a P(S) in each data packet sent across the DTE to DCE interface at the network gateway (*Figure 9-1*) and a P(S) in each data packet sent from DCE to DTE at the network gateway.

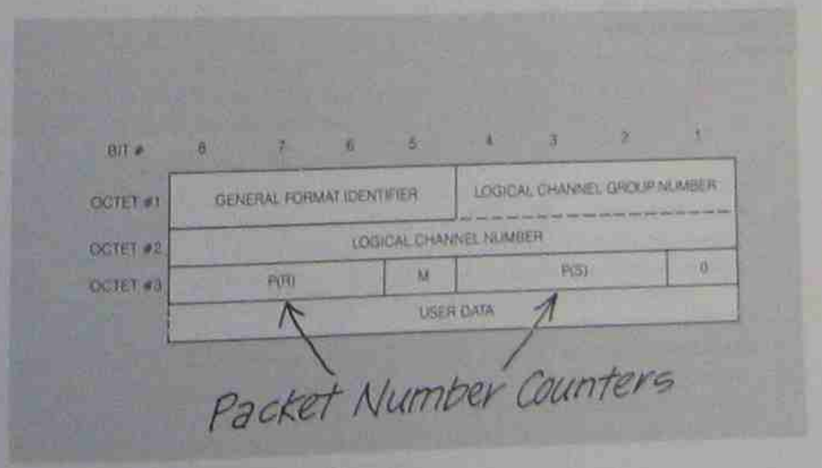


Figure 9-5.
Typical CCITT X.25 Data Packet
(Source: CCITT, Recommendation X.25, CCITT, Geneva, C.H.)

The values of P(S) and P(R) are related within a logical channel at the network level. P(R) is the amount that P(S) is expected to be in the next packet that is coming from the other direction on that particular logical channel.

By definition, P(R) is the amount of the next expected value of P(S) from the other direction on that logical channel (this will be explained below). The values of P(S) and P(R) relate to each other within a logical channel at the network level in exactly the same way that N(S) and N(R) relate to each other on a point-to-point link. The Ns are not related to the Ps, however, because the Ns refer to all packets sent in a link, and the Ps refer only to data packets sent in a particular logical channel within that link.

The use of P(S) and P(R) is illustrated in *Figure 9-6*. A sequence of two-way simultaneous (full-duplex) transmissions of data packets within a logical channel across a DTE/DCE interface at a packet network gateway is illustrated. In this figure, each rectangle represents a packet; the first number in the rectangle is the value of P(S) and the second is the value of P(R) in that packet's header. Inspection of the sequence in *Figure 9-6* will confirm two facts already implied:

1. The values of P(S) in a particular direction in a logical channel proceed in numerical order: 0, 1, 2, 3, etc.
2. The value of P(R) sent in a packet is not updated to i until the entire packet whose value of P(S) is i-1 has been completely received error free. Such update is an acknowledgement of all packets in that logical channel through P(S) = i-1.

The dynamic establishment and disestablishment of a VC using special packets is very similar to that of placing and terminating a telephone call on the public telephone network, as indicated in *Tables 9-3* and *9-4*. Inspection of these tables reveals that seven new packet names have been introduced. An event occurring on one side of the network is usually paired with a complementary event on the other side. For example, a Call Request Packet from calling DTE to DCE at one interface results in an Incoming Call Packet from DCE to called DTE at another interface. *Table 9-5* shows the packet type identifiers (Octet 3 of header) for these pairs of packet types.

Figure 9-6. Values of P(S) and P(R) for a Sequence of Data Packets

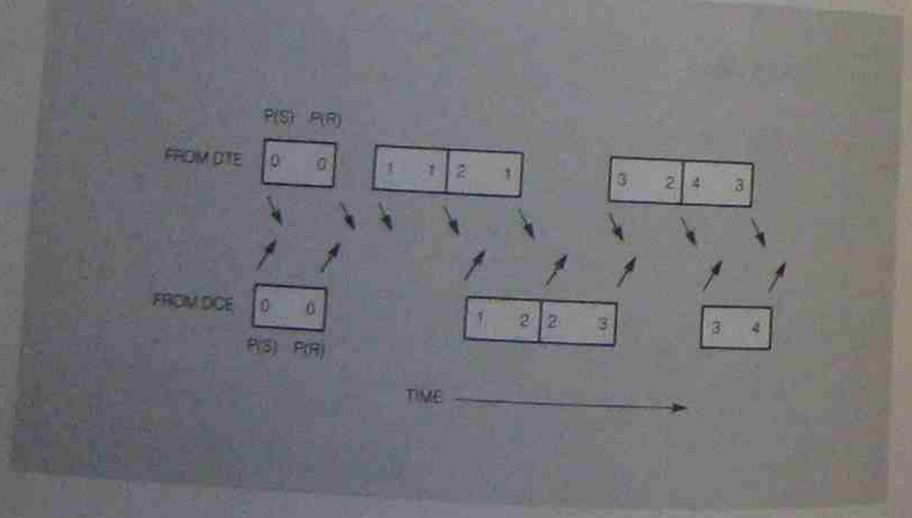


Table 9-3. Network Analogies for Call Establishment

Telephone Network	X.25 Packet Network
Place call	Send "Call Request" Packet
Hear telephone ring	Receive "Incoming Call" Packet
Pick up handset of ringing phone	Send "Call Accepted" Packet
Hear "Hello"	Receive "Call Connected" Packet
Fail to answer ring	Send "Clear Request" Packet*
No answer, busy, or fast busy	Receive "Clear Indication" Packet*
Hang up	Send "Clear Confirmation" Packet

*With cause shown in packet

Table 9-4. Network Analogies After Establishment of Call

Telephone Network	X.25 Packet Network
Hang up to disconnect	Send "Clear Request" Packet*
Hear other person hang up	Receive "Clear Indication" Packet*
Hang up after other person hangs up	Send "Clear Confirmation" Packet
Call disconnected by network	Receive "Clear Indication" Packet*

*With cause shown in packet

Figure 9-7 shows a "normal" sequence of a VC setup, two-way simultaneous data transfer, and terminal-initiated disconnect. Note that the four packet type identifiers of *Table 9-5* are simply passed through the network without the network modifying them.

Figure 9-8 shows what happens when a VC does not complete. If the VC is refused by the called DTE, the called DTE sends Clear Request. If the VC is refused by the network, the calling DTE receives Clear Indication.

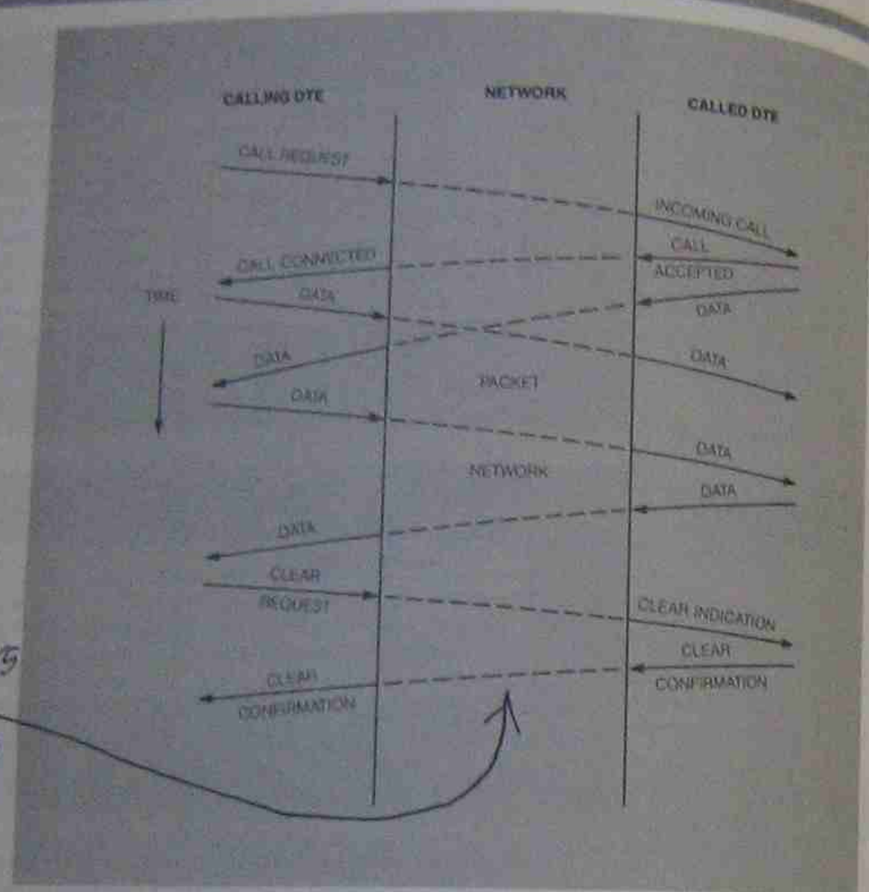
The numbers assigned to the ends of a logical channel do not have to be the same. The intelligent switcher will make the appropriate connections as required regardless of the end numbers.

Call Request/Incoming Call Packets

The headers of the Call Request/Incoming Call pair of packets are the most complex of the various header formats. Not only do they contain the basic three octets shown in *Figure 9-4*, but they also contain destination addressing and facility selection information. In certain cases, they also contain user data.

The logical channel numbers (LCNs) and logical channel group numbers (LCGNs) for PVC service are assigned by the carrier from the ranges of numbers that the carrier decides to designate for PVC service. The numbers assigned to opposite ends of a given logical channel do not have to be the same. Different numbers could be assigned to the opposite ends of the channel, and the intelligent switches within the network would make the appropriate conversions as the packets pass to and fro.

Figure 9-7.
A "Normal" VC Sequence



The packet type identifiers are passed without being changed.

Figure 9-8.
VC Not Completed

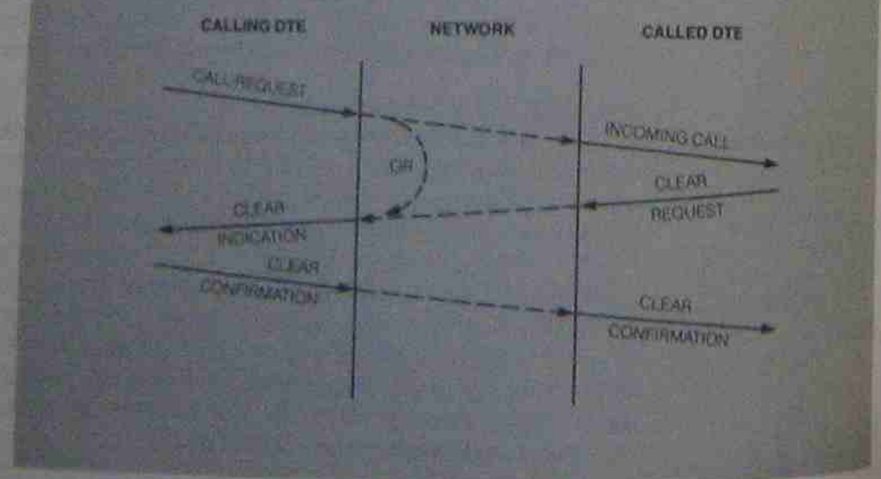


Table 9-5.
Packet Type Identifiers for Call Setup and Clearing

DTE to DCE	DCE to DTE	Octet #3
Call Request	Incoming Call	00001011
Call Accepted	Call Connected	00001111
Clear Request	Clear Indication	00010011
DTE-initiated —	Network-initiated —	
Clear Confirmation	Clear Confirmation	00010111

The logical channel numbers used in VC service will come from a pool of numbers available when the call is made. The numbers at the opposite interfaces are different.

In the case of VC service, the numerical values of logical channels at the opposite interfaces are never the same; the network always converts these numbers to and fro. The channel numbers used are assigned from a pool of numbers available at the time of call set up.

In placing a Call Request, a calling DTE will select the highest logical channel number which does not exceed a limit specified by the carrier and which is not already being used on another VC. The network will then pair the selected logical channel number with that VC from that calling DTE. At the destination end, the network will select the lowest logical channel number which is not less than a limit specified by the carrier and which is not already in use with that called DTE. The network will then pair that logical channel number with that VC to that called DTE. The two DTEs engaged in PVC or VC transactions each never know (or care) what LCGN or LCN is being used by the other DTE/DCE interface.

VC Destination Addressing

The DTE Address field in the Call Request/Incoming Call Packet format contains digits which are handled much like telephone numbers, with up to 14 digits on international VCs. The arrangement of these 14 digits is given in X.121. For domestic calls, fewer digits are required.

There are many other details in X.25 that cannot be covered here and the list of facilities and diagnostic codes continues to grow. The reader is referred to the standard for details.

VALUE-ADDED SERVICES

Enhanced, or value-added, services are provided by some carriers to overcome some of the shortcomings of X.25 that can arise when a terminal does not meet some of the prerequisites for full X.25 use.

X.25 describes a highly flexible basic point-to-point service through a public network. To the user whose terminals 1) operate in the X.25 packet mode, 2) have a significant amount of traffic to send to many locations, and 3) connect via a DTE/DCE serial interface to an X.25 network gateway, X.25 has few shortcomings. For the user who cannot meet all of these criteria, X.25 does have limitations. Enhanced (value-added) service offerings, such as those provided by AT&T Information Systems, GTE Telenet, Tymnet, and others, seek to address these and other matters:

1. Broadcast services (sending the same data at the same time to many receivers).

2. Dial access. (Since the DTE/DCE interface is intended as a two-way simultaneous synchronous gateway, packet-mode access normally cannot be provided via the public telephone network.)
3. Conversion to lower speed start-stop transmission for low-volume terminals.
4. Code and protocol conversion services.

The CCITT has provided guidance with regard to item 3, as will be shown below.

THE X SERIES OF RECOMMENDED STANDARDS

Services and facilities, terminals and interfaces are covered by CCITT sections X.1 through X.39. Network architecture, transmission, switching and all remaining areas are covered by section X.40 through X.290.

As stated earlier, X.25 is part of the "X" series of recommended standards for public data networks being promulgated (made public) by the CCITT. The X series is classified into two categories: X.1 through X.39, which deal with services and facilities, terminals and interfaces; and X.40 through X.199, which deal with network architecture, transmission, signaling, switching, maintenance, and administrative arrangements. From a packet network user viewpoint, the most important X standards are listed below with their titles and brief descriptions.

X.1 - International User Classes of Service in Public Data Networks

Assigns numerical class designations to different terminal speeds and types.

X.2 - International User Services and Facilities in Public Data Networks

Specifies essential and additional services and facilities.

X.3 - Packet Assembly/Disassembly Facility (PAD) in a Public Data Network

Describes the packet assembler/disassembler which normally is used at a network gateway to allow connection of a start-stop terminal to a packet network.

X.20 bis - Use on Public Data Networks of DTE Designed for Interfacing to Asynchronous Duplex V-Series Modems

Allows use of V.24/V.28 (essentially the same as EIA RS-232C).

X.21 bis - Use on Public Data Networks of DTE Designed for Interfacing to Synchronous V-Series Modems

Allows use of V.24/V.28 (essentially the same as EIA RS-232C) or V.35.

X.25 - Interface Between DTE and DCE for Terminals Operating in the Packet Mode on Public Data Networks

Defines the architecture of three levels of protocols existing in the serial interface cable between a packet-mode terminal and a gateway to a packet network.

X.28 - DTE/DCE Interface for a Start-Stop Mode DTE Accessing the PAD in a Public Data Network Situated in the Same Country

Defines the architecture of protocols existing in a serial interface cable between a start-stop terminal and an X.3 PAD.

X.29 - Procedures for the Exchange of Control Information and User Data Between a PAD and a Packet Mode DTE or Another PAD

Defines the architecture of protocols behind the X.3 PAD, either between two PADs or between a PAD and a packet-mode terminal on the other side of the network.

X.75 - Terminal and Transit Call Control Procedures and Data Transfer System on International Circuits Between Packet-Switched Data Networks

Defines the architecture of protocols between two public packet networks.

X.121 - International Numbering Plan for Public Data Networks

Defines a numbering plan including code assignments for each nation.

Interrelationships Between X Standards

Many X standard sections touch on other sections with great detail or add depth to specific areas of concern. In some cases a section will describe conditions that must be met before other sections become applicable.

Figure 9-9 shows the interrelationships between many of these X standards. In summary: X.25 specifies the relationship between a packet-mode DTE and a packet network. X.28 specifies the relationship between a start-stop (asynchronous) DTE and an X.3 PAD which must reside between a non-packet DTE and a packet network. X.29 specifies additional relationships over and above those in X.25 which a packet-mode DTE must satisfy when communicating with a non-packet mode DTE through a packet network and PAD. X.29 also covers the relationships between two PADs when two non-packet mode DTEs are communicating through a packet network.

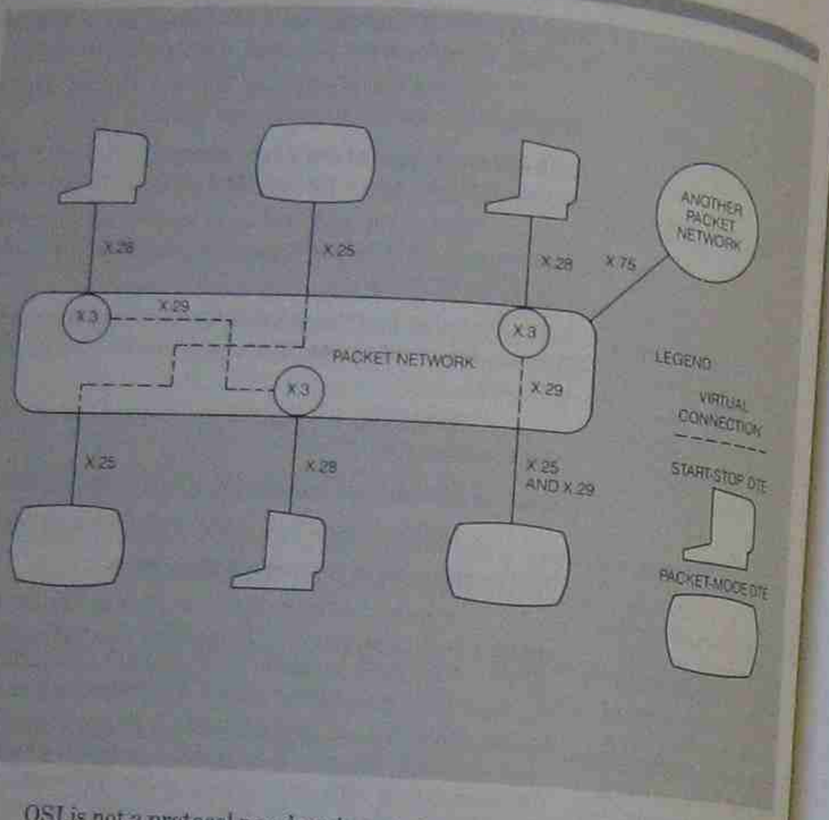
The CCITT envisioned the X.3 PAD as being within the packet network; that is, behind the DCE. The 1980 version of X.3 covers the PAD only for start-stop (asynchronous) terminals, but it leaves open the future possibility of covering other non-packet terminals; for example, IBM BSC. A number of stand-alone X.3 PADs are now marketed for connection of various start-stop and non-start-stop DTEs to packet networks.

OPEN SYSTEMS INTERCONNECTION REFERENCE MODEL

OSI/RM is a reference model for data communication architectures. It has been redrafted by the CCITT and designated CCITT X.200.

In 1983, the International Organization for Standardization (ISO) culminated six years of intensive effort to develop and publish a guideline — a reference model — for describing data communications architectures. The resulting document, International Standard #7498, has been redrafted by the CCITT using its own terminology, and the redraft has been designated CCITT X.200. The generic name for both documents is Open Systems Interconnection Reference Model (OSI/RM) — usually called simply OSI.

Figure 9-9. Relationships Between Some "X" Interface Standards



OSI defines a consistent language and boundaries for protocols so that systems that follow OSI can communicate with one another. However, OSI itself is not a protocol.

OSI is not a protocol nor does it contain protocols. What OSI does is define a consistent language and boundaries for establishing protocols such that systems which abide by its rules should be "open" to one another; that is, be able to communicate.

Decades of frustration caused by incompatibility between competitive and/or geographically distinct communications systems have created such an appetite for the establishment of cooperative systems that there is a virtual stampede toward global standardization of data communications architectures. Many standards committees in several organizations are now engaged in the development of protocols for the various layers in the numerous systems to which OSI will apply.

OSI Layers

OSI defines a complete architecture having seven layers. There is no specific reason why it is seven instead of six or eight; the functions that must be performed just happened to fall into roughly seven groupings. The lowest three layers (physical, link, and network) correspond closely to the physical, link, and packet levels of X.25. Protocols have been and are being developed for other kinds of systems, such as facsimile, integrated voice/data networks, videotex, etc. Even if the protocols were never written, OSI has had a tremendous impact in standardizing the way we view communications.

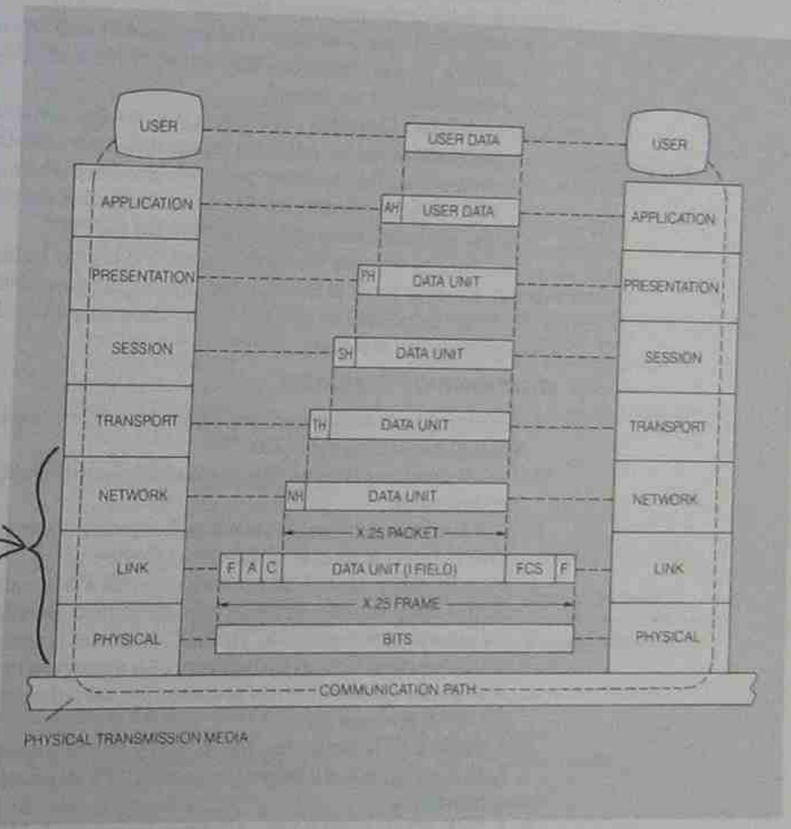
There are seven layers to any architecture as defined by OSI. The three lowest levels correspond roughly to the physical, link and packet levels of X.25.

The official names of the seven layers are, from top to bottom:

7. Application
6. Presentation
5. Session
4. Transport
3. Network
2. Link
1. Physical

Each successive step up the list takes one to a higher level of system supervision. In terms of the protocols, lower levels involve successively more embedded headers. *Figure 9-10* shows how the seven levels relate for a system whose lower three layers correspond to X.25. (In this case, NH is the packet header.) The concept of layers was previouisly discussed in Chapter 8.

Figure 9-10. Seven Embedded Layers of OSI



These correspond roughly to the equivalent layers of X.25.

An Analogy

Whenever people communicate, whether by computer or verbally, they invoke protocols at all seven levels of OSI. Without going into the laborious details and vocabulary of the OSI standard, we shall illustrate the OSI categorizations of the seven levels from top to bottom for a typical telephone call.

The basic requirements to establish, maintain and terminate communications without error are similar whether verbal or computer communications are being used.

7. *Application Layer Concerns* - Am I talking to the right person? Who is paying for this call? Is this the best time to talk, or should I call back later? Does the other party have a pencil and paper to take notes?
6. *Presentation Layer Concerns* - Are we talking the same language and dialect?
5. *Session Layer Concerns* - Can this situation be handled in one call or several? Will other people need to be brought in at different times? Who will control the discussion in a multiparty conversation? Who will re-establish the call if we're cut off?
4. *Transport Layer Concerns* - What is the most cost-effective way to handle this call (or these calls) consistent with priorities? What long distance carrier(s) should be used?
3. *Network Layer Concerns* - Dial the number and listen for call-progress signals. Redial if you get a busy signal or if cut off. Disconnect when the conversation is completed.
2. *Link Layer Concerns* - Talk when you're supposed to and listen when you're supposed to. Ask for a repeat if there is something you don't understand. Tell the other party to slow down if he's talking too fast.
1. *Physical Layer Concerns* - These are the actual sounds being uttered into the mouthpiece and heard from the receiver.

Anyone who has used a telephone in everyday business situations can relate to all seven levels of the reference model just described. By reflecting on X.25, the reader should be able to identify the analogies to the packet, link, and physical layer protocols.

WHAT HAVE WE LEARNED?

1. When protocols are layered, the higher level layers are embedded inside fields of less intelligent layers.
2. Packet switching involves fast store-and-forward computers at each network node.
3. X.25 is an international standard packet switching architecture specifying three layers of protocols at the network gateway.
4. When properly used, X.25 networks provide a fast, reliable, accurate, flexible, and cost effective data communications alternative.
5. X.25 specifies two services: Virtual Call service is analogous to a telephone call; Permanent Virtual Circuit service is analogous to a private line.
6. X.25 doesn't provide for dial access or for start-stop transmission, but enhanced services do.
7. A single X.25 interface can handle many logical channels at the same time.
8. X.25 is but a part of a larger group of CCITT standards on public data networks.
9. X.25/X.121 provides for an extremely large worldwide population of terminals.
10. Open Systems Interconnection is a reference architectural model which is revolutionizing the way we view telecommunications.

Quiz for Chapter 9

1. The electrical state of the control leads in a serial interface is a concern of:
   a. The physical-layer protocol.
   b. The link-layer protocol.
   c. The network-layer protocol.
   d. None of the above.
2. The X.25 standard specifies a:
   a. Technique for dial access.
   b. Technique for start-stop data.
   c. Data bit rate.
   d. DTE/DCE interface
3. The X.25 standard is:
   a. Required for all packet switching networks.
   b. A recommendation of the CCITT.
   c. A complete description of a public data network.
   d. Used by all packet terminals.
4. The X.25 standard for packet networks is analogous to:
   a. PBX standards for a telephone network.
   b. Handset standards for a telephone.
   c. Local loop standards for a telephone network.
   d. Switching standards for a telephone network.
5. The value of the "A" field in a LAPB frame specifies:
   a. The calling DTE for a virtual call.
   b. The called DTE for a virtual call.
   c. The data sink for a VC or PVC.
   d. One of the two ends of a serial interface.
6. A group of packets from a source through an X.25 packet system to a sink:
   a. Arrive in the same order sent for VC, but not for PVC.
   b. Arrive in the same order sent for PVC, but not VC.
   c. Arrive in the same order sent for both VC and PVC.
   d. None of the above.
7. The value of N(S) is related to the value of P(S) as follows:
   a. N(S) − P(S) = number of outstanding packets.
   b. N(S) − P(S) = number of outstanding frames.
   c. N(S) − P(S) = number of active logical channels in a link.
   d. They are not related.
8. A protocol is a set of rules governing a time sequence of events that must take place.
   a. Between peers.
   b. Between non-peers.
   c. Across an interface.
   d. None of the above.
9. The OSI Reference Model defines the functions for seven layers of protocols:
   a. Including the user and communications medium.
   b. Not including the user or communications medium.
   c. Including the communications medium but not the user.
   d. Including the user but not the communications medium.
10. The X.25 standard covers how many OSI layers:
    a. Three
    b. Four
    c. Seven
    d. None
11. Architecturally, link protocols are at a level:
    a. Lower than physical protocols.
    b. Higher than physical protocols.
    c. The same as physical protocols.
    d. None of the above.
12. A data packet is a packet header together with:
    a. A network layer.
    b. An administrative layer.
    c. User data.
    d. A packet switch.

13. A network node joining three or more links is a:
    a. DSE.
    b. DTE.
    c. DCE.
    d. DTE and DCE.
14. The X.25 standard specifies how many separate protocol layers at the serial interface gateway:
    a. 8
    b. 2
    c. 4
    d. 3
15. The LAPB frame structure and the frame structure of SDLC are:
    a. Opposite.
    b. Identical.
    c. Reversed.
    d. None of the above.
16. Establishing a virtual "connection" is functionally equivalent to:
    a. Placing a telephone call prior to a conversation.
    b. Connecting a virtual memory.
    c. Physically connecting a DTE and DCE.
    d. None of the above.
17. In X.25 network layer protocol, the data packets normally contain:
    a. One octet of header plus data.
    b. Two octets of header plus data.
    c. Three octets of header plus data.
    d. Four octets of header plus data.
18. The OSI reference model is:
    a. Worthless.
    b. A protocol.
    c. Not a protocol.
    d. None of the above.
19. Layer 1 of the OSI model is the:
    a. Link layer.
    b. Physical layer.
    c. Network layer.
    d. Transport layer.
20. The applications layer of the OSI model is the:
    a. Seventh layer.
    b. Sixth layer.
    c. Fifth layer.
    d. Fourth layer.

ST. GEORGE COLLEGE OF TAFE

Network Design and Management

ABOUT THIS CHAPTER

The user's terminal is the source and sink of all user data, but good network design and management is usually necessary in order to make data communications networks usable and affordable. By network design, we mean the selection of various circuit parameters and/or the selection and interconnection of various devices to accomplish design goals.

This chapter covers the goals, principles, and tools of efficient network design and management. Many of the administrative principles of network management, such as planning and budgeting, are similar to other kinds of management. The technical aspects to be discussed here include network design goals and throughput timing, networking devices, interconnection, fault isolation, fault correction, and quality assurance.

Any network, no matter how simple or complex, may be incorporated into a larger network. The approach of this chapter is to start with the principles of good network design for basic point-to-point networks, establishing a solid basis for these, and then to discuss some of the technologies for building more complex networks.

NETWORK DESIGN GOALS

To achieve an optimum cost of a data network vs application balance, factors other than equipment cost must enter into the final cost determination.

The greatest motivation for network design is cost. Cost includes a multitude of factors other than the prices of the terminals and networks. For example, the costs of local area networks are usually justified on the basis of higher speed, because the time of the people and computers who use these networks is expensive. In other situations, communications reliability is the basic cost justification. In many cases, the failure of a data communications system to function properly when needed can cost a business far more than the price of the system.

If price were no object, most data networks would be a simple point-to-point data channel. Indeed, if a point-to-point channel can be kept busy at a reasonably high bit rate for transmitting data that's worth sending, then it may be the most cost-effective way to handle a particular data communication task.

On the other hand, many data applications cannot cost-justify dedicated channels because of insufficient quantity and/or priority of the data. In such cases, it is sometimes possible to combine data from more than one source into a single long-haul transmission path by use of multiplexers and other networking devices. While the use of such techniques may allow cost

justification of the application, the addition of equipment is likely to reduce reliability and to increase the time and effort required to find and correct a problem.

The underlying objective of data networking is to strike a satisfactory balance between the accurate, timely, and secure delivery of user data and total cost. The term "satisfactory" implies that the user who is paying for the delivery system is satisfied with it.

Accuracy

Previous chapters have discussed the subject of accuracy in terms of error control. Not all data transmissions are required to be error-free; for example, overseas telex messages and cablegrams which contain human-language messages do not suffer loss of meaning if a word is occasionally misspelled. On the other hand, the transfer of accounting information between financial institutions requires a high level of accuracy.

Timeliness

The timely delivery of data involves: The amount of data to be transmitted, the importance and priority of the information, the average rate of transfer of error free data, and the availability of the network to the user.

Timeliness involves four general aspects which are discussed below:

Amount of Data

Obviously, the amount and priority of data will have an impact on the throughput rate required, but cost must be considered in establishing these parameters. In some extreme situations, the amount of data may be so great that the best way to transfer it will be by physically transporting truckloads of magnetic tape! At the other extreme, a satisfactory "networking" solution might be a simple low-speed transmission over a dialed point-to-point telephone channel.

Priority of Data

As a general rule, the importance and frequency of testing is directly proportional to the importance and priority of the data. An Italian proverb asserts, "Good things cost less than bad ones." While this proverb doesn't apply universally to data communications, it does remind us that the best solution isn't always the cheapest. There are numerous situations, for example in industrial process control and manufacturing, in which a total failure of a data link could cost a company more than $10,000 an hour. In such cases, reliability, testing, and planning for alternative methods obviously are of extreme importance.

Information Throughput Rate

The information throughput rate, properly known as the Transfer Rate of Information Bits (TRIB), will be covered in detail in the next section. Basically, TRIB is the average rate of transfer of the actual error-free bits of user data, not counting overhead bits.

Availability of the Network

Availability of the network is determined by three factors: access time, mean time between failures (MTBF), and mean time to restore service (MTRS). For private lines and permanent virtual circuits, access time is of no concern; for dialed connections and virtual calls, the time delay to establish the call might be significant in comparison with the transmission time if the amount of data is small. MTBF usually refers to the time between "hard" (permanent) failures; but if the error rate becomes high enough, the TRIB can nosedive even though the connection still exists. MTRS may involve either a temporary restoration, as in the case of dialed backup for a private line, or it may simply refer to repair time if no alternative service is available.

Security

As more and more confidential and proprietary information is being transmitted through data communication systems, the security of this information becomes more important.

Security is receiving increasing attention as data communication is used more and more for significant and important matters of everyday life. Many sources of information are now available which cover this area in detail. Besides the physical aspects of security, data communication security involves both privacy and authentication. Privacy refers to secrecy (use of codes) and protection from unauthorized access (use of passwords). Authentication has to do with ensuring that a data message hasn't been tampered with between source and sink, and with verifying that the sender of the message is as claimed.

In certain kinds of transmissions, authentication is more important than any other aspect of networking mentioned so far. For example, a Swiss bank receiving an international telex from another bank containing instructions to transfer funds between accounts will be interested in knowing whether extra zeros have been added to the amount, whether account numbers have been modified, and whether the claimed source of the message is authentic.

In summary, data network design involves a complex series of interrelated judgments involving these and other factors:

1. Cost of the delivery system
2. Priority of the data
3. Response time required
4. Throughput rate
5. Accessibility
6. Reliability
7. Testing
8. Contingency Planning
9. Privacy
10. Authentication

TRANSFER RATE OF INFORMATION BITS

The TRIB is a result of several factors other than the data link bit rate. In almost all cases, the TRIB is less than the bit rate at the serial interface.

As mentioned earlier, TRIB is sometimes called information throughput rate. By definition,

$$\text{TRIB} = \frac{\text{Number of information bits accepted by the sink}}{\text{Total time required to get those bits accepted}}$$

Due to the burst nature of most data transmission, TRIB has meaning only as an average over a period of time. Although the data link bit rate has a lot to do with TRIB, other factors may at times have as much or more influence. For

The major parameters that adversely affect the TRIB are transmission overhead and delays.

example, if the channel is noisy, a higher bit rate may increase the error rate to such an extent that the TRIB actually goes down.

TRIB is almost always less than the bit rate at the serial interface. (In a few sophisticated data compression systems, the TRIB may appear to exceed the actual interface bit rate because the receiver puts out more bits than were actually transmitted.) Two other parameters that usually affect TRIB are transmission overhead and delays. Both of these parameters are closely tied to the coding and blocking of the data and to the protocols used. For example, if the protocol requires that acknowledgement of a block must be received by the sender before the next block is transmitted, there will be round trip delays added to the denominator of the TRIB definition given above. In such instances, block length and delays can have a substantial effect on TRIB. Besides bit rate, the factors that should be considered in estimating TRIB are:

1. Non-information bits sent with the data. These include:
   a. Start and stop bits if asynchronous transmission is used.
   b. Parity bits if ASCII code is used.
   c. Redundant (or "stuffed") zeroes if a bit-oriented link protocol (SDLC/HDLC) is used.
   d. Filler bits used by teleprocessing system utilities to fill out partial data blocks, if any.
2. Non-information characters in the message stream. Depending on the link protocol used, these may include:
   a. Sync characters or flags.
   b. Address characters.
   c. Control characters (STX, ACK, etc.) or control byte.
   d. Error-checking characters (BCC or FCS).
   e. Transparency characters (DLE) and pads.
3. Non-information messages required in the administration of the link protocol. These are the initialization, connect, disconnect, polling and status messages, etc. They usually are not counted in TRIB calculations if the sessions are sufficiently long for them to have a negligible effect.
4. Carrier turn-on delay. This is the period of time between request-to-send from the terminal and clear-to-send from the modem. It is also known as "modem turnaround time." It is required only when a modem is operating with switched carrier on an analog circuit. Normally this situation would apply on a circuit which cannot support the bit rate of 4800 bps. It also applies to at least the remote (secondary) stations, and sometimes to all stations, on a multipoint circuit. The amount of the delay, if any, is programmed into the modem at the time of installation, based on the modem manufacturer's recommendations for the particular circuit type and bit rate. It can be from a few milliseconds to a few hundred milliseconds.

TRIB also is affected by transmission delays, buffer delays, response time computations, number of filler bits, and the transmission error rate.

5. Modem propagation delay. All synchronous modems operating on voice-grade circuits buffer the data both on transmit and receive. Depending on bit rate, the delay can be in the range of 2 to 10 milliseconds (ms) per modem.
6. Circuit propagation delay. Microwave radio signals, including communication satellite signals, and signals in optical fibers travel at near the speed of light. Electrical signals in local wires and cables travel somewhat slower. Also, many local telephone systems contain buffers of various types. As a rule of thumb, estimate 6 ms for local equipment on each end, and 1 millisecond for each 150 miles of cross-country terrestrial circuit. This gives [12 + (miles/150)] milliseconds for terrestrial one-way delay. Depending on latitude and longitude, satellite signals travel 45,000 to 50,000 miles in propagating from one satellite earth station to another; they also are buffered in the satellite and in the earth stations. As a rule of thumb, estimate 350 ms from user-to-user (one way), including terrestrial links between respective earth stations and users.
7. Other propagation delays due to buffering. Most devices used in creating complex networks for data communications store groups of received bits in buffers before retransmitting those bits on the next link. Multiplexors, concentrators, and other types of communications processors are examples of such devices. Depending on the number of bits so buffered and the circuit length, the buffer delays in some systems can exceed all other propagation delays in the transfer of data. Buffer delays are normally included in the equipment specifications provided by the manufacturer.
8. Computation time for response and other interblock time delays. While the time for data calculations normally wouldn't be considered in TRIB, processing time required to check for transmission errors and fabricate an appropriate response should be included since it is a communications function. Also, systems which allow multiple blocks to be sent between acknowledgements often require time gaps or filler bits between successive blocks.
9. Error rate. For blocked data, if the error rate is not large, it is usually sufficient to estimate TRIB by first assuming no errors, then making an adjustment based on the block error rate (BLER). For example, if 1% of the blocks contain an error, the block error rate is said to be 1%. This means that 1% of the blocks will need to be retransmitted; thus the TRIB will be lowered by approximately 1% due to errors.

Example of TRIB Calculation

Problem: How long will it take to transmit 10,000 records a distance of 500 miles every day on a dialed connection at 4800 bps using BSC protocol? There are 80 EBCDIC characters per record, and 2 records per block, thus, there are 5000 blocks containing 160 characters. What is the TRIB?

The first step in calculating TRIB is to determine the propagation delays

Solution: Since the connection is dialed at 4800 bps, the carrier must be switched, which means there will be a carrier turn-on delay in each direction for each block (one for data, one for acknowledgement). From the modem manufacturer's recommendations, this is established as 100 ms. Other delays include:

1. Circuit propagation delay which, for a terrestrial 500 mile circuit, would be about 15 ms.
2. Modem propagation delay which, according to the modem operation manual, is 4 ms.
3. Computation delay which, according to measurements on the actual terminals, is 10 ms for the receiving terminal and negligible for the transmitting terminal.

Assume that each data transmission block includes:

1 leading pad character
4 sync characters
1 STX character
160 data characters (2 records at 80 characters each)
1 EOB or ETX character
2 character lengths of BCC
1 trailing pad
170 characters

The next step is to determine the character count in the data block and acknowledgement block, and the transmission time required at the data rate.

Thus, each block has 170 characters including overhead, but not counting DLE characters that would be sent if transparency is required. If each character requires 8 bits, the transmission time actually needed to send a data block at 4800 bps will be:

$$\frac{170 \times 8}{4800} = 283 \text{ ms}$$

Assume that each acknowledgement block includes:

1 leading pad character
4 sync characters
2 character lengths of ACK0/ACK1
1 trailing pad character
8 characters (all overhead)

The transmit time actually needed to send an acknowledgement block will be:

$$\frac{8 \times 8}{4800} = 13 \text{ ms}$$

The calculation is completed by combining all delay times to obtain round-trip time per block, which results in the total time for all blocks. The total time divided into the actual data bits transferred gives the TRIB value.

Including delays, the *round-trip* time to send 1 block will be:

100 ms near-end carrier turn-on
283 ms data block transmission
4 ms transmit modem propagation
15 ms circuit propagation
4 ms receive modem propagation
10 ms sink terminal calculation
100 ms far-end carrier turn-on
13 ms acknowledgement block transmission
4 ms transmit modem propagation
15 ms circuit propagation
4 ms receive modem propagation
552 ms or 0.552 second round-trip time per block

The total time needed to send 5000 blocks, neglecting retransmissions due to errors, is 0.552 x 5000 = 2760 seconds, or 46 minutes. If we assume that 1% of the blocks will contain at least 1 bit in error, then the time will be lengthened by about 1%. The actual number of information bits transmitted is 160 characters per block x 8 bits per character x 5000 blocks = 6,400,000 bits. If errors are not counted, then:

$$\text{TRIB} = \frac{6{,}400{,}000 \text{ bits}}{2760} = 2319 \text{ bits/sec}$$

Maximizing TRIB

Several actions can be taken that would improve the TRIB. Installing a two-way simultaneous channel might provide the desired improvement, if the added expense of a private line is justified.

It will be noted that the TRIB in the example shown above is less than half of the modem bit rate of 4800 bps! Major factors contributing to this poor result were the modem carrier turn-on delay (modem turnaround) and the fact that acknowledgement of each block had to be received before the next block could be transmitted. (Note that the total of these times accounted for almost 40% of the round trip time.)

Persons who are unfamiliar with the type of calculation just illustrated often assume that the best way to improve throughput is to increase modem speed. To attempt such a remedy in the above example would probably result in lower (not higher) TRIB, since it is likely that the modem turnaround, modem propagation delay, and error rate all would be higher. Not only would the modems cost more, but also the dialing and redialing required to obtain a connection that would operate satisfactorily at a higher bit rate would be both time-consuming and frustrating to the operator — if it would work at all.

One possible approach to improvement of TRIB in the example shown might be to eliminate modem turnaround entirely by installing a two-way simultaneous channel. At 4800 bps, this would require a private line. Comparison of private-line tariffs with those for dialed connections for the distance and amount of data involved will reveal that dialing is considerably less expensive.

Other possible approaches to improving TRIB performance include changing to a different protocol or increasing block length.

Still another approach to TRIB improvement might be to change to a different protocol. SDLC is capable of allowing acknowledgement of up to 7 frames with only one response. However, to change from BSC to SDLC link protocol would require expensive software and probably hardware modifications in both terminals.

The best way to improve the TRIB in this case would be to increase block length. Assuming sufficient buffer capacity in both terminals, an increase of block length from 2 records to 10 records would result in a total round-trip time of 1.619 seconds per block, with only 1000 blocks required to be sent. The total time needed would be 27 minutes, a daily saving of 19 minutes of long distance calling. The new TRIB would be 3953 bps, an increase of 70%.

Caution: Increasing block length always increases the block error rate because there is a greater exposure to the probability of error in a long block. For a block length increase of 500%, the BLER will increase by roughly the same factor. In the example, if the block error rate had really been 1% before (very high for such a short block), it would be about 5% now. This would erode some of the time saving.

Optimum Block Length

Exact matching to the optimum block length often is not practical because of what the blocking program can accomplish or because long-term changes in error rates and delays cause the optimum block length to change.

Figure 10-1 illustrates how block length affects TRIB, all other factors being equal. For very short blocks, the TRIB is low (as illustrated in the calculations above) because overhead is a significant proportion of the total transmission. For very long blocks also, the TRIB is low because there is a significant probability of error within the block, and most blocks have to be retransmitted. Somewhere between the extremes, there is a block length, L_M, which yields maximum TRIB. Whether L_M is optimum depends on several factors besides TRIB:

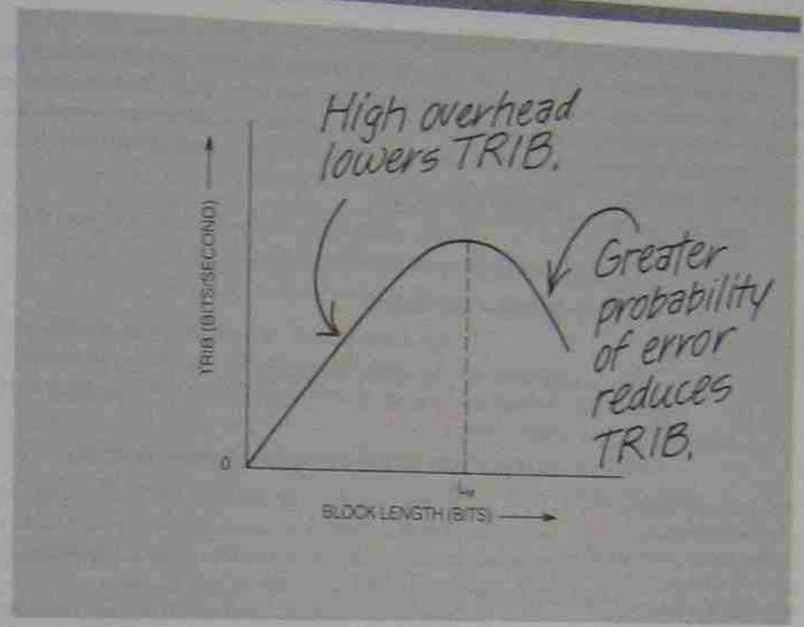
1. If the error rate is low, the value of L_M may be so large as to be impractical, particularly for an interactive data system in which fast response is more important than throughput.
2. Quite often, the memory capacity of the terminals is not sufficient to handle a block length of L_M.
3. The utility programs that actually perform the blocking of data usually won't allow a precise match with L_M.
4. Long-term changes in error rates and delays make precise matching impractical anyway.

Regardless of these considerations, the network designer should make reasonable estimates of TRIB, be aware of the major impairments to faster TRIB, and ensure that the combination of factors is "best" for each situation.

NETWORKING DEVICES

Several types of special-purpose devices exist for improving cost-effectiveness of data networks. Although space does not permit a thorough coverage or explanation, the basic purposes and advantages of the main types are given in the paragraphs that follow.

Figure 10-1.
Block Length (L_M) for Maximum TRIB



Multiplexors

Both the FDM and the TDM have advantages and disadvantages that must be considered to determine which to use in a particular application.

Multiplexors are devices that allow the combination of several data channels independently into one physical circuit. Data streams so combined can be separated and recovered at the opposite end of the system. There are two main types: frequency-division multiplexors (FDM) and time-division multiplexors (TDM). FDM has the advantage that it can be used on multipoint circuits, but the individual channels usually must each be for start-stop data. TDM can be used if all channels are point-to-point and generally allows faster bit rates and potentially more channels than FDM at less cost.

TDMs are of two types: classical (or dumb) TDMs and statistical (or smart) TDMs. Dumb TDMs should be used if the duty cycles of the data channels to be multiplexed are relatively high. A high duty cycle means that the channel is actually carrying bits most of the time. Since dumb TDMs send even idle (marking) bits when a channel becomes inactive, they waste a lot of potential capacity when the duty cycles are low.

Whenever usage statistics indicate a low duty cycle as the normal mode of utilization of most of the channels to be multiplexed, a statistical multiplexor system is usually a good choice. These devices send only data bits (no idle bits), so they can appear to have a higher total bit rate than the actual rate if there is a high percentage of idle bits coming from the terminals.

Modem Sharing Devices

Modem sharing devices allow several terminals to share the serial interface of one modem.

Modem sharing devices (MSDs) do just that: They allow several remote terminals at one site on a common multipoint circuit to share a single modem at that site. MSDs do not multiplex; they simply provide a means of sharing the serial interface of one modem among several terminals.

A useful networking device is the protocol converter, a microprocessor which converts data from one protocol to another.

Line Bridging Devices

Line bridging devices allow two or more analog circuits to be shared as one. Bridges are not multiplexors: They allow one modem to be used with circuits to more than one destination. Telephone companies and users can use bridges to create multipoint analog circuits from point-to-point segments.

Modem Eliminators

Modem eliminators are used to connect two DTEs together directly. They also are called null modems.

Protocol Converters

Protocol converters are microprocessors which convert data operating in one protocol, say start-stop ASCII, to another, say IBM BSC. A dial-up X.3 PAD for converting start-stop to X.25 is an example of a protocol converter.

INTERCONNECTION OF NETWORKING DEVICES

When actually making the physical connections between the various types of networking devices, be sure that all equipment manuals and interface standards are thoroughly understood.

The last section contained only a partial listing of devices intended to interconnect digital serial interfaces. Space does not allow a discussion of the numerous possible alternative requirements in the actual interconnection of leads between devices, terminals, modems, etc.

Probably the best advice that can be given is to study the equipment manuals and the interface standards themselves. Fortunately, all standard digital serial interfaces have been designed so that no actual damage can be caused by interconnecting two devices incorrectly. In the case of uncertainty, a programmable break-out box, (discussed in the next section) can be a handy tool for trying different combinations until the connection works.

In addition to the physical mating of pins and receptacles when connecting two serial interfaces together, particular attention should be given to:

1. Which is the driver (source) of signal and which is the receiver (sink) of signal. Each pin-receptacle pair must contain one of each. For example, pin 2 of an RS-232C connector involves a driver in a DTE and a receiver in a DCE.
2. Stand-alone TDMs are generally schizophrenic; that is, the terminal ports each behave as DCEs communicating with DTEs, and the trunk port behaves as a DTE communicating with a DCE. TDMs generally are manufactured with female receptacles on all ports, both terminal and trunk, requiring male-to-male cables between the trunk port and trunk DCE. When extending such a TDM port by use of a remote channel, special care must be given to control, data, and timing leads. The TDM manual should be read to obtain specific information.

3. Whenever synchronous devices are interconnected, special attention should be given to clocking. Remember that the DCE is normally the source (driver) for both the transmit clock and receiver clock. However, when two DCEs are interconnected, one DCE must accept external clocking, and it must be programmed (optioned) for external clocking, usually by an internal switch setting. The presence of clock pulses on the external clock pin is not sufficient to make a modem accept external clocking.
4. Digital transmission systems normally will not accept external clocking, since their timing comes from a central network clock.

FAULT ISOLATION

Fault detection usually is pretty easy; fault isolation is more complicated. The first question is, "Is the problem in a terminal or in the communications system?" Troubleshooting data terminals is beyond the scope of this book. If the terminals pass self-test procedures, but still will not communicate, chances are good that the problem is in the communications system.

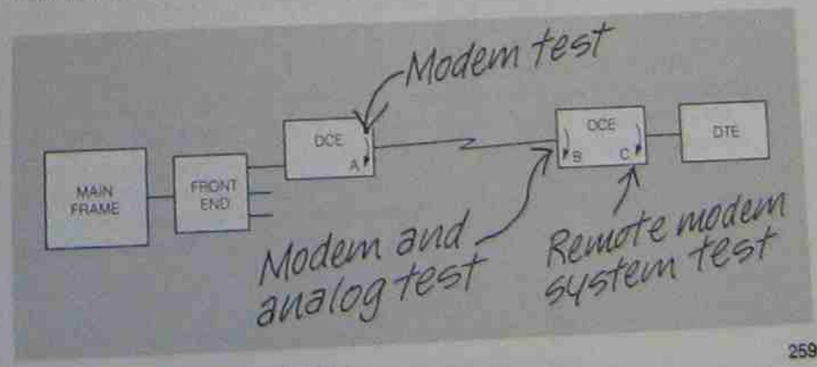
Loopback Tests

In the loopback test, which is a basic data communication fault isolation technique, faults are successively isolated by the process of elimination.

One of the basic techniques for isolation of faults in data communications is the loopback test. In a loopback, the output at the far end of a system or subsystem is connected to the input of the return path. Then the output of the return path is examined in relation to the input to the outgoing path.

Figure 10-2 illustrates how successive loopbacks can be used to isolate a fault on a simple point-to-point data link which uses an analog circuit. The local modem is first self-tested using an analog loopback at its far terminals (point A). If it passes the test, the loop is removed from the local modem. Then the local modem and the analog circuit are tested together using a remote analog loopback at the near terminals (point B) of the remote modem. If that test also appears satisfactory, then the remote analog loop is removed and the entire data link is tested end-to-end round-trip using a digital loopback at the far terminals (point C) of the remote modem. (Loopbacks at points B and C are useful only if the channel is full-duplex.) Using a process of elimination, the location of the fault can be narrowed to either the local modem, the analog circuit, or the remote modem.

Figure 10-2. Loopback Test Points



Loopbacks cannot determine which side of a path, outgoing or return, is faulty, but only the person who has to repair it really needs that information. End-to-end tests can distinguish the direction of the fault, but that often requires at least modest training of a person at the remote site.

Not only should one know what parameters of the faulty device can be tested, but the capabilities, limitations and special requirements of the test system equipment also should be known.

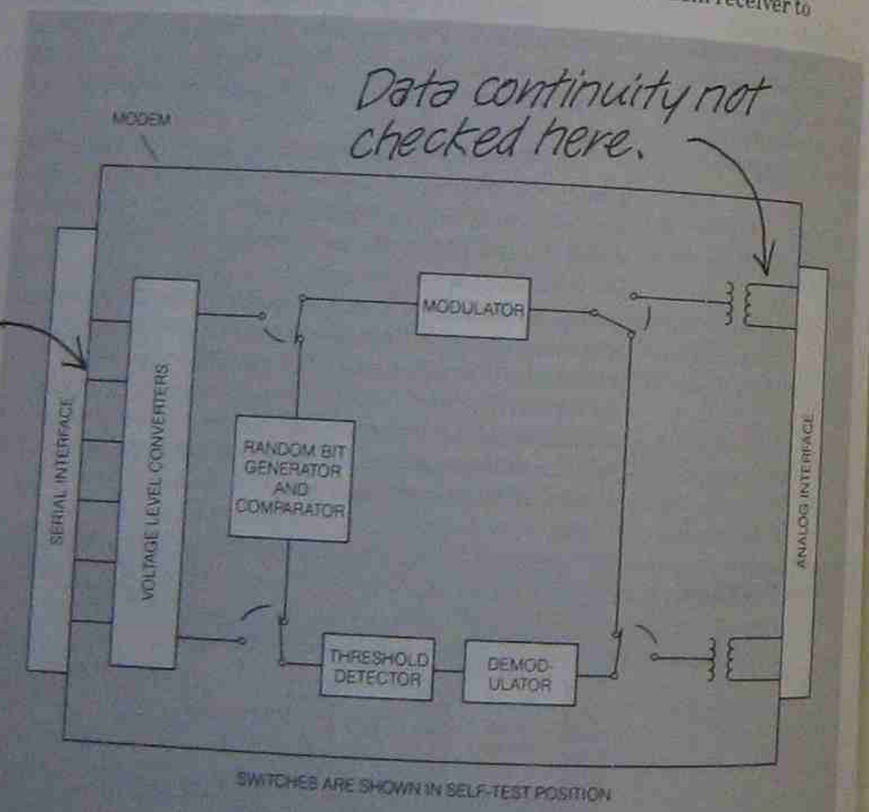
Caution Number One

Know what is being tested when performing a self-test on a device. As with terminals, no self-test ever tests *all* of the circuitry. For example, *Figure 10-3* shows that a self-test of a modem doesn't check data continuity through the line transformers and through the RS-232C interface circuitry. The latter problem can be very treacherous. Whenever practical, use external tests to be sure.

Caution Number Two

Most analog data circuits are designed to operate with a substantial drop in signal level from input to output; for example, 16 dB loss each way in a private-line voice-grade circuit to be used for data. When an analog loopback is performed at the remote end without inserting gain at that point, the input signal for the return path is 16 dB lower than it should be. Such a low-level signal can cause both the analog return path and the local modem receiver to

Figure 10-3. Some Modem Circuitry Is Not Tested in Self-Test



operate improperly, perhaps giving a false indication. Many systems that provide for remote analog loopbacks automatically insert gain in the loopback path to avoid this problem. If the modems used in the test depicted in *Figure 10-2* are not so equipped, then a different procedure should be used. For example, perform a self-test on each modem, then with an assistant at the far end, perform an end-to-end test.

Data Communications Test Equipment

A wide variety of test equipment is available to assist in the detection and isolation of faults in data communication systems.

Amplifier-Monitor

The self-contained battery-operated audio amplifier-speaker-monitor, sometimes called a "watergate box," is readily available at retail electronic supply stores for around $20 or less. It is useful not only for detecting modem carrier and changes in modem output signals, but it's also the easiest way to detect a noisy circuit or to tell whether a modem is putting out tone-dial or pulse-dial signals. It can also be used to monitor call progress and answerback tones on dialed data links.

Breakout Box

The basic specialized data testing device is the breakout box, so-named because it provides access to individual leads in the serial interface between a DTE (such as terminal) and a DCE (such as a modem). Prices range from about $60 to about $250 for breaking out the 25 leads of an RS-232C interface. Price depends on ruggedness, on whether the unit is programmable (leads can be crisscrossed) and whether data activity can be observed directly without a voltmeter. The higher priced units feature built-in battery power, full tri-state LED monitoring of leads, programmability, ruggedness, and DIP switches to open the circuits of individual leads. They may also provide positive and negatve test voltages.

Bit Error Rate Tester

The next step up in acquisition of specialized data test equipment would be a bit error rate tester (BERT). RS-232C BERTs range in price from under $500 to around $3000, depending on whether block error (BLERT) and error-second measurements are to be included, whether start and stop bit insertion/deletion is required (if asynchronous multiplexors are in the circuit), and various automatic indicator features. Some models also feature programmable message generators for checking out teleprinters. High-speed BERT/BLERTs in the multimegabit/second range can cost much more.

Decibel Meter

If analog voice-grade circuits are used extensively, a decibel (dB) meter is a wise investment for checking transmit and receive levels. A basic dB meter costs around $200. Persons considering purchase of a dB meter should consider a combination unit. Excellent portable meters with digital display are available. Such a meter which can read dBm, relative dB, frequency, relative frequency, voltage, current, and resistance costs about $400 to $500. Such an

expenditure is not recommended unless testing on analog circuits is expected to be a common occurrence.

An oscilloscope, used with a breakout box for access, is an important error isolation and testing tool.

Oscilloscope

A multichannel oscilloscope (or scope for short) with x-y inputs can be used in conjunction with a breakout box for observing serial interface signals and/or in conjunction with a modem for detecting certain impairments on analog circuits. Examples of the use of a dual-trace scope in the RS-232C interface would be to measure modem turnaround time, to check for crosstalk interference between timing leads, and to observe degradation of signals in long serial interface cables.

With synchronous modems having signal-space (constellation) pattern generators, an x-y oscilloscope becomes a powerful instrument for detecting phase hits, phase jitter, amplitude hits, harmonic distortion, dropouts, and impulse noise in analog circuits. With other modems, a scope can be used to observe eye patterns for harmonic distortion, phase jitter, and certain types of hits.

Scopes which are suitable for these purposes generally cost in the range of $800 to $3000, depending on compactness, maximum data rate, and number of channels.

A voice-frequency test set permits measurement of frequency response, noise, and distortion of voice grade analog circuits.

Voice Frequency Test Set

For the really serious user of data over voice grade analog circuits, a voice frequency test set is a device which can quantify the major parameters of a voice-grade analog circuit. Standard capabilities include frequency response (amplitude received versus frequency for constant amplitude transmitted), C-message noise in dBrnc, C-message notched noise (S/N) in dB, and audio monitoring. Costs can range from $1500 upward, depending on the degree of automation in the tests and additional test functions provided. While the less expensive units do not provide for measurement of envelope delay distortion (one of the two parameters affected by C-conditioning), large amounts of delay distortion often are accompanied by large amounts of attenuation distortion which is much easier to measure. Voice frequency test sets normally should be used in pairs to provide end-to-end testing.

A data link content monitor monitors and displays signals crossing the serial interface. Some allow specified sequences to be stored for later analysis.

Data Link Content Monitor

Many large users of data communications have painfully discovered that the general purpose computer is not an efficient device for debugging data link and network protocols. Nor is it the best device for bit error rate testing of wideband data channels, T1 multiplexors and the like. The computer time, programmer time, and other expensive resources saved in a single project often can pay for the purchase of test equipment to cover these areas.

One equipment category is the basic RS-232C data link content monitor at approximately $2000 to $6000. This device simply monitors and

displays signals passing across the serial interface. The display is normally in terms of characters, if appropriate, or it can be in binary or hexadecimal in the case of uncoded data. Pricing varies with the number of codes, the type of display, the ability to trap (or freeze) and display specific sequences, the amount of memory, scrolling of trapped memory contents, speed of the device with and without recording in memory, and whether the conditions of various non-data leads in the serial interface are recorded with the data for later analysis.

Since a simulator is actually a special-purpose computer, it can be programmed to simulate a computer port, remote terminal, or network gateway for system development.

Simulator

Monitor devices are excellent for diagnosing protocol problems in data links, but they are not efficient for system development. This function is best handled by a simulator. Such a device can be user-programmed to simulate a computer port, remote terminal, or network gateway in order to "exercise" the hardware/software system being developed. Simulators are actually special-purpose computers utilizing easy-to-learn high-level programming languages. They also can be programmed to operate effectively as content monitors, BERTs, and BLERTs. Prices range from $10,000 to $25,000, depending on protocols supported, speed, serial interfaces supported, amount of memory, type of display, ease of programming and portability. The newer units provide simulation up through the packet level (OSI Level 3) of X.25/X.75 protocols.

In a large complex data network, it may be cost effective to add a centralized network control system to assist in fault detection and isolation.

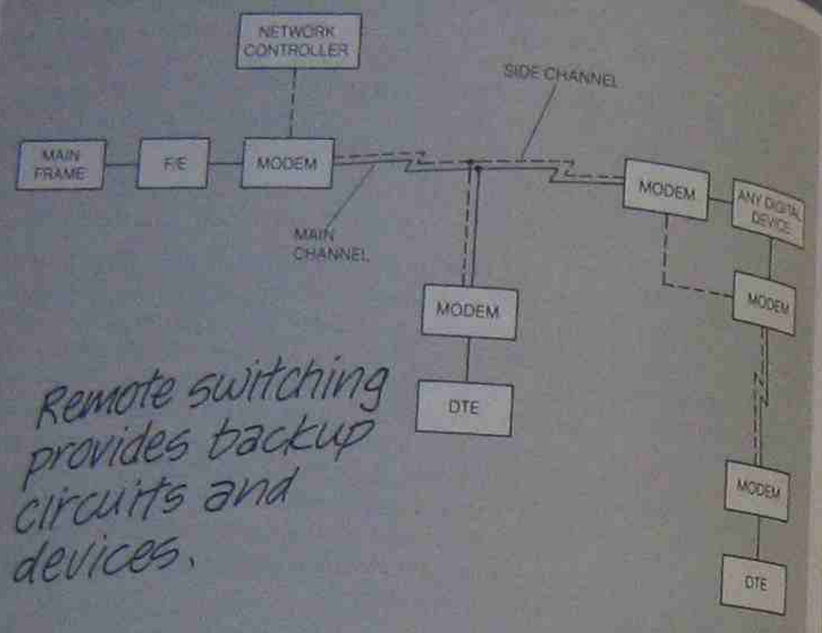
Network Control System

Many companies which operate large complex data networks have found the timely isolation of faults to be so important as to cost-justify the installation of a centralized network control system as depicted in *Figure 10-4*. Such systems commonly use the low end of the frequency spectrum of the voice-grade circuit to carry a special interrogation and reporting "side channel." Such systems add significantly to the cost of each modem in the system.

A modem which is addressed by the central unit can be commanded to perform certain tests and report the results back for display or recording. Such tests as continuity, level of received signal, and condition of an RS-232C lead can be accurately conducted without interrupting or modifying data flow. Other tests, such as self-tests and loopbacks, do cause interruption of data flow. Many of the more sophisticated systems can be user programmed to perform interruptive tests automatically during off-hours when actual interruption of data is unlikely. Test results can be automatically logged and compared with former test results to detect changes in circuit parameters.

Network control systems usually have a scheme for remote switching of circuits and devices to provide backup in the event of failures. The combination of remote testing and remote backup of data communications systems constitutes a powerful fault isolation and correction technique.

Figure 10-4.
Network Control System



The entry of digital electronics into data communications is having a tremendous influence on network control systems.

However, the cost effectiveness of a network control system based on testing and dedicated side channels to carry test information in the analog link must be weighed in view of the inroads being made by digital transmission systems. Digital segments don't support analog side channels. As digital transmission systems replace analog systems over the next few years, this type of control scheme will become less practical. A new generation of network control systems based on digital interrogation and reporting is now beginning to replace the analog systems.

RESTORATION OF SERVICE

Providing reliable service restoration promptly and economically is a challenging task.

Given enough planning and money, prompt and reliable restoration of data communications service is easy, but to make it economical is a little harder.

The planner for data communications backup should consider the following questions:

1. What is the frequency and length of downtime to be expected?
2. Can steps be taken to decrease the frequency and length of downtime without incurring additional costs; that is, are you getting what you already are paying for?
3. What will be the extra costs incurred due to complete outage of the service for the frequency and duration indicated?
4. Can these extra costs be reduced by a reduced-capability backup; for example, backing up a 4800 bps private line using a 1200 bps dialed connection?

5. What are the actual costs of all relevant backup schemes? Include the cost of personnel time as well as the cost of backup. The cost of backup must include the operating costs, such as dialed connection charges, as well as initial investment and installation costs.

Getting the Most for the Money

Existing systems performance often can be improved by closely examining the existing system and established procedures.

Before designing and implementing a backup technique, a review of current procedures might improve the frequency and duration of downtime.

1. Is there ongoing positive communication not only with the people in the user's own organization, but also with suppliers of equipment and services?
2. Do users have a thorough knowledge of the capabilities and limitations of all data communications system facilities, software and hardware at their disposal, particularly the capabilities and limitations of self-contained test features?
3. Are there well-designed restoration procedures, agreed to in advance by all concerned and strictly followed?
4. Are new systems benchmark tested as they are placed in service, and then tested on a regular schedule to detect changes in circuit parameters?
5. Are exceptions (failures) documented, and are repeat offenders discovered?
6. Is there an ongoing in-house training program to ensure that each individual in the organization can properly handle all appropriate parts of the above five points for those portions of the system within his or her purview?

Rapport With Suppliers

Close personal rapport between supplier and customers at all levels is of utmost importance to the smooth, efficient restoration of service of a communications system.

Central to any cost-effective service-restoration program is a firmly established positive rapport with all major suppliers of equipment and services – especially with key individuals such as telephone company (telco) personnel. Many data-link failures can be traced directly or indirectly to a breakdown in people-to-people communications between vendors and users. For example, in the case of an organization like a telephone company with many levels of personnel that deal with the customer, the rapport between supplier and customer must exist at all levels: Data technicians should be on a first-name basis with telco testboard personnel, technician supervisors with testboard supervisors, data communications operations manager with the telco network operations manager, and so on. When such rapport exists, several symptoms will be apparent:

1. Employees on both sides will be more helpful and understanding, especially in high-pressure situations.

2. Information on new testing and quality assurance techniques will flow easier. The user will gain a better understanding of the built-in test capabilities in the vendor's services and equipment; in fact, technicians on both sides will be eager to share their excitement over some new piece of test gear or a new discovery.
3. Better confidence in each other's abilities will emerge and will give rise to more of a supplier-user team approach.
4. The vendor employee will often render assistance above and beyond what is required and expected.

Data communications managers who succeed in cultivating this kind of rapport with their major vendors consider it among their greatest assets. In particular, they are careful to establish and follow good procedures for the testing of systems and proper reporting of problems to the vendor. Inherent in this process is the careful development of an escalation and reporting system.

Escalation and Reporting

Clearly established procedures must be developed to ensure that the proper level of management is involved in the correction of problems in a timely manner.

Through meetings with all concerned, the data communications manager should develop a very specific escalation schedule. This schedule should show exactly who becomes involved at increasing levels of authority, what calls are placed, and at what levels when a particular type of problem fails to show progress toward resolution after a specific time lapse. All persons involved in this schedule should be involved in its development, should "sign-off" on it, should have a personal copy of the finished product, and should have no cause for negative feelings whenever it is invoked.

Before asking for help from higher levels or from the vendor, it is absolutely vital that the user exhaust all means at his disposal to make use of test facilities available to him – especially self-tests. Remember the story about the little boy who cried "wolf"? Nothing can damage credibility and rapport faster than to blame the vendor when the problem is another vendor or in-house.

The personnel involved in system and data subsystem testing should be thoroughly familiar with the capabilities and limitations of the tests and the test equipment.

Personnel responsible for data subsystem and system testing must be thoroughly familiar with self-tests and their limitations, and with the meaning of any other tests for which they have equipment. Troubles should be reported only after such testing has been adequately documented, and escalation should be invoked only after the lack of progress in problem resolution has been documented.

Such documentation of problems and their resolution should be permanently recorded and reviewed with the vendor at regular intervals. This review aids in spotting areas of recurring difficulties and possibly degrading facilities and components, so these can be serviced before a "hard" failure occurs. This is why the benchmarking of newly acquired facilities is important.

Finally, appreciation should be expressed when appropriate. Nothing will make an employee of a vendor work harder for a user than an appreciation letter to his boss for good work done in the user's behalf.

Thought must be given to backup systems to keep priority lines in at least partial operation while repairs are being made. In high-priority situations, a duplicate hot-standby private line is often used.

Backup Techniques

Even the best procedures will not totally eliminate failures of circuits and equipment. When priority of data requires backup during the repair process, decisions must be made in the areas of providing spare parts for on-premise equipment (such as multiplexor boards and power supplies), providing complete standby devices (such as modems), providing an alternate physical communication medium (such as dialed backup), and determining who will do what when a failure occurs.

In high-priority situations, it isn't uncommon to find duplicate hot-standby private lines to backup the regularly used private lines. In such cases, it's important that the backup be provided via diversely routed circuits, even to the extent of using different entry points into the user's buildings.

The majority of backup plans provide for dialed backup, often with reduced capability. The following factors should be considered:

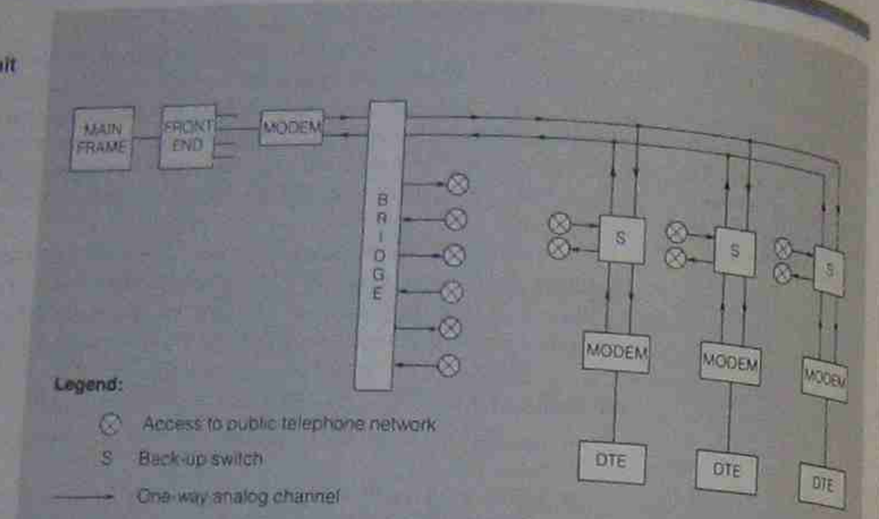
1. Backup on the public telephone network is generally limited to 9600 bps.
2. Depending on several factors, including distance from the modem to the central office, satisfactory dialed connections at 9600 bps may not be attainable, requiring fall-back to 4800 or 7200 bps.
3. If full-duplex backup is required at bit rates greater than 1200 bps, two calls may be required. This doubles the cost of providing backup. Some users provide only half-duplex backup for a full-duplex circuit; however, this type of fall-back may require changing the link protocol option selections whenever backup is invoked.
4. If the circuit to be replaced is multipoint, full backup will require simultaneous calls to each remote site that is affected; that is, six calls to fully backup a 3-drop full-duplex circuit. Some users provide backup of a multipoint line to only one remote site at a time, which reduces the number of dialable lines required in the data center.

Suppliers of network control systems generally have system-integrated provision for remotely controlled backup and switching of circuits, modems, multiplexors, or anything else money can buy. For the less sophisticated user, stand-alone devices are available from a number of suppliers.

Figure 10-5 illustrates a typical stand-alone arrangment to provide remotely controlled full-duplex dialed backup of a multipoint circuit. The central site bridge provides for simultaneous transmission and reception on the private line and each pair of central office business lines. Because of the high cost for usage of a multiplicity of dialed connections, most users prefer manually-dialed backup to ensure control over long-distance charges.

ST. BROOKS TECHNICAL COLLEGE LIBRARY

Figure 10-5. Dialed Backup of a Multipoint Analog Circuit



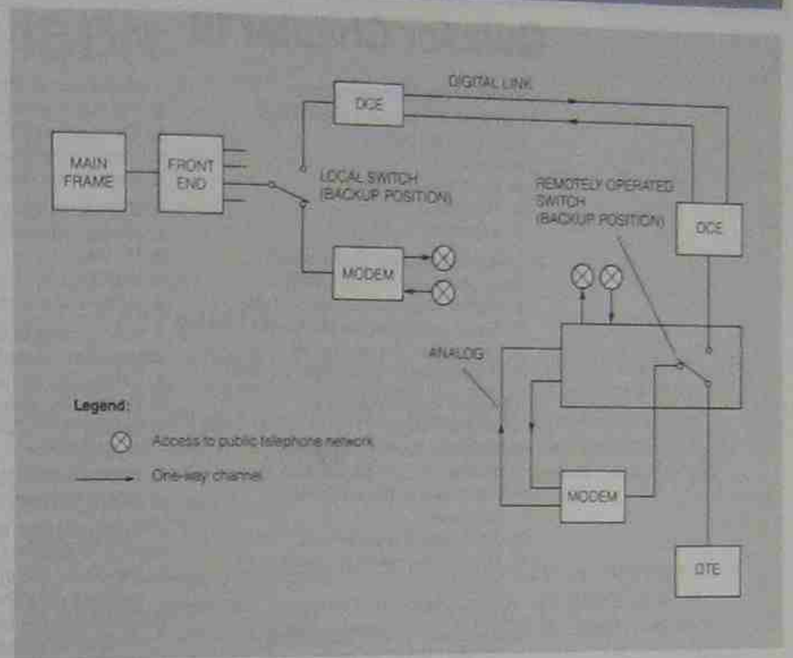
There are several methods that may be used to provide dialed backup service. In high-priority situations, alternate paths should be provided for both the private line and the remote DCE.

There are several schemes for the operation of the "S" switches in *Figure 10-5*. One typical scheme for operation of one "S" switch is as follows:

1. The central site technician places a call on one business line to the appropriate line connected to an "S" switch. The switch "answers" the call and places the call on "hold," starting an abort timer.
2. Then the operator places a second call to the other line. If the call comes in before the abort timer runs out, the switch will answer and connect the modem to both dialed lines; if not, it will abort the first call.
3. The "S" switch will revert back to the private line if either no modem carrier is detected within a specified interval or the carrier is interrupted. More sophisticated types of "S" switches require reception of special coded commands both for switching away from the private line and for returning to it.

Some high-priority systems require the ability to substitute another path for both the private line and the remote DCE. *Figure 10-6* shows such a configuration. The operation is similar to that in *Figure 10-5*, except for the position and arrangement of the switch within the circuit. As indicated in the figure, this arrangement is also useful for backing up a digital link.

Figure 10-6. Dialed Backup of a Digital Link



WHAT HAVE WE LEARNED?

1. The underlying objective of data networking is to strike a balance between the accurate timely delivery of user data and total cost.
2. Any network, no matter how simple or complex, may be incorporated into a larger network. A larger network cannot be optimized until its component parts are individually optimized.
3. Information throughput rate involves many factors besides the modem bit rate.
4. Classical time division multiplexors provide a powerful tool for combining data transmission link requirements into a smaller number of higher speed links. Statistical TDMs may be even more powerful if the duty cycles of individual links to be combined are low.
5. Frequency division multiplexors can be useful if the links involve low-speed multipoint data transmission.
6. Other useful networking devices include sharing devices, modem eliminators, and protocol converters.
7. Loopbacks are a common and simple technique for testing data communications components and networks.
8. Test equipment for data communications falls into two generic categories: digital testing and analog testing.
9. Good network design includes appropriate planning for contingencies and procedures for restoring service.
10. Dialed backup is a common technique for restoring service on analog and digital transmission systems.

ST. GEORGE COLLEGE OF T.A.F.E.

Quiz for Chapter 10

1. Which of the following is not considered a primary objective of data networking?
   a. Security.
   b. Packet switching.
   c. Timeliness.
   d. Accuracy.
2. On a point-to-point data channel using an analog circuit, TRIB is:
   a. Higher than the modem bit rate.
   b. Equal to the modem bit rate.
   c. Lower than the modem bit rate.
   d. Not related to the modem bit rate.
3. The block length that gives maximum TRIB:
   a. Increases as the error rate increases.
   b. Decreases as the error rate increases.
   c. Isn't related to error rate.
   d. Is called the optimum block length.
4. Smart TDMs:
   a. Are always better than dumb TDMs.
   b. Are always better than FDM.
   c. Are the same as modem sharing devices.
   d. May be better than dumb TDMs if the data terminals have a low duty cycle.
5. FDM devices can operate:
   a. On multipoint analog data channels.
   b. More efficiently than dumb TDMs.
   c. More efficiently than smart TDMs.
   d. Like modem sharing devices.
6. Protocol converters are:
   a. The same as multiplexors.
   b. The same as TDMs.
   c. Usually operated in pairs.
   d. Usually not operated in pairs.
7. Null modems are a type of:
   a. Modem eliminator.
   b. Modem.
   c. Multiplexor.
   d. Protocol converter.
8. A basic technique for isolating data communication faults is:
   a. Calling the telephone company.
   b. Performing a loopback test.
   c. Calling the DTE service department.
   d. Simulating the system on a special-purpose computer.
9. The main purpose of a data link content monitor is to:
   a. Determine the type of switching used in a data line.
   b. Determine the type of transmission used in a data link.
   c. Detect problems in protocols.
   d. Measure bit error rates.
10. Full-capability dialed backup of a multipoint full-duplex 4800 bps private line having 4 remote drops requires:
    a. 4 calls.
    b. Either 4 or 8 calls.
    c. 8 calls.
    d. Can't tell.

ISDN

ABOUT THIS CHAPTER

ISDN, which is the acronym for Integrated Services Digital Network, augurs the potential for the development of a universal digital network that provides integrated voice and data on common telephone company facilities. In this chapter, we'll examine the idea behind ISDN, its architecture, and some of the benefits that can be expected from it in the not-too-distant future.

CONCEPT

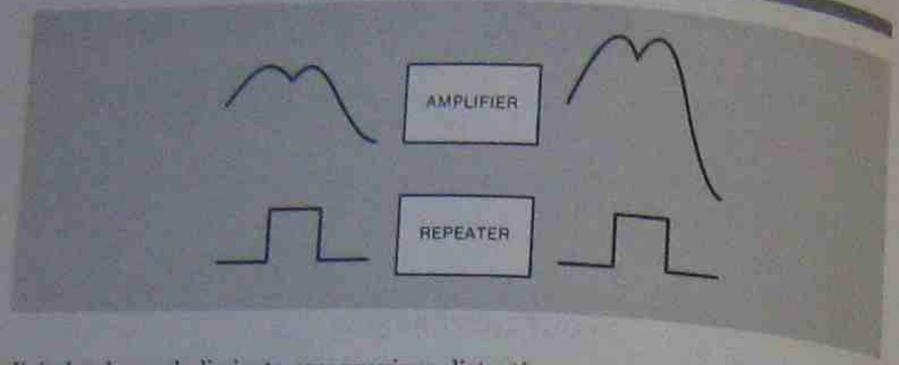
The need to transmit human speech resulted in the development of a telephone system that was originally designed for the transmission of analog data. Although the telephone system satisfied the basic need to transmit human speech, its design required the conversion of digital signals produced by computers and terminals into an analog format for the transmission of digital data. This conversion was awkward and expensive because modems were required at both ends of a telephone channel to do the digital to analog and analog to digital conversions.

A rapid decrease in the cost of semiconductors and the evolution in digital signal processing resulted in the application of digital technology to the telephone network. In the 1960s, telephone companies began to replace with digital switches the electro-mechanical switches located in their central offices. By the early 1970s, several communications carriers were offering end-to-end transmission services in which unipolar digital data from terminals and computers were converted first into a modified bipolar digital format, then through a series of digital repeaters in the network, the data was transmitted to its destination. At its destination, data was converted back to its original unipolar digital format. The unipolar to bipolar signal conversion enabled the telephone company to space repeaters farther apart, reducing the construction cost of the digital network. *Figure 11-1* illustrates the use of amplifiers and repeaters on analog and digital circuits.

The amplifier used on analog circuits amplifies the entire signal to include any signal impairments. The digital repeater, which is also known as a data regenerator, regenerates a new digital pulse, eliminating any distortion to the pulse that occurs as it travels on a digital circuit.

Because analog amplifiers increase the size of an analog signal to include any previous distortion, whereas digital repeaters regenerate a

Figure 11-1.
Amplifiers Versus
Repeaters



digital pulse and eliminate any previous distortion to the digital signal, the error rate on a digital network is significantly lower than that on an analog facility. In addition, the devices required to perform the unipolar to bipolar and bipolar to unipolar signal conversion are much less expensive than modems required to do signal conversion on analog facilities.

By the mid-1980s, most telephone companies had incorporated a large amount of digital technology into their plant facilities so that a significant portion of the lines connecting telephone company central offices transported speech in digital form, although speech continued to be carried in analog form from the subscriber to the central office. At the central office, speech is digitized for transmission over the backbone network of the telephone system. Similarly, at the central office closest to the destination of the telephone conversation, the digitized speech is reconverted into its original analog format, then transmitted to the subscriber's telephone.

The progression of telephone systems technology in the use of digital technology forms the basis for ISDN. Thus, ISDN can be viewed as an evolutionary progression in the conversion of the analog telephone system into an eventual all-digital network, enabling both voice and data to be transported end to end in a digital format.

Besides integrating voice and data, ISDN provides a level of communications capability above that obtainable with conventional analog technology. When voice and data are integrated, subscribers will be able to talk on the telephone and use a computer or terminal at the same time over a common subscriber line. For business, this capability should improve the productivity of office workers and reduce the cost of wiring buildings and offices because there should be no need to install separate wires to each desk for voice and data.

For individual subscribers, ISDN may result in the offering of a series of new functions accessible to their homes over existing telephone wire. Electronic meter reading, slow scan video, electronic mail, and other applications can be expected to be offered to individual subscribers and businesses.

ISDN ARCHITECTURE

Two methods of access to ISDN have been defined: Basic access and Primary access. ISDN Basic access deals with the connection and

operation of individual telephone instruments and terminals to the digital network. Primary access governs the method by which many Basic access subscribers can be connected to the network over a common line facility.

Basic Access

Basic access defines a multiple channel connection that is derived by the time division multiplexing of data on twisted pair wiring. This multiple channel connection links an end-user terminal device directly to a telephone company office or to a local Private Automated Branch Exchange (PABX). *Figure 11-2* illustrates the channel format of the ISDN Basic access method.

Under the Basic access format, up to three channels can be used to provide simultaneous data streams.

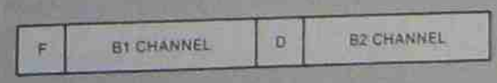
As shown in *Figure 11-2*, Basic access consists of two bearer (B) channels and a data (D) channel. The three channels are multiplexed by time onto a common twisted pair wiring media. Each bearer channel can carry one digitized voice conversation or data stream at a transmission rate of 64 Kbps. Digitized voice in ISDN is encoded by sampling voice 8,000 times per second and encoding each sample in 8 bits, which results in a 64 Kbps data rate. The inclusion of two bearer channels enables Basic access to provide the subscriber with the ability to simultaneously transmit data and conduct a voice conversation on one telephone line or to be in conversation with one person and receive a second telephone call. In the latter case, assuming subscribers have an appropriate telephone instrument, they could put one person on hold and answer the second call, all on one line! With analog technology, two physical lines are required.

The D channel is designed to control the B channels through the sharing of network signaling functions and for the transmission of packet switched data. It is anticipated that the capability to transmit packet switched data will enable the D channel to be used to provide many new applications, including monitoring home alarm systems and reading of utility meters on demand.

Primary Access

Primary access is a multiplexing arrangement whereby a grouping of Basic access users shares a common line facility. Primary access is designed to directly connect a PABX to the ISDN network. This access method eliminates the need to provide individual Basic access lines when a group of terminal devices shares a common PABX that could be connected

Figure 11-2.
ISDN Basic Access
Channel Format



B channels are 64 Kbps each.

D channel is 16 Kbps.

2B + D service is a 144 Kbps data stream.

When framing (F) and multiplexing overhead is considered, the actual data rate of a Basic access channel is 192 Kbps which is the interface transmission rate of the line.

directly to an ISDN network via a single high speed line. Because of the different methods used to multiplex digitized voice conversations between telephone company offices in North America and Europe, two Primary access standards have been developed.

In North America, Primary access consists of a grouping of 23 B channels and one D channel to provide a 1.544 Mbps composite data rate. This data rate is more commonly known as the standard T1 carrier data rate. The D channel in Primary access, and each B channel, operates at 64 Kbps. Multiplying 24 channels by 64 Kbps results in a data rate of 1.536 Mbps, which is precisely 8 Kbps fewer than the T1 carrier's 1.544 Mbps data rate.

The T1 carrier used in North America is not the same as that used in Europe due to different methods of signaling.

A North American T1 carrier consists of 24 digitized voice channels. Each voice channel is sampled 8,000 times per second, and 8 bits are used to encode the digitized value of each sample, Thus, a 64 Kbps data rate is required to transmit a digitized voice conversation. To permit synchronization of the T1 signal, a single bit, known as a framing bit, is added to the data stream. The framing bit represents 24 channels, as illustrated in *Figure 11-3*. One sample of 24 channels of digitized voice is represented therefore by 193 bits. Because the sampling occurs 8,000 times per second, the data rate of a T1 carrier is 193*8,000 or 1.544 Mbps.

In Europe, the T1 carrier consists of 30 digitized voice channels and 2 separate signaling channels. Because each channel operates at 64 Kbps, the resulting T1 carrier data rate is 2.048 Mbps.

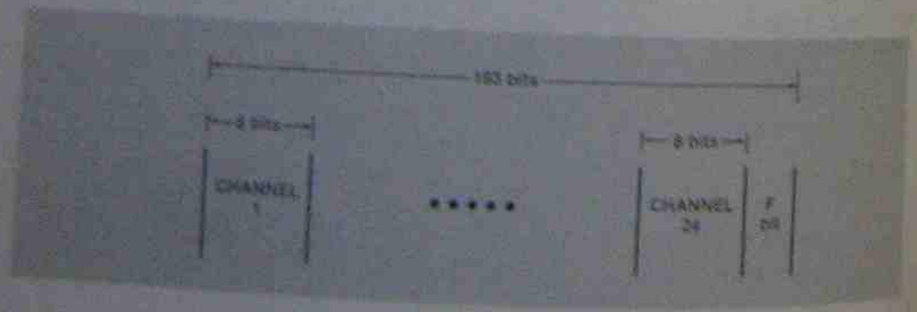
Network Characteristics

The major characteristics of an ISDN network follow and form the basis of business' and telephone companies' desire for the development of this new network.

ISDN integrates voice, data, and video services.
ISDN has a digital end-to-end connection that provides high transmission quality.
ISDN has improved and expanded services because of B and D channel data rates.
ISDN is more efficient and productive.
ISDN offers advances in device connectivity.

The digital nature of ISDN integrates voice, data, and video services, which alleviates the need for subscribers to obtain separate

Figure 11-3. The T1 Frame



facilities for each service. Because an ISDN network is designed to provide end-to-end digital transmission, pulses can be regenerated easily throughout the network, resulting in the generation of new pulses to replace distorted ones.

In comparison, analog transmission facilities use amplifiers to boost the strength of transmission signals, which also increases any impairments in the analog signal. Because regeneration is superior to amplification, digital transmission has a lower error rate and provides a higher transmission signal quality than an equivalent analog transmission facility.

Basic access provides three signal paths on a common line, and ISDN, too, offers to subscribers the potential of improvements and expansion to existing services. For existing services, current analog telephone line bandwidth limitations preclude data transmission rates more than 19.2 Kbps occurring on the switched telephone network. In comparison, under ISDN, each B channel can support a 64 Kbps transmission rate, whereas the D channel can operate at 16 Kbps. In fact, if both B channels and the D channel were in simultaneous operation, a data rate of 144 Kbps would be obtainable on a Basic access ISDN circuit, exceeding by a factor of seven the current analog circuit data rates.

Because each Basic access channel consists of three multiplexed channels, different operations can occur simultaneously without requiring the subscriber to acquire separate multiplexing equipment. Thus, a subscriber could receive a call from one person, transmit data to a computer, and have a utility company read his or her electric meter, all at one time. With the ability to conduct as many as three simultaneous operations on one line, the subscriber gains efficiency and productivity. Efficiency should increase because one line now can support several simultaneous operations, whereas the productivity of the subscriber can increase because of the ability to receive telephone calls and conduct a conversation while transmitting data.

Advances in device connectivity to the ISDN network can be expected to occur in two areas. At the physical interface, an eight-pin modular plug and jack will provide a command interface that will enable devices to be portable between jacks. This will eliminate special cabling, and usually, additional telephone company installation charges for the movement of devices between offices. Digital telephones and other instruments can be expected to have an intelligent reporting capability. When a telephone is moved from one office to another, simply plugging it into a jack will result in it reporting its extension and location to a PBX. This capability will permit employees who are moving within an organization to receive immediately telephone calls at their new location without waiting for the company to reprogram the PBX.

Telephone Equipment and Network Interface

A key element of ISDN is a small set of compatible multipurpose user-network interfaces developed to support a wide range of applications. These network interfaces are based on a series of reference points for different user terminal arrangements that define these interfaces. *Figure 11-4* illustrates the relationship between ISDN reference points and network interfaces.

The Terminal Adapter (TA) can be expected to play a key role in the use of ISDN as it permits equipment with a non-ISDN interface to be connected into an ISDN network.

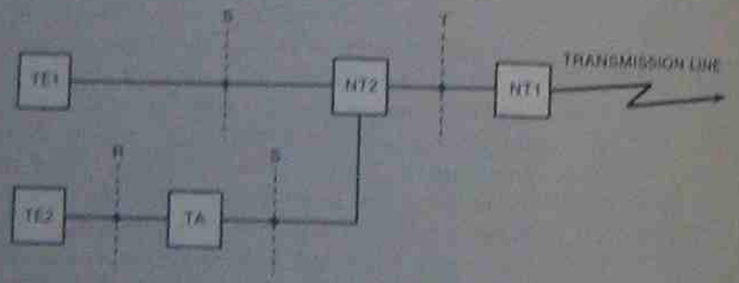
The ISDN reference configuration consists of functional groupings and reference points at which physical interfaces may exist. The functional groupings are sets of functions that may be required at an interface, whereas reference points are employed to divide the functional groups into distinct entities.

The TE (Terminal Equipment) functional grouping is comprised of TE1 and TE2 type equipment. Digital telephones, conventional data terminals, and integrated voice-data workstations are examples of TE equipment. TE1 type equipment complies with the ISDN user-network interface and permits such equipment to be connected directly to an ISDN "S" type interface that supports multiple B and D channels.

TE2 type equipment are devices with non-ISDN interfaces, such as RS-232 or the CCITT X- or V-Series interfaces. This type of equipment must be connected through a TA (Terminal Adapter) functional grouping, which is used to convert a non-ISDN interface (R) into an ISDN interface (S). The TA does both a physical interface conversion and protocol conversion, permitting a TE2 terminal to operate on ISDN.

The NT2 (Network Termination 2) functional group includes devices that do switching and data concentration functions equivalent to the first three layers of the OSI reference model. Typical NT2 equipment can include PABXs, LANs, terminal controllers, concentrators, and multiplexers.

Figure 11-4.
ISDN Reference Points and Network Interfaces



TE1 (Terminal Equipment 1) type devices comply with ISDN network interface.

TE2 (Terminal Equipment 2) type devices do not have an ISDN interface and must be connected through a TA (terminal adapter) functional grouping.

NT2 (Network Termination 2) includes switching and concentration equipment which performs functions equivalent to layers 1 through 3 of the OSI reference model.

NT1 (Network Termination 1) includes functions equivalent to layer 1 of the OSI reference model.

The NT1 (Network Termination 1) functional group is the ISDN digital interface point and is equivalent to layer 1 of the OSI Reference Model. Functions of NT1 include the physical and electrical termination of the loop, line monitoring, timing, and bit multiplexing. In Europe, where most communications carriers are government-owned monopolies, NT1 and NT2 functions may be combined into a common device, such as a PABX. In such situations, the equipment serves as an NT12 functional group. In comparison, in the United States, the communications carrier may provide only the NT1, whereas third-party equipment would connect to the communications carrier equipment at the T interface. Refer to the CCITT I recommendation series for detailed information about ISDN reference points and network interfaces.

By providing a set of standardized network interfaces, both pre- and post-ISDN terminals will be able to share the benefits of using ISDN. Pre-ISDN terminal devices will be connected to ISDN facilities by Terminal Adapters, which will convert a non-ISDN interface (R) into an ISDN interface (S). With several well publicized field trials of ISDN being conducted in the United States and abroad, it is only a matter of time until commercial ISDN offerings become available for selection by many analog telephone subscribers. Similar to the manner in which "plastics" was used as the all-encompassing word of the future in a popular 1960s movie, "ISDN" can be expected to be its replacement for the late 1980s.

WHAT HAVE WE LEARNED?

1. ISDN can be viewed as an evolutionary process because communications carriers have been converting their facilities to digital technology since the 1960s.
2. Access arrangements to the ISDN network include Basic and Primary access.
3. The multiple channels on one ISDN physical line are obtained through time division multiplexing.
4. Multiple channels on one ISDN physical line permit multiple voice or voice and data transmission to occur at the same time on a common circuit.
5. The Basic access D channel on which data is transmitted in packet form can be expected to be used for such new applications as monitoring home alarm systems and reading of utility meters as well as for the transmission of signaling information to control the operation of the B channels.

ST. GEORGE COLLEGE OF T.A.F.E.

Quiz for Chapter 11

1. The data rate of the ISDN Basic access B channel is
   a. 32 Kbps
   b. 64 Kbps
   c. 144 Kbps
   d. 192 Kbps

2. The data rate of the ISDN Basic access D Channel is
   a. 64 Kbps
   b. 16 Kbps
   c. 8 Kbps
   d. 144 Kbps

3. The equation that defines the composition of an ISDN Basic access line is
   a. 2B + D
   b. B + D
   c. B + 2D
   d. 2B + 2D

4. In North America, the equation that defines the composition of an ISDN Primary access line is
   a. 30B + D
   b. 2B + D
   c. B + D
   d. 23B + D

5. The network interface that will permit equipment without an ISDN interface to be connected into an ISDN interface is known as
   a. TE1
   b. TE2
   c. TA
   d. TE

6. The number of channels on which different operations can occur simultaneously on one ISDN Basic access line is
   a. 1
   b. 2
   c. 5
   d. 3

7. Digital telephones and integrated voice-data workstations are examples of what type of ISDN equipment?
   a. TE
   b. TA
   c. NT2
   d. TP

Glossary

Amplifier: A device used to increase the strength of an analog signal.

Analog Signal: A signal, such as voice or music, that varies in a continuous manner. Contrast with digital signal.

ASCII (American Standard Code for Information Interchange): A 7-bit code established by the American National Standards Institute to achieve compatibility between data services. Equivalent to the International ISO 7-bit code.

Attenuation: The difference between transmitted and received power due to transmission loss through equipment, lines, or other communication devices.

Bandwidth: The frequency range between the lowest and highest frequencies that are passed through a component, circuit, or system with acceptable attenuation.

Baseband: The frequency band occupied by a single or composite signal in its original or unmodulated form.

Basic Access: The method whereby individual telephone instruments and terminals will be connected to a digital network.

Baud: A unit of signaling speed equal to the number of signal events per second. Not necessarily the same as bits per second.

Binary Coded Decimal (BCD): A system of binary numbering where each decimal digit 0 through 9 is represented by four bits.

Binary Synchronous Communications (BSC or BiSync): A communication protocol developed by IBM which has become an industry standard. It uses a defined set of control characters and control character sequences for synchronized transmission of binary coded data between stations in a data communications system.

bis: Meaning second in Latin, this term is used as a suffix to denote a secondary version of a CCITT modem standard.

Bit: Contraction of binary digit. The smallest unit of information. A bit represents the choice between a one or a zero value (mark or space in communications terminology).

Bit Rate: The speed at which bits are transmitted, usually expressed in bits per second. Not necessarily the same as baud rate.

Block: A sequence of continuous data characters or bytes transmitted as a unit. A coding procedure is usually applied for synchronization or error control purposes.

bps (bits per second): A measure of the information transfer rate of a data channel.

Buffer: A storage device used to compensate for a difference in rate of data flow, or time of occurrence of events, when transmitting data from one device to another.

Byte: A binary element string operated upon as a unit and usually shorter than a computer word. Eight-bit bytes are common.

Carrier: A signal suitable for modulation by another signal containing information to be transmitted. The carrier is usually a sine wave for analog systems.

CCITT: (Consultative Committee for International Telephone & Telegraph) A standards making body whose recommendations are more closely followed in Europe than North America.

Channel, Voice Grade: A channel, generally with a frequency range of about 300 to 3400 Hz, suitable for transmission of speech or data in analog form. Data transmission rates of 9600 bps may be achieved by modulation techniques that produce a baud rate of 2400.

Character: A letter, figure, number, punctuation or other symbol contained in a message or used in a control function.

Character Set: The characters that can be coded and/or printed by a particular machine.

Code: A set of unambiguous rules specifying the way in which characters may be represented.

Common Carrier: A company that furnishes communication services to the general public and which is regulated by appropriate state or federal agencies.

Conditioning, Line: The addition of equipment to a leased voice-grade channel to improve analog characteristics to allow higher rates of data transmission.

Constellation Pattern: The pattern obtained from plotting the location of each bit combination modulated by a modem through phase and amplitude changes.

Contention: The facility provided by the dial network or a port selector which allows multiple terminals to compete on a first-come-first-served basis for a smaller number of computer ports.

Data Communications Equipment (also Data Circuit-terminating Equipment) (DCE): The equipment that provides the functions required to establish, maintain, and terminate a connection, and provides the signal conversion required for communication between data terminal equipment and the telephone line or data circuit.

Data Terminal Equipment (DTE): A computer or business machine that provides data in the form of digital signals at its output.

dBm: Decibel referenced to one milliwatt. Used in communications circuits as a measure of signal power. Zero dBm equals one milliwatt into a specified impedance, often 600 ohms.

Decibel (dB): A logarithmic measure of the ratio between two powers, P_1 and P_2. The equation is: $dB = 10 \log_e P_2/P_1$.

Delay Equalizer: A corrective network which is designed to make the phase delay or envelope delay of a circuit or system substantially constant over a desired frequency range.

Delay, Propagation: The time required for a signal to travel from one point to another in a component, circuit, or system.

Demodulation: The process of recovering data from a modulated carrier wave. The reverse of modulation.

Dial-Up Line: A communications circuit that is established by a switched circuit connection using the telephone dial network.

Dibit: A group of two bits. In four-phase modulation, such as differential phase shift keying (DPSK), each possible dibit is encoded as one of four unique carrier phase shifts. The four possible states for a dibit are 00, 01, 10, and 11.

Digital Repeater: A data regenerator that, on the detection of the rise of a digital pulse's leading edge, regenerates the pulse.

Digital Signal: A discrete or discontinuous signal; one whose various states are identified with discrete levels or values.

Distortion, Delay: Distortion resulting from non-uniform speed of transmission of the various frequency components of a signal through a transmission medium. Also called group delay.

Distortion, Harmonic: The result of nonlinearities in the communication channel that cause harmonics of the input frequencies to appear in the output.

Distortion, Linear (or Amplitude): An unwanted change in signal amplitude so that the output signal envelope is not proportional to the input signal envelope, but no frequency related distortion is involved.

Echo Suppressor: A device that allows transmission in only one direction at a time. They are inserted in telephone circuits to attenuate echos on long distance circuits. They are not desirable in data communications circuits because they increase the turnaround time.

Equalization: The process of reducing the effects of amplitude, frequency, and/or phase distortion of a circuit by inserting networks to compensate for the difference in attenuation and/or time delay at various frequencies in the transmission band.

Filter: A network designed to transmit electrical signals having frequencies within one or more frequency bands and to attenuate signals of other frequencies.

Frequency Division Multiplexer: A device which divides the available transmission frequency range into narrower bands, each of which is used for a separate channel.

Frequency Response: The change in attenuation with frequency relative to the attenuation at a reference frequency. Also called attenuation distortion.

Frequency Shift Keying (FSK): A form of frequency modulation commonly used in low-speed modems in which the two states of the signal are transmitted as two separate frequencies.

Full-Duplex: Refers to a communications system or equipment capable of simultaneous two-way communications.

Gateway: Hardware and software that permits devices located on a local area network to access the facilities of another network.

Half-Duplex: Refers to a communications system or equipment capable of communications in both directions, but in only one direction at a time.

Handshaking: Exchange of predetermined codes and signals between two data terminals to establish a connection.

Interface: A shared boundary defined by common physical interconnection characteristics, signal characteristics, and meanings of interchanged signals.

Intermodulation Noise: Spurious frequencies, such as sum and difference frequencies, which are the products of frequencies transmitted through a nonlinear circuit.

ISDN: (Integrated Services Digital Network) A future offering designed to provide a universal digital network that will permit the integration of voice and data on a common telephone company facility.

Jitter: A tendency toward lack of synchronization caused by mechanical or electrical changes.

Kilostream: A digital network that operates in the United Kingdom.

Line: 1) A circuit between a customer terminal and the central office. 2) The portion of a transmission system, including the transmission media and associated repeaters, between two terminal locations.

Link: A circuit or transmission path, including all equipment, between a sender and a receiver.

Local Area Network: A communications network that is restricted to a small geographical area, usually within a building or on a campus, and that has cabling normally installed and/or controlled by the organization that operates the network.

Loop, Local: The pair of wires between a customer terminal and the central office.

Loopback Test: A test of a communications link performed by connecting the equipment output of one direction to the equipment input of the other direction and testing the quality of the received signal.

Mark: One of the two possible states of a binary information element. The closed circuit and idle state in a teleprinter circuit. See Space.

Modem (MOdulator/DEModulator): A type of DCE that converts digital data to an analog signal for transmission on telephone circuits. A modem at the receiving end converts the analog signal to digital form.

Modulation: The process of varying some characteristic of the carrier wave in accordance with the instantaneous value or samples of the intelligence to be transmitted. Amplitude, frequency, and phase are the characteristics commonly varied.

Multiplex: To interleave or simultaneously transmit two or more messages on a single channel.

Noise: Random electrical signals, introduced by circuit components or natural disturbances, which tend to generate errors in transmission.

PABX: (Private Automatic Branch Exchange) A device that is installed on a customer's premises and that permits a large number of telephones to automatically access the switched telephone network by using a smaller number of lines connecting the PABX to the telephone company network.

PBX (Private Branch Exchange): Telephone switching equipment dedicated to one customer and connected to the public switched network.

Personal Computer: A computer with processing power based on a microprocessor and designed primarily for use by one person.

Polling: The individual selection of multiple terminals by a controller to allow transmission of traffic to/from all terminals on a multidrop line in an orderly manner.

Port: An interface on a computer configured as data terminal equipment and capable of having a modem attached for communication with a remote data terminal.

Primary Access: A multiplexing arrangement whereby many ISDN Basic access subscribers can be connected to a digital network over a common line facility.

Protocol: The rules for communication between like processes, giving a means to control the orderly communication of information between stations on a data link.

Redundancy: The portion of the total information contained in a message which can be eliminated without loss of essential information.

Repeater: A communications system component which amplifies or regenerates signals to compensate for losses in the system.

Serial Transmission: A method of information transfer in which the bits comprising a character are sent in sequence one at a time.

Slow Scan Video: The process whereby still video images are transmitted frequently to provide the appearance of video motion.

Space: One of the two possible states of a binary information element. The open circuit or no current state of a teleprinter line. See Mark.

Start Bit or Element: The first bit or element transmitted in the asynchronous transmission of a character to synchronize the receiver.

Stop Bit or Element: The last bit or element transmitted in the asynchronous transmission of a character to return the circuit to the at-rest or idle condition.

Symbol: The graphical representation of some idea which is used by people. Letters and numbers are symbols.

TA: (Terminal Adapter) A device that is used to permit equipment with a non-ISDN interface to be connected to that digital network.

ter: Meaning third in Latin, this term is used as a suffix to denote a third version of a CCITT modem standard.

TE1: (Terminal Equipment 1) Devices that comply with the ISDN network interface and can be connected directly to the digital network.

TE2: (Terminal Equipment 2) Devices that do not have an ISDN interface and must be connected by using a TA (terminal adapter) to the digital network.

Terminal Emulation: Software or firmware that enables a personal computer or terminal to duplicate the screen attributes of a terminal.

Time Division Multiplexer: (TDM) is a device that permits the transmission of two or more independent data channels on a single high speed circuit by interleaving the data from each channel on the circuit by time.

Transmission, Asynchronous: Transmission in which each information character is individually synchronized, usually by the use of start and stop elements.

Transmission, Synchronous: Transmission in which the sending and receiving instruments are operating continuously at substantially the same frequency and in which the desired phase relationship may be maintained by means of correction.

Trellis Coded Modulation: A modem modulation process in which one or more redundant bits is added to each group of bits used to generate a signal change. The extra bit or bits permits only certain sequences of signal points to be valid, resulting in a lowering of the error rate.

Turnaround Time: The actual time required to reverse the direction of transmission from sender to receiver or vice versa when using a half-duplex circuit. Time is required for line propagation effects, modem timing, and computer reaction.

UART (Universal Asynchronous Receiver/Transmitter): A device which performs asynchronous communication functions by converting parallel digital output from a DTE into serial bit transmission and vice versa.

Bibliography

A History of Engineering & Science in the Bell System: The Early Years (1875 - 1925), Bell Telephone Laboratories, Inc., 1975.

Bell Laboratories Record, Bell Telephone Laboratories, Inc., November 1980.

Bellamy, John, *Digital Telephony*, John Wiley & Sons, Inc., 1982.

CCITT Recommendations V.29, X.21, and X.25, CCITT, Geneva, C. H.

Doll, Dixon, *Data Communications; Facilities, Networks and Systems Design*, John Wiley & Sons, 1978.

EIA Standard RS-232-C (and others), Electronic Industries Association.

Fike, John L. and Friend, George E., *Understanding Telephone Electronics*, Texas Instruments Incorporated, 1983.

Freeman, R.L., *Telecommunication System Engineering*, John Wiley & Sons, Inc., 1980.

Freeman, R.L., *Telecommunication Transmission Handbook*, 2nd Ed., John Wiley & Sons, Inc., 1981.

General Information — Binary Synchronous Communications, IBM Publication Nr. GA27-3004, IBM Systems Development Division, Publications Center.

General Information — IBM Synchronous Data Link Control, IBM Publication Nr. GA27-3093-2, IBM Systems Development Division, Publications Center.

IEEE Transactions on Communications, IEEE Communications Society, August 1980, Vol. COM-28.

Kreager, Paul, *Practical Aspects of Data Communications*, McGraw-Hill Book Co., 1983.

Martin, James, *Telecommunications and the Computer*, Prentice-Hall, Inc., 1969.

McNamara, John E., *Technical Aspects of Data Communications*, 2nd Ed., Digital Press, 1982.

Members of the Technical Staff, *Transmission Systems for Communications*, Revised 4th Ed., Bell Telephone Laboratories, Inc., 1971.

Members of the Technical Staff and the Technical Publications Department, *Engineering and Operations in the Bell System*, Bell Telephone Laboratories, Inc., 1977.

Notes on the Network, American Telephone & Telegraph Co., 1980.

Owen, Frank E., *PCM and Digital Transmission Systems*, McGraw-Hill, Inc., 1982.

Parameters of Telephone Network Design, IBM Publication Nr. ZZ-11-3201-0, IBM World Trade Corp., Jan. 1977.

Sherman, Kenneth, *Data Communications: A User's Guide*, Reston Publishing Co., 1981.

Stauffer, J.R., *IEEE Journal on Selected Areas in Communications*, Vol. SAC-1, No. 3, April 1983.

Tanenbaum, Andrew S., *Computer Networks*, Prentice-Hall, Inc.

Index

Accuracy of data, 250
American National Standards Institute (ANSI), 10
American Standard Code for Information Interchange (ASCII), 10–12
  terminals, 31–34
Amplifier-monitor, 261
Amplifiers, development of electronic, 61
Amplitude modulation, 78–79
Analog
  carrier systems, 72
  modulation, 78–80
  signals, 71
Apple, 19
ARCNet (Attached Resource Computer Network), 209–12
ASCII. *See* American Standard Code for Information Interchange
Asynchronous modems. *See* Modems, asynchronous
Asynchronous transmission
  electromechanical method, 45–47
  electronic receiving circuit, 47–49
  start-stop, 25–26
  stop bits and timing errors, 49–50
AT&T, 241
  DR-18, 66
  FT3, 64
  FT3C system, 159
  3A-RDS, 66
AT&T Teletype Corporation, 25
Attenuation
  fiber optics and, 154–55
  wire pairs and, 59

Bandwidths
  analog signals and, 71
  digital signals and, 71–72
  effects of, on transmission channels, 70
  fiber optics and limitations on, 153
Batch processing systems, 15
Batch terminals, 34
Baudot code, 8, 10, 171–73
Baud rate, 82, 116
Bell, Alexander Graham, 61, 145
Bell System
  113 type device, 82
  212A and V.22 modems, 82–83
Binary Synchronous Communications (BSC), 179, 186
Binary system, 5–6
Bipolar signaling, 26
BiSync, 180, 181–85
Bit error rate tester (BERT), 261
Bits, 6–7, 56
Block check character (BCC), 183
Bounded medium, 57–58
Breakout box, 261
Broadband networks, 213–14
Bytes, 7

Call request/incoming call packets, 239, 241
Carrier Sense, Multiple Access with Collision Detection (CSMA/CD), 204, 206
Carrier systems, 72–73
CCITT
  International Alphabet No. 2, 10
  interrelationships between X standards, 243
  list of X series standards, 242–43
  V.22 bis, 83
  V.24, 92
  V.26, 128–29
  V.27, 129–30
  V.29, 182, 133, 135
  V.32, 135
  V.33, 136–37
  V.35, 231
  X.21, 108–9, 230–31
  X.200, 243

CCITT
  X.25, 201
  facilities of, 233
  LAPB procedures of, 233–36
  link level and, 231
  network level of, 232
  packet-level procedures of, 236–39, 241
  permanent virtual circuit service of, 233
  purpose of, 230–31
  value-added services of, 241–42
  virtual call service of, 232
Cellular radio systems, 69
Characters, 7
Character sets, 180
Checksums, 174–75
Christensen, Ward, 175
Clustered terminals, 34
Coaxial cable, 61–62
Codes, standard, 7
Commodore, 19
Computer systems
  in the 1950s, 14–15
  in the 1960s, 15
  in the 1970s, 16
  in the 1980s, 17–18
Control Data, 19
Convolutional coding, 177–79
Crosstalk, 61
CRTs (cathode ray tubes), 30–31
  ASCII terminals, 31–34
  display, 39–40
  keyboards, 36–39
  teleprinters versus, 28
Current-loop
  interface, 103–4
  signaling, 26
Cyclic redundancy checks (CRC), 177–79, 183–84

Data communications
  definition of, 1
  description of, 20–21
  elements of, 55
  importance of, 1
Data Communications Equipment (DCE), 20, 91
Data link content monitor, 262–63
Data packets, 237–39
Datapoint Corporation, 209
Data Switching Exchange (DSE), 228
Data Terminal Equipment (DTE), 20, 77, 91
  -DCE interface, 22–23
Data transmission
  framing, 44
  serial and parallel, 42
  timing of, 42, 44
DCE. *See* Data Communications Equipment
Decibel meter, 261–62
DeForest, Lee, 61
Demand assignment multiple access (DAMA), 166
Dibit phase shift keyed (DPSK) modulation, 82
Digital
  carrier systems, 72–73
  modulation, 85–87
  signals, 71–72
Digital Communications Corporation, 85
Digital Data Communications Message Protocol (DDCMP), 191–92
Digital Equipment Corporation (DEC), 34, 138, 187, 191, 203
Digital-to-analog converter, 118
Digital transmission
  line coding and, 137–39
  repeaters and, 140–41
Diphase coding, 138
Displays, 39–40
Dot matrix printers, 29
Double-buffered interface, 49
Double-current signaling, 26
DTE. *See* Data Terminal Equipment
Dvorak, August, 38

Echoplex, 174
Edison, Thomas, 61
Electricity, early uses of, 2–3
Electromechanical method, 45–47
Electronic receiving circuit, 47–49
Electronics Industry Association (EIA), 22, 91, 104

Electrothermal printing, 30
Entropy, 57
Equalizer
  adaptive, 121, 122
  receiver, 121–23
  transmitter, 120
Ergonomics, 36
Error control
  BiSync and, 183–84
  convolutional coding, 177–79
  See also Protocol
Escape (ESC) character, 12–13
Ethernet, 188, 203–6
Extended Binary-Coded-Decimal Interchange Code (EBCDIC), 10

Fault isolation
  loopback tests and, 259–61
  test equipment for, 261–64
Fiber optics, 63–64
  acceptance angle of fibers, 155
  attenuation and, 154–55
  bandwidth limitations and, 153
  construction of fibers, 149–50
  development of, 145–46
  FT3C system, 159
  how fibers are produced, 155
  light detectors, 156
  light sources for, 155–56
  multimode and single-mode propagation, 150–52
  numerical aperture of fibers, 155
  SL underwater, 159, 162
  Snell's Law and, 148
  wavelength multiplexing and, 156, 158
Fleming, John Ambrose, 61
Flow control, 99–101
Framing, 44
Frequency division multiple access (FDM), 163–64
Frequency division multiplexing (FDM), 72, 88
Frequency modulation, 78–79
Frequency-shift keying (FSK), 82
Friend, George, 191
Full-duplex asynchronous private line modem, 110
Full-duplex protocols, 186–92
Full-duplex transmission, 77, 81

Gauss, Carl Friedrich, 3
GTE, 85, 241

Half-duplex asynchronous private line modem, 110
Half-duplex protocols, 179–86
Half-duplex transmission, 77
Handshaking, 77
Hayes Microcomputer Products, Inc., 111–13
Hewlett Packard (HP), 34
High-frequency radio, 65
High-level data link control (HDLC), 188–91
Honeywell, 19

IBM Corporation, 19
  Binary Synchronous Communications, 179
  BiSync, 180, 181–85
  Selectric typewriter, 37
  Synchronous Data Link Control, 191
  3270 series, 35–36
IEEE 802 standards, 217–18
Impact printers, 29
Inductance loading, effect of, 60
Information
  definition of, 56
  form and content of, 20–22
  occurrence of symbols, 56–57
  redundancy in, 57
  throughput rate, 250, 251–56
Injection laser diodes, (ILD), 156
Ink-jet printing, 30
Integrated Services Digital Network (ISDN)
  basic access and, 272–73
  characteristics of, 274–75
  concept of, 271–72
  network interfaces and, 276–77
  primary access and, 273–74
Intel Corporation, 138, 203
INTELSAT VI, 163
Interfaces
  current-loop, 103–4
  double-buffered, 49
  DTE-DCE, 22–23
  ISDN, 276–77
  modem, 77
  protocols and, 169–71
  RS-232C, 22, 92–104
  RS-366, 106
  RS-422A, 104, 105
  RS-423A, 104, 105
  RS-449, 104, 106
  single-buffered, 49
  standards for, 91
  V.24, 92
International Standards Organization (ISO), 10, 188
Open Systems Interconnection Reference Model, 201-2, 243–46
Intersymbol interference, 70
ISDN. *See* Integrated Services Digital Network
Isochronous transmission, 51

Kersey, J. Ray, 188
Keyboards, 36–39

Light-emitting diodes (LED), 156
Line bridging devices, 258
Link Access Procedure Balanced (LAPB), 231
  procedures of X.25, 233–36
Link layer protocols, 227–28
Links, protocols and, 179–83
Local area network (LAN)
  ARCNet, 209–12
  attributes of an ideal, 199
  broadband, 213–14
  comparisons and trends in, 221–23
  definition of, 197
  Ethernet, 203–6
  International Standards Organization model, 201–2
  obstacles to, 201
  standards, 217–18
  token passing networks, 206–9
  versus other techniques, 197–99
  voice/data PBX, 219–20
Longitudinal redundancy check (LRC), 183–84
Loopback tests, 259–61

Manchester line code, 203–4
Marking state, 5
Mathematical Theory of Communication, The, 57
Media, bounded and unbounded, 57–58
Memory-mapped displays, 40
Message blocks, 183
Microwave radio, 66–67
Modem eliminators, 258
Modems, asynchronous
  auto-answer, 109–10
  Bell System 113 type device, 82
  Bell System 212A and V.22 modems, 82–83
  compression and, 83–84
  description of low-speed, 81–82
  full-duplex, private line, 110
  half-duplex, private line, 110
  interfaces, 77
  Pocketized Ensemble Protocol, 84–85
  purpose of, 75–76
  smart, 111–13
  telephone channel restrictions on, 76–77
  V.22 bis, 83
Modems, synchronous
  1200 bps, 126
  2400 bps, 126–29
  4800 bps, 129–31
  9600 bps, 131–33, 135–37
  components of, 116–18, 120–25
  quaternary amplitude modulation, 135–36
  receiver for, 120–23
  standards, 115–16
  terminal control for, 124–25
  transmitter for, 116–17, 120
Modem sharing devices (MSDs), 257
Modulation
  analog, 78–80
  dibit phase shift keyed, 82

digital, 85–87
pulse amplitude, 86, 87
pulse code, 87
pulse position, 86
pulse width, 86
quaternary amplitude, 131, 132
trellis coded, 136–37
vestigial sideband, 131, 132
Morse, Samuel F. B., 3
Multiplexing
description of, 87–88
frequency division, 72, 88
time division, 72–73, 89–90
wavelength, 156, 158
Multiplexors, 257

Network control system, 252–64
Network design goals, 249–51
Networking devices
interconnection of, 258–59
line bridging devices, 258
modem eliminators, 258
modem sharing devices, 257
multiplexors, 257
protocol converters, 258
Network layer protocols, 228
Networks, packet, 229–41
Neutral signaling, 26
Non-impact printers, 29–30
NRZ (Non-Return to Zero) code, 127–28
Nyquist, Harry, 70

On-line processing systems, 15
Open Systems Interconnection Reference Model (OSI/RM), 201–2, 249–68
Oscilloscope, 202

Packet networks
call request/incoming call, 239, 241
data, 237–39
description of, 229
X.25 system, 230–42
Packet switching, advantages of, 229–30
Parallel transmission, 42
Parity, 173–74
Parkinson's Law, 71
PBXs (private branch exchanges), voice/data, 219–29
Permanent virtual circuit service, 233
Personal computers, 19–20
Phase modulation, 78–79
Physical layer protocols, 227
Pocketized Ensemble Protocol Modem, 84–85
Polar signaling, 26
Printers
dot matrix, 29
impact, 29
non-impact, 29–30
serial, 28–30
Protocol, 22
BiSync, 180, 181–85
converters, 258
definition of, 227
digital data communications message, 191–92
elements of a, 171
full-duplex, 186–92
half-duplex, 179–86
high-level data link control, 188–91
interfaces and, 169–71
link layer, 227–28
network layer, 228
physical layer, 227
synchronous data link control, 191
teletypewriter, 171–75
zmodem, 175–76
Pulse amplitude modulation (PAM), 86, 87
Pulse code modulation (PCM), 87
Pulse position modulation (PPM), 86
Pulse width modulation (PWM), 86

Quaternary amplitude modulation (QAM), 131, 132
QWERTY, 37

Raeal-Vadic, 85
Radiolink systems, 66–68
Radio Shack, 19
Radio systems, cellular, 69
Radio waves, paths of, 65
Redundancy, use of, 57
Restoration of service, 264–68
RS-232C interface, 22
connector shape and pin arrangement, 92
electrical and functional parts of, 92
examples of connections for, 97–101
flow control and, 99–101
limitations of, 101–3
pin designations for, 93–96
RS-366, 106
RS-422A, 104, 105
RS-423A, 104, 105
RS-449, 104, 106
Russell, Steven, 191

Satellite Business System, 68
Satellite radiolink systems, 67–68
Satellite transmission
basic technology, 162–63
demand assignment multiple access and, 166
frequency division multiple access and, 163–64
time division multiple access and, 165–66
Scramblers, 118
Security, 251
Serial printers, 28–30
Serial transmission, 42
Shannon, Claude, 57, 70
Signaling elements, 8
Simulator, 263
Sine waves, analog modulation and, 78
Single-buffered interface, 49
Single-current signaling, 26
SL Undersea Lightguide System, 159, 162
Smartmodem, 111–13
Snell's Law, 148
Spacing state, 5
Synchronization, 184–85
Synchronous Data Link Control (SDLC), 191
Synchronous modems. See Modems, synchronous
Synchronous transmission, 50

Telebit Corporation, 85
Telegraph, development of, 3
Teleprinters, 13
definition of, 25
start-stop asynchronous transmission and, 25–26
types of, 26–28
versus CRT terminals, 28
Teletype Corporation, 25
Teletypewriter (TTY), 13
protocols, 171–75
TELEX, 13
Terminal Adapter (TA), 276
Terminal Equipment (TE), 276
Terminals
ASCII, 31–34
batch, 34
clustered, 34
CRTs, 30–34, 36–42
dumb versus smart, 32–33
emulation, 33–34
IBM 3270 series, 35–36
teleprinters, 25–28
user-programmable versus personal computers, 33
Texas Instruments Inc., 19, 28
Time division multiple access (TDMA), 165–66
Time division multiplexing (TDM), 72–73, 89–90
Timeliness of data, 250–51
Timing of transmissions, 42, 44
Token passing networks, 206–9
Trailblazer, 85, 111, 112
Transfer Rate of Information Bits (TRIB), 250, 251–56
Transmission. See under type of
Transparent-text mode, 186
Transponders, 163
Trellis coded modulation (TCM), 136–37
TWX, 13
Tymnet, 241

Unbounded medium, 57–58
Unipolar signaling, 26
UNIVAC, 19
Universal Seven-Part Data Circuit, 20, 97

Vacuum tube, thermionic, 61
Virtual redundancy check (VRC), 183–84
Vestigial sideband modulation (VSB), 131, 132
Video display terminal (VDT), 30
Virtual call service, 232
Voice frequency test set, 262

Waveguides, 62
Wavelength multiplexing, 166, 168
Weeker, Stuart, 191
Western Electric
  201, 120–29
  208, 130–31
Western Union Telegraph Company, 3
Wire pairs, 58–59
Wyse WY-50, 41–42

Xerox Corporation, 138, 203
Xmodem, protocol, 175–76

Answers to Quizzes

Chapter 1
1. d
2. b
3. a
4. c
5. c
6. b
7. a
8. c
9. c
10. b

Chapter 2
1. c
2. a
3. b
4. b
5. c
6. c
7. a
8. a
9. b
10. c
11. c
12. a
13. b
14. c
15. a
16. c
17. b
18. a
19. c
20. a

Chapter 3
1. c
2. b
3. d
4. d
5. b
6. d
7. a
8. c
9. b
10. c

Chapter 4
1. c
2. a
3. b
4. a
5. b
6. c
7. b
8. a
9. c
10. b

Chapter 5
1. d
2. b
3. c
4. c
5. d
6. c
7. a
8. c
9. b
10. c

Chapter 6
1. a and c
2. c
3. b
4. d
5. d
6. c
7. d
8. b
9. c
10. c
11. c
12. b
13. c
14. d
15. c
16. b
17. f
18. a
19. c
20. b

Chapter 7
1. a
2. c
3. c
4. d
5. b
6. b and d
7. a
8. b
9. c
10. d

Chapter 8
1. c
2. a
3. d
4. b
5. c
6. d
7. a
8. a
9. d
10. d
11. a
12. d
13. a
14. d
15. d
16. c
17. b

Chapter 9
1. a
2. d
3. b
4. c
5. d
6. c
7. d
8. a
9. b
10. a
11. b
12. c
13. a
14. d
15. b
16. a
17. c
18. c
19. b
20. a

Chapter 10
1. b
2. c
3. b
4. d
5. a
6. d
7. a
8. b
9. c
10. c

Chapter 11
1. b
2. b
3. a
4. d
5. c
6. d
7. a

HOWARD W. SAMS & COMPANY
HAYDEN BOOKS

Related Titles

Understanding Expert Systems
Louis E. Frenzel, Jr.

Understanding Advanced Solid State Electronics
Don L. Cannon

Understanding Artificial Intelligence
Henry C. Mishkoff

Understanding Automation Systems, Second Edition
Robert F. Farwell and Neil M. Schmitt

Understanding Automotive Electronics, Third Edition
William B. Ribbens and Norman P. Mansour

Understanding Communications Systems, Second Edition
Don L. Cannon and Gerald Luecke

Understanding CAD/CAM
Daniel J. Bowman and Annette C. Bowman

Understanding Digital Electronics, Second Edition
Gene W. McWhorter

Understanding Digital Troubleshooting, Second Edition
Don L. Cannon

Understanding Electricity and Electronics Circuits
David L. Heiserman

Understanding Electricity and Electronics Principles
David L. Heiserman

Understanding Fiber Optics
Jeff Hecht

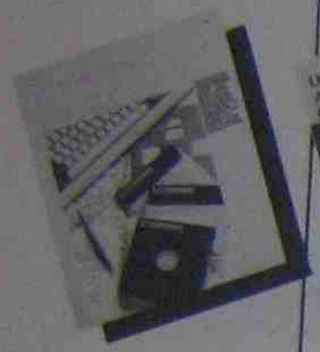
Understanding MS-DOS®
Kate O'Day and John Angermeyer, The Waite Group

Understanding Security Electronics
Joseph J. Carr

Understanding Solid State Electronics, Fourth Edition
William E. Hafford and Gene W. McWhorter

Understanding Telephone Electronics, Second Edition
John L. Fike and George E. Friend

For the retailer nearest you, or to order directly from the publisher, call 800-428-SAMS. In Indiana, Alaska, and Hawaii call 317-298-5699.



Understanding Expert Systems

Louis E. Frenzel, Jr.

This book explains how an expert system can function like an expert or intelligent consultant and answer questions, solve problems, and help make decisions.

This new Understanding Series title is suited for the technically inclined professional or manager interested in artificial intelligence and its subset, expert systems.

With the self-instructional format, readers can gain an understanding of expert systems, how they work, where they are used, and how to develop them. The material is completely illustrated, clear, concise, and application oriented. It also has marginal notes and review questions and answers.

Topics covered include:

- Expert Systems: The Big Picture
- Types of Expert Systems
- Applications of Expert Systems
- Knowledge Representation and Search
- How Expert Systems Work
- Expert System Development Tools
- How to Develop an Expert System
- References: Expert System Tool Vendors, AI Language Vendors, Publications, Glossary, and Quiz Answers

288 Pages, 7 x 9, Softbound
ISBN: 0-672-27065-X
No. 27065, $17.95



Understanding Advanced Solid State Electronics

Don L. Cannon

No single invention has influenced the electronics industry more than the integrated circuit. This volume covers all the major benefits of IC technology for the experimenter, serious hobbyist, and electronics technician in an easily-understood, fully-illustrated text for individualized learning.

It covers the advances in integrated circuits including the latest logic devices, standard cells, gate arrays, and microprocessors.

Basic concepts, progress review of IC fabrication technology, and an update of modern design approaches and IC performance (both digital and linear) are presented in ten chapters.

Topics covered include:

- Integrated Circuit Technology
- Logic Circuits
- Logic Cells and Arrays
- Microprocessors
- Digital Signal Processing
- Graphics Processors
- Communications Processors
- Bit-Slice Systems
- Linear Integrated Circuits
- Interface Integrated Circuits

272 Pages, 7 x 9, Softbound
ISBN: 0-672-27058-7
No. 27058, $17.95



Understanding Artificial Intelligence

Henry C. Mashkoff

Artificial intelligence (AI) provides the potential to model and simulate human intelligence in reasoning and decision-making. The potential impact AI could have on productivity and efficiency is staggering.

This book provides an introduction and basic understanding of this new technology. The book covers definitions, history, expert systems, natural language processing, and LISP machines. You'll also learn the specific requirements of AI hardware and the thrust of AI research.

Topics covered include:

- An Introduction to Artificial Intelligence
- History of Artificial Intelligence
- Expert Systems
- Natural Language Processing
- Amplifying Human Capabilities
- Artificial Intelligence Applications
- Symbolic Processing
- Expert System Development Tools
- LISP Machines
- Continuing Efforts in Artificial Intelligence

258 Pages, 7 x 9, Softbound
ISBN: 0-672-27021-8
No. 27021, $17.95



Understanding Automation Systems, Second Edition

Farnell and Schmall

For the newcomer, here is an in-depth look at the functions that make up automation systems—open loop, closed loop, continuous and semi-continuous process, and discrete parts. This book explains programmable systems and how to use microcomputers and programmable controllers. It surveys the types of robots available and their typical applications.

This is a book for anyone interested in the subject of automation systems, rich with insight and facts, yet written in everyday language that the non-engineer can easily understand.

Topics covered include:

- History and Importance of Industrial Control
- Industrial Control Fundamentals
- Basic Control System Hardware
- Basic Electronic Functions
- Software/Programming Languages
- Continuous Process Control
- Examples of Semi-Continuous (Job Shop) System Control
- Discrete Parts Manufacture Using Programmable Controllers
- A New Dimension—Robots
- Automation Example
- Future Trends and Reliability

280 Pages, 7 x 9, Softbound
ISBN: 0-672-27014-5
No. 27014, $17.95

Visit your local book retailer or call 800-428-SAMS.



Understanding Automotive Electronics, Third Edition

William B. Ribbens

A step-by-step guide to the latest electronic instrumentation and control devices for automobiles, this tutorial clearly explains how electronics technology functions in today's environment. Using this self-paced, individualized learning approach, the reader soon understands the complexity and breadth of automotive electronics.

The book is ideal for the beginning-intermediate level automotive mechanic, industrial training student, or "backyard" automotive repair hobbyist. Providing the necessary background in electronics and automotive basics, the book details electronic instrumentation and explains the microprocessor and its pivotal role in automobile performance. It includes scores of detailed illustrations for such functions as diagnostics, sensors, engine control, and motion control—all presented in an easily understood, easy-to-learn format.

Topics covered include:

- Automotive Fundamentals
- The Systems Approach to Control and Instrumentation
- Electronics Fundamentals
- Microcomputer Instrumentation and Control
- Computer-Aided Diagnostics
- Sensors and Actuators
- Electronic Engine Control Basics
- Typical Digital Engine Control System
- Vehicle Motion Control
- Automotive Instrumentation
- Future of Automotive Electronics Systems
- Glossary
- Answers to Quizzes

302 Pages, 7 x 9, Softbound
ISBN: 0-672-27064-6
No. 27064, $17.95



Understanding Communications Systems, Second Edition

Cannon and Luecke

In an easy-to-understand, self-paced learning format, this book presents a fascinating look at the world of electronic communications and its impact on our everyday lives. The book explores many of the systems that are used every day—AM/FM radio, telephone, TV, data communications by computer, facsimile, and satellite. It explains how information is converted into electrical signals, transmitted to distant locations, and converted back to the original information.

Topics covered include:

- The World of Communications
- Using Electrical Signals for Communications
- Basic System Functions and Conversions
- The Communications Spectrum
- Telegraph and Telephone System
- Radio and Television Systems
- Computer Networks and Systems
- Facsimile
- Satellite Communications Systems

288 Pages, 7 x 9, Softbound
ISBN: 0-672-27016-1
No. 27016, $17.95



Understanding CAD/CAM

Bowman and Bowman

For the newcomer to CAD/CAM, here's an in-depth look at this marriage between engineering design and manufacturing. From multi-color graphs and charts to 3-D depictions of machine parts, you will get the basic concepts and fundamentals underlying applications and future considerations.

Each description of this state-of-the-art communications tool is in everyday language that makes the book ideally suited for self-paced, individualized learning. Chapter quizzes and marginal notes allow for fast learning and increased retention.

The book discusses points to consider when buying a CAD/CAM system, including initial cost justification, purchase, and implementation. Training, system maintenance, and security are also discussed.

Topics covered include:

- CAD/CAM Overview
- Anatomy of a Graphics Workstation
- Input/Output
- Data Highways
- Artificial Intelligence
- The Engineer's Dialogue with the CRT
- CAD and the Personal Computer
- Applications: New Directions for Engineering
- The Third Wave in Engineering (Cost Justification)

304 Pages, 7 x 9, Softbound
ISBN: 0-672-27068-4
No. 27068, $17.95



Understanding Digital Electronics, Second Edition

Gene W. McWhorter

Learn why digital circuits are used. Discover how AND, OR, and NOT digital circuits make decisions, store information, and convert information into electronic language. Find out how digital integrated circuits are made and how they are used in microwave ovens, gasoline pumps, video games, cash registers, and other high-technology products.

Topics covered include:

- Let's Look at a System
- How Digital Circuits Make Decisions
- Building Blocks that Make Decisions
- Building Blocks with Memory
- Why Digital?
- Digital Integrated Circuits
- Mass Storage and Digital Systems
- How Digital Systems Function
- Programmed Digital Systems
- Digital Electronics Today and in the Future

264 Pages, 7 x 9, Softbound
ISBN: 0-672-27013-7
No. 27013, $17.95

Visit your local book retailer or call 800-428-SAMS.



Understanding Digital Troubleshooting, Second Edition

Jack L. Fisher

Digital electronic systems are now made from the simplest, like appliances, to the most complex, like computers. Because of this, more and more repair and maintenance technicians are required to work on digital systems. This book provides an in-depth look at the logic technology used in these systems and troubleshooting techniques and procedures for maintenance. It presents the basic concepts and fundamentals, explains readers to know how to use the latest test equipment, and shows how to apply it to the troubleshooting of digital systems.

Topics covered include:

- Digital Systems Fundamentals
- Digital Systems Functions
- Troubleshooting Fundamentals
- Troubleshooting Logic Problems
- Sequential Logic Problems
- Memory Problems
- Input/Output Problems
- Basic Timing Problems
- Advanced Techniques

272 Pages, 7 x 9, Softbound
ISBN: 0-672-22705-2
No. 27055, $17.95



Understanding Robotics and Electronic Controls

Jack Darrow

Through concise and varied text, the beginner or student learns the elements needed to analyze the robotic control principles of other and electronic gates from basic information to a complex system.

It includes practical illustrations that the reader can relate to the subject as well as multiple choice, end-of-chapter questions with answers in the back. There is material with which to test knowledge and skills, and a glossary of terms, index, and quiz answers.

Topics covered include:

- Understanding Basic Principles
- Reasons of Electrical Energy
- The Simple Electrical Circuits
- D.C. Series Circuits
- D.C. Parallel Circuits
- Combined Series and Parallel Circuits
- Electromagnetism
- What is Alternating Current?
- Calculating Resistance
- Inductance
- R.L. Circuits
- The Effect of Capacitance
- R.C. Circuits
- R.L.C. Circuits
- Transformer Action
- Some Applications of Electronic Circuits
- Appendixes: Glossary of Terms, Index, and Quiz answers

320 Pages, 7 x 9, Softbound
ISBN: 0-672-22692-7
No. 27062, $17.95



Understanding Electricity and Electronics Simplified

David Herrmann

A self-instructional format of modern fundamental principles of electricity and electronics, this book includes numerous illustrations and complete solutions to the examples as well as multiple choice, self-check questions and answers.

For the electronics or electrical technician or student with no prior experience with other texts, this book is the ideal introduction to the basic principles and theorems for easy learning and increased retention.

Topics covered include:

- The World of Electricity and Electronics
- Basic Electrical Circuits
- How to Use Meters
- The Basic Telephone System
- Reading Diagrams
- Understanding Resistors
- How to Solder
- Understanding Transformers
- Understanding Capacitors
- Understanding Diodes and Transistors
- Understanding Integrated Circuits
- How Cathode Ray Tubes Work
- Basic Circuit Action
- Audio Amplifiers
- Radio Transmitters and Receivers
- Television Transmitters and Receivers
- Digital Circuits and Computers
- Appendixes: Glossary of Terms, Index, and Quiz Answers

256 Pages, 7 x 9, Softbound
ISBN: 0-672-22661-7
No. 27061, $17.95



Understanding Fiber Optics

Jeff Hecht

Fiber optics—an inexpensive medium for moving voices or data without electromagnetic interference—has revolutionized the telephone industry.

Jeff Hecht explains in clear, nontechnical terms how fiber optics technology—telephones, computers, LAN, video, and automation. Transmitters, receivers, and coil couplers are also discussed.

Topics covered include:

- Fundamentals and Applications of Fiber Optics
- Types of Fiber
- Cabling
- Transmitters and Receivers
- Connectors, Splicers, and Couplers
- Integrated Optics
- Instrumentation
- Long- and Short-Haul Applications
- Cable TV and Video
- Computers and LANs
- Optical Interconnection
- Data and Aircraft Systems
- Non-Communication Applications

456 Pages, 7 x 9, Softbound
ISBN: 0-672-27066-8
No. 27066, $17.95



The Waite Group's Understanding MS-DOS®

Kate and Hummel
The Waite Group

MS-DOS is a very powerful and versatile operating system with millions of users. This operating system can be explored by beginning programmers in a hands-on approach at the keyboard.

Understanding MS-DOS introduces the use and operation of this popular operating system for those with little previous experience in computer hardware or software. The fundamentals of the operating system such as EDLIN, tree-structured directories and subdirectories, and such advanced features as redirection and filtering are presented in a way that is easy to understand and use.

Topics covered include:

- Organizing Data and Files
- Redirecting Input and Output
- Using the Text Editor EDLIN to Create and Edit Files
- Special Function Keys and Key Combinations
- Creating Batch Files of Often Repeated Commands
- Create and Use Tree Structured Directories

250 Pages, 7 x 9, Softbound
ISBN: 0-672-27067-6
No. 27067, $17.95



Understanding Security Electronics

Joseph J. Carr

This book details how to select, install, and maintain a security system. A self-instructional format is used to describe the theory of operation, design, and applications of electronic security devices and systems.

With this information homeowners, business professionals, and security personnel will be able to safeguard property on a do-it-yourself basis.

Electronics technicians and engineers will be able to apply their knowledge and experience in the field of security. Circuit designs and theory are provided for various control elements of security systems, and the advantages and disadvantages of various systems are discussed.

Topics covered include:

- Electronic Intrusion Alarms
- Electromechanical Detectors
- Photoelectric and Infrared Detectors
- Ultrasonic and Microwave Intrusion Detectors
- Proximity Detectors
- Audio and Visual Monitoring
- Alarm and Signaling Systems
- Access Control
- Holdup and Assault Alarms
- Custom Integrated Circuits
- Computers in Security Systems
- Detection of Objects
- Automobile Protection
- Computer Crime
- Electrical Eavesdropping and Secure Communications
- Personnel Identification

304 Pages, 7 x 9, Softbound
ISBN: 0-672-27068-4
No. 27068, $17.95



Understanding Solid State Electronics, Fourth Edition

Hoeber and Woodruff

This book explains complex concepts such as electronic semiconductor theory, how electronic circuits make decisions, how digital circuits are made. It helps you develop a basic knowledge of semiconductors and solid-state electronics. A glossary simplifies technical terms.

- What Electricity Does in Every Electrical System
- Basic Circuit Functions in the System
- How Circuits Make Decisions
- Relating Semiconductors to Systems
- Diodes: What They Do and How They Work
- Diode Performance and Specifications
- Transistors: How They Are Made
- PNP Transistor and Transistor Specifications
- Thyristors and Optoelectronics
- Introduction to Integrated Circuits
- Digital Integrated Circuits
- MOS and Large Scale Integrated Circuits
- Linear Integrated Circuits

288 Pages, 7 x 9, Softbound
ISBN: 0-672-27012-9
No. 27012, $17.95



Understanding Telephone Electronics, Second Edition

Stranger and Driscoll

Bringing lower costs and enhanced capabilities to telephone communications, digital electronics, semiconductor integrated circuits, microprocessors and microcomputers are rapidly changing today's analog systems to tomorrow's digital systems. This book explains how the conventional telephone system works and how parts of the system are gradually being replaced by semiconductor electronics. What was in an easy-to-understand format, fully-illustrated, and ideal for self-study, the book will explain the concepts and principles being applied to one of your vital communication links.

Topics covered include:

- The Telephone System
- The Conventional Telephone Set
- Electronic Speech Circuits
- A Microcomputer in the Telephone
- Digital Transmission Techniques
- Electronics in the Central Office
- Network Transmission
- Modems—Telephone Service for Computers
- Wireless Telephones

288 Pages, 7 x 9, Softbound
ISBN: 0-672-27018-8
No. 27018, $17.95

Visit your local book retailer or call 800-428-SAMS.

Visit your local book retailer or call 800-428-SAMS.

Sams Understanding Series

Understanding Data Communications

Brings you basic understanding of the subject—in an easy-to-learn format.

$18.95 US

By: George E. Friend, John L. Fike, H. Charles Baker, John C. Bellamy

ABOUT THIS BOOK

Data communications — the communication of information using digital signals — is the lifeblood of our economy. Facts and figures for corporations, telephone network control signals, banking and credit card purchases, newspaper copy sent from a central location to outlying printing stations, and more, dramatizes the impact of this technology. An understanding of what data communications is and what it does may be as necessary in the future as knowing how to dial a telephone. It is something you need to know.

WHAT THIS BOOK CONTAINS

1. An Overview of Data Communications Learn what data communications is and how it got where it is today. What are the codes and what do DTE and DCE mean?

2. Data Terminals Discusses I/O terminals — teleprinters, CRT, ASCII, non-ASCII, batch, and clustered, as well as terminal subsystems. Covers asynchronous and synchronous operation.

3. Messages and Transmission Channels Defines information and explains wire pairs, coaxial cables, waveguides, electronic amplifiers, fiber optics, microwaves, satellites, radiotelephones, and cellular radio systems.

4. Asynchronous Modems and Interfaces How does a modem interface to a DTE and to the telephone line? Covers current loop, RS-232C, -422A, -423A, -499, -366, and CCITT V.24, V.32, V.33, and X.21.

5. Synchronous Modems and Digital Transmission More bits per baud with synchronous modems. Find out about 1200, 2400, 4800, and 9600 bps modems and unipolar, polar, and diphase digital codes.

6. Fiber Optic and Satellite Communications Two new technology transmission systems are discussed. [illegible] Learn about [illegible] operation, [illegible] frequencies and [illegible]

7. Protocols [illegible] Clarifies the [illegible] protocol [illegible] protocol [illegible] are full-[illegible] USC for [illegible] DDCMP.

8. Alternatives in Local Area Networks The ISO model for local area networks is the basis [illegible] to describe contention access (Ethernet), token passing (ARCNet), Broadband networks, LAN standards comparisons, and trends are presented.

9. Architectures and Packet Networks X.25 is used to describe packet switching architecture, and the ISO model is described further in relationship to it. XSeries of recommended standards is included.

10. Network Design and Management What are important network design goals? How can you maximize TRIB? When faults occur, how do you isolate them and restore service quickly? Lists networking devices.

11. ISDN Describes the background architecture, characteristics, and equipment and interface of the Integrated Services Digital Network. Differentiates Basic and Primary Access.

UNDERSTANDING SERIES

Understanding Series books from Sams form a library written for *anyone* who wants to learn easily and quickly more about today's electronic technology and the era of personal computing, its impact on our world, and its application in our lives. Each book is written in bright, clear, down-to-earth language and focuses on one aspect of what's new in today's electronics. Whether you're a serious hobbyist, experimenter, or otherwise involved in electronics technology — or just curious about what goes on inside today's consumer electronics appliances, calculators, and the new microcomputers — there are *Understanding Series* titles you'll enjoy.

HOWARD W. SAMS & COMPANY
A Division of Macmillan, Inc.
4300 West 62nd Street
Indianapolis, Indiana 46268 USA

90000
9 780672 272707
ISBN 0-672-27270-9

TAFE NSW LIBRARY NETWORK
3 5555 07073655 0

TRL, ANALOG, and DATA COMMUNICATION



william sinnema

621.38
SINN

Digital, Analog, and Data Communication

William Sinema

RESTON PUBLISHING COMPANY, INC.
A Prentice-Hall Company
Reston, Virginia

Library of Congress Cataloging in Publication Data

Sinnema, William, 1937–
Digital, analog, and data communication.

Includes index.
1. Data transmission systems. 2. Digital communications. I. McGovern, Tom. II. Title.
TK5105.S558 621.38 82-326
ISBN 0-8359-1301-5 AACR2

Design and editing
by Ginger Sasser

©1982 by
Reston Publishing Company, Inc.
A Prentice-Hall Company
Reston, Virginia

All rights reserved. No part of this book may
be reproduced in any way, or by any means, without
permission in writing from the publisher.

10 9 8 7 6 5 4 3

Printed in the United States of America

Contents

Preface ..... *xi*

Acknowledgements ..... *xiii*

Chapter 1 Introduction to Communication Systems 1

1–1 The Basic Communication System 1
1–1.1 Block Diagram of a Communication System, 2
1–2 Voice Communication 4
1–2.1 The Decibel, 4
1–2.2 The dBm and dBW, 6
1–2.3 Characteristics of Human Speech, 6
1–2.4 Noise, 7
1–2.5 Zero Transmission Level Point and dBm0, 8
1–2.6 Signal-to-Noise Ratio, 9
1–2.7 Volume Unit, 10
1–2.8 Sampling Theorem, 10
1–3 The Telephone Network 10
1–3.1 Modulation and Multiplexing, 11
1–3.2 Echo, 17
1–4 Digital Communication 17
1–4.1 The Binary, Octal, and Hexadecimal Numbering Systems, 18
1–4.2 Encoding of an Analog Signal into a Digital Signal, 20
1–4.3 Time Division Multiplexing, 21
1–4.4 A Digital AM System, 21
1–4.5 Considerations in the Usage of Digital Communications, 23
1–5 Data Communications 24
1–5.1 Communication Codes, 25
1–5.2 Parallel and Serial Data Transmission, 25
1–5.3 Asynchronous and Synchronous Transmission, 27
1–5.4 Universal Synchronous Asynchronous Receiver/Transmitter (USART), 20 mA Loop and Modem, 31
1–5.5 Simplex, Half-Duplex, and Full-Duplex, 32
Problems 33

Chapter 2 The Telephone Network 37
2–1 Introduction 38
2–2 The Telephone Network 38
2–2.1 Hierarchy of Exchanges, 39
2–3 Voice Bandwidth 42
2–4 Instruments 43
2–4.1 The 500-Type Telephone, 43
2–4.2 Touch-Tone Telephone, 47
2–4.3 Touch-Tone Decoders, 48
2–5 Office-to-Office Signaling 50
2–6 Transmission Line Fundamentals 50
2–6.1 Subscriber's Loop, 56
2–6.2 Trunk Circuits, 59
2–6.3 Frequency Division Multiplexing, 61
2–6.4 CCITT Groups, 66
2–7 Switching Facilities 69
2–7.1 Step-by-Step Switch, 71
2–7.2 Cross-Bar Switch, 73
2–7.3 Electronic Switching Systems (ESS), 80
2–7.4 Digital Switching, 84
2–8 Special Telephone Services 90
Problems 94

Chapter 3 Telephone System Performance Requirements and Impairments 97
3–1 Introduction 98
3–1.1 Four-Wire Terminating Set, 98
3–1.2 The Reference Equivalent, 101
3–1.3 Network Stability, 101
3–1.4 Echo, 102
3–1.5 Via Net Loss, 103
3–2 Telephone System Impairments 105
3–2.1 Crosstalk, 107
3–2.2 Intermodulation Noise and Loading, 108
3–2.3 Thermal and Shot Noise, 112
3–2.4 Impulse Noise, 114
3–2.5 Attenuation Distortion, 115
3–2.6 Envelope Delay Distortion, 115
3–2.7 Line Conditioning, 120
3–2.8 Compandors, 122
Problems 125

Chapter 4 Transmission of Data Over an Analog System 128
4–1 Introduction 128
4–2 Bandwidth Limiting of a Pulse Train 129
4–3 20 mA Loop 134
4–3.1 Line Drivers,141 137
4–4 Cabling Precautions 143
4–5 Bauds and Bits 143
4–6 Modems 145
4–6.1 Turn Around and Message Acknowledgement, 146
4–6.2 Modem Modulation Techniques, 148
4–6.3 Theory of Phase Shift Keying, 152
4–6.4 Interface Standards, 162
4–6.5 T103A Data Set, 165
4–6.6 Line Drivers, 174
4–6.7 Modem Chips, 181
4–6.8 FSK Frequency Response, 183
4–7 Universal Synchronous/Asynchronous Receiver/Transmitter (USART) 184
Problems 197

Chapter 5 Analog Pulse Modulation 203
5–1 Introduction 204
5–2 Time Sampling 204
5–3 The Sampling Theorem: Natural Sampling 207
5–4 Sample and Hold: Flat-Top Sampling 215
5–5 Filter Characteristics 222
5–6 Time Division Multiplexing 224
5–7 Pulse Duration and Pulse Position Modulation 230
Problems 236

Chapter 6 Digital Modulation 238
6–1 Introduction 239
6–2 Simple Delta Modulation 239
6–3 Companded Delta Modulation 242
6–4 Pulse Code Modulation 242
6–5 Aperture Time 248
6–6 Amplitude Sampling 250
6–7 Encoders and Decoders 259
6–7.1 Codecs, 267

6–8 24-Channel System Frame Alignment 274
6–9 PCM Terminal Equipment 278
6–10 Line Pulse Format 280
6–11 North American Multiplex Hierarchy 283
6–12 Measurement of Quantization Noise 284
6–13 Introduction to Information Theory 285
Problems 289

Chapter 7 Effects of Noise and Other Distortions on Digital Transmission 296

7–1 Introduction 297
7–2 Quantization Noise 297
7–3 Effects of Gaussian-Type Noise on Digital Transmission 301
7–4 Transient Noise 307
7–5 Peak-to-Average Ratio Test 310
Problems 312

Chapter 8 Pulse Transmission over Bandlimited Systems 315

8–1 Introduction 316
8–2 The Unit Impulse 316
8–3 Frequency Response of an Ideal or Brickwall Filter 318
8–4 The Eye Diagram 324
8–5 Information Capacity 325
8–6 The Raised-Cosine Channel Response 328
8–7 Partial Response (Correlative) Techniques 330
Problems 346

Chapter 9 Coding for Communications 350

9–1 Introduction 351
9–2 Codes 352
9–3 Error Detection and Correction 358
9–4 Cyclic Codes 362
9–5 Line Codes 371
9–6 Channel Throughput and Efficiency 373
Problems 379

Chapter 10 Network Protocol 382

10–1 Introduction 382
10–2 Switching 383
10–3 Protocol Levels 383
10–4 Data Link Control 388
10–4.1 BISYNC, 391
10–4.2 X.25, 392
10–4.3 Implementation, 400 390
10–5 Packet Level 402
10–6 Conclusion 408
Problems 409

Appendices 412

Appendix A. Fourier Series 413
Appendix B. Noise Power Ratio (NPR) 416
Appendix C. Group Delay 417
Appendix D. The XR-567 Touch-Tone Decoder 420
Appendix E. The Double-Balanced Ring Modulator 425

Index 429

Preface

Digital communication systems are being rapidly developed to provide reliable communication for digitized analog signals and data. Such analog signals as voice and television must first be translated into a digitized format for these systems, whereas data can be directly carried. Since data transmission is probably the fastest growing aspect of communications, chiefly because of the advent of the computer and microprocessor, it is important for technologists and engineers to become aware of the transmission techniques involved.

As the analog telephone system is still very much with us, we must of necessity consider some of its major transmission characteristics and how these are utilized to carry digital signals. At the same time, we should consider some of its limitations when carrying digitized signals. This book shall attempt to acquaint the reader with the widespread telephone network and with the principles and applications of the more recent digital communication schemes. It must be kept in mind that it is impossible, in a reasonably sized book, to cover each subject exhaustively. The reader can obtain more information on any particular area from one of a score of engineering texts.

A chapter-by-chapter breakdown of the material presented in this book is as follows. Chapter 1 is a general description of analog and digital techniques, which are all expanded upon in later chapters. It also defines some of the basic units used in telephony. The telephone network is described in Chapter 2, and Chapter 3 deals with some of the more technical details of the telephone network. This material is included to provide background information on the limitations of the network when attempting to transmit either analog or digital signals. Chapter 4 introduces techniques of transmitting data over the present analog system, and Chapter 5 deals with the problem of aliasing when an analog signal is sampled. Chapter 6 considers the modulation methods most commonly employed when digitizing an analog signal, and also covers some practical aspects of PCM terminal equipment on the North American scene. Information theory is introduced as well. In Chapter 7, the effects of noise and other distortions on digital transmission are explored. Chapter 8 deals with the fundamental concepts of pulse transmission over a bandwidth-limited channel. The Nyquist criteria and various methods of reducing intersymbol interference, such as the raised-cosine channel response and partial response, are also considered. Chapter 9 covers many of the codes, as well as error detection and correction techniques, presently employed in the communication industry. Chapter 10 considers some of the protocols presently employed in the communication network when transferring information between computers and computer terminals. Packet switching is also briefly considered.

This textbook is intended as an introductory text to digital and data transmission techniques. A student using this book should have a knowledge of algebra and be familiar with basic logic circuits. An acquaintance with differential and integral calculus would be helpful,

although not essential. The description of the final equations should prove as an adequate interpretation.

When dealing with modern technology and with the computer in particular, it is imperative that engineers and technologists begin to reflect somewhat on the basic problems that have come to be associated with technology and its development. Towards this end, I would like to recommend the book *Technology and the Future: A Philosophical Challenge* by Egbert Schuurman (Toronto: Wedge Publishing Foundation, 1980). It gives particular attention to the basic dichotomies presently at work in the views of modern technology.

Acknowledgements

In the course of writing a textbook of this nature, one inevitably leans heavily upon the bits and pieces of materials developed by others. My efforts have chiefly consisted of organizing the various bits into a more consistent and logical manner, which can be more readily followed and comprehended by the reader. Thus, I owe a debt of gratitude to the many authors in the electronic communication field.

I must thank with appreciation particularly Mr. Tom McGovern for authoring the last chapter on protocols used in modern communication systems. He is presently with the Computer Systems Technology at the Northern Alberta Institute of Technology in Edmonton. I would also like to thank Mr. J. Want for his suggestions on partial response behavior, and Mr. D. Maruszeczka for his aid in discussing the Lenkurt PCM system. Mr. P. Hancock's advice is also appreciated.

I must extend my thanks to Mrs. Liz Sidor, who has been so pleasantly patient in deciphering my handwritten notes when typing the original notes for student use.

Final thanks must go to my wife, Edith, who has so cheerfully endured the role of "writer's widow."

William Sinnema

1

Introduction to Communication Systems

1–1 THE BASIC COMMUNICATION SYSTEM

The purpose of any communication system is to convey a message from one point in space to another point in space. The originating input is frequently called a source whereas the terminating end is frequently referred to as the sink. If the message is understandable, then information has been conveyed from the source to the destination.

The nature of the message can take on a variety of forms. It can be a continuous time-varying quantity such as speech, music, motor bearing temperature, or pipeline pressure. It can be a discrete signal which only takes on fixed levels or values, such as a signal transmitted by a digital computer as a string of pulses or holes punched in a paper tape. It can be a continuous spatial varying quantity such as the image on a TV camera vidicon tube, or it can be some combination of the above.

Because most messages are not electrical in nature, to carry them over an electrical system the message must be converted to an electrical signal. This function is performed by a transducer. Microphones, speakers, card readers, teletypewriters, and printers are all examples of transducers. With the exception of the discussion on the voice signals in a telephone hand-set, all the messages considered in this text are electrical in nature.

1–1.1 Block Diagram of a Communication System

After a message is converted to an electrical signal, the signal may directly modulate a carrier (high frequency signal), or it may be converted into another format by an encoder, for example, into a binary bit stream, before modulating a carrier wave. Encoding is rarely used in analog communication schemes such as AM and FM, but is frequently employed in digital modulation schemes such as delta modulation and pulse-code modulation (PCM). One of the chief advantages of employing the latter type of encoding is that the signal-to-noise ratio, or in binary terms, the error rate, is almost independent of the number of links or repeating amplifiers in a system. Modulation is employed in order to (a) more efficiently launch the radiated wave into space, (b) permit multiplexing, and (c) to improve signal-to-noise ratio. For efficient launching or reception of

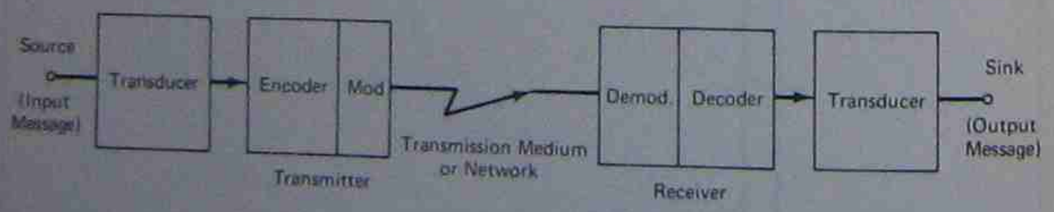


Fig. 1-1 The Basic Blocks of a Communication System

an electromagnetic wave or to obtain what is commonly called matching, the radiating or receiving device (antenna) must be a significant portion of a wavelength in size. This is true since the wavelength (λ) of an electromagnetic wave in free space is related to the velocity of light (C) by the following relation:

$$\lambda = \frac{C}{f} \tag{1-1}$$

where λ is the wavelength in m
C is the velocity of light; 3×10^8 m/s
f is the frequency of operation

For a signal frequency of 100 Hz, for instance, the wavelength would be $\lambda = 3 \times 10^8 \text{ m/s} \div 100 \text{ s} = 3 \times 10^6$ m. If the antenna must be at least one-tenth of the wavelength in size, this would result in a 300 km antenna. By employing modulation, this frequency can be translated upwards so that a reasonably sized antenna can be used. Translating the various modulating or baseband signals upwards in different frequency bands also makes it possible to transmit many signals simultaneously. This is called frequency division multiplexing. It is not uncommon to carry 2,000 voice channels on one carrier with this technique.

As is done in FM, modulation can also be used to improve the demodulated signal-to-noise ratio. The price for this, however, is an increased transmission bandwidth for the same baseband signal—a severe penalty when the spectrum is already fully utilized.

The transmission medium in Fig. 1-1 is not restricted to a microwave link which transmits an electromagnetic wave in free space, but may include a pair of wires, a coaxial line, or an optical fiber. A loss or signal attenuation is experienced in each of these situations, which causes a reduction in the signal-to-noise ratio. For severe losses, repeater or regenerator stations must be inserted to compensate for some of the loss.

The receiver performs the inverse function of the transmitter. It amplifies the weak signal received and demodulates and decodes the signal. The output transducer then converts the electrical signal into its proper form.

Contaminants often affect the quality of a communication system. These contaminants include (1) signal distortion due to imperfect system response, (2) interferences from extraneous signals, and (3) thermal noise from random electron motions in a conducting medium. Within each block of Fig. 1-1, one or several contaminants can creep in to disrupt the system.

A full treatment of noise will be given in a separate chapter.

1–2 VOICE COMMUNICATION

Before considering the nature of voice in speech, first consider the decibel unit that is used to describe the amplitude and frequency response of amplifiers.

1–2.1 The Decibel (dB)

The decibel is a relative power measurement. It does not give the absolute signal level at some point in a system. It is symbolized by dB and defined as

$$\text{dB} = 10 \log (\text{the power ratio}) \qquad (1\text{-}2)$$

Consider the examples of 1 milliwatt (mW) being applied to the inputs of two different amplifiers, resulting in 2 mW and 10 mW at their outputs. In these cases, the gains in dB of the amplifiers (10 log (output power/input power)) are, respectively:

$$10 \log \frac{2 \text{ mW}}{1 \text{ mW}} = 3 \text{ dB}$$

and $$10 \log \frac{10 \text{ mW}}{1 \text{ mW}} = 10 \text{ dB}$$

Please note from these examples that a doubling in power gives a 3 dB gain whereas a power increase of 10 gives a 10 dB gain. Similarly, power ratios of 100, 1000, and so forth result in gains of 20 dB, 30 dB, and so forth. If the power output is less than the power input, the gain is given in terms of negative decibels or often translated as a positive decibel loss. For instance, the gain of the circuit shown in Fig. 1-2 is 10 log (.05/1) = −13 dB. Alternatively, this can be stated as having a loss of 13 dB. If the output and input resistances of a network are identical, the decibel gain can be restated as

$$\text{dB} = 10 \log \frac{(E_{output})^2/R}{(E_{input})^2/R} = 20 \log \left(\frac{E_{output}}{E_{input}}\right) \qquad (1\text{-}3)$$

or $$\text{dB} = 10 \log \frac{(I_{output})^2 R}{(I_{input})^2 R} = 20 \log \left(\frac{I_{output}}{I_{input}}\right) \qquad (1\text{-}4)$$

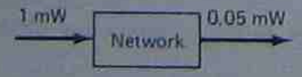


Fig. 1-2 A Negative Gain Network

Even if the resistance levels are not identical, Eqs. (1-3) and (1-4) are often still used. The equations allow for ease in graphically plotting the gain of a frequency-sensitive network such as a filter. Note that ignoring the difference in resistance does not give a true power ratio.

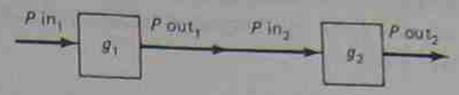


Fig. 1-3 Cascaded Network

In cascading networks, the gains in decibels of the individual networks can be added. This can be easily proven by observing Fig. 1-3 where the individual power gains are

$$g_1 = \frac{P_{out_1}}{P_{in_1}}, \text{ and } g_2 = \frac{P_{out_2}}{P_{in_2}}$$

where $$G_1 = \left(10 \log \frac{P_{out_1}}{P_{in_1}}\right) \text{dB}$$

$$G_2 = \left(10 \log \frac{P_{out_2}}{P_{in_2}}\right) \text{dB}$$

$$G_{overall} = \left(10 \log \frac{P_{out_2}}{P_{in_1}}\right) \text{dB}$$

$$= 10 \log \left(\frac{P_{out_2}}{P_{in_2}} \times \frac{P_{out_1}}{P_{in_1}}\right) \text{dB}$$

since $$P_{in_2} = P_{out_1},$$

Therefore

$$G_{overall} = \left(10 \log \frac{P_{out_2}}{P_{in_2}} + 10 \log \frac{P_{out_1}}{P_{in_1}}\right) \text{dB}$$

$$= (G_1 + G_2) \text{ dB}$$

Example 1-1

Without the use of a calculator, find the approximate power gain of a 13 dB amplifier.

Solution

This is equivalent to a 10 dB amplifier in cascade with a 3 dB amplifier. The overall power gain ratio is therefore 10 × 2 = 20.

Example 1-2

What is the dB loss of the Fig. 1-4 network.

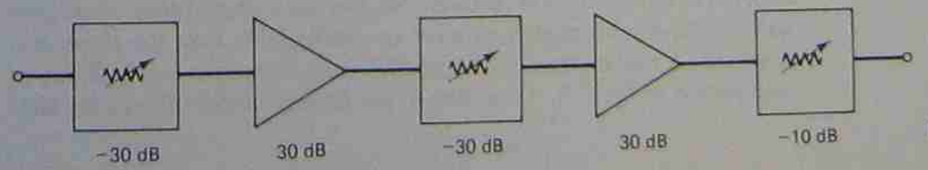


Fig. 1-4

Solution

The net overall gain is $-30 + 30 - 30 + 30 - 10 = -10$ dB or a loss of 10 dB.

1–2.2 The dBm and dBW

The Decibel Referred to One Milliwatt (dBm) gives absolute measure of the signal level at some point in a system. It relates all power levels to a one milliwatt (mW) level.

$$\text{dBm} = 10 \log \frac{Power}{1 \text{ mW}} \tag{1-5}$$

A power level of 1 mW for instance represents 0 dBm. An absolute power level of 2 mW represents 3 dBm. Similar to the dBm, the Decibel Referred to One Watt (dBW) is an absolute decibel unit using one watt as a reference. It is frequently used in microwave applications.

$$\text{dBW} = 10 \log \frac{Power}{1 \text{ W}} \tag{1-6}$$

Thus, 1 mW represents 0 dBm, or

$$10 \log \frac{10^{-3}}{1} = -30 \text{ dBW}.$$

Similarly 1 W represents 0 dBW or 30 dBm.

When adding or subtracting two signals stated in dBm's or dBW's, it is important to convert back to milliwatts or watts and add the latter. For example, a 13 dBm (20 mW) signal added to a 10 dBm (10 mW) signal is equivalent to 30 mW, or 14.8 dBm. On the other hand, the addition of two equal power levels doubles the output power or, equivalently, exceeds either one by 3 dB. Thus, 15 dBm + 15 dBm = 18 dBm.

1–2.3 Characteristics of Human Speech

Human Speech consists of a complex assortment of frequency components ranging from about 100 Hz to 8000 Hz. The powers emitted vary from about 0.01 to 5000 μW giving a dynamic range of around 57 dB. The major portion of energy content exists in the low-frequency range, as shown in Fig. 1-5. As far as understanding is concerned, however, the low frequencies (up to about 700 Hz) add very little intelligibility to the signal.

Intelligibility is a measure of how well a message is received over an imperfect communication medium. Studies have shown that the higher speech frequencies contribute more to intelligibility than the lower frequencies. A curve showing the relative importance for intelligibility is also shown in Fig. 1-5. According to the figure, bandwidth can be con-

served and a good information transfer still obtained by limiting the bandwidth of the transmission medium. The standard CCITT (Comité Consultaif International de Téléphone et de Télégraphie—International Consultative Committee for Telephony and Telegraphy)* bandwidth of a voice channel is 300 to 3400 Hz. Many telephone subsets reduce this even further to about 500 to 3000 Hz.

1–2.4 Noise

Because the ear is not equally sensitive to all frequencies, any noise appearing in telephony does not have the same interfering effect for each frequency. In order to measure the relative interfering effect of noise, some form of frequency weighting is necessary to equalize the effects.

By performing subjective experiments, many standard weighting curves have been adopted that depend upon the telephone subsets used in the experiments. The weighting adopted by CCITT, known as psophometric weighting, uses an 800 Hz reference frequency and is expressed in terms of pW (10^{-12} W). The units are expressed as pWp (p for psophometric weighting) or dBmp.

A uniform 300 Hz to 3400 Hz band of noise, such as white noise, results in a weighted power that is 2.5 dB less than the unweighted power. Thus, if the white noise power in a 300 Hz to 3400 Hz band is

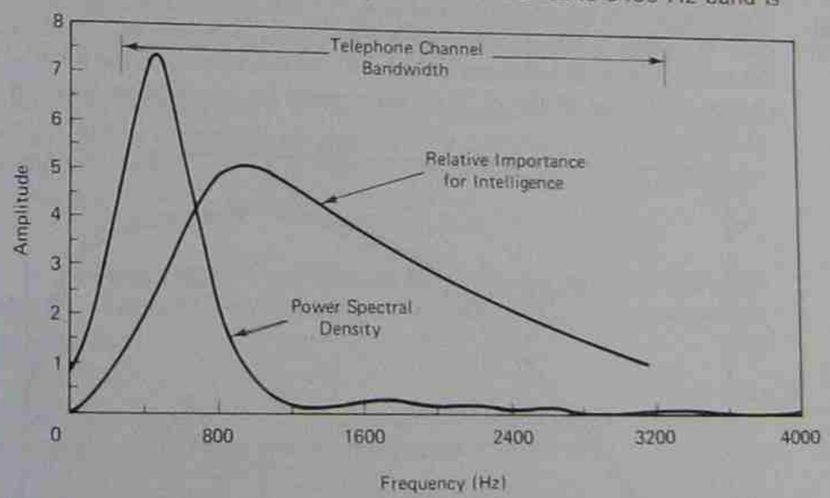


Fig. 1-5 Frequency Characteristic and Power Spectral Density of Human Speech Adapted from "A comparative study of various quantization schemes for speech encoding," P. Noll, *Bell System Tech. J.* vol. 54, no. 9, (Murray Hill, N.J.: Nov., 1975), pp. 1597–1614. © 1975, American Telephone and Telegraph Company. Reprinted with permission.

* See Sec. 1-3 for more information

y dBm, the psophometric noise is (y − 2.5) dBmp. For a 0–4 kHz band of white noise, the unweighted noise power is reduced by 3.6 dB.

In North America different weightings are employed, all using 1000 Hz as a reference. Ignoring the older dBa and dBrn weightings, which were used in the Western Electric type 144 and F1A handsets, the newer 500-type handset displays the C-message weighting. (See Fig. 1-6). In order to maintain a positive number on the relative interfering effects, a 1 pW (−90 dBm) level at 1000 Hz is used as a reference. By denoting this level, that is −90 dBm as 0 dBrnc, all noise measurements appear as positive dBrnc's. White noise over a 300 to 3400 Hz bandwidth, when C-weighted, results in about a 1.5 dB reduction. Thus, a white noise signal at 0 dBm produces a −1.5 dBm, or 88.5 dBrn, C-weighted signal. If the white noise power is −60 dBm, the C-message-weighted noise is −60 + (90 − 1.5) = 28.5 dBrnc.

Conversion between the European and C-message North American units is given by dBmp = dBrnc − 90. Figure 1-6 gives both the CCITT psophometric and the C-message weighting curves.

1–2.5 Zero Transmission Level Point (0TLP) and dBm0

In dealing with telephone systems, it is convenient to define some point as the zero transmission level and to measure the signal and noise levels at other points in the system relative to it. In North America this reference point is located at the output of the toll exchange (see Sec. 2–2.1). If in this instance a signal has experienced a 12 dB loss relative to the toll exchange, it will be at the −12 dB level point or at a relative level of −12 dB.

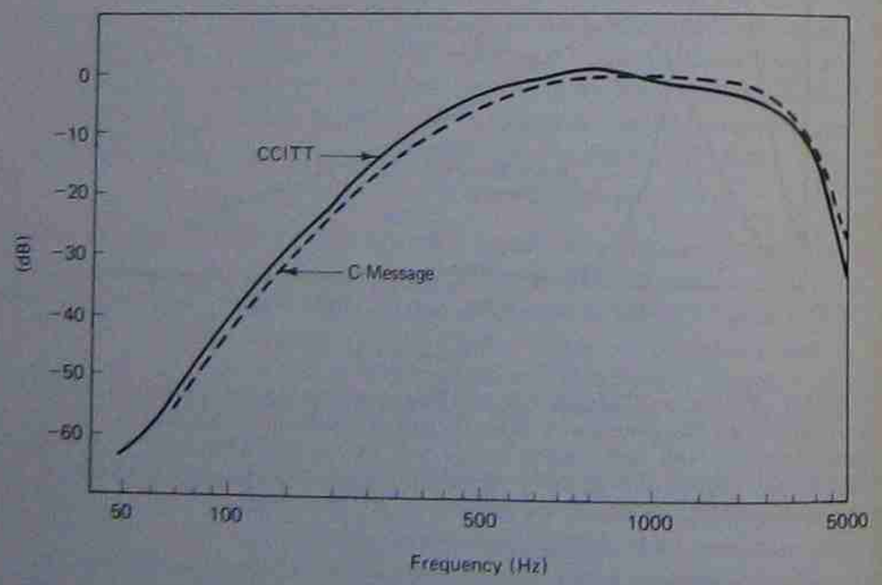


Fig. 1-6 Psophometric and C-Message Weighting Curves

When referring signals back to the zero transmission level point, the units are expressed in dBm0. The dBm0 unit indicates what the power would have been at the zero transmission point. For instance, if the signal power of 11 dBm is observed at a point with a relative level of −12 dBr, then the corresponding power at the 0TLP is 23 dBm0.

Conversely, if the signal power at the 0TLP is 6 dBm0, then at a relative level point of 7 dBr, the signal power will be 13 dBm.

$$\text{In general,} \quad \text{dBm} = \text{dBm0} + \text{dBr} \tag{1-7}$$

When applying a test tone to a system, the tone level at the 0TLP is set at 0 dBm.

1–2.6 Signal-to-Noise Ratio (S/N)

The term *signal-to-noise ratio* is frequently used to express in decibels the amount by which a signal level exceeds the noise level. If a signal level of 20 dBm experiences a noise level of 5 dBm, then the $(S/N)_{dB} = S_{dBm} - N_{dBm} = 20 - 5 = 15$ dB. For voice communication a S/N rate of 30 dB is considered satisfactory, whereas for video the minimum level should be 45 dB.

Example 1-3

What is the signal-to-C-message-weighted-noise ratio at the 0TLP of a system when the signal power is 0 dBm (1 mW) and the unweighted white noise level is −20 dBm spread over a 3.1 kHz bandwidth?

Solution

The signal power level is 0 dBm. The unweighted noise is −20 dBm which, when C-message-weighted over a 3.1 kHz bandwidth, gives a weighted noise of −20 dBm − 1.5 dB = −21.5 dBm. Therefore $S/N_{C\text{-weighted}} = 0 - (-21.5) = 21.5$ dB.

The terms dBrnc0 and dBm0p refer to noise power measured in dBrnc's or dBm0's, respectively, at the 0TLP with C or phosphometric weightings.

Example 1-4

What is the signal-to-C-message-weighted-noise ratio when the signal is at a level of −9 dBm0 and the noise level is 48 dBrnc0?

Solution

Referring all levels to the 0 dB TLP, the signal has a level of −9 dBm0 and the C-weighted noise has a level −90 + 48 = −42 dBm0. Thus, the $S/N_{C\text{-weighted}} = -9 - (-42) = 33$ dB.

If the noise level was unweighted at 48 dBrnc0, then the weighted noise would be decreased by about 1.5 dB, giving an $S/N_{C\text{-weighted}}$ ratio of 34.5 dB.

1–2.7 Volume Unit (VU)

In order to measure the power level, or volume, of broadcast programs and certain types of speech and music, a volume unit (VU) is employed. If a simple decibel meter is used to monitor the program power level, the indicating needle will attempt to follow every fluctuation of the power level and will be difficult to read. To overcome this difficulty, a dc millimeter with a slow response time is used to measure VU's. When a steady sine wave is suddenly applied to such a VU meter, the indicating needle will move to within 90% of the steady state value in 0.3 s and overswing the steady state value by less then 1.5%.

A standard volume indicator is calibrated to read 0 VU when connected to a 600 Ω circuit carrying 1 mW of sine wave power within the 35 to 10 kHz band. It will not, however, indicate the true power of complex waves, but read some value between the average and peak value.

VU indicators are frequently used to monitor speech and music levels. Excessive levels can cause distortion in loudspeakers, handsets, and tape recorders; overmodulation in radio transmitters; or intermodulation in amplifiers.

1–2.8 Sampling Theorem

Any band-limited signal does not need to be continuously transmitted in order to retain all its information content. As will be more fully explained in Chap. 7, no information is lost if such a signal is sampled at regular intervals of time at a rate equal to or higher than twice the highest frequency component. The signal can be reconstructed by passing it through a perfect or ideal low-pass filter.

Take, for example, a voice signal that has an upper frequency component of 4 kHz. By sampling it at a rate of 8 kHz as shown in Fig. 1-7, the sampled signal can be transmitted and reconstructed without any loss of information. The reconstructed signal will naturally experience some delay.

1–3 THE TELEPHONE NETWORK

Throughout North America, Europe, Australia, New Zealand, Japan, and several other nations, telephone hook-up is readily available to most households and offices. With almost 150 million telephones in North

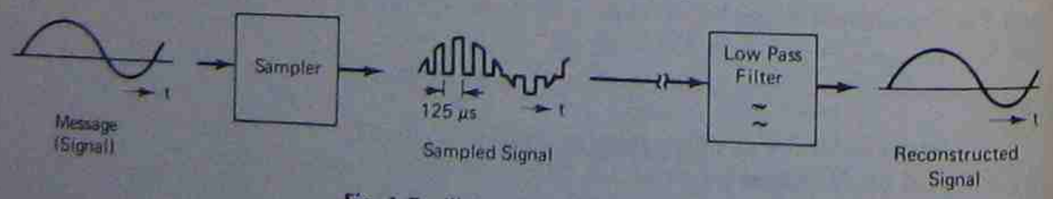


Fig. 1-7 Illustration of Sampling

America and an equal number in the rest of the world, the telecommunications industry is intricately woven in the fabric of our society. The present world-wide switched public telephone system must be considered whenever a new communication service for the public-at-large is contemplated. Large, overlapping communications networks are extremely uneconomical.

In addition, any changes or improvements in the existing network must be compatible with the overall system if they are to be introduced economically. Huge plant investments cannot be quickly written off, and plants cannot be replaced in a short time span. Just to replace telephone sets in every home would involve billions of dollars and many hours of installation time.

Thus, just as with monophonic FM and stereo FM, or black-and-white and color TV, improved communications technology must operate within the limits of old equipment. For example, one limitation of the present telephone set is the bandwidth range of 300 Hz to 3400 Hz (see Fig. 2-6). The present telephone sets cannot be used if operated below or above this frequency range.

Communication networks vary from country to country and from area to area within a country in terms of both the tariff structures and the type of ownership (private, public, or some mixture of the two). Public telephone networks, because of their monopolistic nature, are normally subjected to some form of regulation. In order to permit interconnection with networks located in different countries, each nation's network must meet some internationally set standards.

In Canada about 70% of the telephones are privately operated by Bell Canada, 16% are provincially or municipally owned, and the rest are operated by small independent common carriers. A communication common carrier is a company that offers communication facilities to the public. In the United States the privately-owned American Telephone and Telegraph Company (AT&T) operates over 80% of the telephones. AT&T owns Western Electric Company and most of the stock in the Bell-operated companies. AT&T also owns one of the world's largest research organizations, the Bell Telephone Laboratories. In the U.S. there are approximately 2800 common carrier companies, whereas there are about 1600 in Canada.

In order to formulate international standards in the telecommunications industry, the International Telecommunications Union (ITU) was organized under the auspices of the United Nations. The ITU evolved from the International Telegraph Union, which was established in 1865 for a similar purpose.

Within the ITU there are three major divisions.

1. The Consultative Committee on International Telegraphy and Telephony (CCITT)
2. The Consultative Committee on International Radio (CCIR)

3. The International Frequency Registration Board (IFRB)

The major responsibility of the latter division is the allocation of radio frequency spectrum to various users. The CCIR makes technical recommendations related to radio, and the CCITT makes technical recommendations related to telephony and telegraphy.

The standards set forth by the ITU permit the interconnection of various types of national communication systems, including telephone, telegraph, data, and television. Although most nations have accepted the ITU standards, variations do occur, more in North America than in Europe. This text will refer to both the CCITT and North American Standards.

In North America the manufacturers of electronic components and equipment have formed a trade association known as the Electronics Industries Association (EIA). This association prepares comprehensive standards, many of which eventually become internationally recognized. For a listing of EIA standards and other engineering publications, write:

Electrical and Electronic Manufacturing Association
One Yonge Street, Suite 1608
Toronto, M5E 1R1

or

Electronic Industries Association
2001 Eye Street N.W.
Washington, D.C. 20006.

In the United States, common carriers with interstate facilities are regulated by the Federal Communications Commission (FCC). FCC authority also includes amateur and commercial radio, television and citizens band radio.

1–3.1 Modulation and Multiplexing

The common carrier most frequently transmits information through wire conductors (wire pairs and coaxial cable) and microwaves. The receiver and transmitter are usually hooked to wire carrier systems which in turn are interconnected by a microwave link. The message or signal to be transmitted is the baseband signal. The baseband signal can be a speech signal produced by a telephone, a series of pulses produced by a teletype, or, for wider bandwidth systems, a video signal produced by a television camera. The baseband signal is put on a carrier in a process called modulation. The modulation process in some ways modifies an otherwise continuous sinusoidal carrier signal. The three basic forms of carrier modulations are

a. Amplitude modulation (AM)
b. Frequency modulation (FM)
c. Phase modulation (PM)

In amplitude modulation, the amplitude of the carrier varies in proportion to the amplitude of the baseband signal. In frequency and phase

modulation, the frequency and phase departures of the carrier vary linearly with the amplitude of the baseband signal. Since frequency and phase are not independent quantities, there is always a phase shift with a frequency shift and vice-versa. Figure 1-8 illustrates these modulation waveforms.

The modulation process permits the message to be more efficiently transmitted and, in addition, through FDM allows several messages to be transmitted simultaneously. This can be more clearly understood by considering the frequency spectra of the various modulation schemes when a baseband signal ranging from a frequency of f_1 to f_2 modulates the carrier with a rest or unmodulated frequency of f_c. A table illustrating the various frequency spectra, as well as the bandwidth required in order to permit transmission of the modulated signal, is given in Table 1-1. Variations on the three systems are also included.

As illustrated, the bandwidth (BW) occupied by the modulated signal depends upon the modulation scheme used. For AM, the bandwidth is twice the highest baseband frequency. For single sideband (SSB), the bandwidth is nominally equal to the baseband bandwidth. SSB suffers from poor low-frequency response since removal of one of the sidebands with real filters results in some erosion of the lower-frequency components due to the finite roll-off of the filter skirts. This will be discussed in more detail later when the low-frequency components in SSB generation

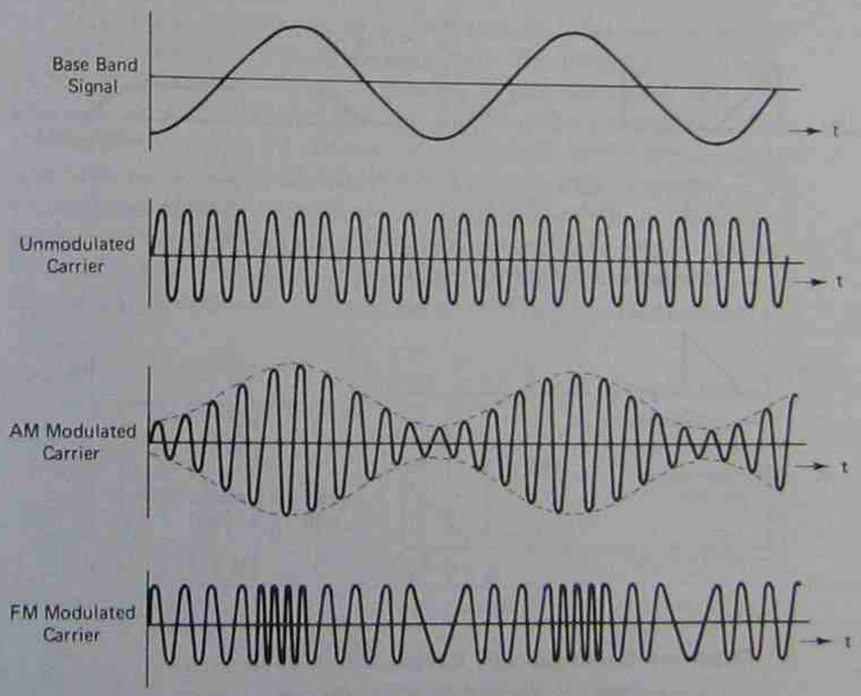


Fig. 1-8 AM and FM Carrier Frequency Waveforms

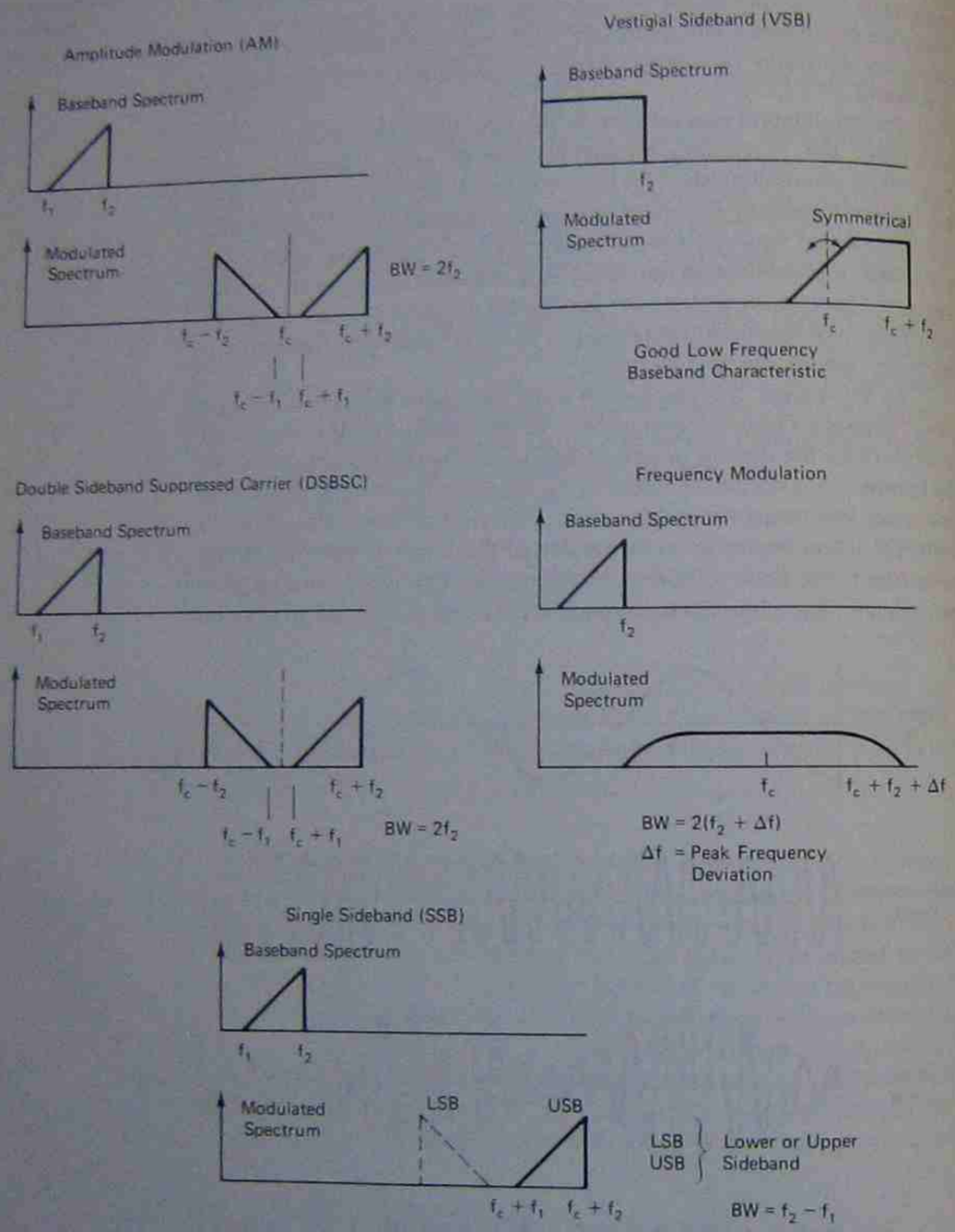


Table 1-1 Spectra of AM, DSBSC, SSB, VSB, and FM Modulation Schemes

are of significance, as in video and data transmission. The vestigial sideband can be used when a remnant of the removed sideband is retained.

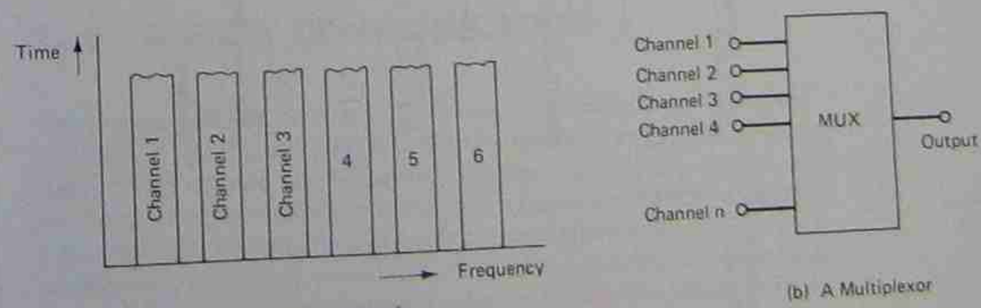
Frequency modulation is very inefficient in its use of spectrum space, but has a greatly improved performance in a noisy environment. The bandwidth occupied is approximated by $2(f_2 + \Delta f)$, where Δf represents the peak frequency shift of the carrier from its rest frequency. This occurs when the baseband signal is at its maximum amplitude.

In order to increase the number of channels that can be carried over a single cable pair, several channels can be stacked so that each channel occupies a unique portion of the frequency spectrum. This process is referred to as frequency division multiplexing and is graphically illustrated in Fig. 1-9. As a result, several channels can be simultaneously transmitted over one physical circuit, with each signal occupying a different frequency slot, or portion of the total available bandwidth.

To illustrate how telegraph signals, voice signals, and television signals can be simultaneously transported from one point to another, consider a simplified but not untypical FDM system as shown in Fig. 1-10.

Because telegraph machines transmit at very low speeds, the bandwidth requirement is minimal for each telegraph channel. OFF and ON pulses, transmitted at 75 or 150 b/s, must first be converted to a narrow band voice frequency signal by the VFCT block. These narrow band signals are then stacked to allow up to 18 telegraph channels on one voice frequency channel.

The voice channel can be stacked with 23 other voice frequency channels to obtain a grouping of 24 channels. This grouping can be transmitted by an N-carrier system. Arranging 25 or more such groups of 24 channels results in 600-plus channels which can be transmitted by the L carrier. Since a single TV channel requires a large bandwidth, television can take up a bandwidth equivalent to 1200 voice channels. A TV channel can readily be carried on an L carrier. The Bell L3 xoaxial cable system, for instance, can carry 660 voice channels and one TV channel. By further upward modulation in frequency, thousands of voice channels can be transmitted over a microwave or satellite link.



(a) Channel Allocation

(b) A Multiplexor

Fig. 1-9 Frequency Division Multiplexing (FDM)

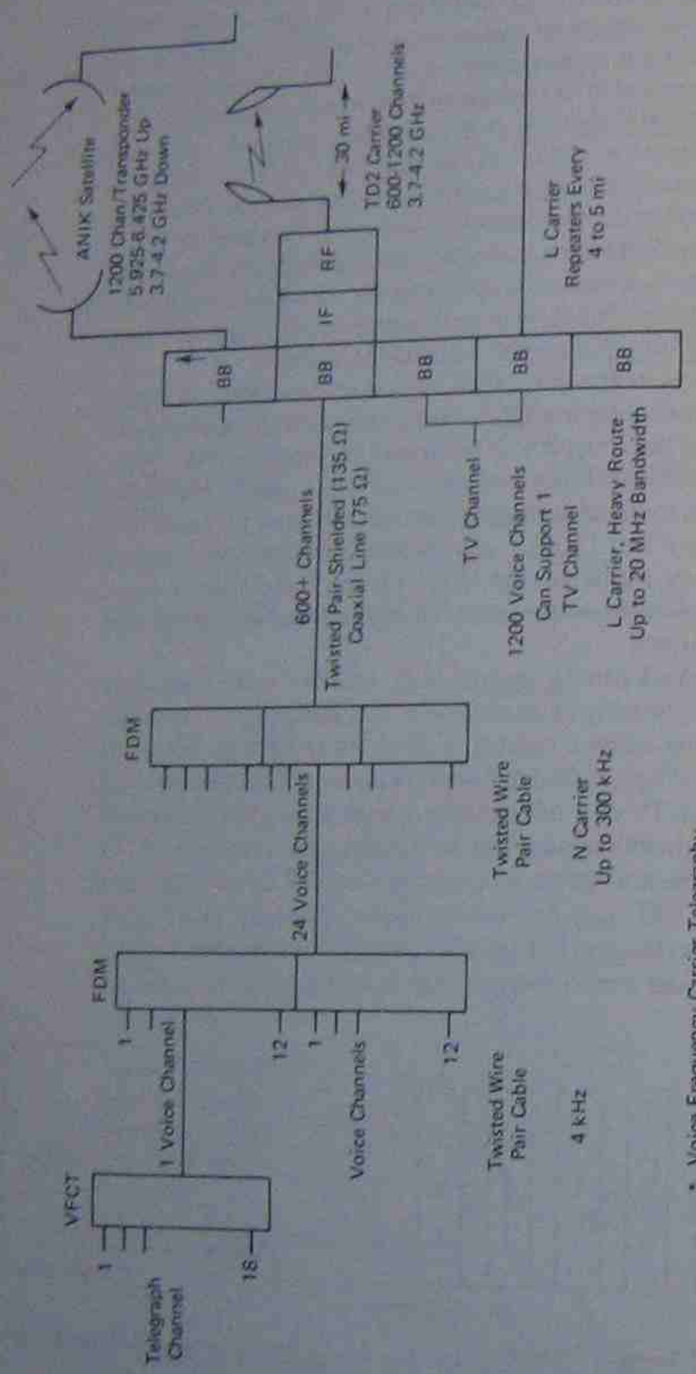


Fig. 1-10 Signal Transmission on a Telecommunication Facility

To recoup a single voice or telegraph channel, the process illustrated in Fig. 1-10 must be reversed.

1–3.2 Echo

Of one particular annoyance to a telephone user is the return of the talker's voice. This echo, or reflection of the voice, is a result of an impedance mismatch in the transmission system. The degree of annoyance experienced from an echo is a function of the loudness and delay time of the echo. For reasonably short delays, sufficient attenuation can be added to the circuit to the satisfaction of the talkers. Table 1-2 gives the approximate attenuation required to meet the echo tolerance of the average subscriber.

Since too much attenuation will cause excessive loss of the desired voice signal to the called subscriber, there is a limit on the amount of attenuation that can be inserted to reduce the echo to a tolerable level.

In North America, an echo suppressor should be used when the mean sound trip delay exceeds 45 ms. The CCITT sets the limit at 50 ms. The voice echo suppressor is a voice-activated attenuator which heavily reduces the reflected signal energy. As shown in Fig. 1-11, when A begins to talk, Echo Suppressor 1 is activated, effectively blocking the reflected signal. The echo suppressor is deactivated about 10 ms after A stops talking. B can override the echo suppressor by interrupting loudly.

Table 1-2 Tolerance to Talker Echo

One-way Propagation Time (ms)	Approximate Echo Path Loss for Average Subscriber (dB)
10	11
20	18
30	23
40	27
50	31

1–4 DIGITAL COMMUNICATION

The previous section dealt with the transmission of a continuously variable analog signal. Such a signal has an infinite number of possible levels since it can assume any value between the "rails." This section will briefly consider the digital transmission of a signal when discrete separate pulses having a finite number of levels are transmitted. Most commonly, a binary system of pulses is used. Only one of two conditions may exist at any particular time, as exemplified by the various nomenclatures in Table 1-3. The two conditions represented by the digits 1 and 0 are termed binary digits, or bits.

Table 1-3 Nomenclature of Binary States

Application	Boolean Nomenclature	
Binary numbers	1	0
Transistor	HI	LO
Signal	SPACE	MARK
Transistor	ON	OFF
Signal	Plus	Minus

1–4.1 The Binary, Octal, and Hexadecimal Numbering Systems

The base of the binary number system is 2, and the system consists of the two digits 0 and 1. The placement or position of the binary number within a sequence of digits determines the weighting that the bit represents. In general, the sequence $a_3 a_2 a_1 a_0 \cdot a_{-1} a_{-2} a_{-3}$, where a_n represents a 1 or a 0, has a weighting of:

2^0 for the a_0 bit
2^1 for the a_1 bit
2^2 for the a_2 bit
2^3 for the a_3 bit
and 2^{-1} for the a_{-1} bit
2^{-2} for the a_{-2} bit etc.

For any other numbering system, the base value of 2 changes to the new base. For the octal numbering system where the base is 8, the weightings change to 8^1, 8^2, and so on.

Example 1-5 Convert the binary number 1011.01 to its decimal equivalent.

Solution

$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-2}$
$= 8 + 2 + 1 + .25 = 11.25$

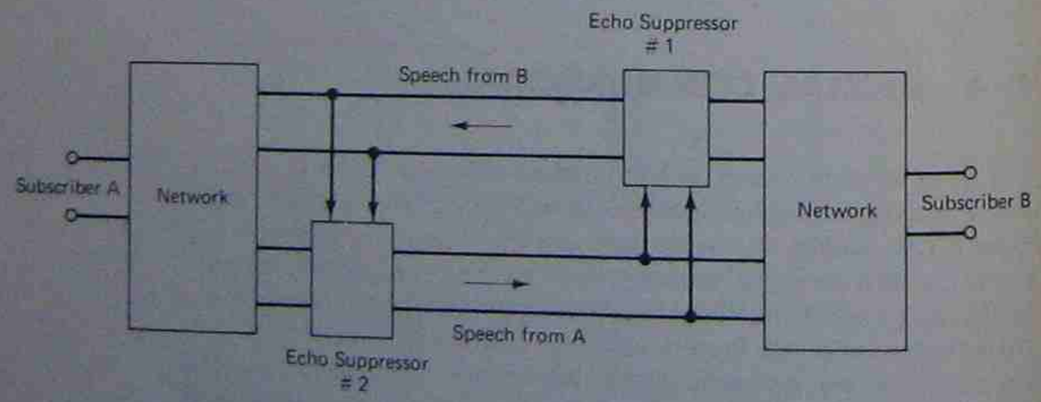


Fig. 1-11 Echo Suppression

Convert the octal number 652.7 to its decimal equivalent. **Example 1-6**

Solution

$6 \cdot 8^2 + 5 \cdot 8^1 + 2 \cdot 8^0 + 7 \cdot 8^{-1}$
$= 426.875$

Table 1-4 gives the decimal, binary, octal and hexadecimal numbers up to $20_{\text{base 10}}$.

Table 1-4 Binary, Octal, and Hexadecimal Numbering Systems

Decimal Base = 10	Binary Base = 2	Octal Base = 8	Hexadecimal Base = 16
0	0	0	0
1	1	1	1
2	10	2	2
3	11	3	3
4	100	4	4
5	101	5	5
6	110	6	6
7	111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F
16	10000	20	10
17	10001	21	11
18	10010	22	12
19	10011	23	13
20	10100	24	14

As can be observed from Table 1-4, a grouping of three binary digits can be represented by an octal number. Thus, to convert a binary number to its octal equivalent, break it up into groups of three digits starting at the base point and working in both directions. Leading zeros can be added or trailing zeros can be added after the base point to make a group of three.

Convert 10011.0111 to octal grouping. **Example 1-7**

Solution

10|011.011|1
adding leading or trailing zeros
010|011.011|100
octal equivalent 2 3 . 3 4

Similarly, hexadecimal numbers can be converted to binary and vice-versa by breaking the binary number up into groups of four.

Example 1-8 Convert 10011.0111 to hexadecimal grouping.

Solution

1|0011.0111

hexadecimal equivalent 1 3 . 7

Conversion between octal and hexadecimal requires the use of binary as an intermediary.

Example 1-9 Convert 372.3 octal to hexadecimal

Solution

3 7 2 . 3

Step 1 372.3 octal = 011|111|010.011

Step 2 0|1111|1010.0110 = FA.6

When dealing with microprocessors or computers it is much easier to read the binary numbers off in octal or hexadecimal form than to use binary throughout. The PDP-11 minicomputer, for instance, uses octal notation, and the 8080 microprocessor uses hexadecimal.

1–4.2 Encoding of an Analog Signal into a Digital Signal

An analog signal can be converted to a digital signal by converting the signal level to a binary equivalent. For example, a signal varying from 0 to 8 mV can be converted to a binary number using Table 1-4. All voltages between 0 and 1 can be represented by 000, all voltages between 1 and 2 by 001, and so on to a maximum of 111. In turn, the binary digits can represent the presence of a pulse in the case of a 1 digit and the presence of no pulse in the case of a 0 digit. This digital technique is known as pulse code modulation (PCM). By this technique only the transmission of pulses or the absence of pulses takes place. The amplitudes of the pulses are not critical, since only their presence or absence needs to be noted.

In analog transmission, any noise on the system is amplified along with the signal, and the noise accumulates as it passes through many amplifiers, or repeaters. In digital transmission, however, each amplifier or regenerator first acknowledges whether a 1 or a 0 is received and then retransmits a cleaned-up pulse. Thus, any noise on a pulse is removed and the distortion does not accumulate. Of course, if the noise is excessive, an error can occur in the detection of a pulse, and the error will continue as the pulse train moves down the system.

1–4.3 Time Division Multiplexing (TDM)

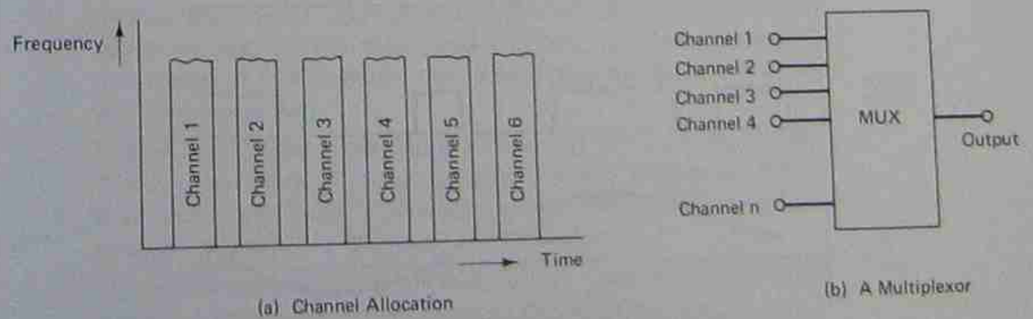
As illustrated in Fig. 1-9, frequency division multiplexing assigns each of several signals to a narrow bandwidth slot and transmits all channels simultaneously. In time division multiplexing, each of the several signals is transmitted sequentially.

At any instant, only one signal is transmitted. As illustrated in Fig. 1-12, Channel 1 occupies the output channel during the first time slot. During the second, Channel 2 occupies the entire bandwidth of the output channel. During the thirteenth time slot, Channel 1 may be transmitted again, and so on. Sampling must be done at a rate high enough to prevent losing any significant information of the input signal. For an analog signal, each signal must be sampled at least twice as fast as its highest significant frequency component. As a result, the bandwidth requirements of the higher speed circuits can become quite large.

1–4.4 A Digital AM System

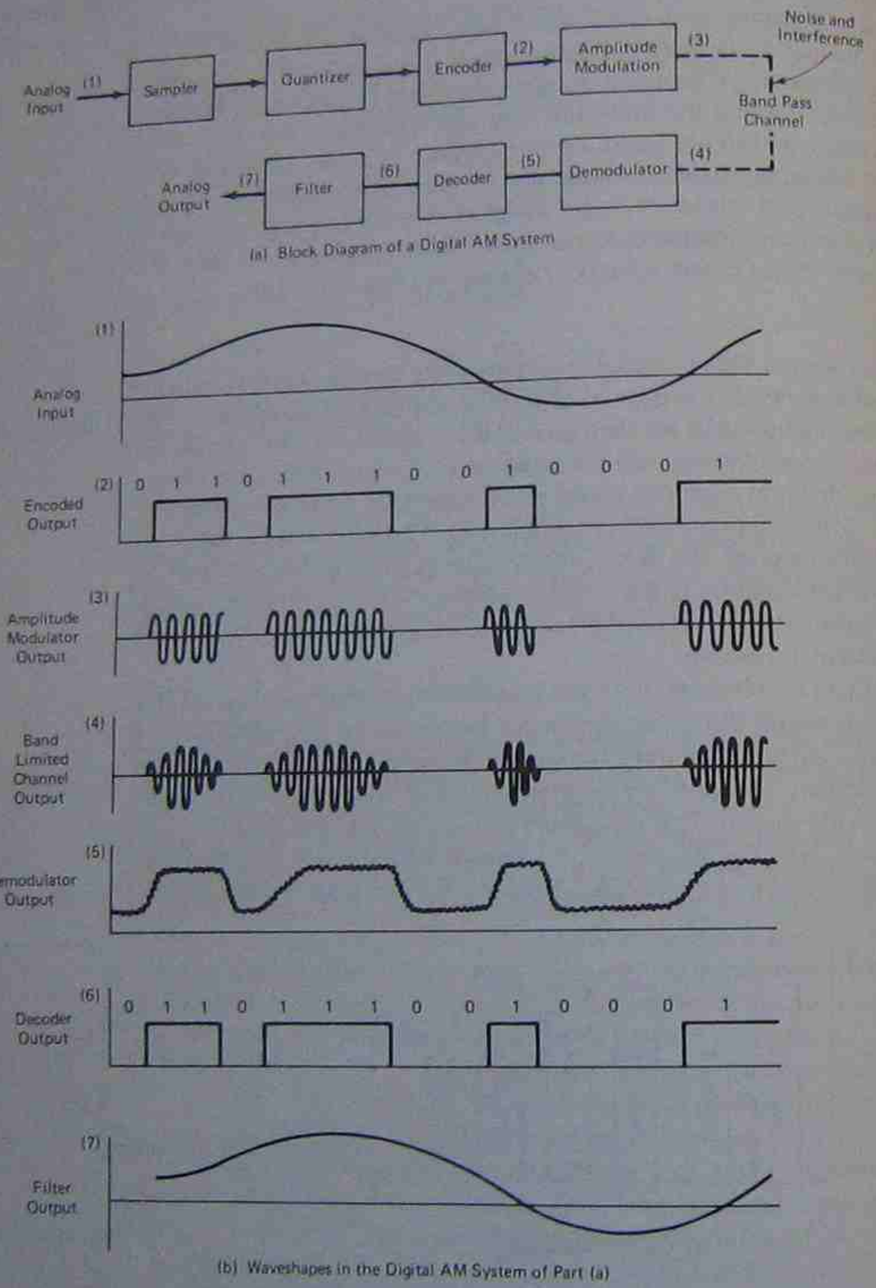
In a very simple digital modulation system the analog signal is initially sampled to generate a sequence of pulse amplitudes, as illustrated in Fig. 1-13. These amplitudes are then quantized, or represented by a level that most nearly corresponds with the actual level. These quantized levels are next encoded into a suitable digital signal which in turn amplitude modulates a carrier that can be transmitted over a bandlimited transmission path. At the receiver, the signal is demodulated and hopefully decoded to a good approximation of the original message. Poor filtering plus quantization noise, and interference entering the system can cause degradation of the original message.

As can be observed from the sketched waveshapes in Fig. 1-13(b), significant signal distortion can occur before errors are introduced. If errors are kept to a minimum, the overall performance of a system can be made virtually independent of the number of multisections or regenerators in the system. That is, if errors in a system are kept within reasonable limits over a given section, the overall performance of a link is almost independent of the distance between the terminals. Noise does not



(a) Channel Allocation (b) A Multiplexor

Fig. 1-12 Time Division Multiplexing (TDM)



(a) Block Diagram of a Digital AM System

(b) Waveshapes in the Digital AM System of Part (a)

Fig. 1-13 A Digital Amplitude Modulation System

accumulate since the regenerators at the end of each section clean up the signal. This nonaccumulation of noise makes digital systems more immune to noise than analog systems and gives them a great advantage in long-distance transmission, which requires many repeaters.

1–4.5 Considerations in the Usage of Digital Communications

Digital communications have several distinct advantages over analog transmission systems, but there are definite disadvantages to this system as well. Because of the heavy investments in analog facilities, it will be many years before digital becomes the major technique in communication systems. In fact, if it were not for the many analog earth stations already in existence, satellite communications would probably now be almost wholly digital. The following lists provide a good summary of the advantages and disadvantages of digital communication.

Advantages of Digital Communications

1. Transmission quality is almost independent of the distance between terminals. The error rate is virtually unaffected by distance due to the regeneration, or cleaning up, of the signal at the regenerators.
2. A mixture of traffic ranging from telephony and telegraphy to data and video information can be easily carried.
3. The capacity of certain existing transmission systems can be increased. For instance, the capacity of a single channel over a cable pair can be increased by employing digital multiplexing equipment whereby several signals are combined into a higher rate signal without any significant impairment. Similarly, the capacity of a satellite communication system can be increased by going to Time Division Multiple Access (TDMA). In either case, the intermodulation problems encountered in a frequency division multiplexed (FDM) system become non-existent, as only one channel in a digital system is being amplified at any particular time. Thus, the satellite amplifier can be heavily driven without the danger of introducing intermodulation noise. In the case of FDM, all channels are simultaneously transmitted, each taking up a limited amount of bandwidth space. In time division multiplexing (TDM), only one channel is transmitted at any particular time, but it takes up the entire bandwidth of the system.
4. Digital cable systems are more economical for 15 to 40 km distances, particularly if cable capacity is nearly exhausted. Rather than laying new cable or providing new ductways, it is more economical to convert to digital on the old cable.
5. Digital communications lends itself to such novel facilities as cryptography, storage, and other forms of digital processing.
6. Digital communications is more suitable for the newer types of transmission media such as light beams in optical fibers and circular

or helical waveguides operating around 30 to 100 GHz. By providing a great deal of redundancy and employing special processing techniques, digital transmission also makes it possible to communicate over long distances—from space vehicles, for example.

7. Digital signal characteristics are convenient for electronic switching in which groups of digits are selected to be switched in turn onto various highways. This is called packet switching.

Disadvantages of Digital Communications

1. One of the major disadvantages of a digital system is its large bandwidth requirement. An audio channel which may normally require an analog bandwidth of 4 kHz will require 64 kHz when transmitted by a PCM system. This can be reduced by signal processing and the employment of special modulation techniques but at the expense of increased costs and complexity.
2. Time division digital transmission is not compatible with the frequency division analog transmission in current use. Both cannot be carried simultaneously as can be done, for instance, with monaural and stereo FM. It is possible to transform analog into digital, and vice versa, by using converters. This is done on digital trunks that are presently in operation.

Because of the basic incompatibility, however, it is very difficult to gradually change over from analog to digital transmission. Unless digital systems become highly economical or demand becomes high, it will be some time before an international satellite digital system becomes viable. Over 100 nations cannot be expected to replace all the earth stations, which cost about $20 million each, with more efficient digital equipment.

In the future, the flow of traffic in communication systems can be expected to become increasingly more digital. Because of the heavy capital investment in the present analog system, the existing telephone network facilities will continue to be used as the main transmission medium. This limits the signal spectrum to the audio frequency range of the telephone channel, or 200 to 3400 Hz. Future discussions in this text will be restricted to this constraint when considering practical applications of digital techniques.

1–5 DATA COMMUNICATIONS

Data communications is primarily concerned with the movement of data from one location to another. The term *data* refers to alphabetical, numerical, or special purpose characters which, when appropriately grouped, constitute some word or message. This data can be taken from,

or delivered to, various types of equipment. It can be sent to some instrument, such as a temperature controller. It can come from a source document, such as an inventory or payroll or from a storage medium, such as paper or magnetic tape or punched cards. It can also come from some register within a computer. Pulses, or digital transmissions, are used in the transmission of such signals because most digital computers can directly handle them. By using communication facilities that can handle digital data, interconnected digital computers can directly process the data.

Data communications is rapidly expanding as more and more businesses and agencies employ computer-based communication systems. Typical business operations that use data communications are

1. reservation services; for example, airlines, trains, and the like
2. mail service
3. inventory control
4. industrial control
5. computer time-sharing
6. automatic utility metering
7. packet switching

1–5.1 Communication Codes

Many different codes are employed in communications to represent alpha-numerical characters and special characters, such as line feed, carriage return, space, and bell. This section will consider only the older, 5-channel baudot code (Fig. 1-14), which is used in telegraphy, or in teletypewriters and paper tapes.

By using a FIGS or LTRS shift key, a dual definition of each of the 5-bit combinations is represented. The shift code that was most recently used determines whether a character is upper or lower case.

Because 5 bits give 2^5 or 32 unique codes, with the use of the shift key the number of unique combinations increases to 64. Of these, 58 are used as character combinations. The remaining six are used as function codes: LTRS, FIGS, space, carriage return, line feed, and blank.

A mark is represented by a direct current pulse and a space by a no-current condition. Thus, the characters can be transmitted from one point to another over a single pair of conductors.

1–5.2 Parallel and Serial Data Transmission

A code character can be transmitted by the current pulse in either parallel or serial fashion. In parallel data transmission, each code element is transmitted simultaneously. For the 5-level baudot code, five pairs of lines must interconnect the receiver and transmitter, as illustrated in Fig. 1-15.

In serial data transmission, each code element is sent sequentially as shown in Fig. 1-16. Only one pair of wire conductors is required for connecting the receiver to the transmitter.

For long distance transmission, the serial method is preferred because the line costs are more reasonable. However, serial transmission is slower than parallel transmission since the several code elements must be transmitted in sequential order instead of simultaneously. Parallel transmission is usually employed when computers are located in the same proximity.

Character		Mark Position				
Lower Case	Upper Case					
A	–					
B	?					
C	:					
D	$					
E	3					
F	!					
G	&					
H	#					
I	8					
J	'					
K	(					
L	)					
M	.					
N	,					
O	9					
P	0					
Q	1					
R	4					
S	Bell					
T	5					
U	7					
V	;					
W	2					
X	/					
Y	6					
Z	"					
Letters (Shift to Lower Case)						
Figures (Shift to Upper Case)						
Space						
Carriage Return						
Line Feed						
Blank						

Fig. 1-14 Channel Baudot Code Format

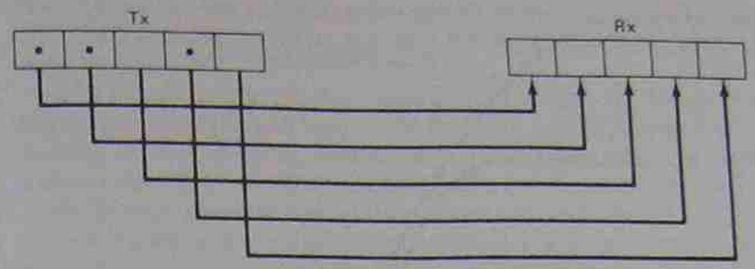


Fig. 1-15 Parallel Data Transmission

To give some idea of distances, which depend upon the signaling speed and error rate permitted with an EIA interface, typical speed versus distance curves are sketched in Fig. 1-17. Since the Electronics Industrial Association RS-232 specification only requires operation up to 50 ft (15 m), these rates violate this standard. However EIA interface cables are normally adequate for longer distances.

1–5.3 Asynchronous and Synchronous Transmission

Serial data transmission can be broken down into two separate classes of transmission: synchronous and asynchronous.

In asynchronous serial data transmission, each character has its own synchronizing information to inform the detector of the first and last bit location of the character and therefore can be sent at varying rates. Each information character is often synchronized by the use of stop-and-start elements, and the length of the time gap between characters is usually not fixed. Asynchronous transmission can be easily generated and detected by mechanical equipment. However, it is slower than synchronous transmission and is not as efficient in code transmission because of the added synchronizing stop-and-start elements.

As an example of asynchronous serial transmission, consider the transmission of the letter, or character, J in the 5-level baudot format shown in Fig. 1-18. The data bits of each character are preceded by a start bit and terminated by a stop interval. The stop bit interval is 1.42 times the length of the other intervals. Most electromechanical equipment still uses the 1.42, or 2-units stop code, but some of the newer equipment, such as the model 37 teletype, uses only the 1-unit stop bit.

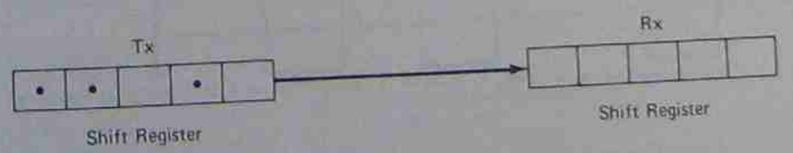


Fig. 1-16 Serial Data Transmission

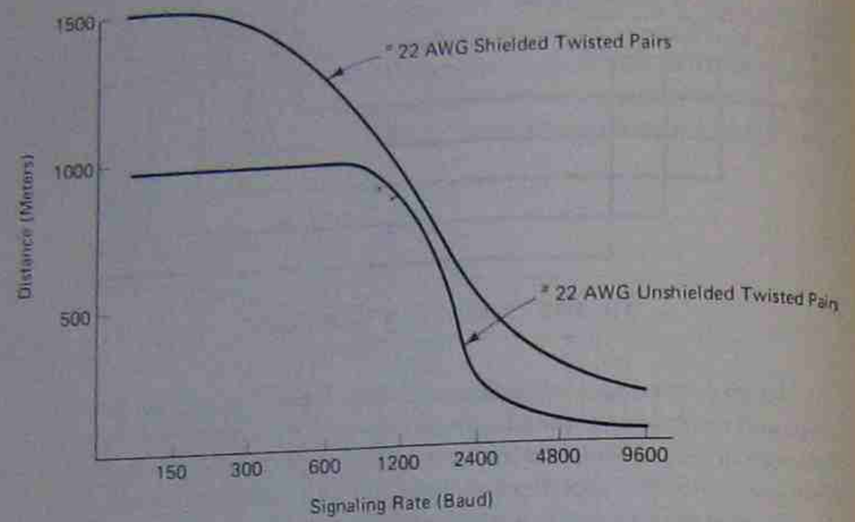


Fig. 1-17 Speed Versus Distance Performance Employing EIA Interface

In the idle condition the transmission holds in the 1 state, or MARK, position; as soon as a 0 state, or SPACE, is detected, the receiver begins to detect the incoming character. After 5 bits, a MARK, or stop condition, should be detected, and the receiver is ready for the next character. Because the start-and-stop bits take up a significant amount of time, the information throughput is less than it would be were they not present.

A block diagram of a typical asynchronous serial receiver is shown in Fig. 1-19. The purpose of the receiver logic is to assure that incoming pulses are sampled at or near each pulse center. This sampling tends to minimize any errors due to pulse distortion.

The clock, which is generally a crystal oscillator, is divided down to provide pulses at 8 times the input pulse rate. By initially setting the divide-by-8 counter to 4 (100), the overflow count, 8 (000), can be used to strobe the incoming pulses into the shift register character buffer. At the end of a character, the start pulse will overflow and set the character flag.

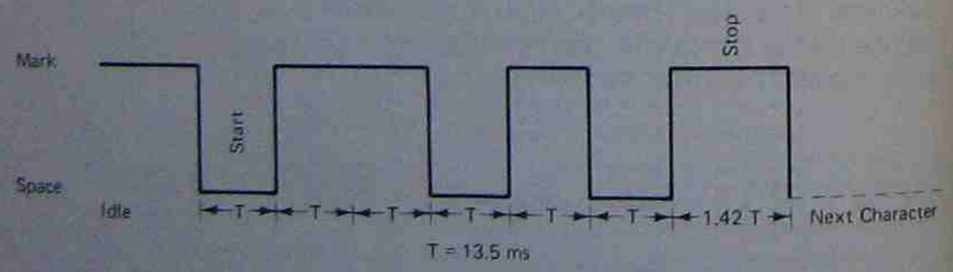
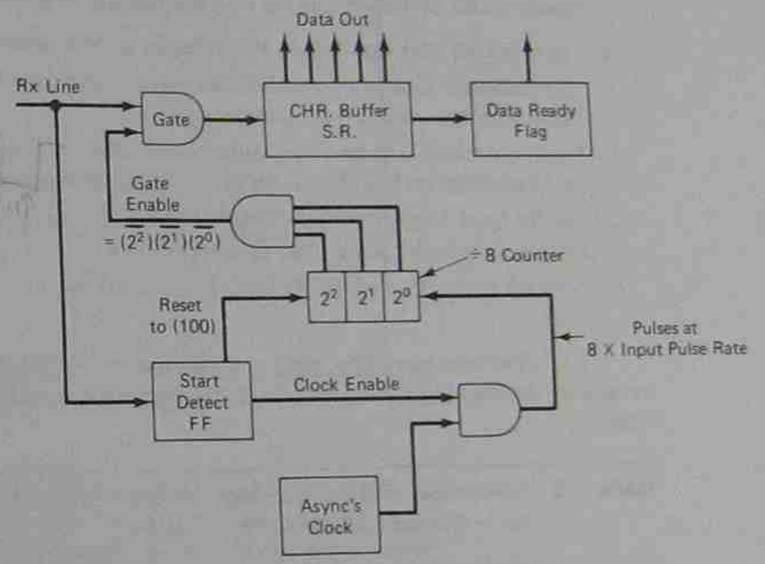


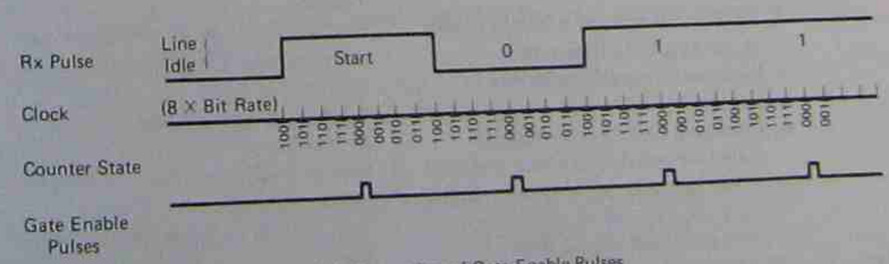
Fig. 1-18 Baudot Code for the Letter J

This flag can be used to reinitiate the system. At the beginning of a character, the incoming start pulse sets off the code detector flip-flop, which starts the clock.

In the above system, the "end-of-code" is detected by counting incoming pulses. This can result in the Rx falling out of step with the Tx if an extra bit, such as noise, gets into the system. In such a case, all the following information would be erroneous. The system should therefore include some pulse-width-testing logic which can separate the wide stop bit from the others and assure re-synchronization at the end of each signal element.



(a) Block Diagram of an Asynchronous Receiver



(b) Generation of Gate Enable Pulses

Fig. 1-19 A Serial Asynchronous Receiver

Synchronous transmission is not made up of start-and-stop pulses, but uses a common clock pulse at the transmitting and receiving end (and a few framing pulses) to achieve synchronization and character pulse identification. The receiver can recognize a unique code in the received bit stream which allows it to lock in on an incoming bit stream and to identify the characters. Synchronous transmission tends to have a larger net data throughput than asynchronous transmission since it does not require the special start-and-stop bits. For a 5-bit code, asychronous transmission requires from 6 to 7 bits per character whereas synchronous transmission requires slightly more than 5. Thus, asynchronous transmitters transmit about 33% fewer characters than synchronous transmitters for the same bit speed.

Synchronous operation can be characterized by the following:

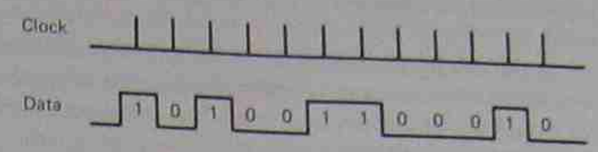
a. There are no start-stop bits to synchronize each character.
b. Information is sent in an entire block, which can consist of many signal elements with no separations.
c. The entire block is framed by unique codes which indicate to the Rx the beginning, end, and so forth of blocks of information.
d. The Rx must know the code length and be able to recognize the unique codes which are used to control the Rx.
e. Every bit in the Tx and the Rx must be synchronized to a common clock.

In synchronous operation, timing signals are provided in some form along with the data bits. These indicate to the Rx when it must sample the data.

Table 1-5 Advantages and Disadvantages of Synchronous and Asynchronous Serial Data Transmission

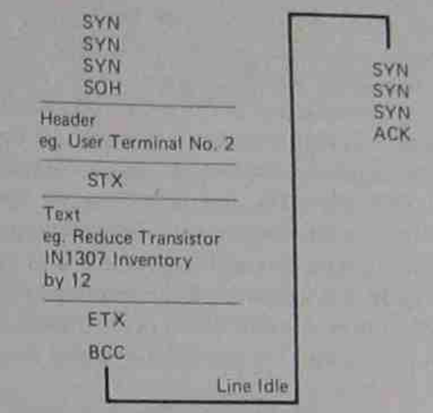
Asynchronous	Synchronous
1. Easily generated mechanically	1. High efficiency with no synchronizing bits required (block framing is required)
2. Can easily drive mechanical equipment	2. Not as sensitive to distortion because timing is included along with data
3. Suitable for manual operation, such as pecking on a typewriter	3. Bit loss can cause whole message to be erroneous
4. Requires two tracking clocks	4. Special codes must be detected for timing and logic control
5. Sensitive to distortion	5. Not suitable for mechanical equipment
6. Inefficient due to start-stop bits	6. Tends to be costly in terms of equipment because of the more complex circuit required
7. Speed limited due to above and due to the fact that pulse-width margin must be allowed to accommodate distortions and timing error	

For example:



(a)

(f) A typical message in BISYNC format may look like



(b)

where SYN is the synchronous idle control character
SOH is the start of header control character
STX is the start of text control character
ETX is the end of text control character
BCC is the block check character (error check) on the header, STX, Text and ETX characters.
ACK is the acknowledgement that the text was properly received.

The advantages and disadvantages of serial asynchronous and synchronous transmission are shown in Table 1-5.

1–5.4 USART, 20mA Loop and Modem

To connect a digital computer with a parallel *n*-bit output (where *n* is typically a multiple of 8) to a single communications line, a Universal Synchronous Asynchronous Receiver Transmitter (USART) is employed. Often, a UART is the only available option. This device either transforms the parallel *n*-bits into a serial bit stream or converts the serial bit stream into *n*-parallel bits. (The USART is often considered part-and-parcel of a

computer facility, as shown by the dotted enclosure in Fig. 1-20.) Some of the options available on a commercial LSI USART will be considered in a later chapter.

In order to directly transmit serial data over a limited distance of up to several hundred meters, sources capable of delivering around twenty milliamperes of current are traditionally used. This is known as a 20 mA loop. Many of the common teletypewriters, for instance, switch the 20 mA currents by sweeping a carbon brush over commutator bars. If at least 18 mA of current is present, the contacting surfaces are kept reasonably clean. Figure 1-20 illustrates, in block diagram form, the use of a 20 mA loop when connecting a UART to a teletype (TTY).

For much longer distances, when the telephone system could be most conveniently employed, the modem is used to transform the binary bit streams into a signal which conforms to the telephone voice bandwidth. Modem is a contraction of the terms modulator and demodulator. A modem thus permits digital signals to be sent over the telephone analog communications facilities. Modems are frequently referred to as data sets by telephone companies. The interface between the modem and the computer/terminal facility is internationally standardized. When connecting a modem to data processing equipment in North America, the EIA standard RS-232-C is followed. This conforms to the CCITT's Recommendation V.24, concerning definitions of the interchange circuits. These definitions will be discussed in much more detail in a later chapter.

1–5.5 Simplex, Half-Duplex, and Full-Duplex

Transmission between two modems can be characterized in one of three ways: simplex, half-duplex, or full-duplex. In North America, the terms have the following meanings:

Simplex operation indicates transmission in one direction only; that is, the modems either receive or transmit. Modems rarely operate in this mode.

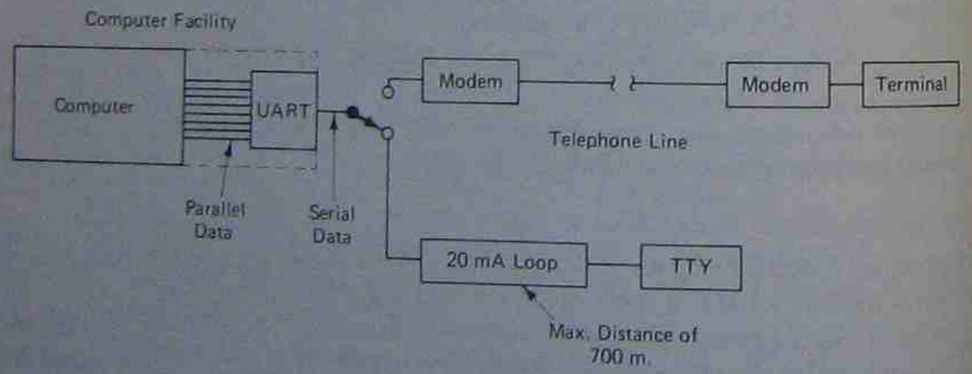


Fig. 1-20 Transmission of Data To and From a Computer To a Teletypewriter or Over a Telephone Line

Half-duplex operation indicates transmission in either direction, but only in one direction at a time. Modems may switch from transmit to receive and vice versa. This will result in some lost data transmission time due to turn-around time.

Full-duplex operation indicates transmission in both directions simultaneously.

These operations are graphically illustrated in Fig. 1-21. Data transmission requires the use of half- or full-duplex operation. After detecting a message, this technique allows the receiver to send a positive or negative acknowledgement back to the transmitter informing the transmitter whether an error-free or an incorrect message was detected. Errors in the detected message are generally unacceptable. In such situations, the negative acknowledgement is usually interpreted as a request for retransmission of the block that was received in error.

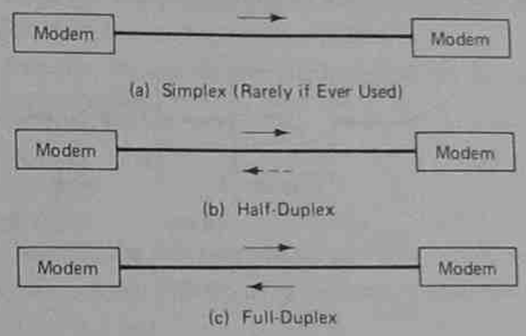
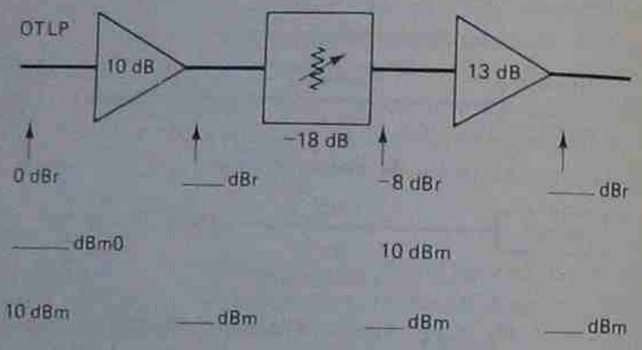


Fig. 1-21 Modes of Modem Operation

PROBLEMS

1. a. State three purposes for modulation.
   b. Describe the three basic forms of carrier modulation; that is, AM, FM and PM.
2. Calculate the physical length of a half wavelength antenna ($\lambda/2$ dipole) at frequencies of 1 MHz, 100 MHz and 10 GHz.
3. a. Calculate the power and voltage gains of a +3 dB and a +10 dB amplifier.
   b. If the power level at the input of a 13 dB attenuator is 2 mW, what is the power level at the attenuator output? Convert the input and output power levels to dBm.

4. Individually, two noise sources produce 6 dBm and 3 dBm of noise. What will the accumulative noise level be in dBm?
5. a. Describe the frequency characteristics of human speech.
   b. Give the rationale for employing weighting networks when monitoring noise levels.
6. a. What is the level in dBm of a 1000 Hz 0 dBrnc tone? Of a 90 dBrnc tone?
   b. What is the noise level in dBrnc of a 0 dBm, 0–3 kHz white noise signal?
7. a. What is the zero transmission level point in a system?
   b. What is the definition of dBm0?
   c. What is the level in dBm0, of a +35 dBm tone at 18 dB relative transmission level point?
   d. What is the definition of dBm0p?
8. Fill in the blanks for the circuit shown.



9. a. Complete the following table:

$$dBm = 10 \log \frac{P(watt)}{10^{-3}}$$

$$dBrn = 10 \log \frac{P(watt)}{10^{-12}}$$

Power (watts)	dBm	dBrn
10^{1}	40	130
10^{0}		
10^{-3}		
10^{-6}		
10^{-12}		

   b. If the power represents white noise levels over a 300 to 3.4 kHz bandwidth, what are the dBrnc levels in (a)?
10. If the output level of a repeater is at −5 dBr, what is the noise level at the zero transmission level point if the noise from the system is −60 dBm?
11. a. Convert 50,000 pW0p to dBm0p.
    b. Convert 2000 pW0p to dBm0P.
12. What is a volume unit, and when does it represent true power?
13. a. State the Sampling Theorem.
    b. What is the minimum sampling rate required to retain all the information within the baseband of 300 to 3.3 kHz?
14. a. Why are international standards in the telecommunications industry desirable? Are there undesirable aspects as well? What are they?
    b. What are the functions of the CCITT and EIA organizations? What do these mnemonics represent?
15. a. Describe the bandwidth employed by the following modulation schemes:
        1) AM
        2) SSB
        3) VSB
        4) FM
    b. Describe the difference between TDM and FDM.
16. State the two ways of reducing the echo problem.
17. a. What is a bit?
    b. Convert 1101.101 to its decimal equivalent.
    c. Convert 63 octal to its binary equivalent.
    d. The flag bits in the High Level Data Link Control (HDLC) are given by 01111110. Convert this to hexadecimal.
18. Explain why overall system performance in a properly designed digital system is almost independent of the number of regenerator stations.
19. Draw a simple block diagram of a digital SSB modulation system.
20. List the advantages and disadvantages of converting an analog communications system to a digital communications system.
21. a. What is data communications?
    b. How many unique combinations can a seven-element code have?
22. a. Compare the speed of parallel and serial transmission.
    b. Based on the answer to (a), why is serial transmission employed?

23. Describe the difference between synchronous and asynchronous transmission.
24. a. What is the function of a modem?
    b. What is the function of a USART?
    c. What is a 20 mA loop?
25. Describe simplex, half-duplex, and full-duplex transmission.

2

The Telephone Network

2–1 INTRODUCTION

Although special data transmission networks for carrying high speed data are gradually being installed, it will be some time before computer users are not dependent upon the existing telephone networks. This chapter gives a general view of the existing network, which was primarily designed to carry voice traffic, and outlines its limitations in transmitting data.

2–2 THE TELEPHONE NETWORK

Connecting several telephones could conceivably involve directly linking every telephone to each of the others. No switching would be required, and there would be no blocking of calls because of a busy line. The disadvantage with such an interconnecting scheme would be the extraordinarily large number of lines required for the number of telephones in existence.

For example, Fig. 2-1 shows that 10 lines would be required for the direct interconnection of 5 telephones, and 15 lines would be required to connect 6 telephones. In general, for n telephones, $n + n - 1 + n - 2 \cdots + 1 = n(n-1)/2$ lines are required. In other words, for a typical community with 10,000 telephones, almost 50 million lines would be required. These numbers become all the more unrealistic when considering the direct interconnection of all the telephones in the world.

In consequence of the large number of direct lines that would be required, switching networks have been devised. The design of switching systems takes into account the fact that not all telephones will be in demand at any particular time. Therefore, less than a full complement of switches is required. In peak demand periods, such as on Christmas and

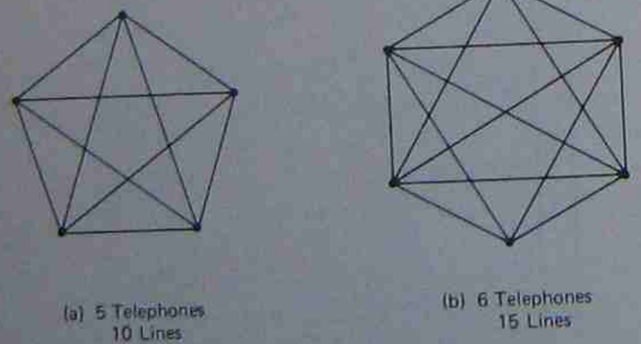


Fig. 2-1 Direct Interconnection of Telephones

Mother's Day, switching facilities become inadequate, and some blocking occurs. This problem can be reduced by employing alternative routing plans. For such a purpose, a hierarchy of switching offices is set up.

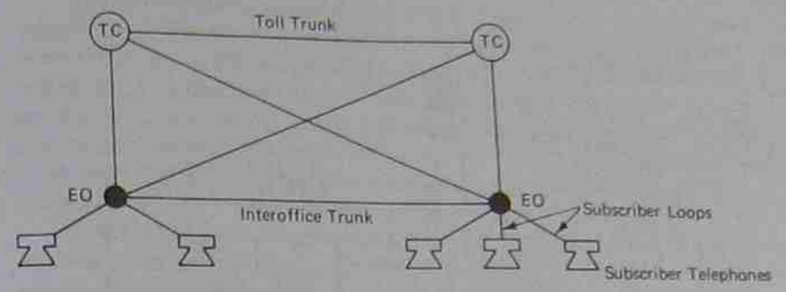


Fig. 2-2 A Two-Level Switched Network

2–2.1 Hierarchy of Exchanges

First, consider a very elementary but typical system with only two levels of switching: an End Office (EO), or local central office, and a Toll Center (TC). As illustrated in Fig. 2-2, a call on such a switching system could take one of several paths. It might be routed from one central office directly to the other central office, which would be the shortest or preferred route. If that line or trunk was busy or not available, it could go to a higher level office, or the toll center(s).

By employing wide-band trunk lines that accommodate hundreds of voice channels with suitable switching stations, a large number of subscribers can be interconnected. The drawback is the possibility that blocking will occur.

In the telephone network, three more toll centers are employed for a total of four levels of toll offices. In North America these are identified as Class 1, 2, 3, and 4 offices. Figure 2-3 illustrates this hierarchy.

Figure 2-4 indicates the locations of the higher control switching points in the United States and the southern part of Canada. Note that Canada has only two Class 1 offices, one in Montreal and one in Regina.

The longest route that any signal can take is to go all the way up and back down the hierarchy. Some of the many possible alternative routes

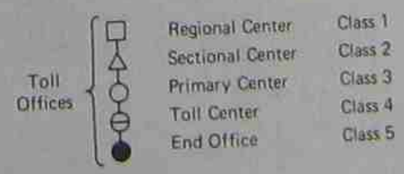


Fig. 2-3 Switching Center Hierarchy

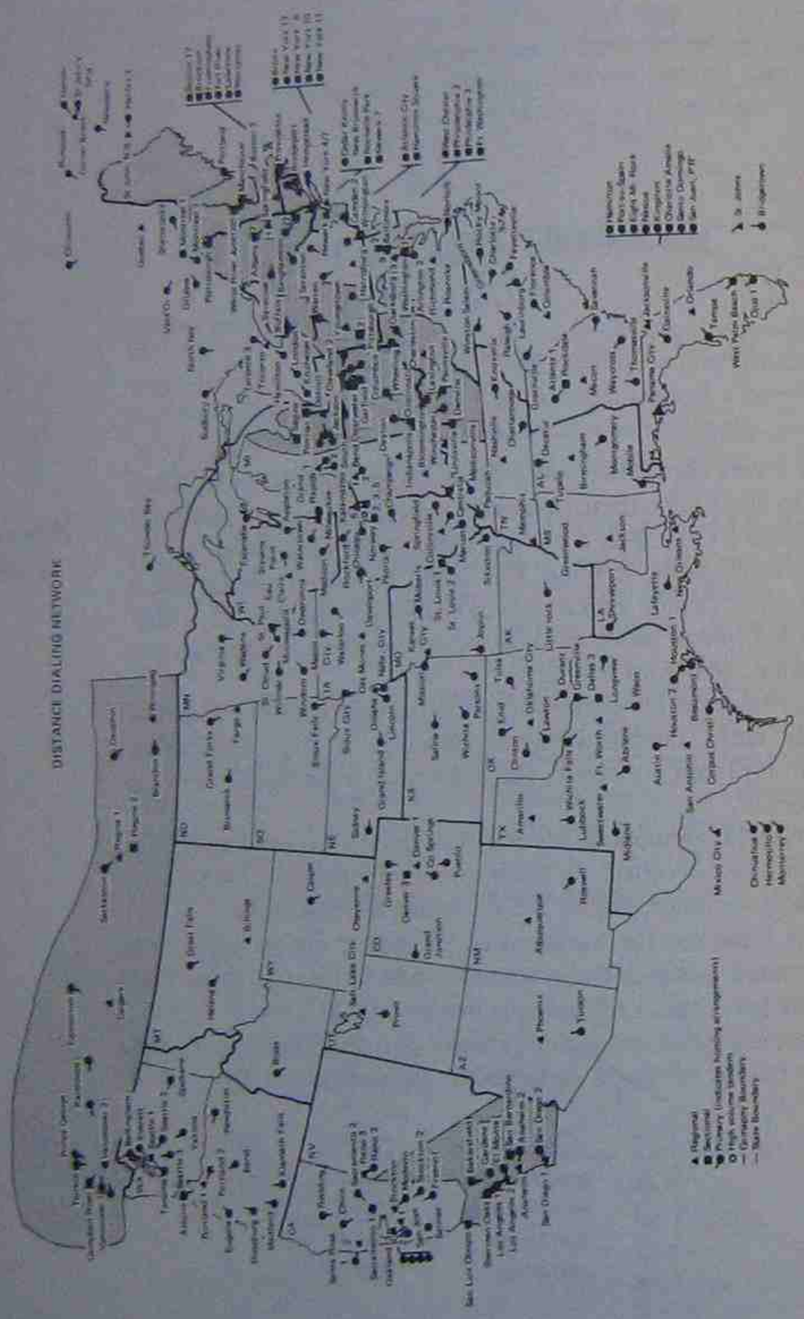


Fig. 2-4 Control Switching Points in the United States. (Reproduced with permission of A.T. & T. Co.)

are indicated in Fig. 2-5 by the dotted trunk lines. The actual choice of path taken at any particular switching office is determined by the availability of the various trunks. It is generally desirable for the final route taken to be the shortest one.

The tandem office shown in Fig. 2-5 is employed only in localized regions where there are a large number of nearby end offices to be interconnected with no toll charges. Rather than run a large number of interoffice trunks, a tandem switching office is incorporated into the network. A private switching center, called a Private Branch Exchange (PBX), can also be connected to the local end office. In this case, an operator using cords with jack plugs or an automatic switch connects the various telephones within a building complex to the local trunk line or lines.

Each subscriber's telephone is connected to the end office by a pair of wires enclosed by a sheath containing many other wire pairs. In addition to signaling (dial pulses, ON or OFF HOOK, etc.) and message transmission, the local loop must also carry the necessary dc power for the operation of each telephone instrument since it is uneconomical to service and replace local batteries. The trunk circuits tend to be of heavier gauge in order to carry the greater transmission and signaling requirements. All long distance trunks are "four-wire," consisting of two "two-wire circuits," one for each direction. These high capacity circuits also contain many repeaters, or amplifiers, to make up for the transmission loss of the signals.

In North America as well as other parts of the world, each end office can handle up to 10,000 telephones. Thus, each telephone can be uniquely identified by a 4-digit code, or the last four digits of a telephone number. For long distance dialing, three more digits are added to identify the end office. This allows for 1000 end offices (10 × 10 × 10). For international calls, even more prefixes are added. To reach Boston from Canada, for instance, a subscriber would dial

1	617	458	6132
Station to Station	Area Code	End Office Code	Subscriber's Number

For overseas calls, an international access code, 011, is dialed first. The area code and end office code are then dialed, using a numbering plan dependent upon the country. To reach Bonn, West Germany from Canada, a subscriber would dial

011	49	2221	613243
Station to Station	(Germany) Country Code	Bonn	Subscriber's Number

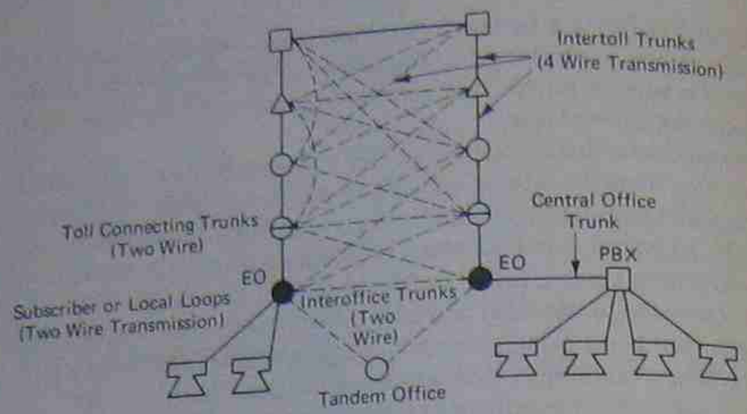


Fig. 2-5 Typical Routing Plans Available to a Call

2–3 VOICE BANDWIDTH

Information transmitted over the telephone system can be in the form of either analog or digital input signals. That is, the originating signal can either vary over a continuous range of amplitudes or rest at definite discrete amplitudes. Since telephone companies insist that the range of frequencies remain within approximately 300 to 3400 Hz, analog carrier techniques have been almost exclusively used. Because of the heavy capital investment in such equipment, it will remain in use for many years to come.

Most of the information of the human voice is concentrated in the 300 to 3400 Hz range of frequency (see Fig. 1-5), and it has been found that a person's voice is recognizable and intelligible if the channel bandwidth is limited to this frequency band. A typical frequency response curve of a telephone circuit is shown in Fig. 2-6.

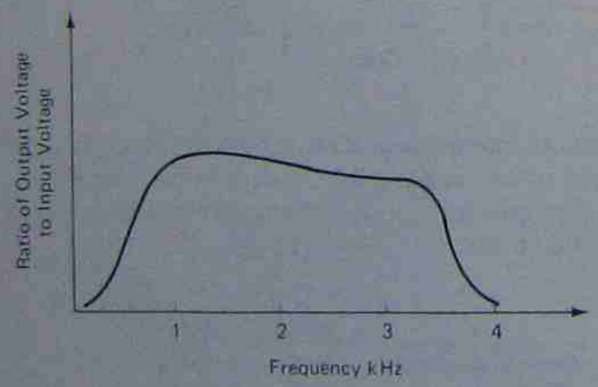


Fig. 2-6 Frequency Response of a Telephone Circuit

Because of the vastness of the telephone network in both size and structure, it is useful to break it down into more workable components. Although there is much interdependence, the network can be divided into the following three major divisions:

1. Instruments: These are the devices which the subscriber attaches to the local loop for transmitting and receiving signals. The only instrument considered in this chapter is the telephone handset. The modem will be dealt with in Chap. 3.
2. Loop and Trunk Circuits: The loop circuit consists of the wire pairs that connect the subscriber to the local switching office. Unless a subscriber is on a party line, each subscriber has his own pair of wires. The trunk circuits interconnect the switching centers. Each trunk can carry many voice channels and can consist of such widely varying transmission media as wire pairs, coaxial lines, and microwaves.
3. Switching Centers: The facilities which interconnect the various telephones.

2–4 INSTRUMENTS

The telephone handset is by far the most familiar appliance in the transmission path. It is a moderately complex piece of equipment and must perform the following functions:

a. Informing the subscriber of an incoming call by ringing.
b. Informing the local office that a call has been originated, answered, or completed.
c. Relaying to the local office the number that has been called.
d. Converting sound waves to electrical waves and vice versa.
e. Providing the proper amount of sidetone, or the sound in the speaker's ear due to his own voice, to the receiver. A slight amount of sidetone is desirable since without it the telephone sounds dead. Too much sidetone causes the speaker to lower his voice, produces ear fatigue, and reduces the intelligibility of the signal due to the reduced signal-to-noise ratio.

The telephone is most commonly connected to the end office through a single pair of wires called the subscriber's loop. Current is supplied to the telephone over the loop pair.

2–4.1 The 500-Type Telephone

The schematic diagram of a model 500-type telephone set (Fig. 2-7) will permit a greater appreciation of the operation of a telephone.

The telephone transmitter (microphone) detects the pressure variations due to sound and transforms them into an electrical signal by causing carbon granules to undergo a resistance change in response to the variations in the applied pressure. The variations in the current from the dc power supply, which represent the speech signal, are then transmitted to the local office. In the telephone receiver, a diaphragm is mechanically coupled to an armature which is driven by an electromagnet. The received current variations applied to the electromagnet cause the armature and the diaphragm to vibrate. Audible sound is thereby produced in the receiver.

The varistors V_{11} and V_{22} in Fig. 2-7 are non-linear silicon carbide devices whose resistance decreases with increased current. This tends to provide automatic regulation of the signal. Because loop lengths between the end office and each subscriber may differ, the transmission loss will

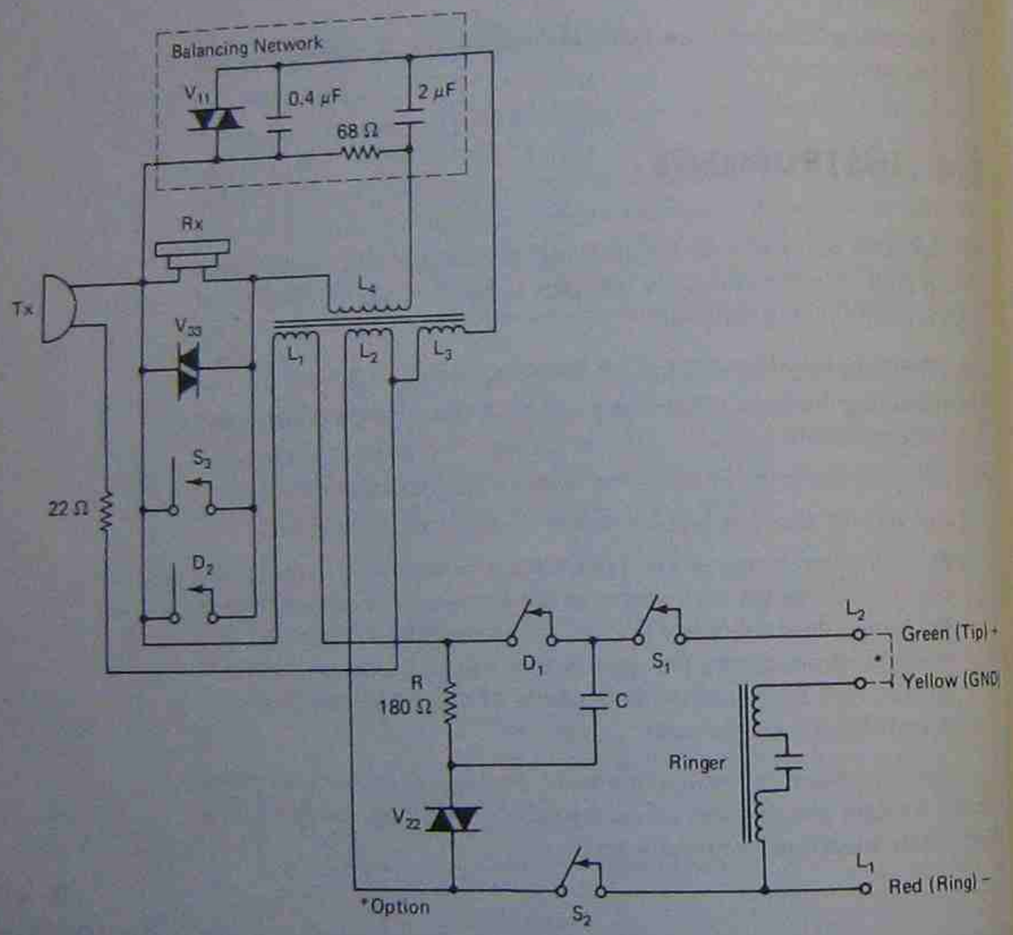


Fig. 2-7 500-Type Telephone Set

also vary. To maintain almost constant transmit and receive levels, these varistors tend to reduce the signal level on short loop lengths by producing a larger shunt loss at the telephone terminals. When the loops are long, the lower line current will cause the varistors to have greater resistances and therefore less shunt loss. The impedance of this type of telephone normally varies between 600 and 900 Ω, depending upon the dc line current. The varistor V_{33} is inserted to suppress dial pulse clicks in the receiver.

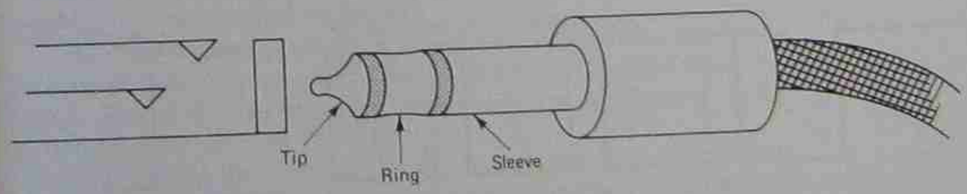
In order to keep the sidetone to a reasonable level, (or to keep the signal in the receiver while transmitting), a hybrid arrangement is used. It has the ability to divert power from the transmitter into the line and not into the receiver, and at the same time is still able to receive power from the line to the receiver. The hybrid consists of the inductor coils L_1 to L_4 and the balancing network. While transmitting, the voltage induced in L_4 cancels most of the voltage across the balancing network. The small voltage that remains at the receiver is used as sidetone. Capacitor C and resistor R suppress the high frequency interference components of the dial pulses to nearby radio receivers.

Usually, the L_2 lead is at dc ground, which causes a negative dc voltage with respect to ground to exist on the L_1 lead. The L_1 and L_2 terminals, commonly known as the Ring (R) and Tip (T) leads, form the connection to the local subscriber loop.

The jack and plug of Fig. 2-8 illustrate the rationale for the T and R (and sleeve S) designations.

By applying a 20 Hz, 85 to 100 V ac signal to the T and R leads, the telephone rings. The current drawn while the ringing occurs can be monitored by the telephone company to determine the number of telephones connected to a line.

Going OFF HOOK (lifting the receiver off its cradle) causes the switches S_1 and S_2 to close and S_3 to open. The normal open-circuited line voltage of 48 V drops to about 5 V due to the line impedance experienced by the dc current now flowing. The telephone is ready to receive or transmit a message (the carbon microphone now experiences a dc bias), or to dial a telephone number. The latter is performed by interrupting the dial pulse switch (rotary dialing) at a 10 Hz rate. For instance, when the



TRS Designation

Fig. 2-8 TRS Designation

digit 5 is dialed, the contacts D_1 for the dial pulse open and close 5 times to interrupt the dc current while the contacts D_2 for the dial shunt stay closed to prevent the dial clicking pulses from being heard in the receiver. After the dial returns to its resting position, the contacts for the dial shunt open. A typical dial pulse train, representing a dialed digit "4", is shown in Fig. 2-9.

The time delay between digits, called the interdigit time, must be of sufficient duration to allow the end office equipment to determine when the dialed digit is complete. In the step-by-step (S × S) office, where selecting switches are directly and sequentially controlled by the dialed pulses, about 600 ms of interdigit time is required. In common control offices, where the digits dialed are stored in a register for a brief period of time, the minimum interdigit time is cut in half to about 300 ms.

The bridged capacitor circuit at the central office shown in Fig. 2-10(a) supplies pattern current to the subscriber loop through the high impedance relay windings to prevent grounding of the voice frequencies. The capacitor C permits voice signals and ringing current to transfer through the network while blocking the dc current. Repeating coils (Fig. 2-10(b)) can also be employed to perform the same functions. These use magnetically coupled transformers to pass the voice and ringing signals through the network and also provide dc to the subscriber loop. The ballast resistors limit the current flow in short loops. Capacitor C bypasses the voice signal around the relay. The relay or dial pulse detector in the central office repeats the signaling or dialing pulses and passes them on to the switching machine.

The break intervals of dialed pulses exceed the make intervals because of the initial charging of the bridged ringer capacitors (C) in the central office when the dialed contacts open. The charging current causes a slight delay in the release of the pulsing relay in the central office. The break-to-make ratio is in the order of 60/40.

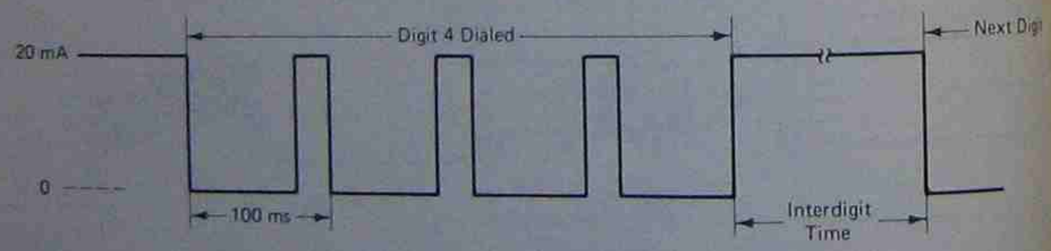
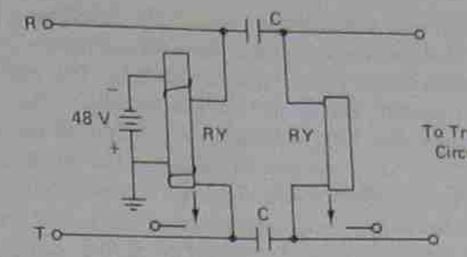
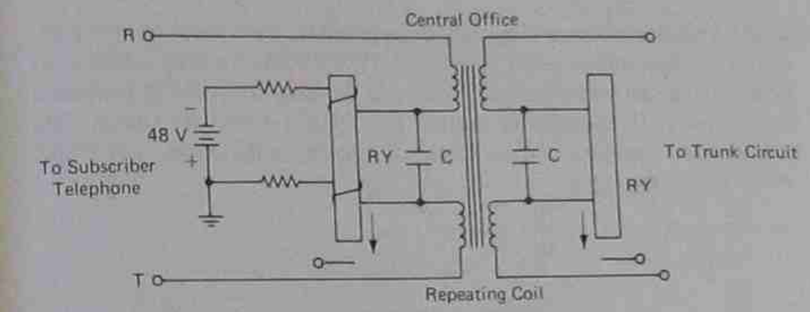


Fig. 2-9 Pulse Train When a "4" is Dialed from a Rotary Dial Telephone



(a) Capacitor Bridge



(b) A Central Office Repeating Coil

Fig. 2-10 Transmission Bridges

2–4.2 Touch-Tone Telephone

The pushbutton mechanism is another increasingly popular method of signaling. This method involves sending a pair of audible frequencies down the line—a different pair of frequencies representing each digit dialed. Known as two-tone or multi-tone signaling and touch-tone dialing, this technique can rapidly transmit a sequence of digits. If the central office employs dial pulse equipment, these touch-tone digits are stored temporarily in a register and then outpulsed sequentially after being decoded to dc pulses. A common central office can accept either dialed pulses or tone signals, converting either into a signal which can be used by common control equipment.

Of the frequency pairs assigned for each digit, one comes from a high-band frequency group and the other from a low-band frequency group. The dual-tone frequency assignments are determined by the switching cross points of Fig. 2-11.

Seizure and line disconnect still remain under dc loop control by the hook-switch. Touch-tone signaling is about ten times faster than dial pulsing and is less prone to error since there is little distortion of voice frequencies even on the longer subscriber lines. The high resistance and capacitance of the longer local loops severely distort the dc pulsing.

Figure 2-12 shows a schematic of a 182E touch-tone telephone. When one of the pushbuttons is pressed, the intersecting row and column make common contact to selected inductor windings. This establishes the correct tuned, or resonant, circuit for the pressed digit. The tone signal begins to appear only after the pushbutton is pressed all the way down and starts operating the COM switch. At this time, the negative supply is applied to the transistor collector which allows a current to flow through the transistor, thus causing a shock excitation of the tuned circuits. The COM switch also inserts a 5100 Ω resistance in the receiver circuit, causing the subscriber to hear a heavily attenuated tone signal.

2–4.3 Touch-Tone Decoders

Touch-tone decoders can be readily designed for tone and frequency decoding. A complete receiver such as a MITEL CM8828 is also available as a package. It can be connected directly to the telephone line to provide a CMOS- and TTL-compatible output. It is ideally suited for use in data transmission, mobile radio and control system applications, and PABX central office signaling.

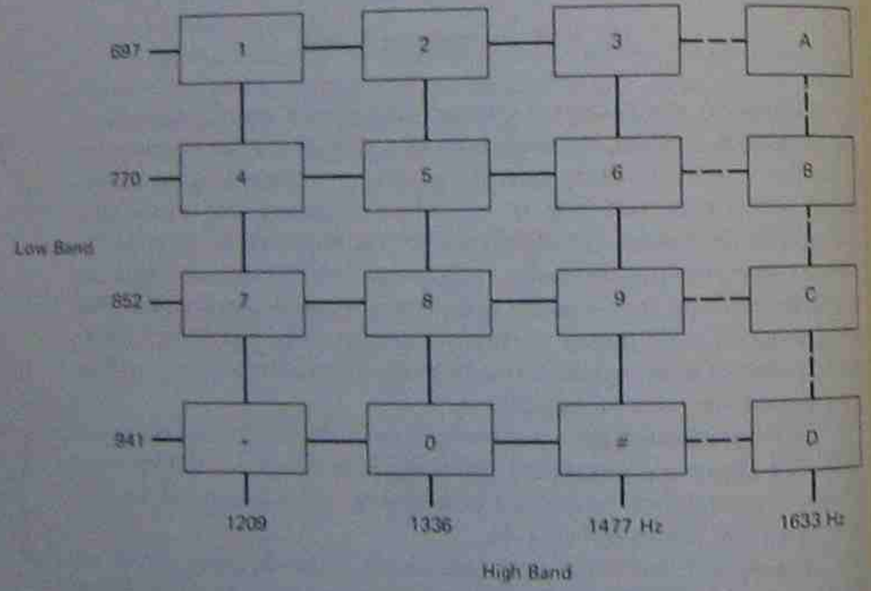


Fig. 2-11 Touch-Tone Dialing Frequency Assignments. The dashed tones are reserved for special purposes such as military and security.

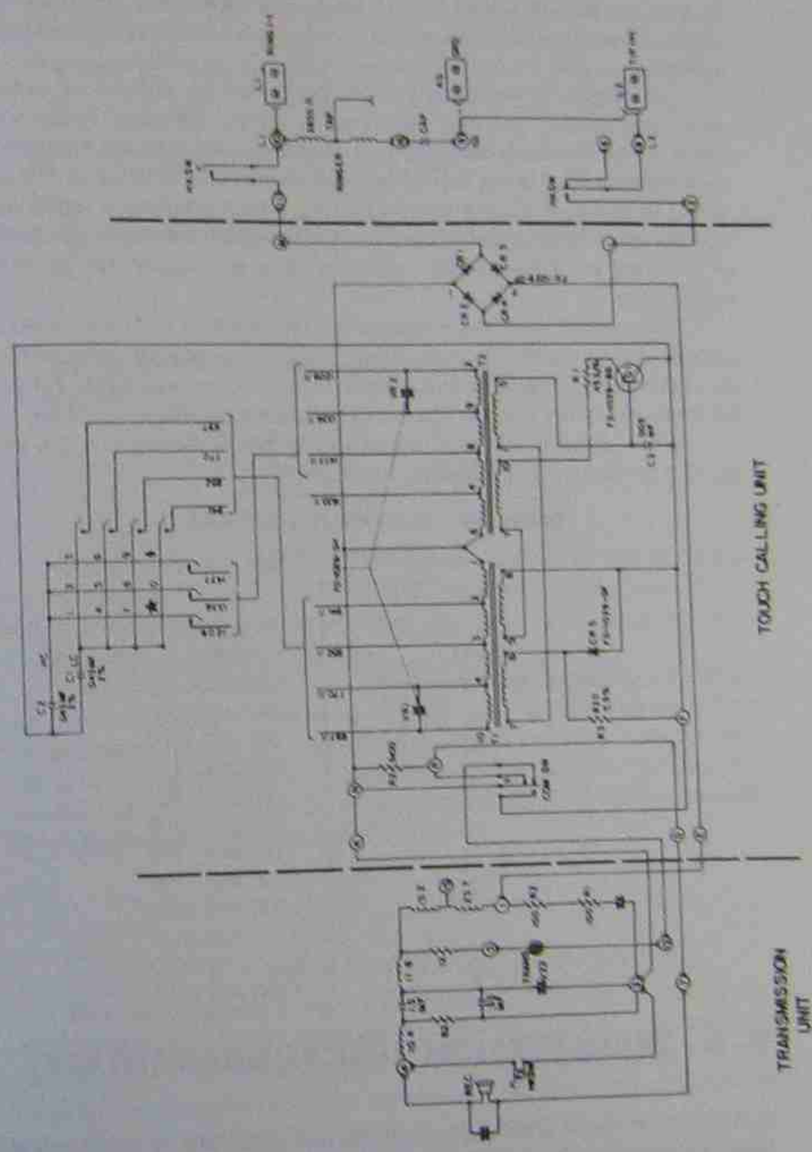


Fig. 2-12 Schematic Diagram for 182E Telephone with Touch Calling. (Courtesy GTE Automatic Electric Incorporated)

2–5 OFFICE-TO-OFFICE SIGNALING

In North America, signaling is relayed from office-to-office in an independent fashion. In such types of signaling, Office A sends routing information to Office B which in turn sends this information, possibly in a different form, to C and so on. This technique is known as point-to-point signaling.

Point-to-point signaling can be performed by in-band or out-of-band frequencies. In-band signaling, in which signaling frequencies occur within the voice bandwidth, is in general use in North America. It runs the danger of being fooled by loud sounds which occur at the signaling frequencies. Out-of-band signaling, which employs a signal outside the voice bandwidth (at 3825 Hz, for example), has the disadvantage of requiring extra bandwidth. However, it is not susceptible to voice interference.

Although dc signaling similar to that described in Sec. 2-4.1 is occasionally used in inter-office signaling, multi-frequency signaling, which uses two out of six possible frequencies, is more common. Table 2-1 lists the frequency pairs used in North America for inter-office signaling.

When a toll channel is not in use in North America, a 2.6 kHz signal is continuously present in the channel.

Table 2-1 North America Office-to-Office Signaling Frequencies

Signal	Frequency Pairs
Start of digit transmission	1100 and 1700
1	700 and 900
2	700 and 1100
3	900 and 1100
4	700 and 1300
5	900 and 1300
6	1100 and 1300
7	700 and 1500
8	900 and 1500
9	1100 and 1500
0	1300 and 1500
End of digit Transmission	1500 and 1700

2–6 TRANSMISSION LINE FUNDAMENTALS

In order to properly understand the behavior of telephone lines, a brief overview of transmission line theory is necessary. For a full development of the theory, refer to an introductory text on transmission line theory.

Any transmission line, electrically speaking, consists of four distributed constants: a distributed inductance, a distributed capacitance, a distributed series resistance, and a distributed shunt conductance. These constants are illustrated in Fig. 2-13.

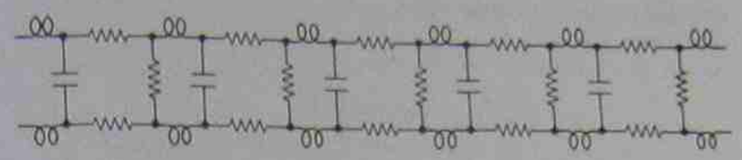


Fig. 2-13 Schematic Representation of the Distributed Constants of a Transmission Line

The four constants are represented on a unit length basis by the following symbols:

L-series inductance	H/unit length
R-series resistance	Ω/unit length
C-shunt capacitance	F/unit length
G-shunt conductance	S/unit length

R represents the resistance or the imperfection of the conductor and is frequency-dependent because of skin effect. G represents the imperfection or dielectric loss of the insulating medium. For modern dielectrics, it is often assumed to be negligible.

The impedance experienced by a wave travelling along a transmission line, commonly known as an incident wave if travelling away from the generator end and a reflected wave if travelling away from the load end, is called the characteristic impedance (Z_0) of the line. If a line is terminated with its characteristic impedance, no reflected wave is present. The characteristic impedance is given by:

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \qquad (2\text{-}1)$$

where $\omega = 2\pi f$ and f is the frequency of operation.

The rate at which a wave travels along a line and the amount of attenuation it experiences depends upon the propagation constant (γ) of the line. This propagation constant, when separated into its real and imaginary components, is symbolized by $\alpha + j\beta$, where α is known as the attenuation constant and β is known as the phase constant. The rate of attenuation is determined by α, which has units of nepers per unit length. To convert nepers to decibels, multiply by 8.686. The phase velocity of

the signal is determined by β, where $v_p = \omega/\beta$. In general:

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{2-2}$$

In order to understand the behavior of the signal on a transmission line, the characteristic impedance and propagation constant in several frequency regions must be examined.

a. dc ($\omega = 0$)

$$\left.\begin{aligned} \alpha &= \sqrt{RG} \quad \text{very small} \\ Z_0 &= \sqrt{\frac{R}{G}} \quad \text{very high} \end{aligned}\right\} \text{since } G \text{ is small} \tag{2-3, 2-4}$$

b. Low frequencies ($\omega L \ll R$); G may be neglected for lines with good insulation.

$$Z_0 = \sqrt{\frac{R}{j\omega C}} \tag{2-5(a)}$$

$$= \sqrt{\frac{R}{2\omega C}} - j\sqrt{\frac{R}{2\omega C}} \tag{2-5(b)}$$

The latter expression shows that the characteristic impedance appears capacitive at low frequencies and that both the resistive and capacitive reactance components decay with an increase in frequency. This dependence on frequency illustrates why it is so difficult to maintain a match when a low-frequency wideband signal is employed.

$$\gamma = \sqrt{R \cdot j\omega C}$$

$$\alpha + j\beta = \sqrt{\frac{R\omega C}{2}} + j\sqrt{\frac{R\omega C}{2}} \tag{2-6}$$

The latter expression shows that attenuation (α) increases with frequency. The R term in the radical represents both the dc resistance and the incremental resistance due to skin effect. The velocity at which an ac wave and its accompanying electric and magnetic fields are propagated can be evaluated by the expression for the phase velocity, $v_p = \omega/\beta$. Thus, at low frequencies the phase velocity can be approximated by

$$v_p = \frac{\omega}{\sqrt{\frac{R\omega C}{2}}} = \sqrt{\frac{2\omega}{RC}}$$

c. Very high frequencies $\omega L \gg R$, $\omega C \gg G$ with G negligible anyway.

$$Z_0 = \sqrt{\frac{L}{C}} \tag{2-7}$$

Note that the characteristic impedance is purely resistive at high frequencies. It is this impedance that manufacturers usually quote when specifying the characteristic impedance of a line, i.e. 50 Ω, 75 Ω, and so forth.

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)\omega C} = \sqrt{j\omega L\left(1 + \frac{R}{j\omega L}\right)j\omega C}$$

$$= j\omega\sqrt{LC}\left[1 + \frac{R}{j2\omega L}\right]$$

using the binominal expansion.

Thus, $\alpha + j\beta = \frac{R}{2}\sqrt{\frac{C}{L}} + j\omega\sqrt{LC}$

At the higher frequencies the attenuation constant flattens out to $R/2\sqrt{L/C}$, whereas the phase velocity flattens out to $1/\sqrt{LC}$. Typical frequency plots of $Z_0 = R_0 + jX_0$, $\gamma = \alpha + j\beta$, and phase velocity are shown in Fig. 2-14.

If the attenuation constant of Eq. (2-2) is minimized with respect to the distributed inductance (L), it is found that for:

$$L = \frac{RC}{G} \tag{2-8}$$

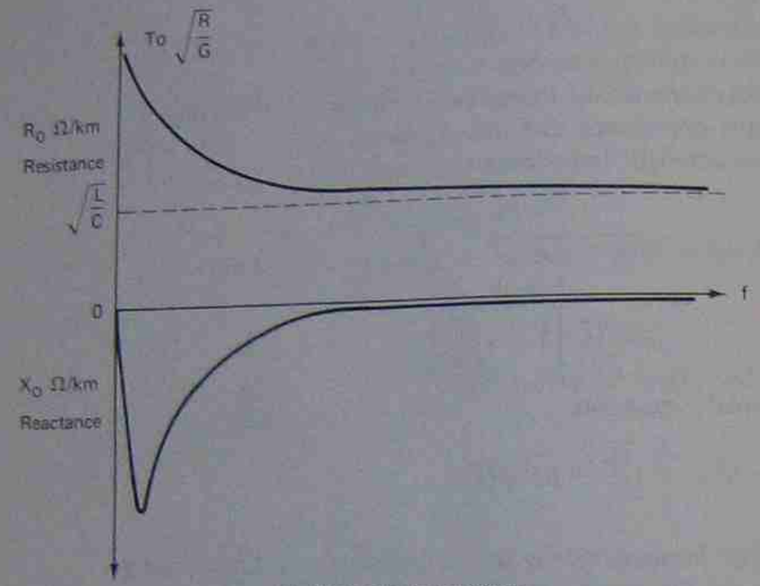
$$\alpha = \sqrt{RG} \tag{2-9}$$

The distributed inductance for an open two-wire or coaxial transmission line is normally much less than the requirement of Eq. (2-8), resulting in a greater than necessary attenuation. By increasing L or by loading the line such that Eq. (2-8) holds, the following line characteristics are achieved:

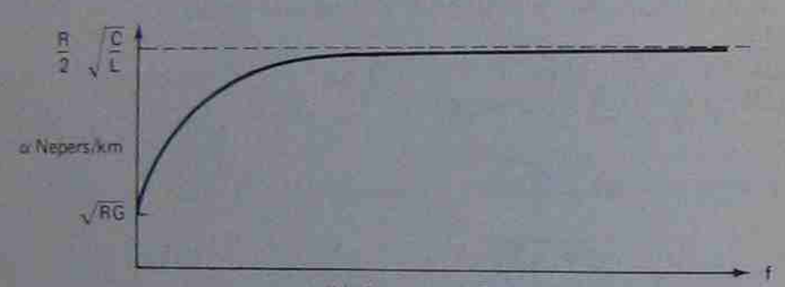
a. Minimum attenuation of $\alpha = \sqrt{RG}$
b. No phase distortion as $v_p = 1/\sqrt{LC}$, which is frequency-independent.

Because L is increased, the phase velocity is significantly slower for the loaded line than if it were not loaded.

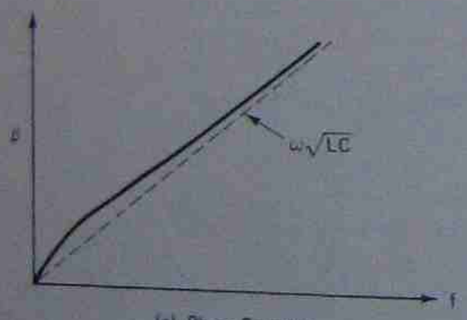
Loading can be achieved by wrapping a highly permeable material, such as iron wire, helically around the conducting core. This method is



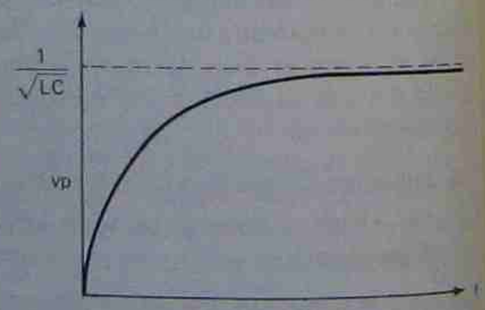
(a) Characteristic Impedance



(b) Attenuation Constant



(c) Phase Constant



(d) Phase Velocity

Fig. 2-14 Characteristic Impedance, Attenuation Constant, Phase Constant, and Phase Velocity vs. Frequency Plots of a Typical Transmission Line.

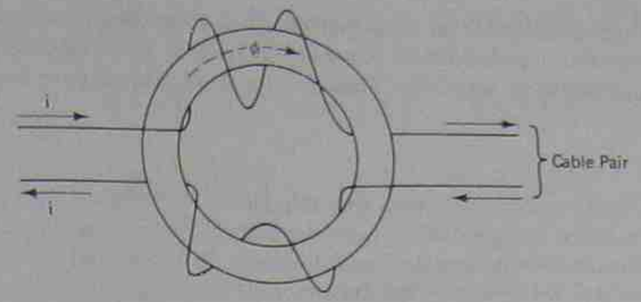


Fig. 2-15 Loading Coil Winding. Flux is enhanced by both currents.

used on submarine cables where the high expense can be justified. On land line systems, lumped inductances, as shown in Fig. 2-15, are added to the line at regular intervals, causing the line to act much like a low-pass filter. The effect of lumped loading is similar to the continuously loaded core up to a critical cutoff frequency. Beyond the cutoff frequency, the attenuation increases sharply, as shown in Fig. 2-16. Using 88 mH loading coils at 1.6 km spacing results in a cutoff frequency of about 3.5 kHz.

The propagation velocity of the signal on a loaded line also tends to be more constant than for an unloaded line, but has the disadvantage of being much lower. This flatness results in very little delay distortion, but the longer delay is of real concern when reflections or echos are present on the line. The latter tend to be very annoying if the delay is in the order of several milliseconds. Typical transmission velocities along a loaded wire pair are around 30,000 km/s at voice frequencies. This results in a delay of 0.03 ms per km or 30 ms for a 1000 km run.

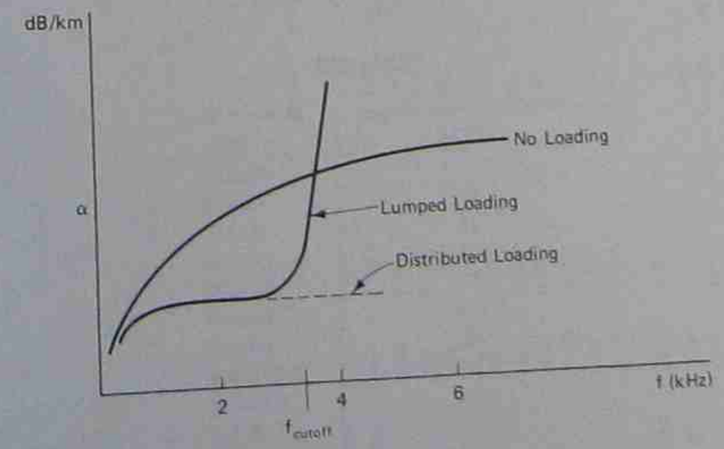


Fig. 2-16 Effect of Loading Coils on Attenuation

Loading coils must be removed from a line if high frequencies are to be transmitted. Baseband data transmission rates of 300 k b/s are generally obtainable on subscriber lines as long as all loading coils are absent.

2-6.1 Subscriber's Loop

The subscriber loop connects the telephone handset to the local switching office. It is generally a two-wire circuit which carries signals in both directions simultaneously, i.e., full-duplex. Its length and minimum conductor size are limited by the maximum loop resistance permitted for either supervisory signaling or transmission attenuation of the voice signal, depending upon which exceeds the allowed degradation first. For step-by-step switching stations, the loop resistance is determined by the signaling and dial pulsing requirements and is limited to 1200 Ω. For crossbar and ESS offices, the resistance is limited to 1300 Ω. If the resistance of a telephone handset is 50 Ω, 1250 Ω are left for the actual subscriber line.

The amount of signal attenuation that is allowed in the local loop depends on the overall national transmission plan in use. The maximum allowable loss is about 8 dB, with about 3.5 dB the average.

Table 2-2 shows the typical resistances of copper conductor employed in local loops. Calculating the maximum loop length that can be used if the resistance is to be kept below 1250 Ω is quite straightforward. For any increase of three gage numbers, the cross-sectional area of the conductor is halved, which causes a doubling in resistance.

For an AWG #22 loop, the maximum loop length determined by the signaling limit would be:

$$l = \frac{1250}{105.96} = 11.8 \text{ km}$$

Table 2-2

American Wire Gage AWG#	Ω/km of Loop at 20°C
12	10.420
14	16.568
16	26.352
18	41.896
19	52.828
20	66.601
21	83.990
22	105.960
23	133.596
24	168.438
25	212.402
26	267.782

Table 2-3 Approximate Characteristics of Standard Types of Paper-Insulated Exchange Telephone Cable Circuits. (Courtesy of Howard W. Sams Inc., Indianapolis, Indiana. From *Reference Data for Radio Engineers*, 6th Edition, p. 35-12.)

Wire Gauge (AWG)	Code No.	Type of Loading	Loop Mile Constants C (μF)	Loop Mile Constants G (μmho)	Propagation Constant Polar Mag	Propagation Constant Polar Angle (deg)	Propagation Constant Rectangular α	Propagation Constant Rectangular β	Mid-Section Characteristic Impedance Polar Mag	Mid-Section Characteristic Impedance Polar Angle (deg)	Mid-Section Characteristic Impedance Rectangular Z_R	Mid-Section Characteristic Impedance Rectangular Z_{XR}	Wavelength (miles)	Velocity (miles per second)	Cut-off Freq (hertz)	Attenuation (dB per mile)
26	BST	NL	0.083	1.6	—	—	—	—	910	—	—	—	—	—	—	2.9
	ST	NL	0.069	1.6	0.439	45.30	0.307	0.310	1067	44.5	719	706	20.4	20 400	—	2.67
24	DSM	NL	0.085	1.9	—	—	—	—	725	—	—	—	—	—	—	2.3
	ASM	NL	0.075	1.9	0.355	45.53	0.247	0.251	778	44.2	558	543	25.0	25 000	—	2.15
		M88	0.075	1.9	0.448	70.25	0.151	0.421	987	23.7	904	396	14.9	14 900	3100	1.31
		H88	0.075	1.9	0.512	76.28	0.130	0.495	1150	14.6	1122	292	12.7	12 700	3700	1.13
		B88	0.075	1.9	0.684	81.70	0.099	0.677	1532	8.1	1515	215	9.3	9 270	5300	0.86
22	CSA	NL	0.083	2.1	0.207	46.92	0.207	0.213	576	43.8	416	399	29.4	29 400	—	1.80
		M88	0.083	2.1	0.447	70.27	0.106	0.434	905	13.7	880	214	14.5	14 500	2900	0.92
		H88	0.083	2.1	0.526	80.11	0.0904	0.519	1051	9.7	1040	177	12.1	12 100	3500	0.79
		H135	0.083	2.1	0.644	83.50	0.0729	0.640	1306	6.3	1300	144	9.8	9 800	2800	0.63
		B88	0.083	2.1	0.718	84.50	0.0689	0.718	1420	5.3	1410	130	8.75	8 750	5000	0.60
		B135	0.083	2.1	0.890	86.60	0.0549	0.890	1765	3.3	1770	102	7.03	7 050	4000	0.48
19	CNB	NL	0.085	1.6	—	—	—	—	400	—	—	—	—	—	—	1.23
	DNB	NL	0.066	1.6	0.188	47.00	0.128	0.138	453	42.8	338	308	43.7	45 700	—	1.12
		M88	0.066	1.6	0.383	82.42	0.0505	0.380	950	8.9	939	146	16.6	16 600	3200	0.44
		H88	0.066	1.6	0.459	84.60	0.0432	0.459	1137	5.2	1130	103	13.7	13 700	3900	0.38
		H135	0.066	1.6	0.569	86.53	0.0345	0.570	1413	4.0	1410	99	11.0	11 000	3200	0.30
		H176	0.066	1.6	0.651	87.23	0.0315	0.651	1643	3.3	1640	95	9.7	9 700	2800	0.27
		B88	0.066	1.6	0.641	86.94	0.0342	0.641	1565	2.8	1560	77	9.8	9 800	5500	0.30
16	NH	NL	0.064	1.5	0.133	49.10	0.0868	0.1004	320	40.6	243	208	62.6	62 600	—	0.76
		M88	0.064	1.5	0.377	85.88	0.0271	0.377	937	4.6	934	76	16.7	16 700	3200	0.24
		H88	0.064	1.5	0.458	87.14	0.0238	0.458	1130	2.8	1130	55	13.7	13 700	3900	0.21

1060 hertz.
In the third column of the above table the letters M, H, and B indicate loading-coil spacings of 9000 feet, 6000 feet, and 3000 feet, respectively, and the figures show the inductance in millihenries of the loading coils used. NL indicates no loading.

As discussed earlier, the attenuation experienced by a signal on a transmission line is determined by the electrical characteristics of the line as well as the signal frequency. In addition, the resistance of the line is slightly temperature-dependent. Any moisture in a coaxial line has a very detrimental effect on the transmitted signal.

Table 2-3 gives losses of standard types of paper-insulated cable circuits. In order to obtain longer loops for a given total attenuation, the conductor diameter can be increased, which would involve going to a lower AWG gage, or inductive loading can be used. Going from an unloaded ASM AWG #24 wire to an unloaded DNB AWG #19 wire, one can increase the loop length by a factor of 2.15/1.12 or 1.9. Loading the line by adding series-lumped inductances at fixed intervals can further increase the loop length. Inserting 88 mH inductors at 6000 ft intervals increases the loop length of loaded versus unloaded DNB cable by a factor of 1.12/.38 or 2.9. As can be further seen from Table 2-3, loading greatly decreases the velocity of propagation, which results in greater signal delays. Loading also has a tendency to increase the characteristic impedance of the line. For local loops, 19H88 and 22D66 cable is most frequently used. The first two digits represent the wire gage, the letters H and D indicate loading-coil spaces of 6000 and 4500 ft, and the last two digits indicate the inductance in millihenries.

In order to obtain even longer subscriber loops, amplifiers or range extenders can be added. Range extenders increase the battery voltage on the loop to extend the signaling range. In addition, they often contain amplifiers to extend the transmission loss limits.

For very long loop lengths, the exact method used to connect the individual subscriber to the switching center depends upon the distance between the two and the number of subscribers. For a few subscribers at distances up to 50 km from the exchange, a distributed carrier system such as radio can be used. In this case, several subscribers on a multiparty line or on private lines are connected to the remote station. The central station is located at the exchange as shown in Fig. 2-17.

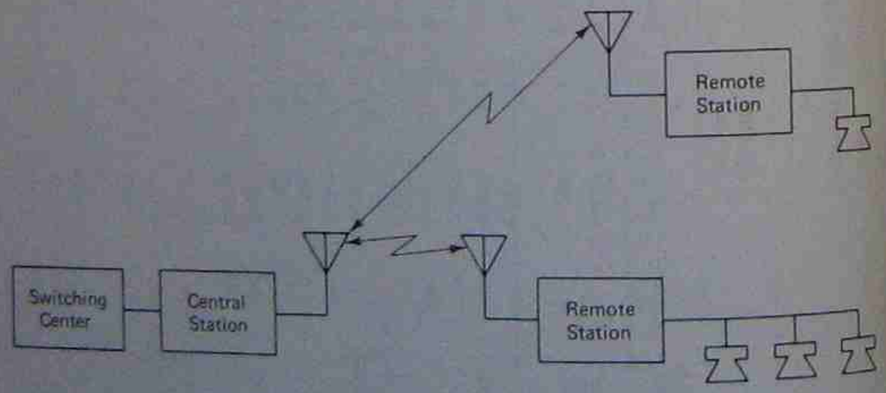


Fig. 2-17 Subscriber Radio System

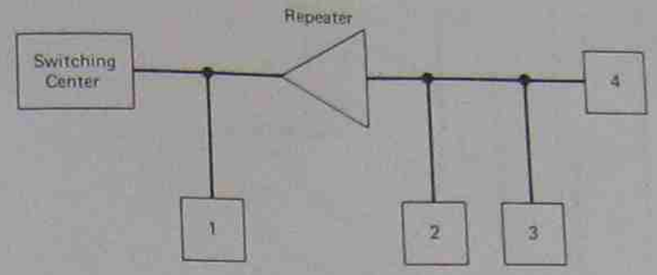


Fig. 2-18 Distributed Carrier System

When there are a small number of subscribers along a rather long route, it can be convenient to allocate a slightly different carrier frequency to each channel. Up to six channels can be placed on a single wire pair, as shown in Fig. 2-18. Such a system is often referred to as a distributed carrier.

When there are clusters of subscribers in a remote region, it is more convenient to go to a lumped carrier system. In such a system, up to a few hundred subscribers are digitally multiplexed onto a common highway to the central station. Such a system is shown in Fig. 2-19.

Figure 2-20 graphically illustrates what type of subscriber loop system would be most applicable in a given situation. For a large number of subscribers near a switching center, standard cable pairs can be used. For slightly longer distances, loaded cables or amplifiers or some similar device needs to be used. For a combination of long loop lengths and a large number of subscribers, remote centers using digital lines are the most attractive. The boundaries illustrated by the dotted lines are, in practice, rather fuzzy.

2–6.2 Trunk Circuits

Beyond the local loop and local central office are the toll-connecting trunks which connect calls to the toll office. Toll offices are, in turn, inter-

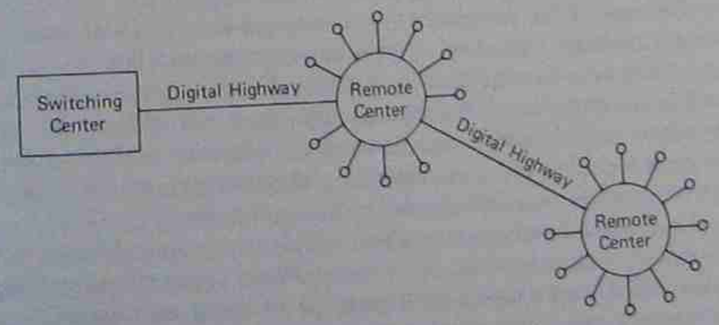


Fig. 2-19 Lumped Carrier System

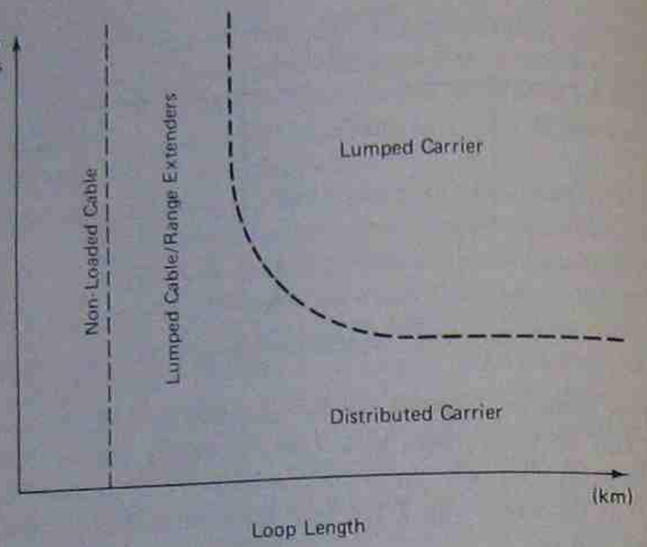


Fig. 2-20 Optimum Loop System

connected with intertoll trunks. Because each trunk circuit must carry several channels simultaneously, the demands placed on these circuits in terms of frequency and phase response are much more severe than on the local loops. All intertoll trunks and many of the toll-connecting trunks are four-wire circuits; that is, there are two pairs of wire for each circuit, or one pair for each direction. In the local loop, on the other hand, a two-wire or single-wire pair must carry the single telephone call in both directions. Four-wire circuits readily permit full-duplex operation. A two-wire circuit can be used for duplex by splitting the frequency band into two separate regions, one for transmitting and one for receiving; however, this results in reduced transmission or information rates.

Table 2-4 uses single diagrams and brief descriptions to illustrate the various types of trunk circuits in use. Each channel is assumed to cover a bandwidth of 300 to 3300 Hz. The crosstalk refers to a message on one wire pair being coupled to an adjacent pair.

Because of the present heavy investment in both local loops and trunk circuit cable, any new or alternative communication modulation scheme will have to employ cable systems already in service. Severe demands have been placed on cables by implementation of the more recent schemes with their higher frequencies. A consideration of the harmful effects on the cables and a consideration of some measures that can be taken to improve the cable characteristics is in order.

One major concern with cable, and also with optical fibers, is the increased attenuation caused by moisture. Water vapor diffusing through a cable sheath can cause a 50% increase in signal attenuation. In the more severe cases, it can result in the loss of wire pairs due to corrosion.

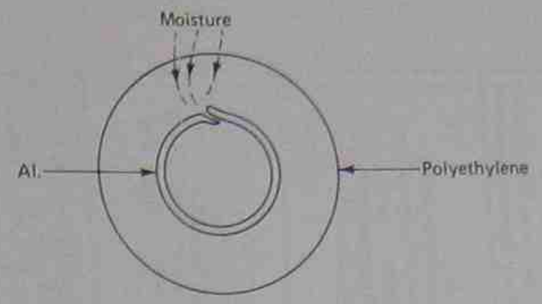


Fig. 2-21 A Bonded Jacket

Moisture can creep into cables through sheath damage due to incorrect cable handling procedures, lightning-causing pin holes, sheath discontinuity in spliced areas, rodents, and so forth. Some penetration occurs even with the best of protected jackets, so that cable moisture gradually increases over the years. The moisture problem is reduced by using polyethylene jackets bonded to aluminum, as shown in Fig. 2-21, or, in the case of air core cable, by pressurizing the cable with dry air or by filling the cable with a petroleum jelly compound. As the latter tends to flow out of cables that are placed in the sunlight, pulverized polyethylene is added to give a more putty-like consistency.

Figure 2-22 shows additional cable faults that can cause signal degradation.

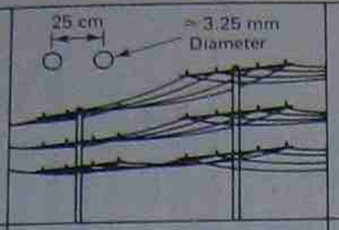
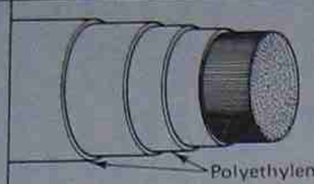
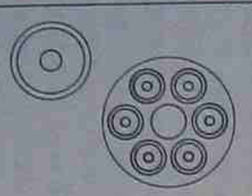
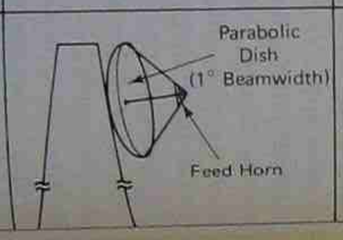
2–6.3 Frequency Division Multiplexing (FDM)

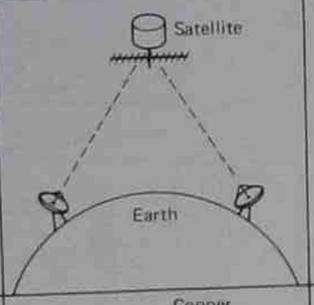
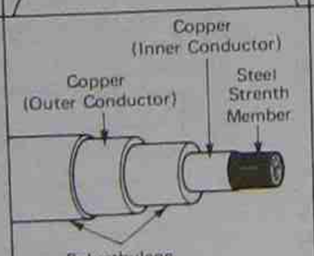
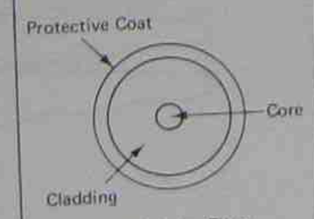
In order to increase the utilization of a transmission line, several telephone conversations can be sent down the line simultaneously. By stacking several low-frequency signals into a higher spectrum location through repetitive modulation, multiplexing is achieved. Multiplexing is the name given to the technique which allows a communication circuit to carry more than one channel.

For example, since open-wire transmission lines can be used up to frequencies of approximately 300 kHz and twisted wire pairs up to frequencies of 120 kHz, several *voice* channels, each 4 kHZ wide, can be frequency-spaced so that all channels can be carried together. This 4 kHz bandwidth includes the nominal 300 to 3400 Hz audio bandwidth along with the guardband which is required due to the finite roll-off of the bandpass filters. A common form of frequency division multiplexing (FDM) is the standard CCITT group which contains 12 voice channels over the 60 to 108 kHz frequency range, as shown in Fig. 2-23.

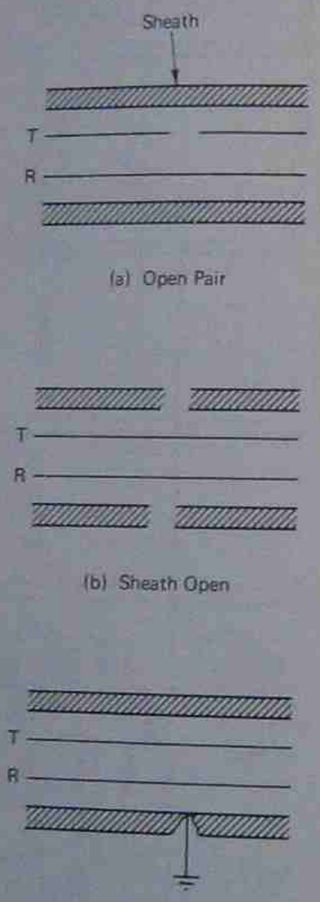
In order to understand how such a group is formed, the technique of generating and detecting single-sideband (SSB) must be considered first. SSB is most often obtained by initially generating the double side-band suppressed carrier (DSBSC) with a balanced modulator and then filtering

Table 2-4 Types of Trunk Circuits

Transmission Media	Construction	Distance Between Repeaters	No. of Channels on a Single Line or Carrier	Operating Frequency Range	Notes
Open Wire		40 km	12	Up to 160 kHz	Bulky and Unsightly. Affected by Weather Conditions, i.e. Large Leakage with Wet Insulators. Severe Crosstalk. High Radiation Losses at the Higher Frequencies. Resistance Increases with Frequency Due to Skin Effect. Large Distance Between Amplifiers.
Twisted Wire Pair Cable		3 to 6.5 km	12–120	Up to 1 MHz	Subject to Severe Crosstalk Because of Nearness of Conductors. Resistance Increases with Frequency Due to Skin Effect.
Coaxial (Tube)		3 to 65 km	L5 Carrier 10,800 per Tube 108,000 per Cable	3 to 60 MHz	No Radiation Losses. Negligible Crosstalk. Large Number of Channels can be Sent Over a Single Tube.
Microwave		30–50 km	TD-2 Carrier 600–1200 per Radio Channel 10,800 per Route	3.7 to 4.2 GHz 5.925 to 6.425 GHz etc.	Fewer Amplifiers Required than Coaxial System. Affected by Rain when Frequency Exceeds 10 GHz. Large Number of Channels can be Sent Over a Single Antenna. High Velocity of Propagation, Minimizing Delay Times. Can Cause Radio Interference.

Transmission Media	Construction	Distance Between Repeaters	No. of Channels on a Single Line or Carrier	Operating Frequency Range	Notes
Geosynchronous Satellite		Direct	600 per Transponder (12 Transponders per Satellite)	Intel Sat IV Uplink 5.925 to 6.425 GHz Downlink 3.7 to 4.2 GHz	Subject to Interference from Terrestial Link and to Interfere with Terrestial Link Long Delay Times – Up to 270 ms Large and Thus Costly Earth Transmitting Antennas Required Large Distance and Large Area Coverage. Economical Receive Only Earth Stations. Much Lower Cost per Channel than Submarine Cable for Transatlantic Communications.
Submarine Cable		18 km	4000	28 MHz max. Frequency	Employ Time Assignment Speech Interpolation (TASI) While a User is Silent, Even Between Words, the Channel is Transferred to Another Talker.
Optical Fibre		4 to 30 km	Several Hundred Voice Channels	1 micron Wavelength 1 Ghz-km	Difficult to Splice in the Field. Cannot be Used to Carry DC Power. Not Subject to Interference or Tapping.

out one of the sidebands with a sideband filter. DSBSC consists of both sidebands with the carrier removed, whereas SSB consists of only one sideband with the carrier removed. SSB is very efficient in terms of spectrum utilization since it contains only one sideband as opposed to two in the AM and DSB cases. The extra sideband in DSB is not wasted power since it provides a 3 dB improvement in S/N ratio at the receiver. Neither SSB or DSB wastes the large power in the carrier which occurs in AM



(a) Open Pair

(b) Sheath Open

(c) Grounded Sheath on Buried Cable (Due to Sheath Perforation)

Fig. 2-22 Common Cable Faults

transmission, or in power which contains no information. The mathematical expression for a DSB output is:

$$k\, m(t) \cos \omega_c t \qquad (2\text{-}10)$$

where $m(t)$ represents the modulating signal.
$\omega_c = 2\pi f_c$ represents the carrier angular velocity.
k represents the circuit constant.

In order to handle this mathematically, consider a sinusoidal modulating signal of $m(t) = \cos \omega_m t$. Using Eq. (2-10), the following is obtained:

$$k \cos\ \omega_m t \cos \omega_c t = \frac{k}{2} \cos (\omega_c + \omega_m)t + \frac{k}{2} \cos (\omega_c - \omega_m)t \qquad (2\text{-}11)$$

$$\cos \omega_1 t \cos \omega_2 t = \frac{1}{2} \cos (\omega_1 + \omega_2)t + \frac{1}{2} \cos (\omega_1 - \omega_2)t \qquad (2\text{-}12)$$

This DSB signal contains two sidebands, one at $\omega_c + \omega_m$ and one at $\omega_c - \omega_m$. In general, if the modulating signal frequency spectrum of the audio signal is like that represented by the triangle in Fig. 2-24(a), the output frequency spectrum from the balanced modulator appears as that shown in Fig. 2-24(b).

The two sidebands are called the upper and lower sidebands. The upper sideband (the one to the right of f_c) remains erect, whereas the lower sideband is inverted. If only the lower sideband is transmitted (SSB), the receiver needs to invert the sideband for proper detection. SSB is obtained by passing the DSBSC signal through a sideband bandpass filter with sharp cutoff characteristics at the passband edges. Even with filters that have an extremely fast frequency roll-off, it is impossible to transmit signals with low frequency information with SSB. Either some of

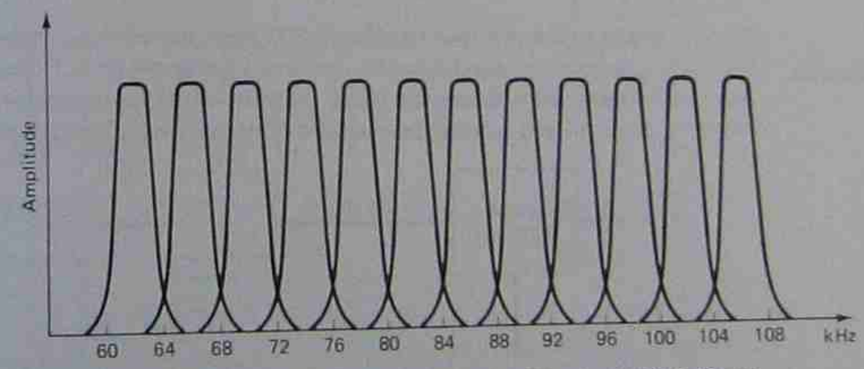
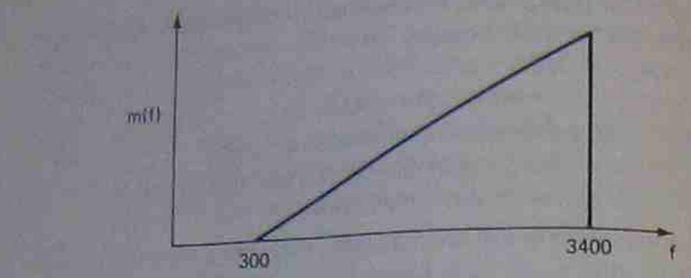
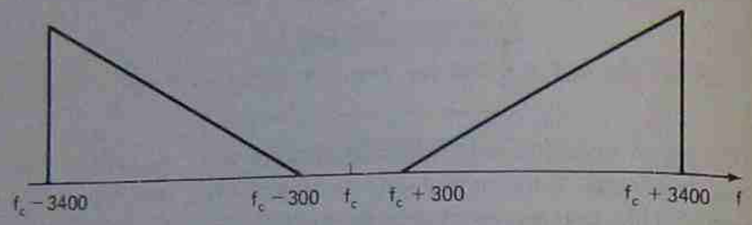


Fig. 2-23 Frequency Division of Standard CCITT Group



(a) Modulating Signal Spectrum



(b) DSB Spectrum

Fig. 2-24 Double Sideband Spectrum

the adjacent sideband still gets through or some of the desired low frequencies are rejected by any real filter. Because both data and video contain low frequency information, it is undesirable to employ SSB in such situations. The block diagram of SSB generation is shown in Fig. 2-25.

A practical double-balanced modulator circuit is described in Appendix E.

2–6.4 CCITT Groups

The discussion in the previous section on SSB generation should make it relatively easy to understand how the standard 12-channel CCITT Group is formed. Figure 2-26 shows the block diagram of the equipment employed to transmit and receive the frequency division multiplexed signal.

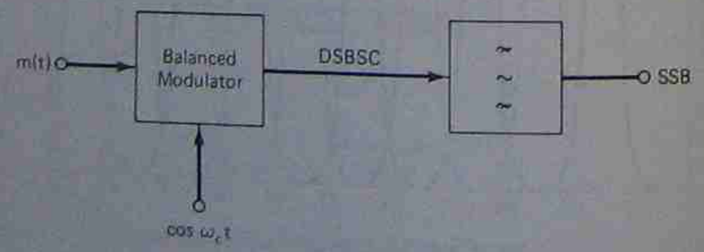


Fig. 2-25 Generation of SSB

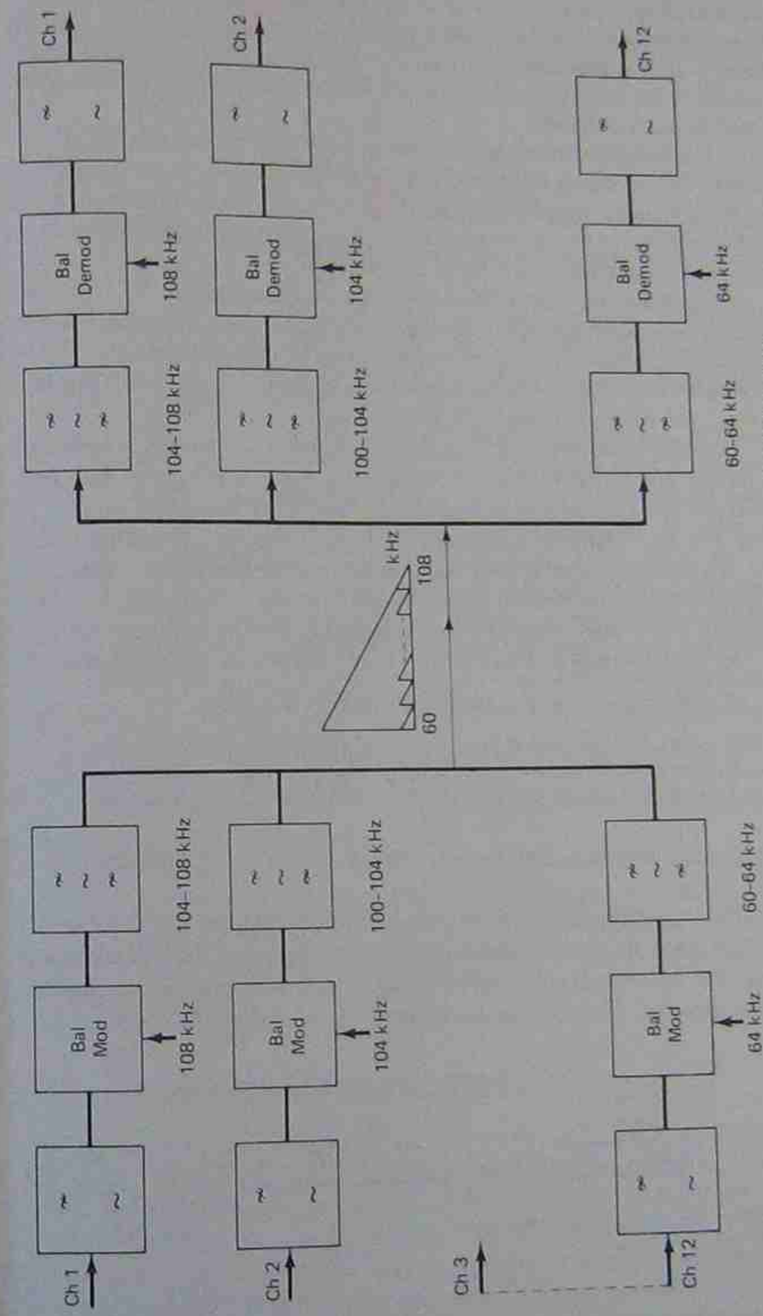


Fig. 2-26 Formation and Detection of the Standard CCITT Group

The low-pass filters are inserted at the sending end to ensure that each channel is limited to well within the 4 kHz band so as not to interfere with adjacent channels when all the channels are combined. A guard band is also inserted between the channels to minimize the interference between channels. Overall, about 3.3 kHz of useable bandwidth is allocated to each channel.

The balanced modulators form a double sideband signal with no carrier. The bandpass filters reject the upper sidebands, which results in a neatly packaged transmitted signal of 12 channels within the 60 to 108 kHz range. On the receiver end, the various bandpass filters allow only their respective channels to pass through.

The balanced demodulators, which are identical to the balanced modulators, cause the sum and difference frequencies to appear at the output. The difference frequencies are chosen by the final low-pass filter, causing an inversion of the received channel spectrum. This is illustrated in Fig. 2-27 where Channel II is extracted.

In order to accommodate more traffic, these 12-channel groups can be further combined to form a supergroup. A supergroup can be combined to form a mastergroup and so on. The CCITT hierarchy is given in Table 2-5. Such standard combinations allow modularity of equipment as well as common international modulation and demodulation schemes. Multiple modulation has the disadvantage of introducing more noise and distortion each time modulation and demodulation takes place.

In North America, a variation is made above the supergroup level.

10 supergroups = 1 mastergroup (600 channels)

6 mastergroups = 1 jumbogroup (3600 channels)

3 jumbogroups = 1 jumbogroup multiplex (10,800 channels)

12 super-mastergroups can also be combined to give 10,800 channels.

The maximum shift in frequency permitted by CCITT is ± 2 Hz. To maintain such frequency accuracy, a pilot signal is often transmitted along with the speech signal which is used by the receiver for frequency synchronization. The receiver oscillator is phase-locked to the incoming

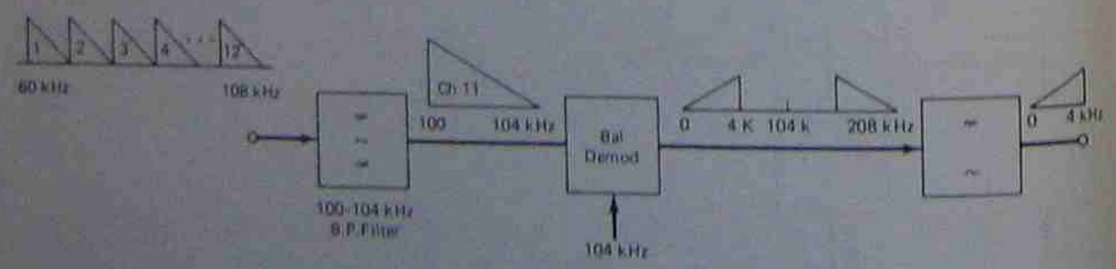


Fig. 2-27 Frequency Spectrum for Channel II in the Receiver

Table 2-5 CCITT Standard for Multiplex Groups

	CCITT Standard	Number of Voice Channels
12 telegraph channels or 24 telex channels	= 1 channel	1
12 channels	= 1 group	12
5 groups	= 1 supergroup	60
5 supergroups	= 1 mastergroup	300
3 mastergroups	= 1 super-mastergroup	900

pilot tone. If the pilot is lost for some reason, the system continues to operate but without synchronization.

To obtain the various local oscillator frequencies illustrated in Fig. 2-26, a master frequency generator can provide a 4 kHz signal to drive a harmonic generator. Various narrow-band filters select the desired harmonics to act as the local oscillator signal.

On international circuits, the end-to-end frequency tolerance should be better than 2 Hz. A frequency synchronizing pilot is normally used to maintain this accuracy. With some of the recent master oscillators, stabilities are such that synchronizing pilots are not required. In order to maintain signal level regulation, level regulating pilots are initiated on each group, super group, master group, and so on. These pilot signal amplitudes are monitored when manually adjusting amplifier and attenuator gains. To maintain signal levels on a dynamic basis because of variations in amplifier gains and line losses as a result of aging and temperature changes, automatic gain control circuits also monitor pilot signal levels. Signal levels must be tightly controlled to prevent poor signal-to-noise ratios at reduced signal levels and excessive intermodulation noise from overdriving the circuits. The latter can be more fully appreciated after circuit loading is considered.

Incoming pilot levels in the multiplex-receive equipment are also monitored to detect equipment failure. If the pilot drops outside the level regulating range (usually about ± 5 dB), an alarm is triggered. In the absence of a pilot, the channel is assumed to have failed and a standby channel unit is brought into operation.

2–7 SWITCHING FACILITIES

Switching machines are used to interconnect any pair of subscribers in the public telephone system. Within a radius of about 10 km from each end office, up to 10,000 subscribers can be interconnected directly through the local exchange. In larger population centers, several end offices are interconnected by inter-exchange trunks to provide expanded service on a toll-free basis. For larger distances, toll centers are provided

through toll connecting trunks. The term *trunk* stems back to the days when the telegraph lines followed the railway trunk lines. When a toll office is used, a record of the call is kept so that appropriate billing of charges can be made.

When a subscriber picks up the telephone handset, current flows in the local loop informing the local office that the handset is OFF-HOOK. A line-finder switch in the local office searches for the subscriber that is calling and connects to that line. The switching equipment is then ready to receive the number to be called. Depending upon the type of switching machine used, the connecting path can be established in one of several ways:

a. Step-by-step switch.
b. Cross-bar switch.
c. Electronic switch.

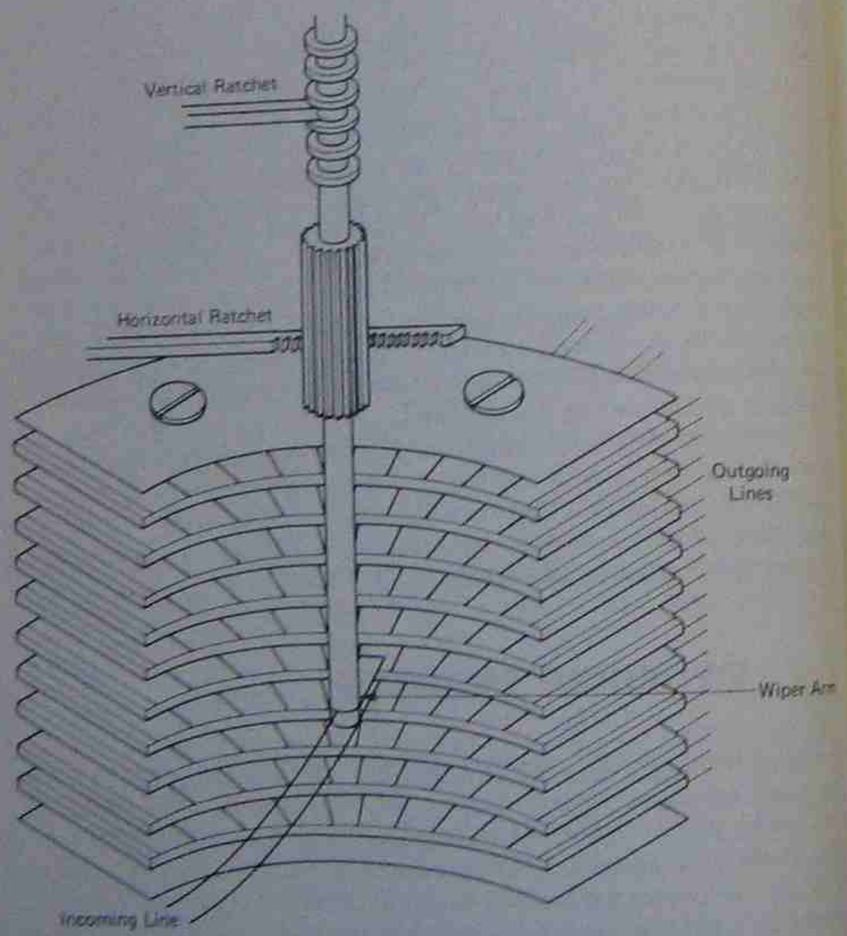


Fig. 2-28 Stepping Switch

2–7.1 Step-by-Step Switch

The first automatic telephone switching systems in the world employed the Strowger or step-by-step switch. Many of these switches, shown in Fig. 2-28, are still installed. The Strowger switch, named after its inventor, consists of a wiper arm which makes contact with any one of the 100 pairs of contacts located in a cylindrical fashion around the shaft axis. A vertical stepping magnet operates a ratchet which determines to which row the wiper arm will be attached, and a horizontal stepping magnet operates another ratchet which determines to which column the wiper arm will be attached. The horizontal rest position is determined by the number of received pulses if the switch is used as a connector, or by the detection of an unoccupied set of contacts if it is used as a selector.

The step-by-step switch (S × S) is a progressive control switch in which each selecting switch is directly controlled by the digits dialed. As illustrated in Fig. 2-29, the first digit dialed causes the wiper arm of the first selector to advance to the vertical level dialed. The arm then rotates until it finds an unoccupied set of contacts. The subscriber is now connected to the second selector switch. After the second digit is dialed, the wiper arm of the second selector is raised to the corresponding level and then rotated to a free set of contacts. Each digit dialed causes another selector to set up the proper electrical path. The final or connector switch takes on two digits, one for the vertical level and one for the horizontal position to connect to the desired called subscriber. In addition, the connector:

a. Causes ringing to be applied to the telephone.
b. Supplies "talk" battery.
c. Supplies supervision.
d. Tests for subscriber "busy" condition.

After going OFF-HOOK and closing the loop relay, the line finder switch is activated and automatically connects to the subscriber contacts. It remains connected as long as loop current flows and drops out after the loop is broken. When dialing 3678150, as illustrated in Fig. 2-29, the first digit, 3, causes the first selector to step up three levels as the three pulses are received. The switch then automatically hunts for an unoccupied set of contacts along that row. Since each contact on a selector is connected to the wiper arm of a succeeding selector switch, the second selector is ready to receive the next digit dialed, 6. As the number is dialed, a connecting path is progressively established through the switching system until finally the called station is reached. The final connector switch takes on two digits, one for the vertical and one for the horizontal positions of the called subscriber. The first three digits indicate the exchange requested; the last four digits locate the subscriber within that exchange.

Other options available are shown in Fig. 2-29. For instance, if the first digit is a 0, an operator is called. If it is a 1, direct distance dialing (DDD) equipment is called. For the relative merits of S × S switching, refer to Table 2-6.

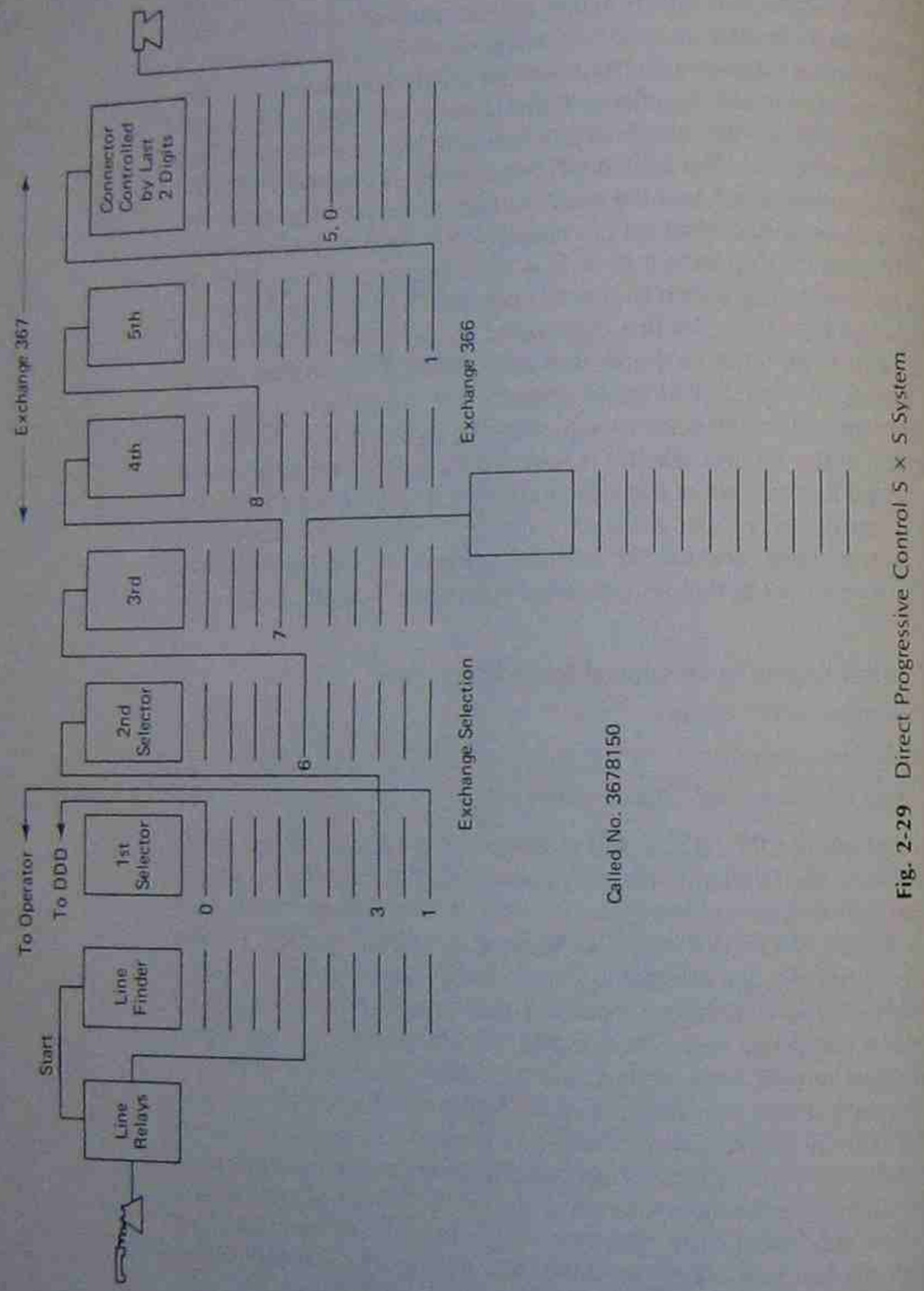


Fig. 2-29 Direct Progressive Control 5 × 5 System

Table 2-6 Relative Merits of Direct Progressive Control Switching

Advantages	Disadvantages
• Signaling is relatively simple and easy to understand. • Inexpensive for small number of subscribers. • Post-dialing delay is negligible. • Easy to add more switches for expansion.	• Speed controlled by subscriber and limited by the characteristics of the switching machine's signal repertory. • Number scheme controls routing of the call. • High level of impulse noise due to current switching. • High level of impulse noise generated by the physical vibrations in neighboring equipment due to mechanical movements of wiper arms. • Maintenance is costly.

2–7.2 Cross-Bar Switch

In common control switching systems, the portion of the equipment that receives the dialed address information and sets up paths is called the common control equipment. Because the common control equipment is used only for the time period when the call is set up, it can be released to deal with other connections once a connection is established, and many subscribers can share the same switching logic equipment. This is different from the step-by-step system in which the circuit is tied up during the duration of the call. Only after all the dialed digits have been received and stored does the common control network begin determining the best path through the network. This results in reduced blocking probabilities as compared to the step-by-step system in which the call is sequentially set up as the digits are dialed. In the latter case, blocking of some following stage is undetected in advance, causing an increased risk of a busy circuit.

The cross-bar switch forms the basic switching mechanism in common control cross-bar switching and consists of electrical contacts arranged in a grid format. A crosspoint, or contact point, is formed when a particular row and column are activated. The row, or horizontal location, is selected by operating the appropriate select magnet, and the column, or vertical location, is selected by the hold magnet. A typical cross-bar switch consists of 10 horizontal rows and 20 vertical columns with a set of four contacts at each crosspoint, as illustrated in Fig. 2-30.

As shown in more detail in Fig. 2-31(a) and (b), operation of the select bar causes the finger to move up or down, which causes closure of the corresponding upper or lower sets of contacts when the holding bar is activated. Upon release of the select bar, only the set of contacts under the vertical holding armature remains trapped until the holding magnet is

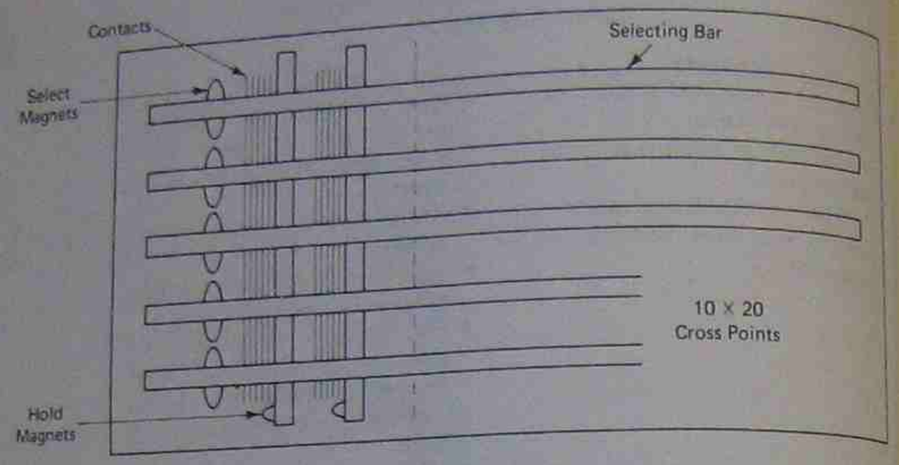


Fig. 2-30 Cross-Bar Switch Illustrating Select and Hold Magnets

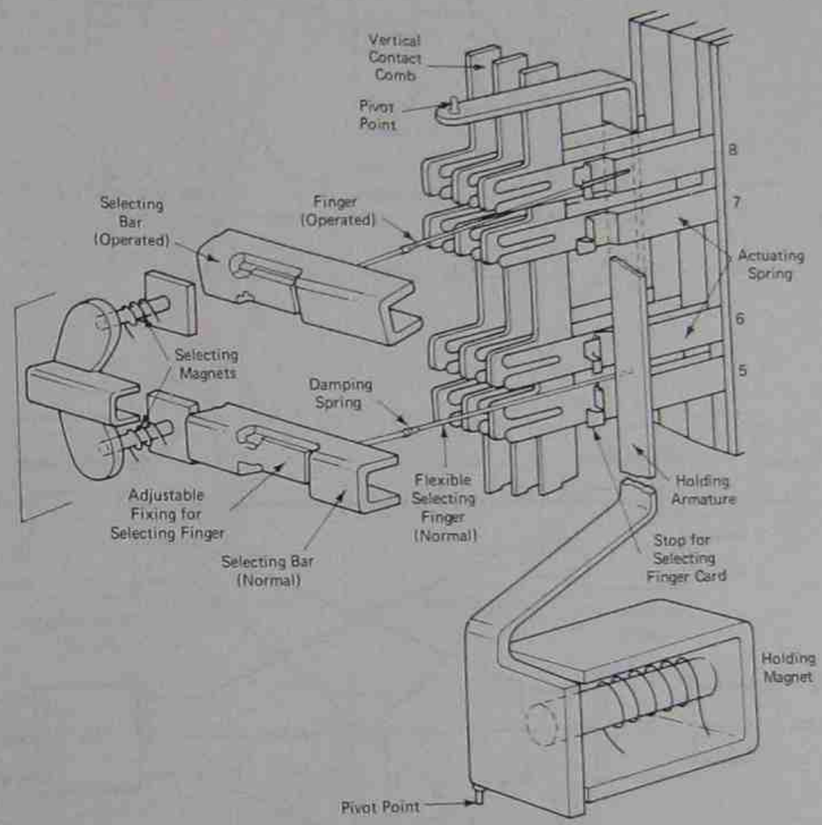
also released. In this manner, up to 20 independent contact closures can be maintained.

Any telephone system must have a number of parallel selection stages in order to handle several simultaneous calls. If all of the selection stages are available to all of the calling lines, then the system has full availability. Generally a system has limited availability to keep costs within reach, while still assuring that all the called lines have access through certain sections of an exchange. This limited availability causes some of the calls to be blocked. To more fully comprehend how blocking is reduced, consider the simplified 3 × 3 cross-bar switch shown in Fig. 2-32(a) and (b).

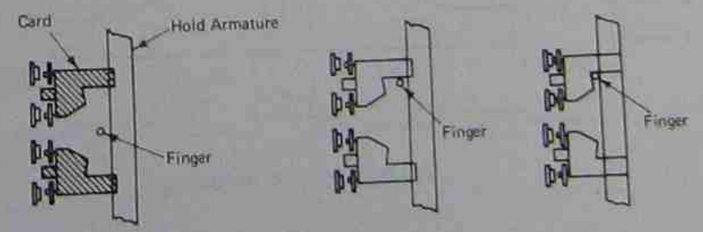
Telephone switching networks are made up of combinations of such blocks (10 × 10 crosspoints are the most common) to allow for access to a larger number of subscribers. Figure 2-33 illustrates a two-stage network made up of nine individual switches. With the interconnections shown, Outlet 1 has only one link to Inlet 1. If Link 1 is occupied, all of the three outlets on the first secondary switch cannot be connected to any of the inlets of the first switch of the primary stage, even if they are free. This latter condition is described as blocking. This arrangement allows a given inlet to be connected to nine outlets rather than to three if only a single switch was used.

The blocking can be reduced by going to a three-stage switch, as shown in Fig. 2-34. Whereas in the two-stage switch Outlet 1 has only one link to Inlet 1, a three-stage switch has three links.

To consider the reduction in the number of crosspoints required in a typical telephone exchange, consider the four-stage switch shown in Fig. 2-35.

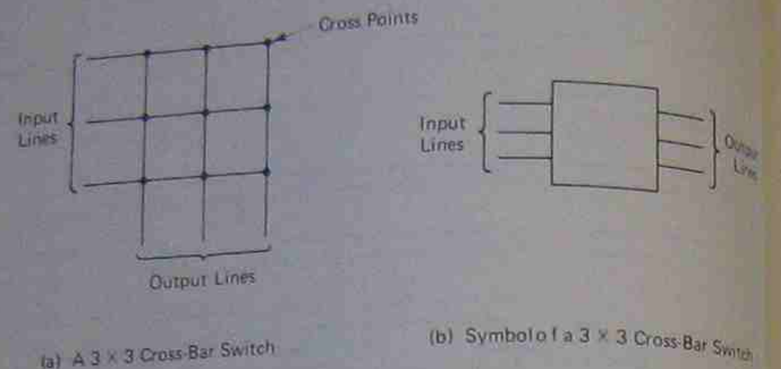


(a) Principle of Operation of Crossbar Switch



(b) Finger Action

Fig. 2-31 A Cross-Bar Switch (Courtesy of Northern Telecom.)



(a) A 3 × 3 Cross-Bar Switch

(b) Symbol of a 3 × 3 Cross-Bar Switch

Fig. 2-32 Symbol of a Single Switch

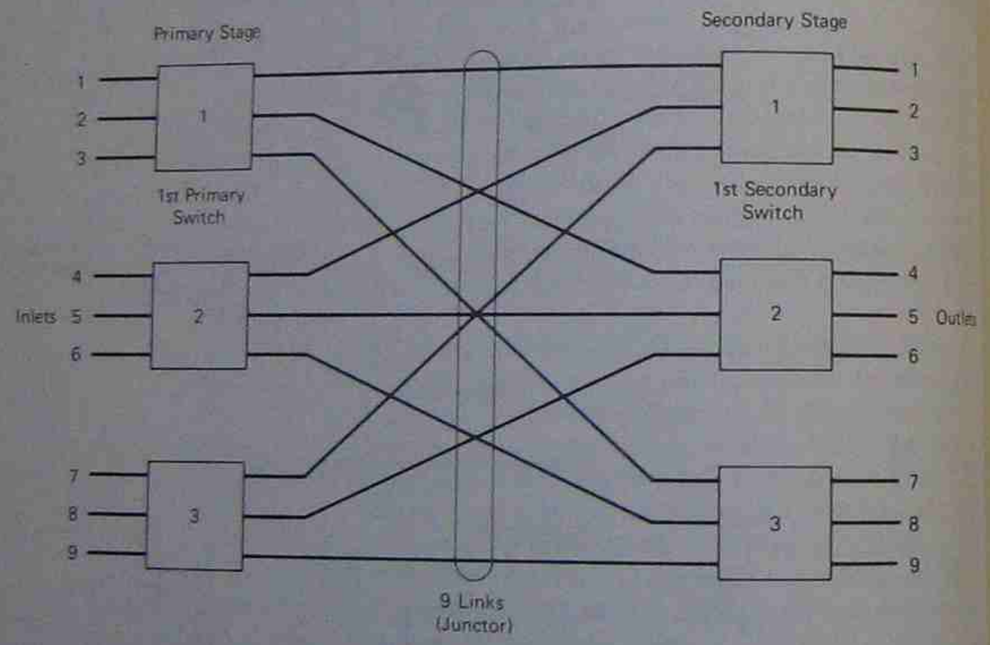


Fig. 2-33 A Two-Stage Switch

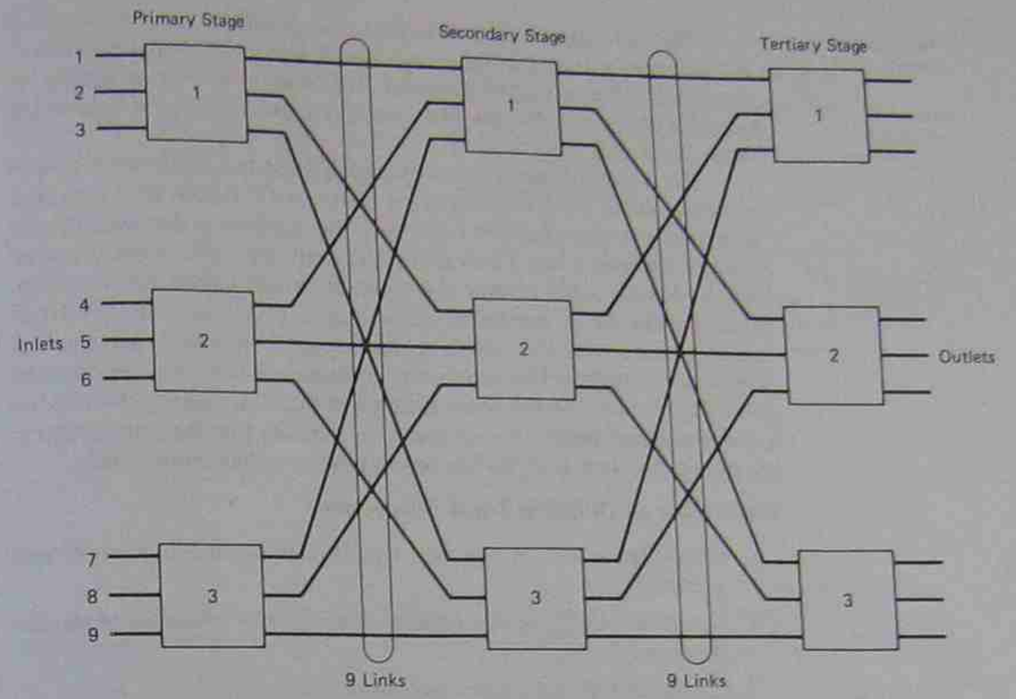


Fig. 2-34 A Three-Stage Switch

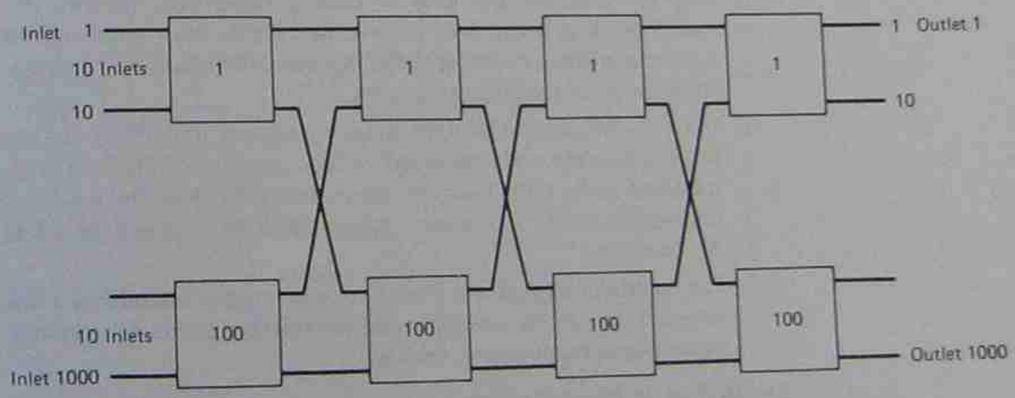


Fig. 2-35 A Four-Stage Switch

The four-stage switch has 10,000 cross-points per stage for a total of 40,000 cross-points. If a single stage switch was used, a total of one million cross-points would be required. The amount of wire-wrapping required is phenomenal in the first case and impractical in the case of the single-stage switch.

The establishment of a call through a cross-bar exchange is shown in the simplified block diagram of the Northern TELECOM SA-1 cross-bar switching machine. (Refer to Fig. 2-36.) The machine is designed for use in an unattended Class 5 end office and employs a two-stage cross-bar switch network under control of a group of circuits called the controller. It can handle, on an expandable basis, up to 1500 lines. The SA-1 can connect a subscriber to a toll switching center on an interswitch trunk or over a carrier system. The connection procedure is best illustrated in step form. The numbers on the block diagram of Fig. 2-36 refer to the step sequence outlined below. The dotted lines indicate that the connection is temporary and lasts only for the set-up time of establishing a path.

Establishing an Outgoing Trunk Connection

1. Controller seized by the line circuit. Calling line is located and identified.
2. Controller selects an idle register. (Calling lines and class of service are stored in register.)
3. Controller selects line-link switches.
4. Controller connects register to calling line via the trunk switch connector and line-link switches.
5. Register furnishes dial tone to calling subscriber, notifying the subscriber that it can now receive dial digits. After receiving the connection office code digits, the register determines the particular trunk group of the distant location.
6. Via a by-link path around the outgoing register link (ORL), an idle trunk is selected and connected. Subsequently, a connection is established in the ORL, releasing the by-link path. After the final four digits of the called subscriber's station are dialed, the register seizes the controller.
7. The controller through the trunk switch connector establishes a link between the calling subscriber and the selected trunk. All common control equipment is then released.

Establishing an Incoming Connection

1. The incoming call seizes the incoming trunk circuit and bids for an idle register through the by-link connection. The incoming register link (IRL) is set up and the by-link path opened. Only one incoming seizure can be served and one by-link path established at a time. Incoming trunks are given various priorities, such that toll switched

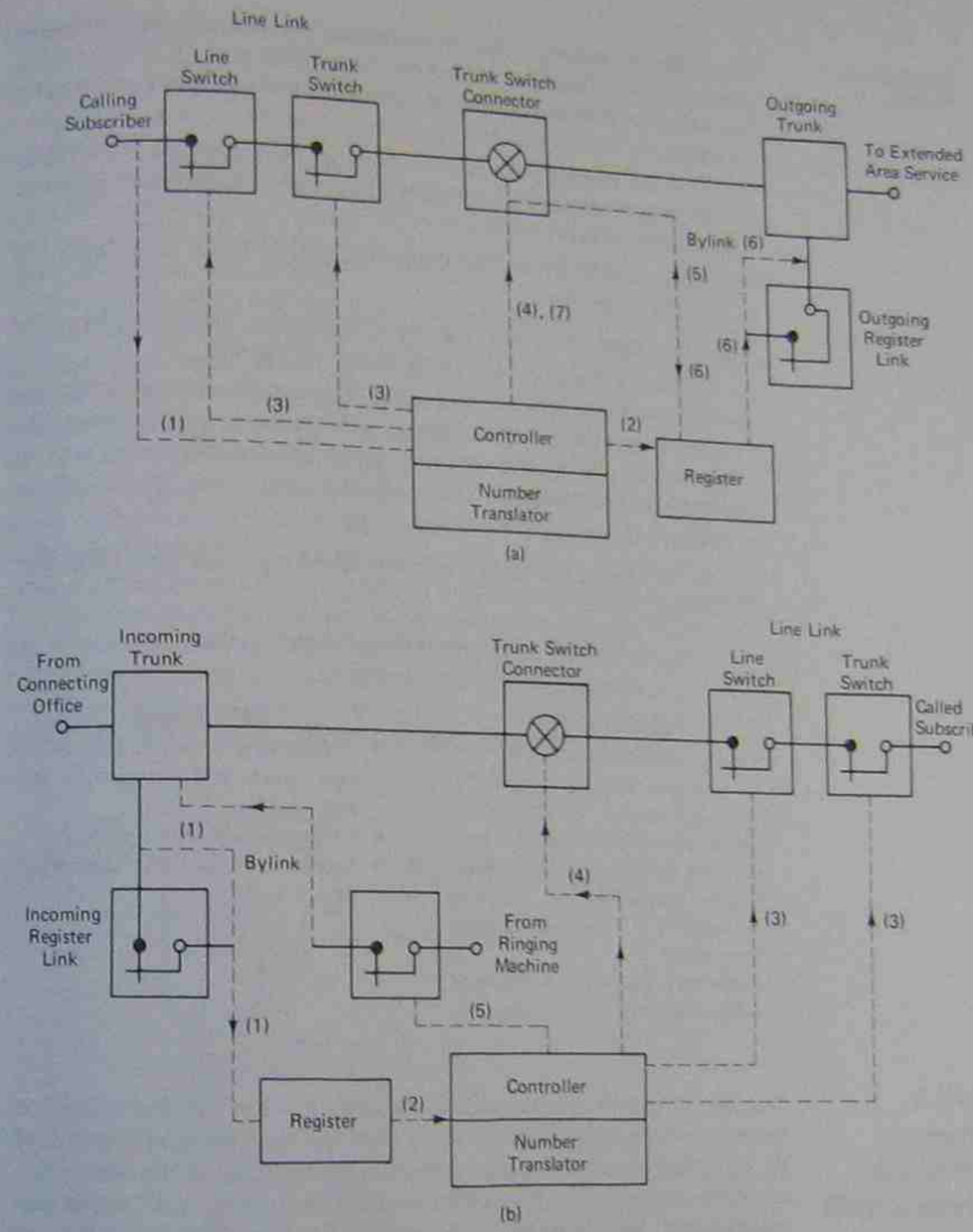


Fig. 2-36 Establishment of an (a) Outgoing Trunk Connection, (b) Incoming Trunk Connection

traffic, for instance, has first preference when competing with other types of traffic. The trunk with the highest priority is served first when it requests connection with other traffic arriving at the same time.

2. After the register records the incoming digits (typically four digits), the register seizes the controller and transmits the number and trunk location to the controller.
3. The controller locates the called line and performs a line-busy test on the called line.
4. If the called line is idle, the line link between the trunk and called line is connected via the trunk switch connector.
5. The controller sets up the proper ringing to the called subscriber line. When the controller has completed the ringing function, the controller, the register, and the register link are released to wait for another incoming call. The main line connection remains under control of the incoming trunk.

For the relative merits of common control cross-bar switching, refer to Table 2-7.

Table 2-7 Relative Merits of Common Control Cross-bar Switching

Advantages	Disadvantages
• Faster switching than S × S. • Only a few common control circuits are needed for connection set-up by a given office—typically six for a 10,000 line office. • Lower probability of blocking than with S × S because switches look ahead. • Elementary diagnostic techniques are available for fault-finding.	• A fault in a common control circuit causes the removal of several circuits. • Relatively high start-up cost. • More complex to understand than S × S.

2–7.3 Electronic Switching Systems (ESS)

In order to handle the increasing volume of traffic on the telephone system, it is necessary to move away from the slow switching speeds of the electromechanical Strowger and cross-bar switches to the more modern electronic switch, which uses computer technology with stored program control (SPC). A system such as the Bell Laboratories No. 1 ESS can handle up to 100,000 subscriber lines. In this and similar ESS's, a set of software programs instructs a high-speed processor on establishing a connection. A stored program resides in a semi-permanent memory that cannot be erased by circuit errors or equipment malfunction. It contains an instruction set which informs the processor what actions to take under a given system condition. It also contains such information as the location

of each subscriber and trunk line, class of service, alternate trunk routes, and so forth. In order to change this information, which is typically stored as magnetized or demagnetized spots (1's and 0's) on an aluminum sheet, the memory card can be temporarily removed and altered by a separate machine.

In addition to the semi-permanent memory, a temporary ferrite sheet read/write memory keeps track of such transient actions as the state of the subscriber and trunk lines, the digits dialed, and so forth. It is wiped clean after a call is completed.

A very elementary block diagram of an ESS is shown in Fig. 2-37.

To determine whether a subscriber is ON or OFF HOOK, a saturable-core ferrod sensor is used in the line circuit. It consists of two identically-wound solenoid coils around a ferrite rod. When current flows in the line and through the solenoid coils, the ferrite saturates. In addition, a pair of sensing wires are threaded through two holes drilled in the ferrite rod. When an interrogate pulse is sent down one of the sensing wires, with no dc current flowing in the line, a readout pulse is detected in the other sensing wire. If a dc is flowing in the line, the ferrod is saturated, no transformer action is possible, and no readout pulse is detected. The scanners scan these ferrods to determine the line and trunk status, and this information is stored in the temporary memory. Each ferrod sensor is scanned about 10 times per second or every 100 ms.

The crosspoint switches employed in the switching network are similar in character to the cross-bar switch, except that they employ much

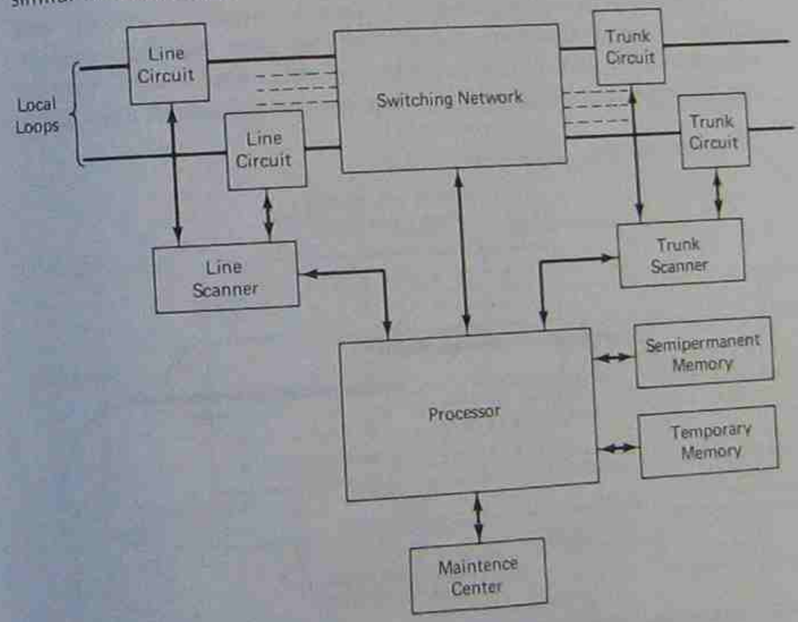


Fig. 2-37 ESS Block Diagram

faster reed switches called ferreeds. One such sealed reed switch is shown in Fig. 2-38.

The two reeds, made from a magnetic material, are firmly supported at the opposite ends of a nitrogen-filled glass tube. A small gap is normally maintained at the free ends. The glass tube is surrounded with a coil winding which, when energized, causes an induced N-pole at the free end of one reed, and an induced S-pole at the free end of the other reed. The free ends are therefore magnetically attracted, causing closure of the relay contacts. They can operate in the order of a fraction of a millisecond. Rows of ferreed switches are stacked vertically to form a switching block in the switching network.

The central processor or control is much faster than any of the external switches in the network since it must translate all the information given to it by the line and trunk scanners, memories, and maintenance center into a program of action. It then forwards this to the same equipment. Most of the instructions given to the processor are related to maintenance rather than call processing in order to assure a high standard of reliability. To achieve the generally accepted standard of no more than two hours outage in a 40-year time span, much time is spent on diagnosing faults. As soon as a fault occurs, an alternative working configuration must be found and the faulty unit taken out of service. A diagnostic routine is used to pinpoint the exact fault location. The technician makes the repair and rechecks the unit with the diagnostic routine. After passing the test, the unit can be put back into service.

Some electronic manufacturers, such as Northern Telecom and Nippon Electric Company, have combined the electromechanical cross-bar switch with a stored program common control system. A central processor still controls the system and determines the actions to be taken.

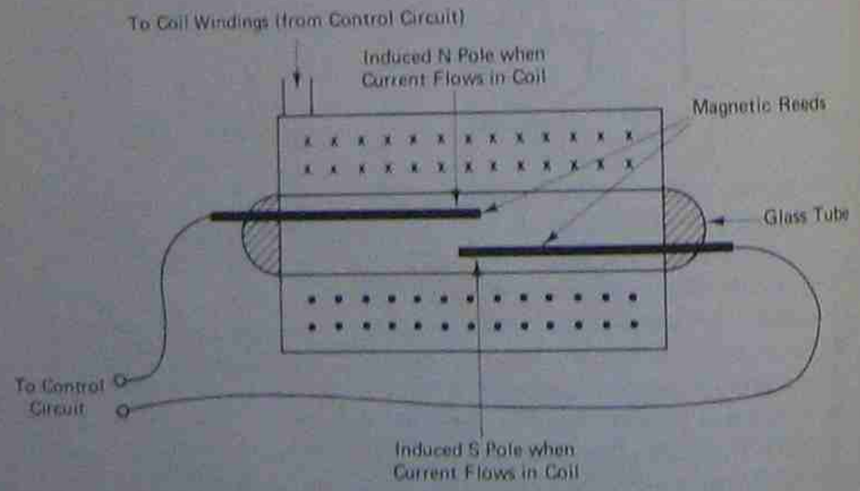


Fig. 2-38 Sealed Reed Switch

The implementation of these actions is carried out by the electromechanical line and trunk link networks, originating and terminating junctors, markers, and so forth. The marker operates and releases the network crosspoints under the order of the central processor unit. It does not make any decisions of its own as in the common control cross-bar system.

The block diagram of the Northern Telecom's SP-1 ESS system, which employs this technique, is illustrated in Fig. 2-39. The signal distributor is similar in function to the line scanner. It takes orders from the central processor to perform such actions as operating a trunk relay and informing the processor of any originating trunk calls. The Link Network is a cross-bar switching matrix which concentrates the incoming traffic via

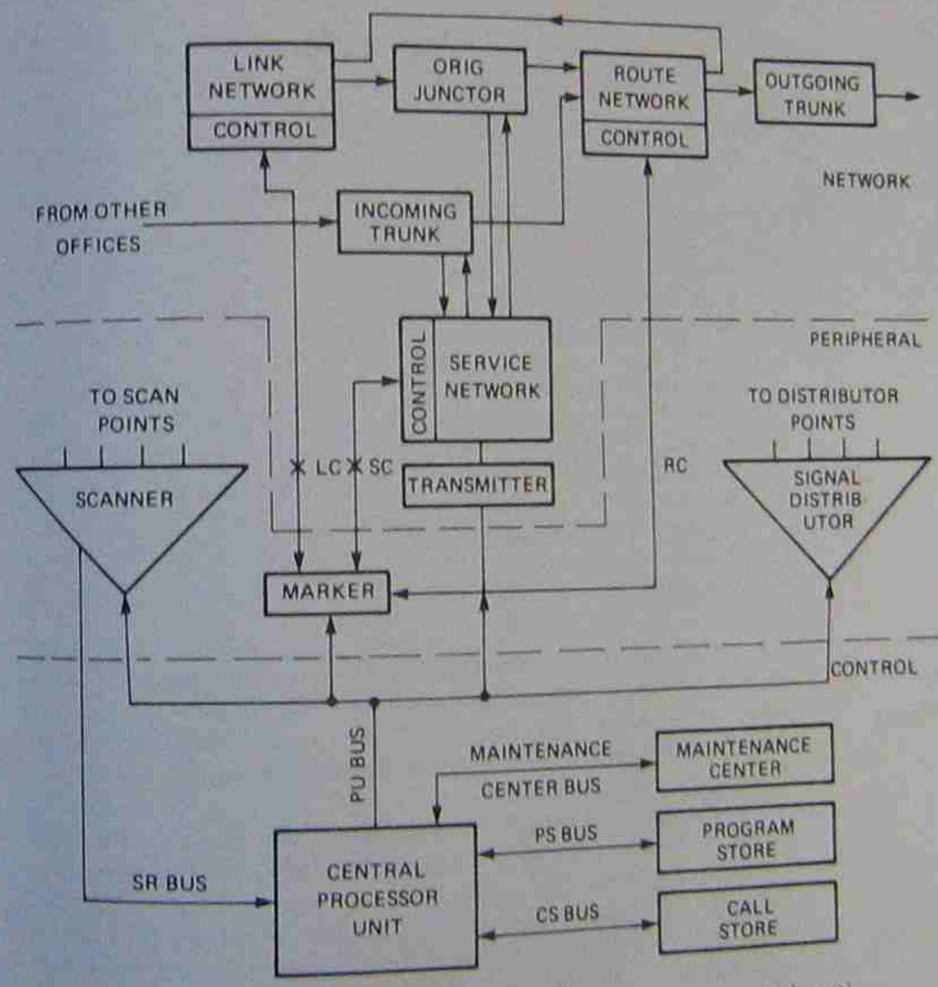


Fig. 2-39. SP-1 ESS Functional Blocks (Courtesy Northern Telecom)

the originating junction to the Route Network cross-bar switching matrix. The originating junction provides the access to the route network and to the service network for calls originating from subscriber lines via the link network.

For the relative merits of Electronic Switching refer to Table 2-8.

Table 2-8 Advantages and Disadvantages of ESS

Advantages	Disadvantages
• Up to 1000 times faster to establish a connection than by electromechanical system. Once digits have been dialed, digital switches and control logic operate in fractions of microseconds. • Automatic routing and fault-record producing facilities. • Extensive diagnostic routines. • Offers many classes of service, e.g.: a. Rerouting of a dialed call to another number (call forwarding). b. Modification of the way in which a call is processed; e.g., coin box call collection and restriction on PBX lines to local calls. c. Modification of the way in which a call is charged; e.g., free calls and wide area dialing. d. Special dialing codes; e.g. abbreviated dialing, call forwarding, permission of toll calls. • Small space requirements. • High reliability.	• Expensive for small system. • Catastrophic in case of failure.

2–7.4 Digital Switching

In an analog switch such as a cross-bar, a physical connection or path is established between the input and output for the duration of a call. The advent of solid-state switches which can make connections at rates exceeding millions of times per second has made it feasible to interleave connections in both space and time.

If a message is sampled often enough, which according to the Sampling Theorem must be at a rate of at least twice the highest frequency component in the message, no information is lost in the sampled signal and the original signal can be accurately reconstructed from the sample. In an elementary pulse code modulation (PCM) system, such as that illustrated in Fig. 2-41, an analog-to-digital converter translates each sampled amplitude into a binary coded 8-bit word. These 8 bits are trans-

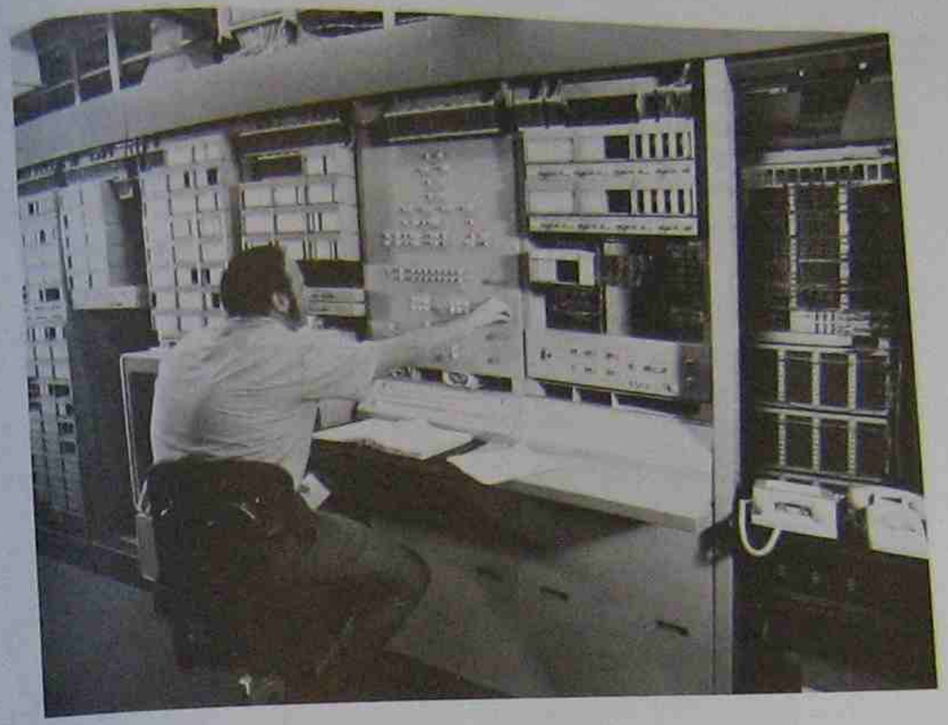


Fig. 2-40 ESS No. 1 Display and Master Control Panel, Plymouth, Michigan. (Reproduced by permission of A. T. & T Co.)

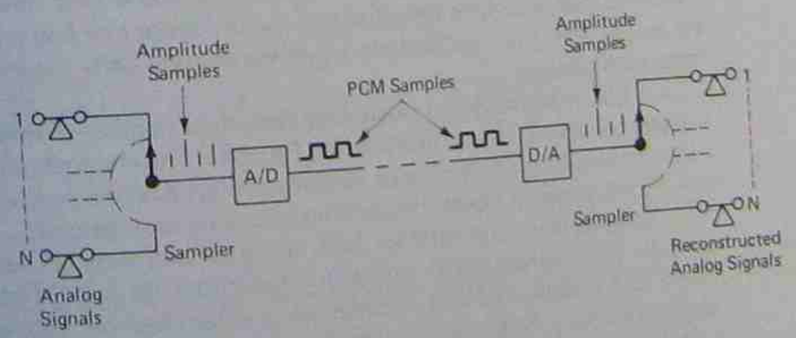


Fig. 2-41 A Basic PCM System

mitted in a small time interval called a time slot. For voice transmission in which the bandwidth is limited to a range of 300 Hz to 3.4 kHz, a sampled rate of 8 kHz is used, which slightly exceeds the minimum sampling rate of $2 \times 3.4\ k = 6.8$ kHz. The time taken to sample all of the N inputs, or the collection time for all N inputs, is called a frame, as illustrated in Fig. 2-42. This will be dealt with in much more detail in Chap. 6, Sec. 6–4.

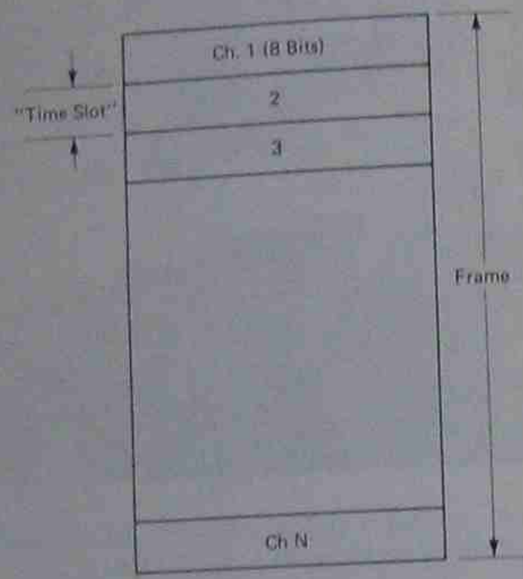


Fig. 2-42 N-Channel Frame with a Nominal 8 Bits per Channel

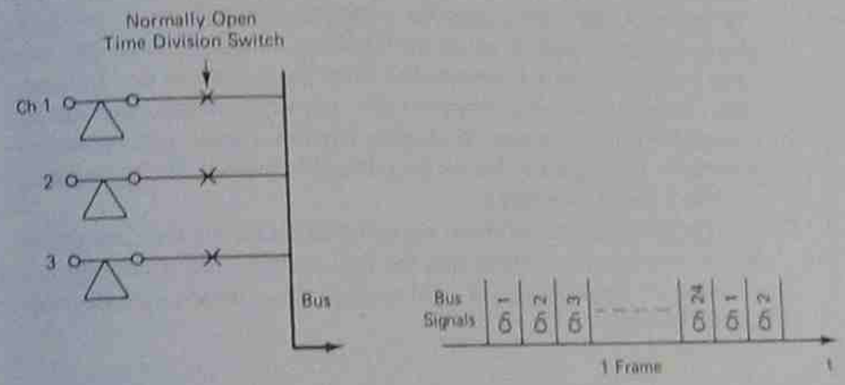
Two types of switches are used in digital switching—space and time. In space switching, each input time slot is switched to the correct output. For the brief time slot interval, the proper connection between the input and output is held. Such a space switch could well be called a time-multiplexed space switch. In time switching, the 8 bits within each time slot are stored in memory and read out at some later time. Time switching allows for time delays so that the processor can find appropriate empty time slots, particularly when transferring calls between two buses or highways.

A bus or highway is a single line carrying a number of time interleaved channels (TDM), as illustrated in Fig. 2-43. The channels are cyclically sampled by the time division switch and placed on the highway. Figure 2-43(b) shows a typical consecutive sampling arrangement.

To understand the need for both space (S) and time (T) switches, consider the requirement to connect a subscriber on one 24-channel single bus system, called a module, to a subscriber on another 24-channel module. As an example, see Fig. 2-44 and assume that Channel 6 in Bus 1 must be interconnected with Channel 9 in Bus 2.

To accomplish this connection, it is necessary to find an empty or idle pair of matching time slots on the two buses. If, for instance, Channel 6 on Bus 2 is busy, connecting Bus 1 to Bus 2 during the Channel 6 time slot interval would cause interference between the Channel 6's in the two modules. For this example, assume that Channel 20 is idle in both modules. The following operations could then be followed to establish a suitable linkage.

1. Delay Channel 6 in Bus 1 for 14 channels, so information resides in Time Slot 20. (Time switching)
2. Connect Bus 1 to Bus 2 in Channel 20. (Space switching)
3. Delay Channel 20 for 13 channels; $20 + 13 = 33 - 24 = 9$ in a 24-channel system (Time switching). This assures that information emanating from Bus 1 Channel 6 is coincident in time with Bus 2 Channel 9. The arrows in Fig. 2-44 illustrate the switching procedure.



(a) Time Multiplexed Bus or Highway

(b) 1 Bus Frame

Fig. 2-43 Illustration of Bus or Highway

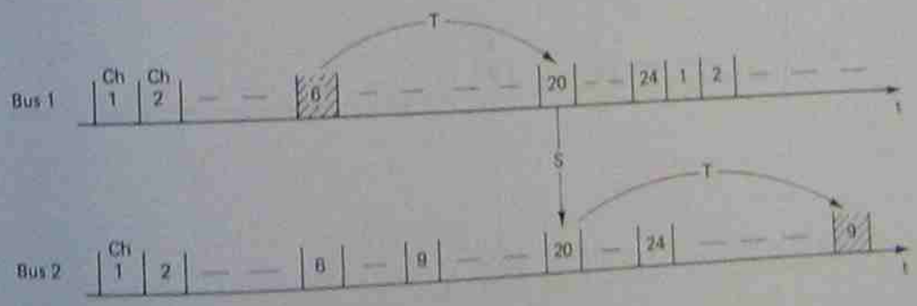


Fig. 2-44 Time-Space-Time (TST) Switching

Since this method has the restriction that both idle slots must occur at the same time during the same frame, it is easy for a blockage to occur. A more flexible approach is to add a buffer to each bus so that an intermediate message can be processed and delayed. This can be illustrated by using an example similar to the previous one in which Channel 6 in Bus 1 is trying to communicate with Channel 9 in Bus 2. See Fig. 2-45.

1. Connect Channel 6 in Bus 1 to Buffer Time Slot 6. (Space switching)
2. Delay information in Buffer Time Slot 6 to Time Slot 9. (Time switching)
3. Connect Buffer Time Slot 9 to Bus 2. This is also where Channel 9 resides in time.

To visualize the operation of the buffer store and to understand how time delay can be accomplished, consider a buffer with 24 time slots, each of which can hold 8 bits. Even though the incoming messages are stored in a certain cyclic sequence, they can be read out in a different cyclic sequence, depending upon the switching instructions. Figure 2-46 shows such an example in which the incoming Message 1 is written into Time Slot 1, Message X is written into Time Slot X, and so on. The contents can be read out in a different cyclic sequence which need not be consecutive and may appear to be quite random. Delays can be readily obtained by reading out a channel on a delayed time slot, such as reading Channel X out in Time Slot Y.

For simplification purposes, we will omit including the connection store memories when symbolizing the time switch. The following is a more detailed block diagram of a TST configuration, which includes both T and S switches.

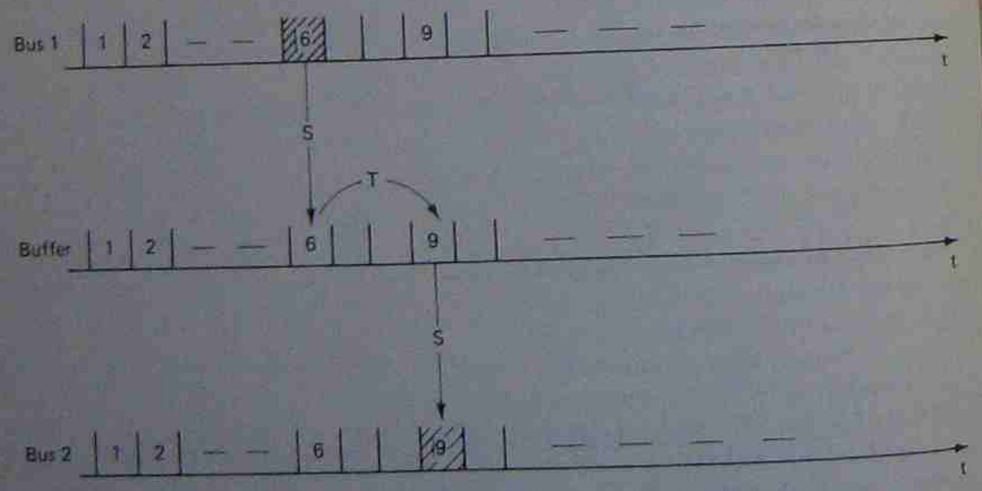


Fig. 2-45 Space-Time-Space (STS) Switching

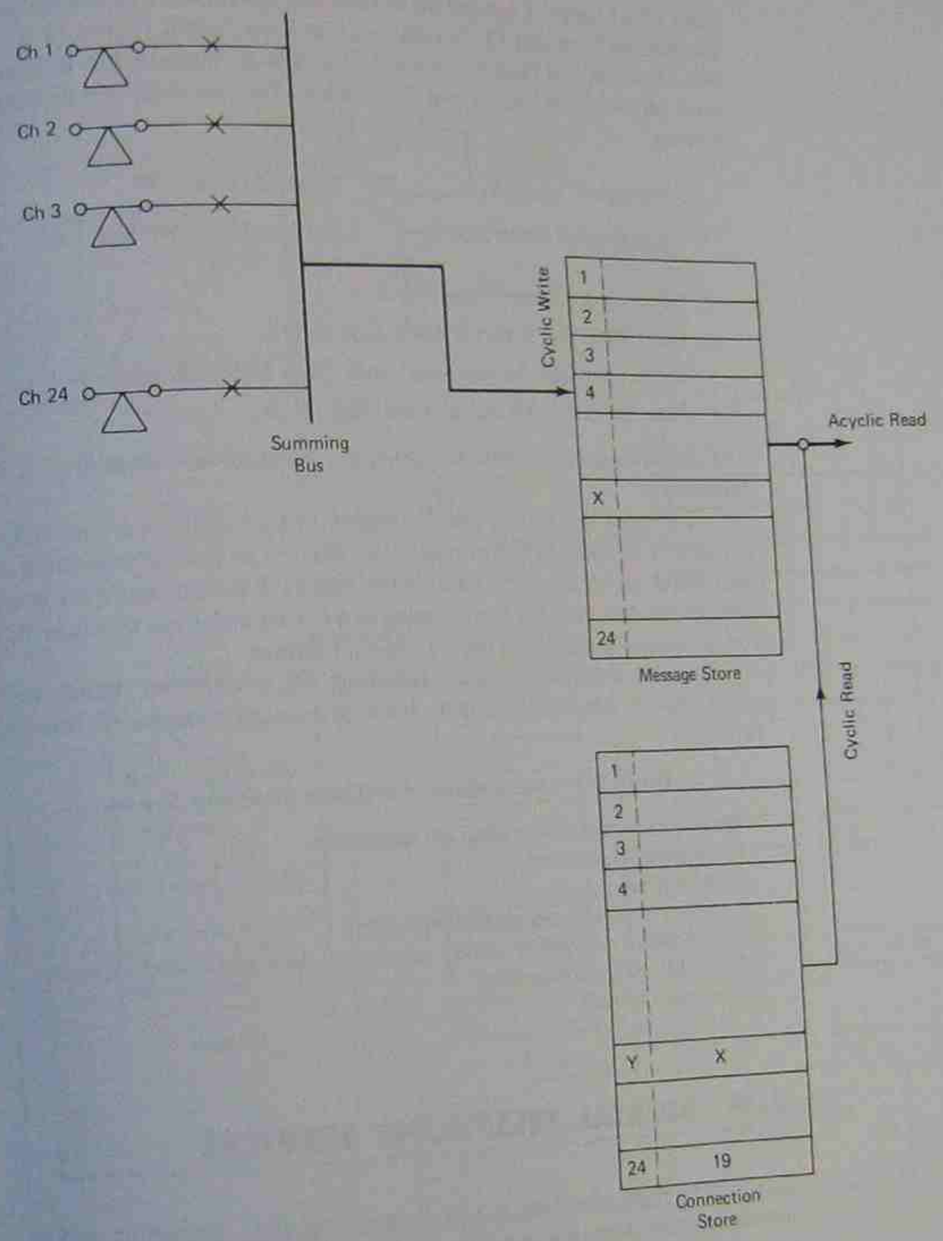


Fig. 2-46 Operation of a Time Switch

In the block diagram shown in Fig. 2-47, assume Line 15 in Module 1 wants to communicate with Line 23 in Module N. Also assume that Line 15 in Module 1 is assigned to Time Slot 3 and Line 23 in Module N is assigned to Time Slot 13. As illustrated on Time Switch 1, Line 15 is read into Time Slot 3. If Time Slot 3 and 13 are busy on network Bus B, Line 15 must be transferred to an idle slot such as 19. The steps are written as follows:

Module 1 Line 15	TO	Module N Line 23
(assigned to Time Slot 3)		(assigned to Time Slot 13)

1. Write Line 15 into Time Slot 3.
2. Read Time Slot 3 out at Time Slot 19 (T).
3. Write Time Slot 19 Module 1 into Time Slot 19 Module N (S).
4. Read Time Slot 19 out at Time Slot 13 (T).

The digital-to-analog converter sends the message to Telephone 23 in Module N.

The input signal may also be digital. In such cases, no conversion circuitry is needed. Synchronization is required to permit separation of the control signals from the information signals. It is also necessary to assure proper data rates for multiplexing to a higher order bus structure that has an integral multiple of the 24-channel format.

The advantages of digital switching are impressive, particularly when a system goes totally digital. Some of these advantages are listed in Table 2-9.

Table 2-9 Advantages of a Digital Switching System

Easier to maintain than analog; self diagnostics.
Requires less floor space.
Dissipates less power.
Handles more traffic per equipment frame.
Can multiplex a variety of sources; e.g. video, data, audio, and facsimile.
Allows for ready cryptography.
Eliminates crosstalk.

2–8 SPECIAL TELEPHONE SERVICES

To meet the many demands of voice, television, data, facsimile, and telemetry transmission, the common carriers have developed many communication services. The following tree shows some of the services offered by the Trans-Canada Telephone System (TCTS), a consortium of the major common carriers in Canada. Although other systems may use

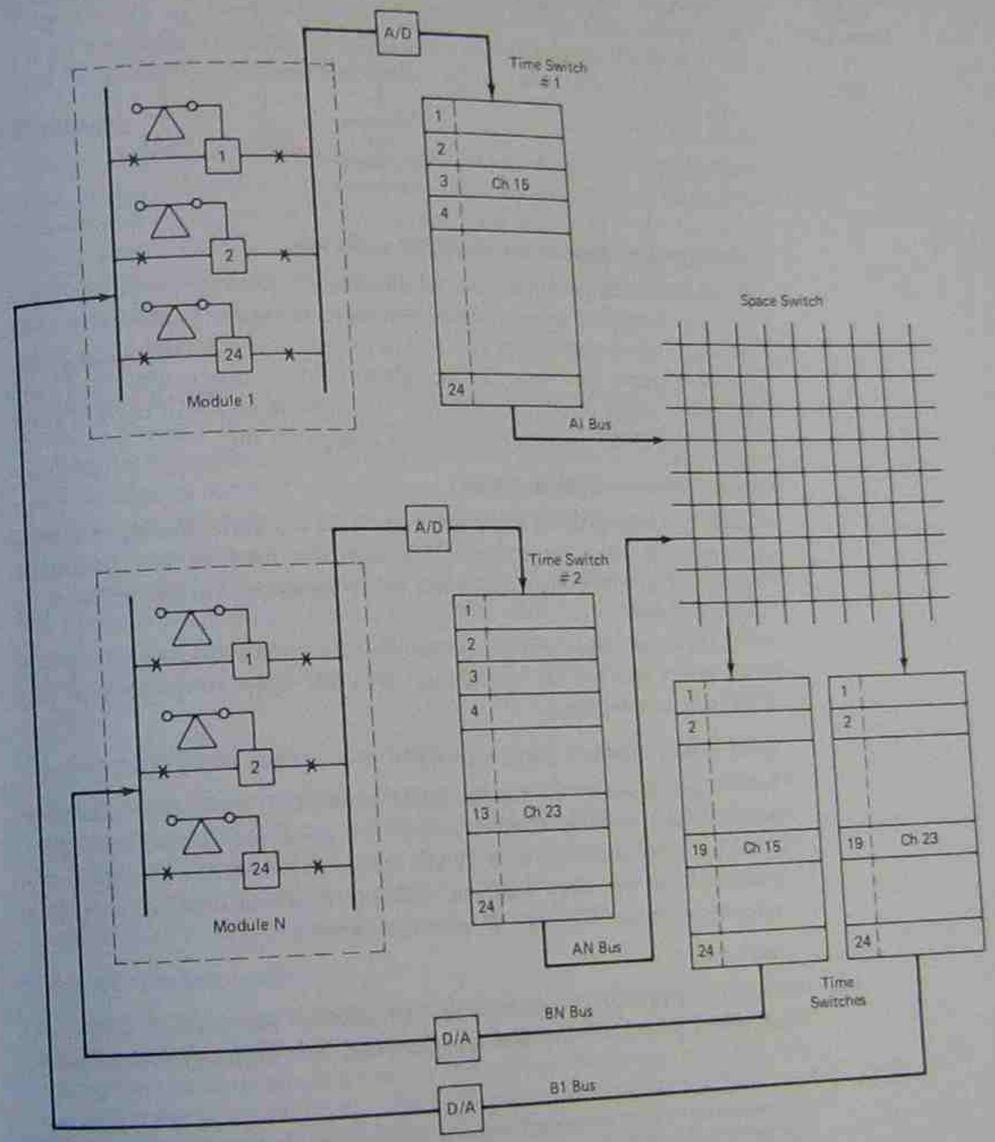
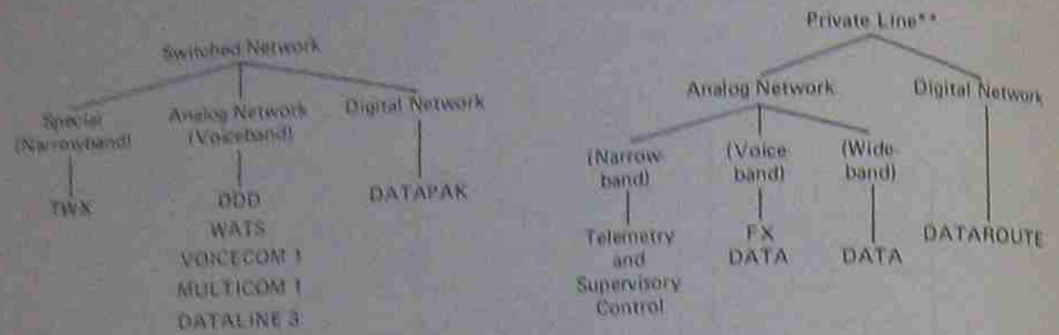


Fig. 2-47 A TST Block Diagram

slightly different descriptive terms, the services listed are representative of those offered by other communication companies.*



Teletype Exchange Services (TELEX and TWX)

Teletype Exchange Service provides direct-dial point-to-point, or conference, connections between teletypewriters. No voice transmission is involved. Data is transmitted at a 110 b/s rate. Either a Model 33 or 35 terminal is used—the Model 33 for light to medium usage, and the 35 unit for heavy usage (multi-copy forms). The eight-level ASCII code is used, which has seven information bits and one parity bit.

Direct Distance Dialing (DDD)

Direct Distance Dialing is the most common network offering. It is used primarily for voice transmission. Through this method, any telephone subscriber can be connected to any other telephone on the network by direct dialing.

Data sets allow for the transmission of data over the two-wire half-duplex facilities at speeds up to 1200 b/s, asynchronous and 2400 b/s, synchronous.

Wide Area Telephone Service (WATS)

WATS operates over the regular DDD network. It gives the subscriber access to any dialable telephone within a specified or restricted geographical area identified as a WATS zone. A flat rate is charged, depending upon the zone size and whether it is full-time service or a measured-time service, e.g. 10 hours per month.

Voicecom 1

Voicecom 1 provides voice-only service between specified locations on the DDD network. Connection is established by a preprogrammed automatic dialer. Only one digit needs to be dialed. An exclusion feature can be added to prevent access to the line by any other extensions.

* For more information about communication services offered by common carriers, contact personnel in the computer communications group in your area.

** A private line consists of a channel or circuit supplied to a subscriber for his exclusive use.

Multicom 1

Multicom 1 provides voice and data communication between specified locations on the DDD network. As in the case of Voicecom 1, connection is established by a preprogrammed automatic dialer. Data sets are required for data (up to 2400 b/s). The voice feature is generally used to ensure the proper establishment of a connection.

Dataline 3

Dataline 3 provides data transmission from remote originating data stations to a centralized data terminal. The central terminal, such as a time-shared computer, may process the data and return it to the originating station. Calling is permitted only from the remote originating data station. Automatic dialing is provided. Bit rates of 2400 b/s, synchronous and 1200 b/s, asynchronous are permitted.

Datapak

Datapak is a nationwide, public, packet-switching network. It consists of switching nodes at various centers interconnected with high speed 56k b/s digital transmission trunks. Transmission facilities are only allocated after a data or control package has been assembled. Normal service handles data packets of up to 256 octets (2048 bits) and provides an average network transit delay of 0.49 s.

Non-packet-mode terminals such as teletypewriters are connected to Datapak by Network Interface Machines (NIM's). Packet-mode terminals use the Standard-Network-Access Protocol (SNAP). SNAP establishes clears calls and, in general, manages the flow of data along the network.

Telemetering and Supervisory Control

These channels are designed to transmit meter readings from remote locations to a central point (telemetering). If in addition, a control signal can be sent out to the remote station from the central station to change some operation, supervisory control is possible. Typical pulse rates of 30 and 55 baud – dc pulse transmission are employed.

Foreign Exchange (FX)

Foreign exchange provides a dedicated line between a customer location and a remote telephone exchange. Toll-free calls can be made to any telephone subscriber attached to the remote exchange, and any telephone subscriber can contact the customer's location toll-free. Voice and data or both can be transmitted at a rate of up to 1200 b/s, asynchronous and 2400 b/s, synchronous. This is a very economical method of transferring information when a customer has many calls to make to subscribers in a remote exchange area.

Data

Voice-grade data channels on a private line can transmit up to 9600 b/s, depending upon the conditioning of the channel. Two- or four-wire circuits can be made available and data sets are required. More than two stations located in different exchanges can be interconnected (multipoint), or more than two stations within the same exchange can be interconnected (multidrop).

Wide-band data channels can transmit at a rate of up to 50k b/s and higher. Full-duplex operation is used and a separate voice-grade circuit is used for voice contact, supervision, and so forth.

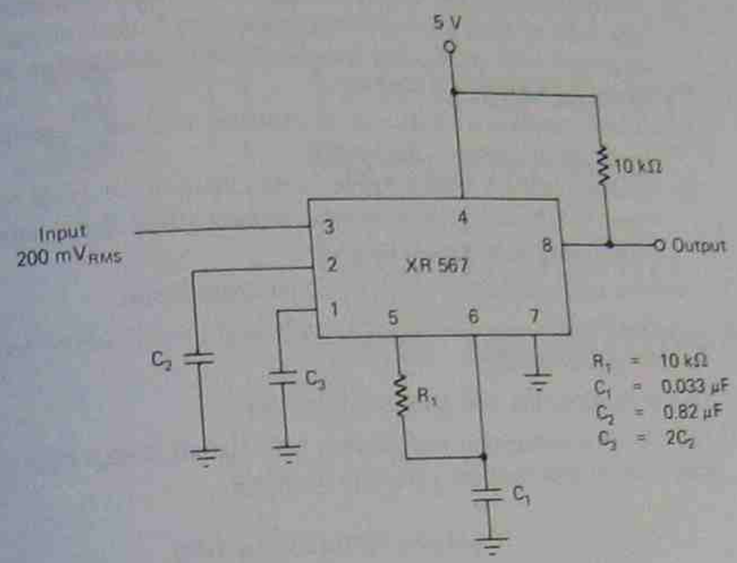
Dataroute

Dataroute on a private line provides full-duplex data transmission at rates of up to 50k b/s. It can provide multipoint and multidrop service. (See Data). Voices cannot be carried. For customers outside the Dataroute serving area, access can be made through a dialed-up or private line. Such would be the case if the subscriber was located outside of Victoria or Vancouver in British Columbia.

PROBLEMS

1. a. What are the two major advantages of switching networks?
   b. For direct interconnection of 10 telephones, how many lines would be required?
2. a. Sketch out the switching center hierarchy.
   b. What is a PBX?
3. What range of frequencies is usually considered the standard for telephone transmission.
4. What are the chief differences between loop and trunk circuits?
5. What are the major functions of the following in telephony?
   a. sidetone
   b. varistor
   c. rotary dial
   d. hook switch
   e. capacitor bridge or repeating coil
   f. four wire terminating set
6. What is the difference between touch-tone signaling and dial pulsing?

7. a. Determine the center frequency and capture range of the decoder shown.



   b. If the input signal is at f_o, what is the output voltage level? If the input frequency is outside the capture range of the decoder, what is the output voltage level?
8. What is in-band and out-band signaling, and what is the disadvantage of each?
9. Define the following characteristics of a transmission line:
   a. characteristic impedance
   b. distributed line constants
   c. attenuation constant
   d. phase constant
   e. neper
   f. phase velocity
10. How does the transmission line loss vary with frequency?
11. a. How is a line physically loaded?
    b. State two significant advantages of line loading.
    c. What affect does line loading have upon propagation velocities?
    d. What affect does lumped loading have on the attenuation characteristics of a line?
12. a. If a subscriber's loop length is limited by a maximum signaling resistance of 1300 Ω, what maximum length (in km) can be

achieved with AWG #24 and #19 lines? Assume a telephone handset resistance of 50 Ω.

b. If a subscribing loop length is limited by the signal attenuation, say 8 dB at 1000 Hz, what maximum length (in km), can be achieved with #24 ASM and #19 DNB non-loaded wire? (Refer to Table 2-3.)

c. If both conditions in (a) and (b) must be met, what condition limits the maximum loop length?

d. If H88 loading is added to the cable circuit of (b), what maximum length (in km) is achievable when 8 dB of signal attenuation is permitted at 1 kHz?

13. List three cable faults that result in signal degradation.

14. Describe the difference between a distributed carrier system and a lumped carrier system.

15. Define multiplexing and give its advantages.

16. The following expression represents a DSB signal. Sketch both the time domain and frequency domain response.

$$e = 10 \cos (2\pi\ 10^3 t) \cos (2\pi\ 10^6 t)$$

17. a. What is the function of a balanced modulator?
    b. Give a block diagram which employs the filtering technique in generating SSB.
    c. What is the purpose of a guard band between channels?

18. a. What is the standard group size?
    b. How many channels does a jumbo group carry?
    c. What is the maximum oscillation frequency drift permitted by the CCITT in carrier transmission?
    d. Give three uses of a pilot signal in carrier transmission.

19. List three major types of telephone switching equipment and give the major advantages and disadvantages of each.

20. a. What is a space switch?
    b. What is a time switch?
    c. What is a bus, or highway, in telecommunications jargon?

21. List the advantages of digital switching.

3

Telephone System Performance Requirements and Impairments

3–1 INTRODUCTION

In order for the telephone systems to satisfy customers, certain performance criteria must be met. This chapter will consider some of the fundamental transmission objectives that must be met for the systems to operate properly. It will also consider some reasons for signal-to-noise degradation as the signal passes through the system.

Most telephones are connected to the local switching center by means of a two-wire circuit. Both the received and transmitted signals travel over the same electrical transmission path within the same frequency band. If any gain needs to be added to compensate for transmission loss on the cable, it is not feasible to use a bidirectional amplifier, as shown in Fig. 3-1, since such a configuration must use a gain of less than unity for stable operation.

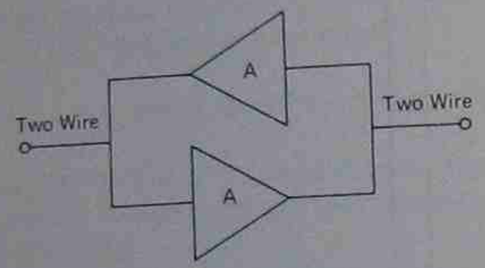


Fig. 3-1 Bidirectional Amplifier

3–1.1 Four-Wire Terminating Set

One technique to obtain reasonable gain of up to 10 dB on two-wire trunks is to use a negative impedance repeater. Such a regenerative feedback amplifier causes an effective negative resistance to appear across the line, resulting in a net power gain. It readily passes dc signaling, maintains a signal path in case of failure, and tends to be inexpensive. Generally, not more than one can be used on a route.

For transmission beyond the local subscriber loop, the receive and transmit paths are usually separated to give two one-way paths which result in a total of two wire pairs, or four wires. The device designed for such a splitting and recombining effort is called a four-wire terminating set, or hybrid transformer. It employs two cross-coupled transformers with identical windings, as shown in Fig. 3-2.

A signal applied to the RETURN line will generate equal voltages across each winding, but because of the antiphase in the upper windings, nothing will appear in the GO line. Equal signals will appear in the two-wire line and balance circuit. Similarly, any signal applied to the two-wire line will split evenly between the GO and RETURN lines. The portion appearing in the RETURN line will be absorbed in the output stages of the RETURN amplifier, whereas the signal appearing in the GO

line will be passed on by the GO amplifier. This division of signals causes the signal appearing in the GO amplifier to experience a 3 dB loss. Similarly, a signal applied to the RETURN input undergoes a 3 dB loss when appearing at the two-wire line port. Because of the non-ideal nature of transformer windings, an additional 0.5 dB insertion loss is experienced in practice, resulting in an overall 3.5 dB signal loss as it traverses through a hybrid transformer. In actual practice, the hybrid is also non-symmetrical. The loss is higher from the return to the 2 W line than from the 2 W line to GO.

To achieve negligible feedthrough from the RETURN to GO lines, the impedance of the balance network must be made equivalent to the characteristic impedance of the two-wire line. However, because a line characteristic impedance tends to be frequency-dependent at the lower frequencies and because of the variations between the many local loops that may be switched to the hybrid, it is impossible to achieve perfect balance with a fixed balance network. This results in some signals being coupled between the GO and RETURN paths. The loss experienced between the GO and RETURN paths is called balance-return loss and is given by:

$$B = 20 \log \left| \frac{Z_N + Z_L}{Z_N - Z_L} \right| \qquad (3\text{-}1)$$

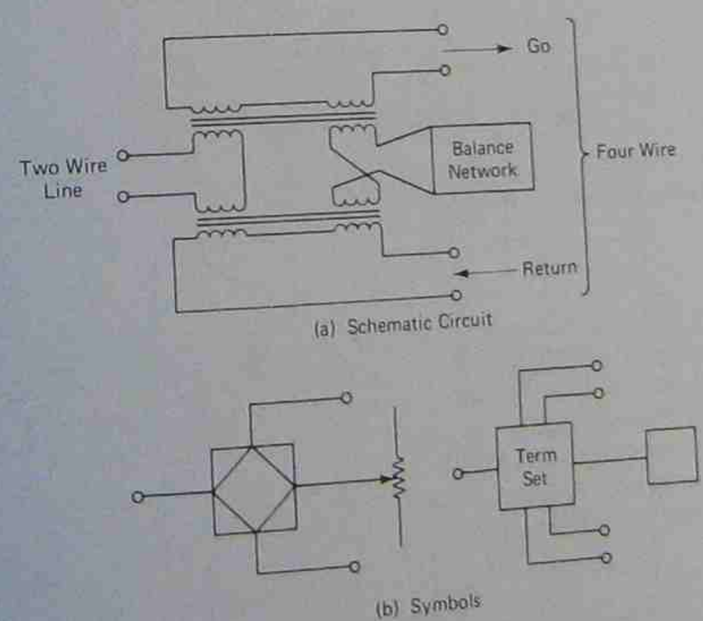


Fig. 3-2 Four-Wire Terminating Set

where Z_N and Z_L are the balancing network and line impedances respectively. This equation ignores any imbalance in the transformers. Note that by careful matching where $Z_N = Z_L$, the balance-return loss goes to infinity. In fact 60 to 70 dB is about the highest practically available. Typically, B is in the order of 50 dB. If the line appears as an open or short circuit, then B goes to 0 dB.

Figure 3-3 illustrates the losses through a hybrid transformer.

When considering the stability of a network, the balance-return loss is taken across the entire frequency band from 0 to 4 kHz. The minimum value of B tends to govern the stability. When dealing with echos, however, an average value of B (B_e) over the 500 to 2500 Hz frequency range is used.

A complete long-distance telephone network is illustrated in Fig. 3-4. The dashed section can be made up of a multiplexed microwave system, a multichannel carrier on a cable, etc.

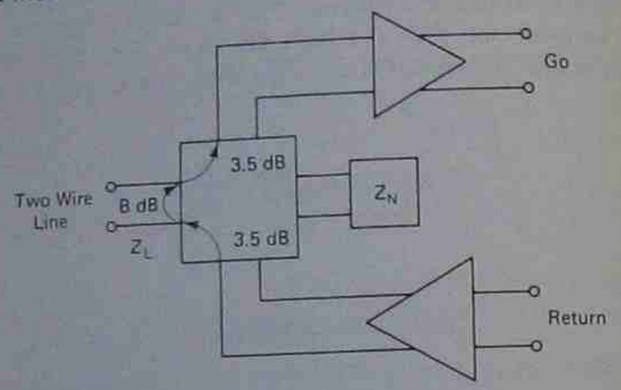


Fig. 3-3 Losses Within a Hybrid Transformer

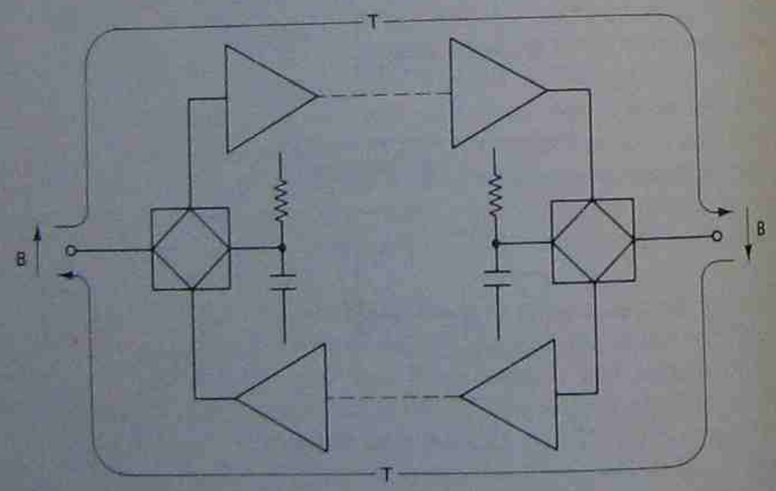


Fig. 3-4 Typical Telephone Network

3–1.2 The Reference Equivalent

In setting up a system of international standards that will allow different national systems to be interconnected at a level of quality satisfactory to customers, it is necessary to:

a. Define the sending and receiving levels.
b. Allocate maximum permissible noise contributions.
c. Allocate losses and propagation times.

To aid in achieving these objectives, the International Telecommunications Union has a reference system in Geneva called NOSFER (Nouveau Système Fundamental pour la détermination des équivalents de Reférence) against which telephone companies can compare their telephone subsets. The rating against this NOSFER standard is known as the reference equivalent. The transmitter and receiver can be rated separately and are called:

Transmit reference equivalent, tre

Receive reference equivalent, rre

A positive reference equivalent indicates that the circuit is less efficient or more objectionable to a subscriber than the laboratory standard. Adding the tre and rre produces an overall reference equivalent, ore. CCITT recommends that the maximum limits for the reference equivalent on an international connection be:

20.8 dB tre

12.2 dB rre

for 97% of the connections made.

This leads to a maximum overall reference equivalent of 33 dB. With an ore of 32 dB, the British post office has found 9.7% of the subscribers to be unsatisfied. Thus, an ore of 33 dB would prove unsatisfactory to about 10% of its subscribers. Common carriers desire to maintain as low a reference equivalance as possible for a higher percentage of subscriber satisfaction.

3–1.3 Network Stability

Consider the total loss around the telephone network of Fig. 3-4. If T represents the loss between the two-wire end points, which includes the hybrid losses (typically 7 dB), the total loop loss will be $2T + 2B$ dB for symmetrical sending and return paths. If this loss is 0 dB, the gain of the network is 1, and the circuit will oscillate or "sing." Even with some loss, the circuit may still ring, much like talking into a hollow steel barrel. To avoid such a condition, it is desirable to maintain a total loop loss of at least 6 dB. Thus:

$$2T + 2B \geq 6 \text{ dB}$$

or

$$T \geq 3 - B \text{ dB} \qquad (3\text{-}2)$$

If a shorted, or open, condition is allowed to exist on the two-wire side of the hybrid so that B is 0, Eq. (3-2) states that the T loss must be at least 3 dB. In any real system, T varies considerably due to improper alignment and the effects of temperature on amplifier gains and line losses. Thus, T must undergo a loss considerably in excess of 3 dB to assure a low probability of singing. Even with the high loss required through the network to minimize singing, the ore obtained remains below the 33 dB minimum permitted. Of more concern, however, is the echo problem on long lines.

3–1.4 Echo

When a person speaks into a telephone system, an echo or reflected signal can occur if there is an impedance mismatch in the system. Such an echo can be extremely annoying and will worsen with increased delay and loudness. As shown in Fig. 3-5, the speaker's echo experiences a delay of $2(t_1 + t_2)$ seconds and a loss of

$$(B_e + 2T) \text{ dB}. \tag{3-3}$$

As the delay increases, more subscribers become dissatisfied with the telephone conversation. More attenuation needs to be added in the echo path in order to give improved customer satisfaction. It has been found that for a 20 ms one-way propagation delay, 50% of the customers are satisfied if 17.7 dB are added to the echo path. For a 50 ms one-way delay, 30.9 dB need to be added for 50% customer satisfaction.

As can be observed from these results, delays beyond 50 ms require even more attenuation, thus exceeding the 33 dB ore allowed. What can

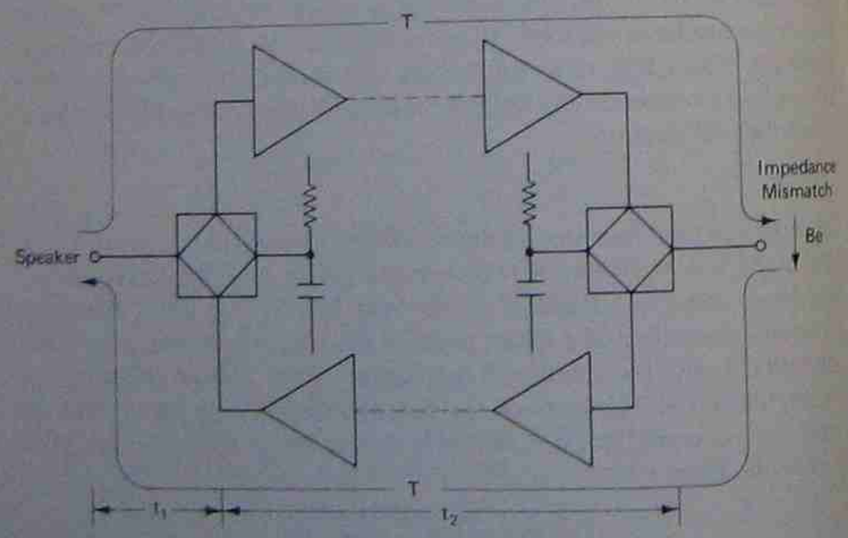


Fig. 3-5 Echo Path

be done in the case of a synchronous satellite link which has about a 270 ms one-way propagation delay? Echo suppressors are usually added to the system. Generally, echo suppressors are used for near round-trip delays exceeding 45 ms. These electronic devices are inserted in the reflected path to suppress the echo, and are activated by a speech burst from the remote end, as shown in Fig. 3-6. When the person speaks, the echo suppressor is activated, and a 60 dB attenuator is placed in the return line. If someone begins transmitting from the opposite end, the suppressor is removed (the break-in circuit is not included), and a similar echo suppressor is inserted at the opposite end.

Because a syllable or two of speech must be detected before an echo suppressor is turned ON or OFF, significant delays and some loss of signal are experienced when echo suppressors are reversed in direction. Since such delays are unacceptable with data communication, echo suppressors are disabled throughout such transmission.

3–1.5 Via Net Loss

In designing a transmission system, a national transmission plan is normally drawn up for each country. Such a plan sets the standards regarding singing, echo levels and delays, and overall reference equivalent, so that a given percentage of subscribers will give the connection a fair or better rating. In North America a variable loss plan (VNL) is employed to meet these objectives.

One method to meet the ore and still meet the singing and echo within tolerable limits is to increase the balance-return losses. Increasing the B terms in Eqs. (3-2) and (3-3) allows a decrease of the T terms, thereby improving the ore of the connection. To achieve larger balance-return losses requires improved matching at the hybrid. This can be done by more careful adjustment of the balance network and careful control of the subscriber loop. Despite taking all reasonable precautions, reflections

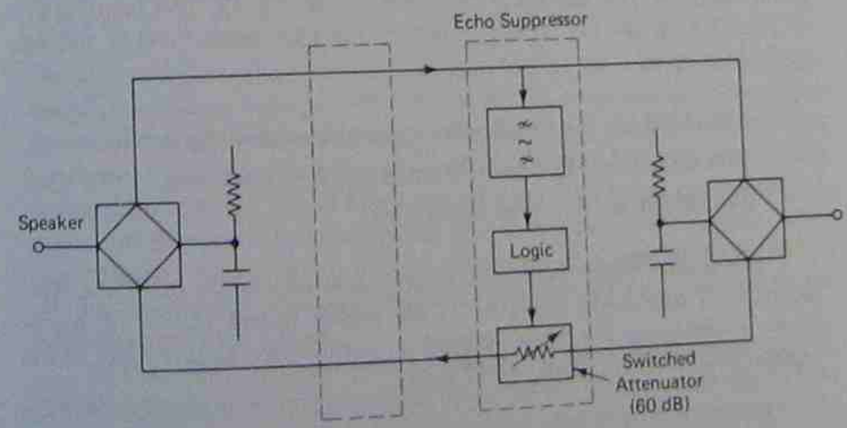


Fig. 3-6 Echo Suppressor

or echos are always present. Since echo suppressors are used for round trip delays exceeding about 45 ms, sufficient loss must be present in the loop to reduce the echo problem for delays less than this. The maximum loss in decibels at which a circuit can operate without objectionable singing or echo is known as Via Net Loss (VNL). In North America each toll trunk circuit is designed to give a loss of:

$$\text{VNL} = 0.1 \times \frac{2l}{v} + 0.4 \text{ dB} \tag{3-4}$$

where l = circuit length – km
v = velocity of propagation – km/ms
l/v represents the one-way path delay in ms
0.1 represents the number of dB loss that must be inserted per ms of delay to maintain tolerable echo level. For the echo round trip this must be multiplied by a factor of 2.

The one-way loss, or overall connection loss, between two end-offices consists of n links in tandem that can be extrapolated from Eq. (3-4).

$$T = 0.2\frac{l}{v} + 0.4n + 4 \text{ dB} \tag{3-5}$$

where T is the overall connection loss
n = number of links
l/v = total one-way path delay in ms

The additional 4 dB come from adding an extra 2 dB between the end-office and toll center to control singing when a round trip delay approaches 0 ms. This assures a minimum loss of 4.4 dB. The factor $0.2/v$ in Eqs. (3-4) or (3-5) is known as the via net loss factor (VNLF) and has typical values of

0.0015 dB/mi for a high velocity carrier system (L-carrier),
0.03 dB/mi for H-88 loading on #19AWG wire, and
0.01 dB/mi for two-wire open wire.

Consider a single trunk which has a one-way delay of 11 ms. This results in a round trip delay of 22 ms which leaves (45 – 22) ms or 23 ms of delay for the other intertoll and end-office to talk links in the system. The VNL in a delay of 22 ms is

$$\text{VNL} = 0.1 \times 2 \times 11 + 0.4 = 2.6 \text{ dB}$$

The corresponding intertoll length (one-way) for an L-carrier link can be calculated by substituting the VNLF for the $0.2/v$ terms in Eq. (3-4).

$$l = \frac{\text{VNL} - 0.4}{\text{VNLF}} = \frac{2.6 - 0.4}{0.0015} = 1467 \text{ mi}$$

Figure 3-7 shows a simplified North American transmission plan. Note that all end-office to toll center trunks operate with a minimum loss of 2 dB. To maximize the balance-return loss, these trunks also are carefully matched to the toll exchange.

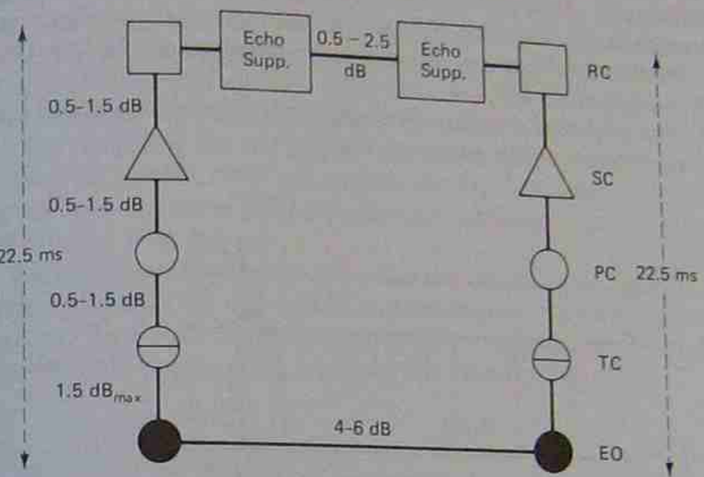


Fig. 3-7 Simplified North American Transmission Plan

3–2 TELEPHONE SYSTEM IMPAIRMENTS

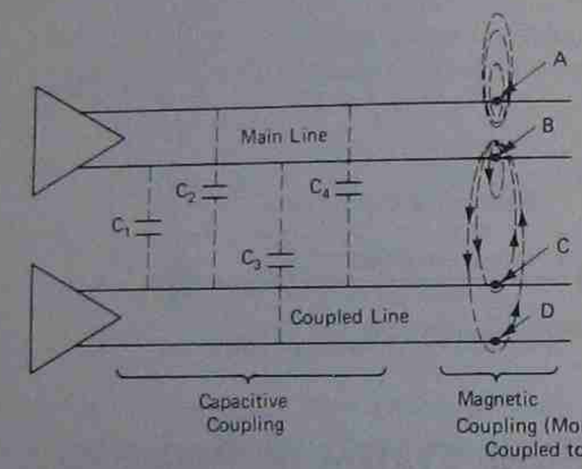
The acceptability of the telephone system to the user is largely dependent upon information impairments suffered during transmission. Any degradation in the signal-to-noise ratio seriously affects the listener's opinion regarding the satisfactory operation of the system. Although the various types of impairment are at times grouped into two major categories, i.e. noise and distortion, it is not always easy to allocate every impairment to one category or the other. In such cases, noise is referred to as effects coming from extraneous services, and distortion is directly related to the transmission of the signal.

Noise, such as crosstalk and thermal, cannot be removed by any special operation. Intermodulation noise, which is very much dependent upon the signal amplitude, cannot be eliminated once it is present in the system. Measures can be taken, however, to minimize the introduction of

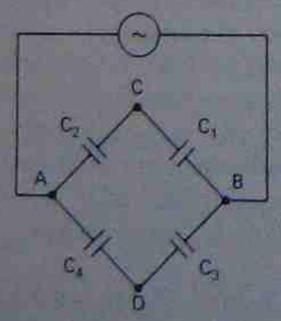
such noise by reducing the signal levels or reducing the number of channels carried by the system. Distortions, such as attenuation and phase-versus-frequency distortion, can be removed by the use of equalizers, or devices which compensate out the nonlinear attenuation and phase characteristics.

In the design of a telephone network, one must make certain that the noise contributed by the various segments of the system do not exceed those set by the CCITT. For long distance transmission, the noise power is allocated on a per length basis. For example, on long distances up to 2500 km, a cable or radio-relay system can contribute up to 3 pW/km plus 2300 pW for multiplexing equipment. Radio systems operating at the very long distance of 2500 km may contribute up to 10,000 pW of noise power. These recommendations can be found in the CCITT White Book 111.

The following sections will discuss more serious noise and distortion contributors on the telephone system.



(a) Capacitive and Inductive Coupling



Imbalance of Bridge as $\frac{C_1}{C_2} \neq \frac{C_3}{C_4}$

(b) Equivalent Bridge Circuit of Capacitive Couplings

Fig. 3-8 Illustration of Capacitive and Inductive Coupling

3–2.1 Crosstalk

Because of the close proximity of adjacent wire pairs in a cable, signals are easily induced into neighbouring pairs by either capacitive or inductive coupling. This phenomenon, illustrated in Fig. 3-8, results in "crosstalk."

Because capacitive coupling tends to be more severe at higher signal frequencies where capacitive reactance is decreased, this capacitive coupling is of major concern when applying digital signals with their fast rise and fall times to cable installed for analog transmission. The amount of crosstalk tolerated by the average listener establishes an upper limit on the number of wire pairs in a cable that can be used for digital transmission. Crosstalk that occurs near repeater stations is termed Near End Crosstalk (NEXT). In this region, the large amplitude signals from an amplifier output are coupled to an adjacent wire pair receiving a low-level signal, resulting in significant interference. Since NEXT is due to the pick-up of signals travelling in opposite directions, it can be reduced by physically separating the GO and RETURN wire pairs.

Far End Crosstalk (FEXT), as illustrated in Fig. 3-9, is due to coupling of signals travelling in the same direction and is of major concern for long closely spaced wire pairs.

To reduce NEXT, particularly with the digital cable, screens are inserted to separate the GO and RETURN line pairs to opposite sides of the screen. The screen acts as an electromagnetic shield, greatly reducing the crosstalk. Examples of such screens are shown in Fig. 3-10.

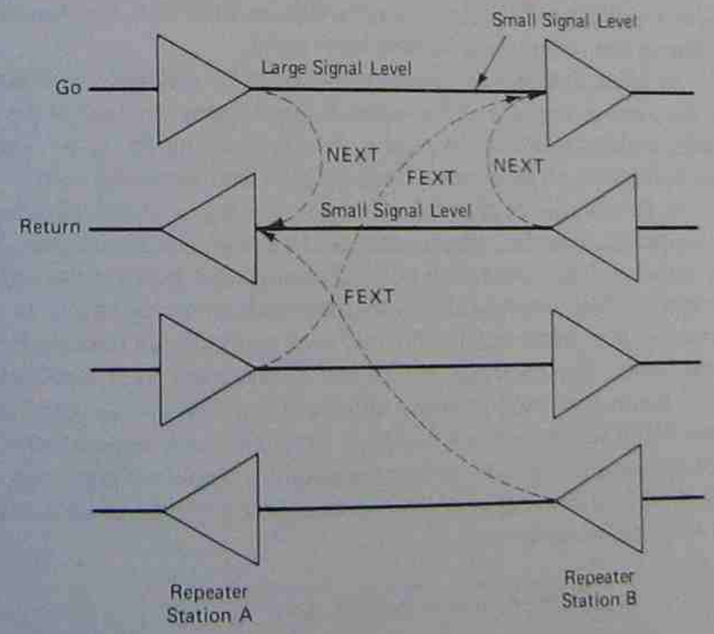


Fig. 3-9 Crosstalk

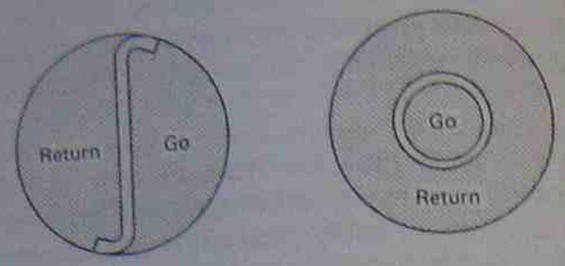


Fig. 3-10 Screened cable

3–2.2 Intermodulation Noise and Loading

When several speakers are on a system, the average power is greater than it would be if there were only one speaker. With a second speaker added to a system, the average power can be expected to increase by 3 dB (10 log 2). For N talkers, each operating over a different frequency band added to a system, the average power will be increased by 10 log N dB, where N represents the number of talkers (channels). To simulate typical loads on a multichannel FDM system, CCITT recommends that white noise at the zero relative level point in the system (usually taken at the toll center in North America) be used. The white noise power levels are:

$$P_{avg} = -15 + 10 \log N \text{ dBm0} \qquad \text{for } N \geq 240 \qquad (3\text{-}6a)$$

$$P_{avg} = -1 + 4 \log N \text{ dBm0} \qquad \text{for } 12 \leq N < 240 \qquad (3\text{-}6b)$$

where N represents the number of channels and dBm0 is the power level in dBm at the system zero relative level point.

As more speakers or channels are added to a system, it is obvious that the average signal level increases. A good approximation of the magnitudes of these levels in a typical system is given by Eq. (3-6). This formula includes the power contibuted by pilot and signaling losses.

As the number of channels increases, the signal levels increase and the amplifiers can be driven into an overload, or non-linear, state. This results in intermodulation noise. Although it is beyond the scope of this text to deal theoretically with intermodulation products in a N-channel system, some insight into how such products are obtained would be profitable. The example in this text is restricted to a two-tone test system commonly used in single sideband transmitters. In general, the output/input voltage relationship of an amplifier may appear somewhat like that shown in Fig. 3-11, where the output-to-input-voltage ratio tends to drop off for the higher voltages. The transfer function can be expressed as a MacLaurins series:

$$v_o = a_1 v_i + a_2 v_i^2 + a_3 v_i^3 + \cdots \qquad (3\text{-}7)$$

where the nonlinear or higher order terms represent distortion.

Intermodulation products are obtained when two or more tones of an SSB signal encounter a non-linear transfer function containing more than just square law terms ($a_2 v_i^2$).

$$\text{Let} \qquad v_i = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t \qquad (3\text{-}8)$$

where ω_1 and ω_2 are "in-band" information tones in the original SSB signal.

$$\text{Let} \qquad v_o = a_1 v_i + a_2 v_i^2 + a_3 v_i^3 \qquad (3\text{-}9)$$

Substituting Eqs. (3-8) and (3-9) and using Eq. (2-12), we obtain:

$$\begin{aligned} v_o = {} & \frac{a_2}{2}(A_1^2 + A_2^2) + a_1[A_1 \cos \omega_1 t + A_2 \cos \omega_2 t] \\ & + a_2\left[\frac{A_1^2}{2}\cos 2\omega_1 t + \frac{A_2^2}{2}\cos 2\omega_2 t + A_1 A_2 \cos(\omega_1 + \omega_2)t\right. \\ & \left. + A_1 A_2 \cos(\omega_2 - \omega_1)t\right] + a_3\left[\left(\frac{3A_1^3}{4} + \frac{3}{2}A_1 A_2^2\right)\cos \omega_1 t\right. \\ & + \left(\frac{3A_2^3}{4} + \frac{3}{2}A_1^2 A_2\right)\cos \omega_2 t \\ & + \frac{A_1^3}{4}\cos 3\omega_1 t + \frac{A_2^3}{4}\cos 3\omega_2 t \\ & + \frac{3}{4}A_1^2 A_2 \cos(2\omega_1 + \omega_2)t + \frac{3}{4}A_1 A_2^2 \cos(\omega_1 + 2\omega_2)t \\ & \left. + \frac{3}{4}A_1^2 A_2 \cos(2\omega_1 - \omega_2)t + \frac{3}{4}A_1 A_2^2 \cos(2\omega_2 - \omega_1)t\right] \end{aligned} \qquad (3\text{-}10)$$

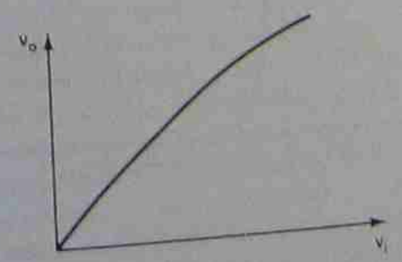


Fig. 3-11 Amplifier Gain Characteristic

Generally the two tones are adjusted to give equal amplitude sidebands, i.e. $A_a = A_b$. Under such conditions the output spectrum appears similar to that shown in Fig. 3-12.

The two frequencies on either side of the desired fundamental tone (f_a and f_b are the intermodulation products due to third order distortion (IM):

$$2f_a - f_b \quad \text{and} \quad 2f_b - f_a \tag{3-11}$$

If the audio two-tone test frequencies fed into the SSB transmitter are 1 and 1.6 kHz, the desired upper sideband components will be:

$$f_a = f_c + 1 \text{ kHz}$$

$$f_b = f_c + 1.6 \text{ kHz}$$

where f_c is the unmodulated carrier frequency.

The intermodulation products then will be:

$$2f_a - f_b = 2f_c + 2 \text{ kHz} - (f_c + 1.6 \text{ kHz}) = f_c + .4 \text{ kHz}$$

$$2f_b - f_a = 2f_c + 3.2 \text{ kHz} - (f_c + 1 \text{ kHz}) = f_c + 2.2 \text{ kHz}$$

Note that the intermodulation products are separated from their nearest tones by the amount of the separation between the original tones. Both intermodulation products are in-band and once present cannot be removed.

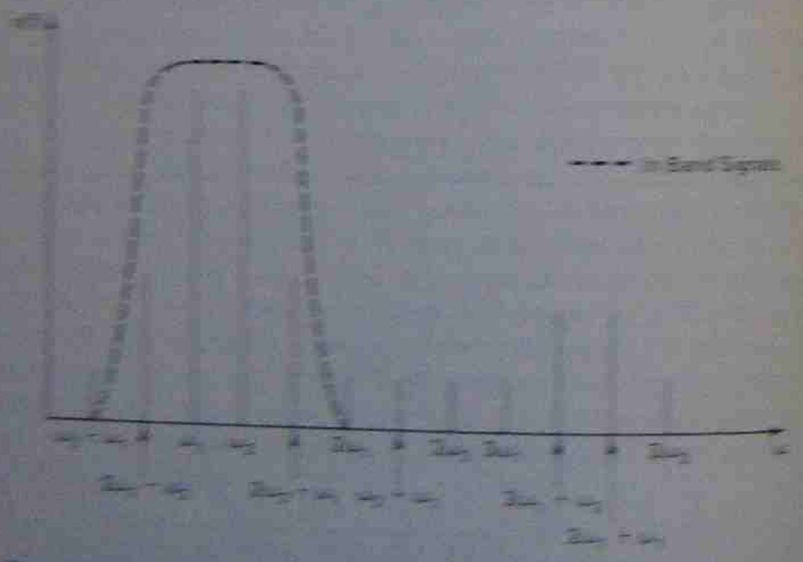


Fig. 3-12. Frequency Spectrum as a Result of Intermodulation Distortion

Another important feature to note is the rate of increase in the amplitudes of the intermodulation terms as compared to the desired sideband components. Again, for equal amplitude side tones, the amplitudes of the third order intermodulation products are proportional to A^3 whereas the desired sidebands have amplitudes proportional to A ($a_3 \ll a_1$). Thus, if both fundamental components are doubled or increased by 20 log $2A_1/A = 6$ dB, the intermodulation components are increased by a factor of 8 or 20 log $(2A^3/A^3 = 18$ dB.

For every 1 dB increase in the fundamental component, a 3 dB increase occurs in the intermodulation component. Single sideband transmitter maximum output power ratings are usually specified in terms of how far the intermodulation product must be below the power level of one of the two-frequency test signals. In Radio Standards Specification 123, Department of Transport, Ottawa, Canada, the maximum RF power output rating for transmitters delivering less than 100 W is that power developed by the transmitter when the intermodulation product is 26 dB below the power level of one of the two-frequency test signals.

Again considering an FDM system, as the number of channels on a system is increased, the average power to the amplifier increases, thus increasing the likelihood of overdriving the power amplifiers. This causes an increase in intermodulation noise. Intermodulation noise is measured by applying white noise of suitable level (see Eq. (3-6)) to the FDM system and keeping one channel-wide gap free from the applied noise. A band reject filter can be used for this purpose. At the receive end of the system, the intermodulation noise creeping into the channel gap is monitored.

Measurement of Intermodulation Noise

The measurement of intermodulation noise can be most conveniently made by applying white noise loading. Consider a noise loading set which measures the intermodulation and thermal noise (explained in next section) in terms of a noise power ratio (NPR). The noise power ratio, expressed in decibels, is defined as the ratio of noise in the test channel with all channels loaded with white noise to noise in the test channel with all channels except the test channel fully noise loaded. To measure the NPR of a system, white noise of a suitable level (see Eq. (3-6)) is passed through a band-stop filter to provide one "quiet" telephone channel within the baseband signal. The baseband signal with the quiet slot is applied to the transmitter. The resulting noise at the receiver output, which occurs within this quiet slot, is a measure of the thermal and intermodulation noise. Figure 3-13 illustrates the NPR measurement technique.

The channel signal-to-noise ratio or test tone level-to-noise ratio in terms of NPR is shown in Eq. (3-12). The derivation of this equation is explained in Appendix B.

$$\left(\frac{S}{N}\right)_{dB} = NPR - (P_{avg})_{dBm0} + 10 \log \left(\frac{\text{original base bandwidth}}{\text{channel slot bandwidth}}\right) \quad (3\text{-}12)$$

where NPR is the noise power ratio
P_{avg} is given by Eq. (3-6).

3–2.3 Thermal and Shot Noise

Two other major sources of noise in an electronic system are:

a. Thermal noise due to the random motion of free electrons within a conductor, and
b. Shot noise due to the random fluctuations of discrete carriers arriving at a collector or plate.

The available thermal noise of a device at temperature T over a bandwidth B is given by:

$$P = kTB \quad (3\text{-}13)$$

where B is the bandwidth
T is the absolute temperature
k = Boltzmann's Constant = 1.38×10^{-23} J/°K

Equation (3-13) is valid for all frequencies below that where quantum-mechanical effects occur. When dealing with optical lasers, for instance, the quantum effect is predominant because of the discrete photon nature of light signals.

When a repeater or amplifier's input resistance is at room temperature (T_o = 290°K) rather than at some effective radiation resistance temperature, as in the case of a satellite receiver, the total noise at its output is given by:

$$P_o = n_f g\, kT_o B \text{ watts} \quad (3\text{-}14)$$

where n_f is the repeater noise factor
g is the repeater power gain
k = Boltzmann's constant = 1.38×10^{-23} J/°K
T_o = 290°K
B is the system bandwidth

This noise factor includes all noise due to processes within the repeater as well as room temperature thermal noise generated at repeater input. Note that the noise factor is an indication of the noise contributed by the repeater. If $n_f = 1$, the noise of the amplifier output is merely the noise at its input (kT_oB) multiplied by the power gain of the amplifier. The n_f can be measured and is usually given by the manufacturer of the equipment. It can be expressed in decibels as $N_f = 10 \log n_f$ dB. Sometimes a noise temperature (T) is given that is related to the noise factor by:

$$n_f = 1 + \frac{T}{T_o} \quad (3\text{-}15)$$

Equation (3-14) can also be expressed in dBm's as

$$P_o = N_f + G - 174 + 10 \log B(\text{dBm}) \quad (3\text{-}16)$$

where G is the amplifier gain in dB's.

Repeater gains are normally designed to make up line losses, which results in a nominal 0 dB gain or loss over a complete repeater section, as shown in Fig. 3-14. In identical 0 dB gain sections, the total noise contributed will be mP_o. Equation (3-16) then becomes

$$P_o = N_f + G - 174 + 10 \log B + 10 \log m(\text{dBm}) \quad (3\text{-}17)$$

where N_f = repeater noise figure (dB)
G = repeater power gain in dB
B = repeater bandwidth
m = number of repeater sections
P_o = output noise in dBm

Thermal and shot noise tend to be independent of signal levels, but an increase in signal levels will cause a decrease in the effect of thermal noise. On the other hand, an increase in the signal level tends to increase the intermodulation noise. The goal in a communications system is to find an optimum signal level that will give a minimum amount of total noise. Figure 3-15 shows the optimum level at which all amplifiers should be operated.

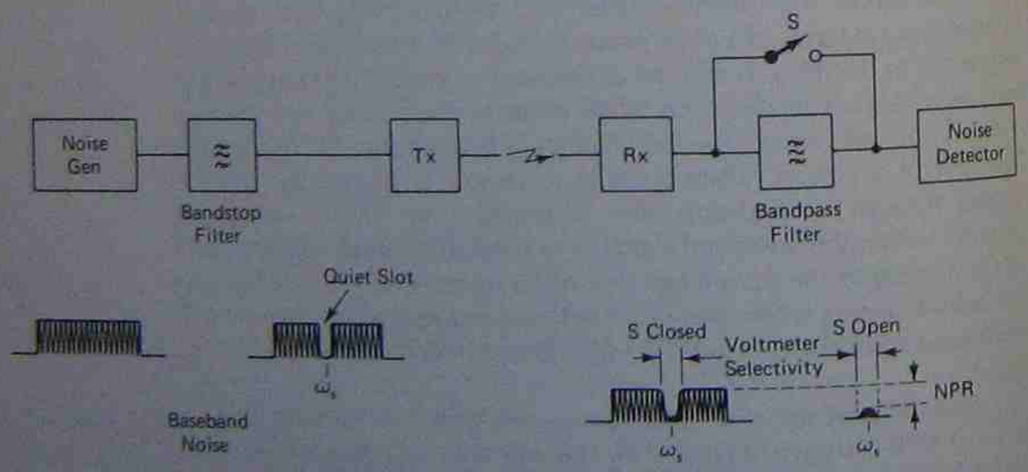


Fig. 3-13 Intermodulation Noise Measurement

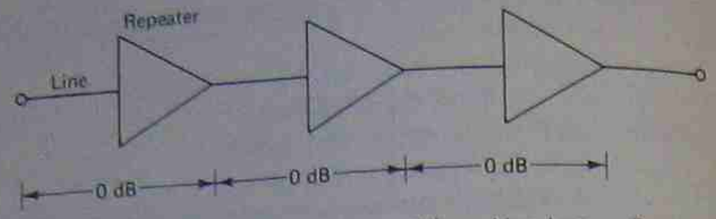


Fig. 3-14 Illustration of Repeater Sections Where Line Loss = Repeater Gain

Data transmission causes heavier line loading than voice communication. Data is usually transmitted at a maximum constant bit rate, operating the system at full capacity. In voice communication, the activity factor is typically down to 0.25, which represents the fraction of the system capacity being actively used at a given time. (The maximum bit rate has an activity factor of 1.) The activity factor in voice communication tends to be low because listening and talking are usually shared on a 50-50 basis and there are punctuation pauses during a conversation. In addition, a voice communications system is designed so that, on the average, less than the full complement of channels is in operation. By not fully utilizing all channels, blocking of calls is kept reasonable. If all N-lines are normally kept busy, all new calls would of necessity be turned away.

In order to cope with data transmission, data is usually permitted up to about 25% of the channels, and the rest is allocated to voice communications. Exceeding the 25% limit causes excessive intermodulation noise. In addition, data channels should not be near the group band edge where the filter group delays become highly non-linear.

3–2.4 Impulse Noise

Impulse noise, as the name suggests, is an impulsive or spikey form of interference. It is heard as a crackle and is generally not too objectionable in voice communication since it rarely causes the message to be misunderstood. It is of serious concern, however, in data communication since it can cause clusters of errors in a data bit stream.

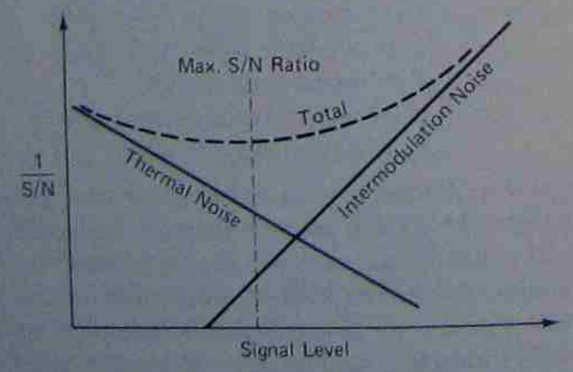


Fig. 3-15 Optimum Signal(S)-to-Noise(N) Relationship

Impulse noise appears in telephone systems chiefly because of the electric mechanical switches in the exchanges. These switches can cause vibrations to occur over entire equipment racks, inducing any contacting surfaces to create noise spikes. Impulse noise can also be the result of faulty soldering or dirty relay contacts. It can also originate from external sources, such as power lines and electrical storms.

3–2.5 Attenuation Distortion

As noted in Sec. 2–6, the attenuation of a transmission line is not flat with frequency but tends to vary as the square root of the frequency at the lower frequencies. Attenuation distortion is caused as the high frequency components in a signal experience greater attenuation than the lower frequencies as the signal travels down a transmission line.

Fortunately, this type of attenuation can be flattened out by the use of an attenuation equalizer. If the electrical network has a frequency loss that is complementary to that of the line, the net result is a flat frequency response. Inductive loading of cable also tends to flatten out the attenuation curve if the cutoff frequency is not exceeded. See Fig. 3-16.

3–2.6 Envelope Delay Distortion

As mentioned in Sec. 2–6, the velocity of signal propagation over a transmission line tends to vary with frequency. Because of this variation, the different frequency components in a signal arrive at the receiving end at different times. This delay distortion is a problem when bandpass filtering is used since the delay at the band edges is quite different than at the midband frequency points.

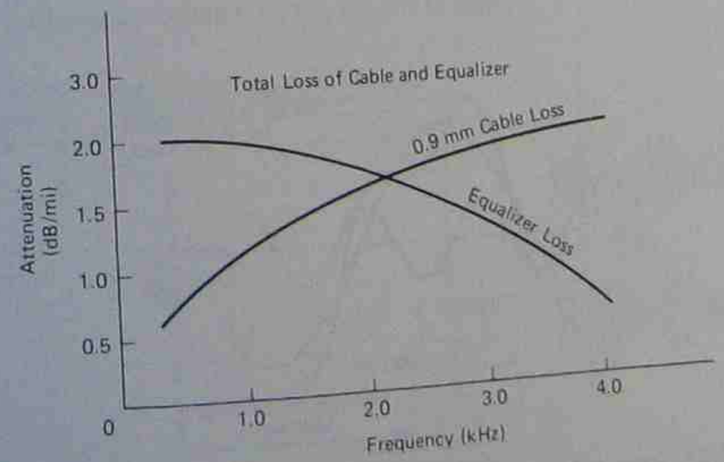


Fig. 3-16 Attenuation Distortion of a 0.9 mm Cable

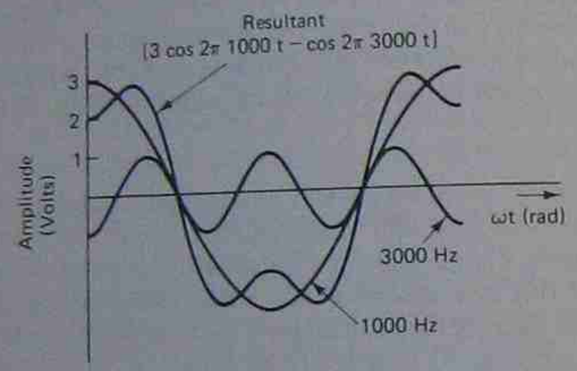
Sharp filters cause greater delay distortion of frequencies approaching the band edge than filters with a more gradual roll-off. Loaded cables also show a greater delay distortion than non-loaded cables.

To illustrate the effects of envelope delay, consider a 1000 Hz square passing through a 5 mile length of 22H88 loaded cable. Because of the low-pass filtering of the telephone facility, only the 1000 and 3000 Hz components will be present. As can be found from almost any reference on Fourier series (or as shown in Appendix A), the input signal can be expressed as

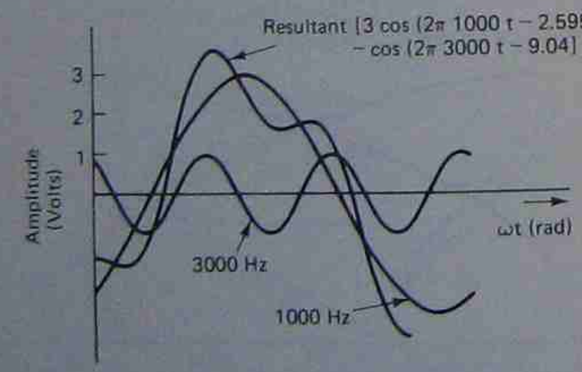
$$e_i = 3 \cos 2\pi 1000t - \cos 2\pi 3000t \qquad (3\text{-}18)$$

This waveform is plotted in Fig. 3-17.

From Table 2-3, it can be noted that the phase shift for 22H88 cable is 0.519 radians/mi at 1 kHz. At 3 kHz, this phase shift increases to 1.808 radians/mi.



(a) Input



(b) Output

Fig. 3-17 Input and Output Waveforms on a 5 Mile Section of 22H88 Cable. (Attenuation Neglected)

For a cable circuit 5 miles long, the 1 kHz component will be shifted by 2.595 radians and the 3 kHz component will be shifted by 9.04 radians. Neglecting any attenuation effects, the output voltage can be expressed as:

$$e_o = 3 \cos (2\pi 1000t - 2.595) - \cos (2\pi 3000t - 9.04) \qquad (3\text{-}19)$$

This is plotted in Fig. 3-17(b). The distortions in the output waveform can be readily seen when compared with the input waveform.

Group or envelope delay can be measured by an amplitude-modulated signal, as shown in Fig. 3-18. The modulation frequency (ω_m) is constant and kept much lower than the carrier frequency (ω_c). As shown in Appendix C, the group delay is given by

$$T_g = \left.\frac{db}{d\omega}\right|_{\omega_c} \qquad (3\text{-}20)$$

where b is the total phase shift experienced by the envelope as it travels down the line. If the modulation frequency (ω_m) is kept small enough, b will be linear (see Fig. 3-19(a)) within the $2\omega_m$ frequency band, and Eq. (3-20) can be approximated by:

$$T_g = \left.\frac{\Delta b}{\Delta \omega}\right|_{\omega_c} = \frac{\Delta b}{\omega_m} \qquad (3\text{-}21)$$

In an actual measurement system, the modulation of frequency is 41.66 Hz as recommended by the CCITT for the 200 Hz to 200 kHz range. This is a compromise frequency since lower frequencies require an excessively high resolution phase meter, and higher frequencies distort the non-linearity required by the Eq. (3-21) relation.

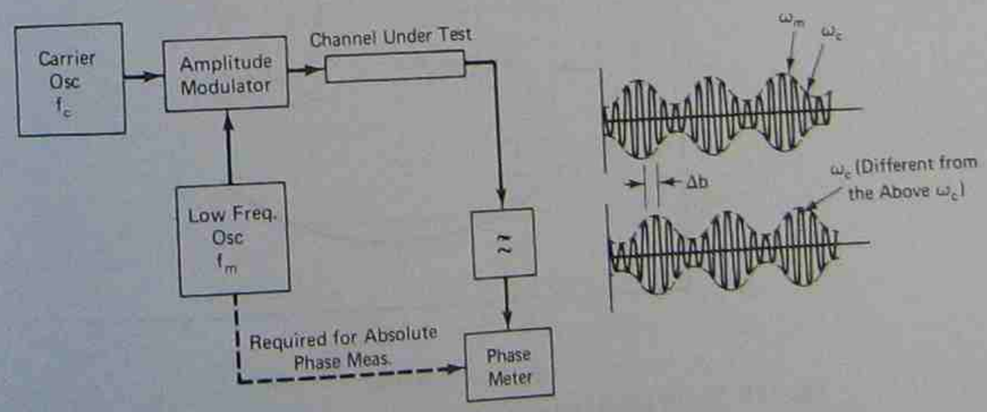


Fig. 3-18 Measurement of Group Delay

After the AM signal passes through the network shown in Fig. 3-18, the output is demodulated and compared with the original modulating signal. As the systems measure the difference in phase shift between the two sidebands, the group delay can be computed as

$$T_g = \frac{\Delta b}{2\pi(f_m)}$$

By varying the carrier frequency ω_c, the entire group delay curve can be obtained across the bandwidth of the system.

Measuring the absolute delay between the sending and receiving ends is difficult to determine because of the large number of 2π radian phase shifts the signal undergoes. It is more common to measure the relative envelope delay. Relative delay uses a specific frequency as a reference, usually the one with the minimum delay, and measures the delay at the other frequencies by taking the difference between its delay and the reference frequency delay. Thus, from Fig. 3-19(b), the relative group delay at f_c is equal to

$$\Delta t_g = T_g(\omega_c) - T_g(\omega_r) \qquad (3\text{-}22)$$

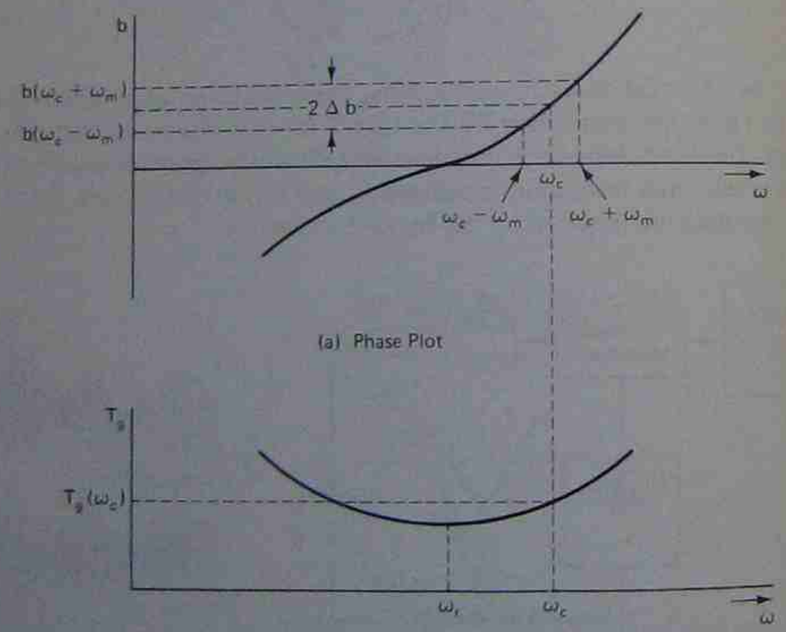


Fig. 3-19 Group Delay Measurement Using an AM Signal with Modulation Frequency f_m.

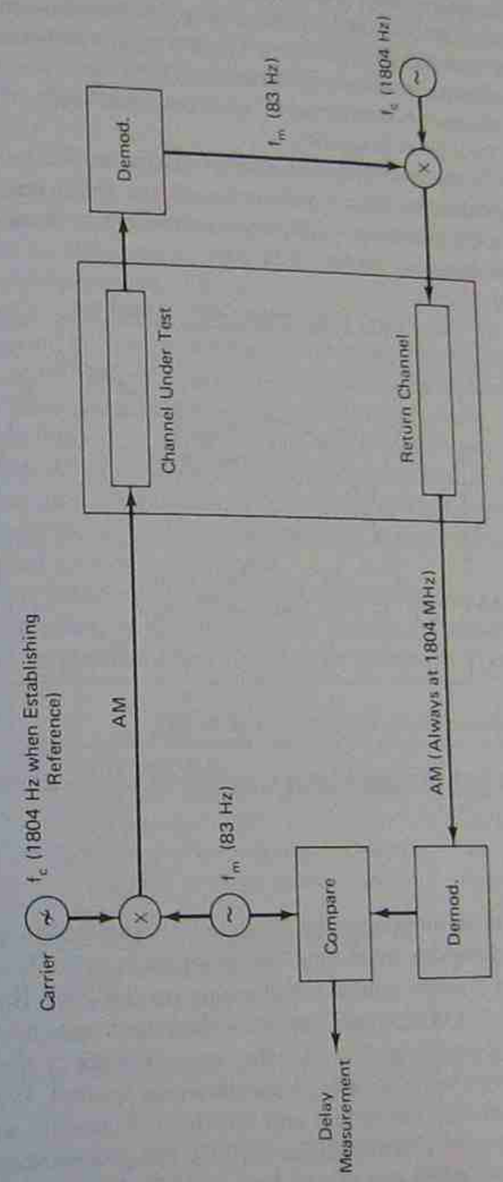


Fig. 3-20 Envelope Delay Measurement System

Another technique of measuring envelope delay requires a four-wire circuit, as shown in Fig. 3-20. The test equipment transmits an AM signal whose carrier frequency, f_c, can be varied. At the receiver end, the envelope information is detected and transmitted back to the test instrument at a fixed carrier frequency. Any delay in the return loop will be constant and can be zeroed out. The test equipment compares the phase of retransmitted envelope with the original modulation signal, and the resulting delay can be monitored over the bandwidth of the channel by varying the carrier frequency.

As in the case of attenuation distortion, envelope delay distortion can be reduced by inserting delay equalizers. Delay equalizers have opposite, or complementary, phase characteristics to those of the transmission path, as shown in Fig. 3-21.

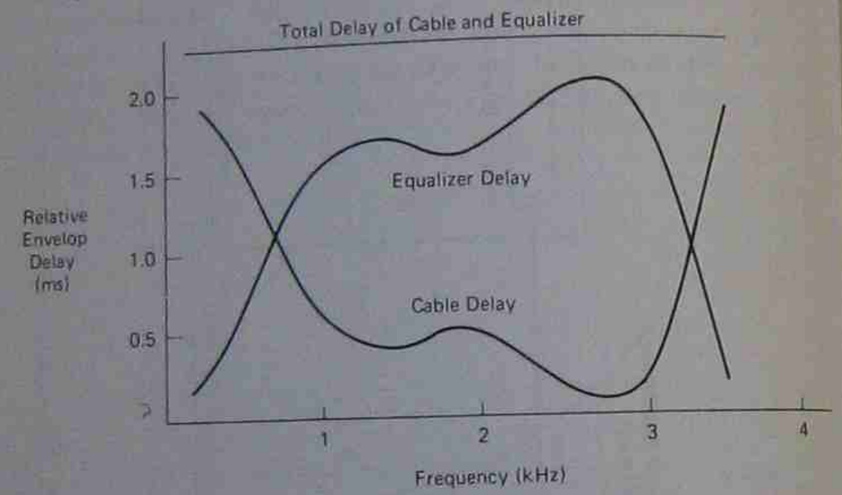


Fig. 3-21 Envelope Delay Correction of a Voice Circuit Cable

3–2.7 Line Conditioning

One method of improving the passband characteristics of a telephone channel and thereby increasing the information capacity of a telephone system is to provide special conditioning on the line. This requires the rental or leasing of a private line rather than the more common switched line or dialed-up connection since the latter are more or less randomly selected and can have no special conditioning applied. Private lines are more permanently connected and can be individually conditioned to provide improved channel characteristics. Other advantages and disadvantages of switched and private lines are listed in Table 3-1.

In order to obtain high-speed data communications on a voice-grade line, the attenuation distortion, envelope-delay distortion, the signal-to-noise ratio, and the harmonic distortion must be controlled. The

first two are controlled through C-conditioning and the latter by the more recent D-conditioning.

The commonly used 3002 unconditioned voice grade private line has certain attenuation and envelope-delay distortions as shown in Fig. 3-22(a). Attenuation distortion can be somewhat compensated for by employing circuit conditioning or equalization. This usually consists of several tandem networks near the receiving end of the line. The net result is a more flattened amplitude response, as shown in the C2-conditioned line of Fig. 3-22(b).

The delay equalizer operates in much the same fashion. The human ear tends to be insensitive to delay, and therefore, envelope delay distortion has not been of traditional concern. Delay distortion is of significance, however, for data transmission when pulses are employed.

In carrier multiplex systems in which several channels form a group, the channels which are nearest the group band edge suffer the most severe delay distortion because of the highly nonlinear characteristics of the filter near the cut-off regions. Data transmission channels should not be allocated near these group edges but near the group center.

D-conditioning is a more recent option and not available throughout North America. It limits harmonic, or nonlinear, distortion. Nonlinear distortion is irreversible and cannot be compensated out once it is present. Once the signal has been altered by nonlinear impairments, the original signal cannot be re-created.

Attenuation delay equalizers are generally designed in a bridged-T or lattice configuration to compensate out distortions occurring in the passband. Operational amplifiers can also be employed in conjunction

Table 3-1 Switched vs. Private Lines

	Switched Lines	Private or Dedicated Lines
Advantages	Can access millions of other telephones/machines. Can access other computers if present computer is under repair or overloaded. Can use alternate line facilities in case of line failure. Low cost if usage is low. Can handle up to 9600 b/s 1600 baud.	Always available. Can be conditioned. More efficient in time as no time is spent on busy signals and handshaking. Can handle higher line speeds because of conditioning (less line noise and distortion) and no signalling frequencies to avoid.
Disadvantages	Limited to line speeds of 9600 bits/s. Lines can be busy. Subject to larger error rates due to impulse noise. Contain echo suppressors which interfere with data communications if not disabled.	More costly for low usage

with suitably designed passive networks to obtain the desired passband delay characteristics. Typical circuits are illustrated in Fig. 3-23.

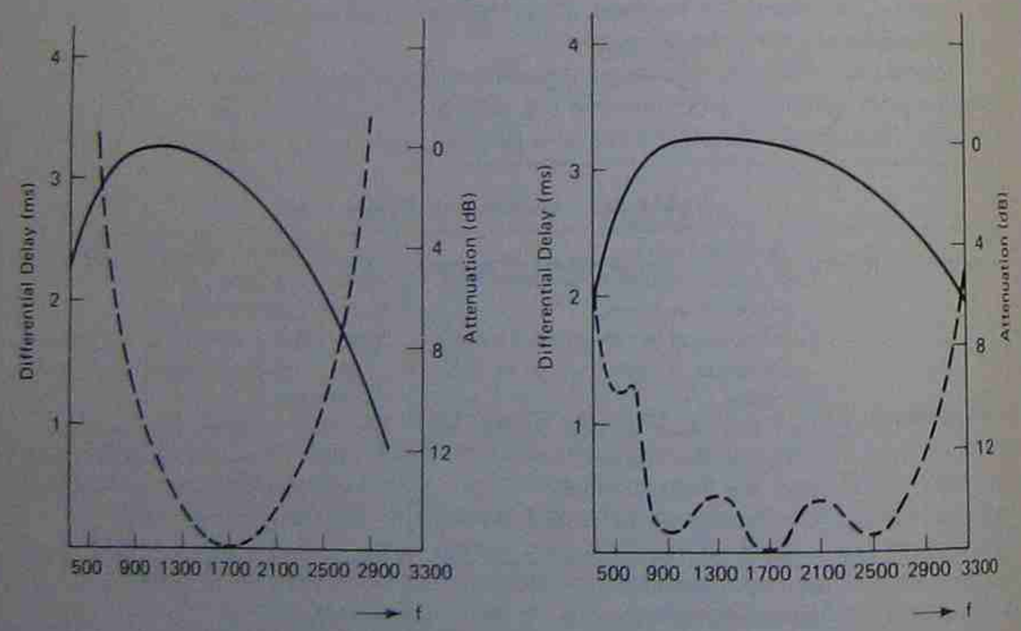
Many high speed dial-up modems now also incorporate automatic equalizers of various types. In this way, 9600 b/s can be obtained over the switched telephone network.

3–2.8 Compandors

To reduce the dynamic range of a signal, compression is applied to the signal to slightly reduce the maximum signal amplitudes and significantly raise the minimum signal amplitudes. The intensity *range* of the input signal, normally 60 dB, is reduced to a range of 30 dB. At the receiving end, an expandor restores the signal to its original dynamic range. The joint combination of the COMPressing-expANDING process is commonly called companding. Companding a signal results in:

a. Improved signal-to-noise ratios, and
b. Reduced peak powers with less likelihood of overloading.

To understand the reason for the reduction in noise contributed by the system, consider Fig. 3-24. Because the weaker signals at the receiver undergo the greatest attenuation, the low-level noise is attenuated more than



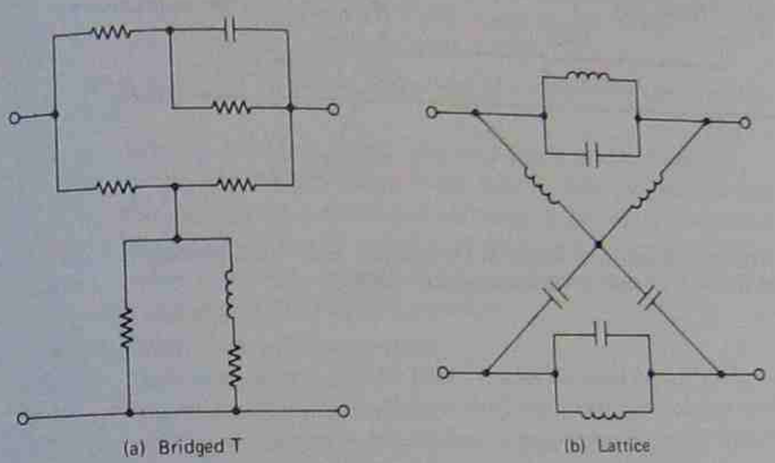
(a) Basic 3002 Line

(b) C2 Conditioned Line

Fig. 3-22 Typical Line Attenuation and Delay Characteristics
———— attenuation
– – – – – – – – delay

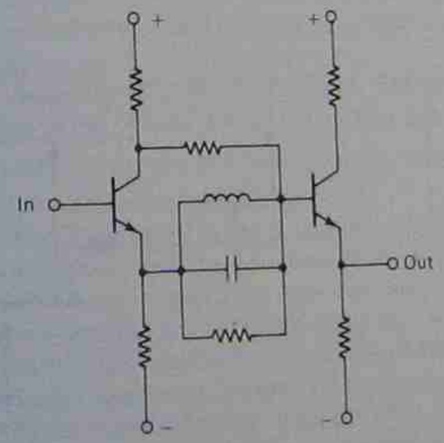
even the weakest signal. This results in an improved signal-to-noise ratio as compared to that of the non-companded system.

Compandors respond to signal level changes with a syllabic time constant, that is in the order of 20 ms. Instantaneous companding is not normally employed since it requires excessive bandwidths. A simplified diagram of a compressor and expandor circuit is shown in Fig. 3-25.



(a) Bridged T

(b) Lattice



(c) Operational Amplifiers

Fig. 3-23 Attenuation and Delay Compensating Networks

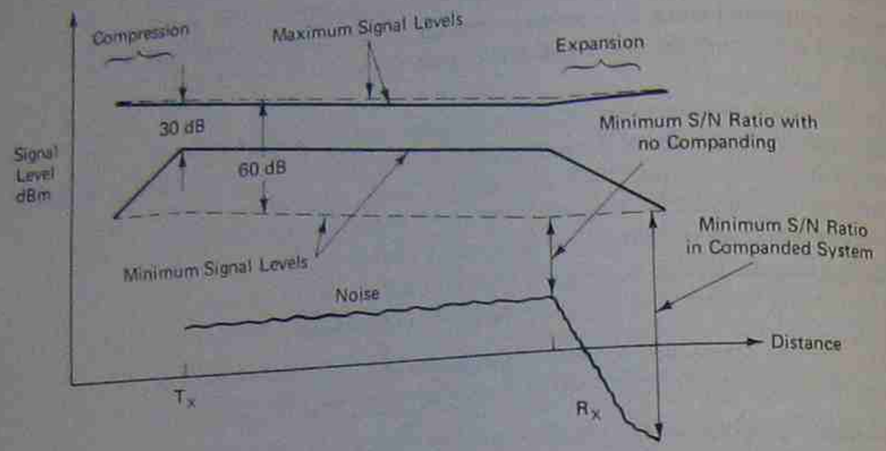
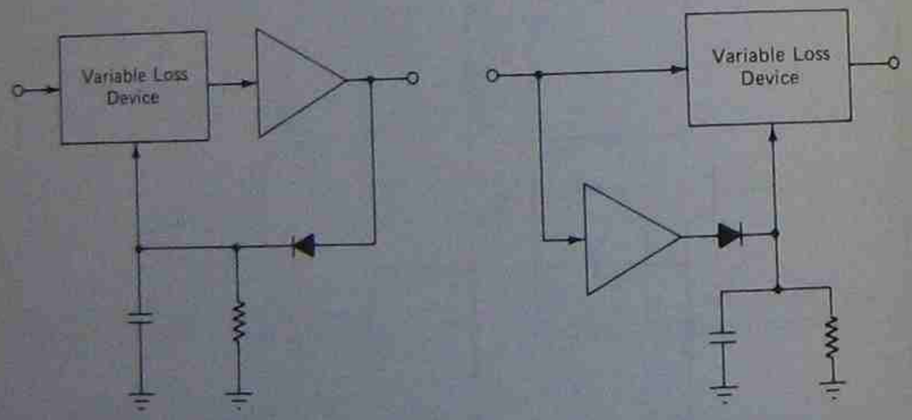


Fig. 3-24 Signal-to-Noise (S/N) Ratios in Companded and Non-Companded Systems. The dotted lines refer to non-companded system.



(a) Compressor (b) Expandor

Fig. 3-25 Compressor-Expandor Circuits

PROBLEMS

1. a. Why is it so difficult to match the balance network in a four-wire terminating set to the line characteristic impedance?
   b. What is the characteristic impedance of #22 CSA NL line at 1000 Hz? (See Table 2-3.)
   c. What is the balance-return loss of a 1 kHz signal applied to the return arm of a four-wire terminating set when the balance network impedance is 600 Ω?
   d. If a signal experiences 3.5 dB loss as it traverses through a hybrid transformer (refer to Fig. 3-3), what residual power appears in the GO arm when a 0.1 mW signal appears in the RETURN arm under the conditions given in (c)?
2. What does the reference equivalent of a telephone circuit represent?
3. a. What is a singing circuit, and how can it be prevented?
   b. In the telephone network of Fig. 3-4, the balance-return loss of the east circuit is 40 dB and that of the west circuit is 50 dB. If the two hybrid coils each have a transmission loss of 3.5 dB in either direction, calculate the total repeater gain if a 6 dB gain margin against singing is provided.
4. a. What is an echo suppressor?
   b. How long a loop of #19 DNB NL can be used before an echo suppressor must be included? With the same line loaded with H88, how long may it be before a suppressor must be added? (Refer to Table 2.3 for velocity data.)
5. a. What does the Via Net Loss of a circuit represent?
   b. Find the Via Net Loss required for a 1000 mi L-carrier trunk circuit.
   c. If the 1000 mi L-carrier is the only link between end offices, what is the desired overall connection loss?
6. What causes crosstalk, and how can it be reduced?
7. a. What causes intermodulation noise?
   b. In Eq. (3-10), what is the second harmonic term?
   c. When monitoring the harmonic products and intermodulation products on a spectrum analyzer, how can one determine which product is a harmonic and which is an intermodulation term? (Note: Varying the amplitudes of the test tones results in different amplitude charges of the unknown products.)
   d. Why are intermodulation products often of more concern than harmonic, sum and difference products, etc.?
8. The NPR in a 12-channel CCITT group is measured at 25 dB follow-

ing the CCIR procedure of measuring signal-to-intermodulation noise under busy conditions. (See Eq. 3-12). What is the intermodulation S/N ratio?

9. a. What causes thermal and shot noise?
   b. What thermal noise power is available from a 100 kΩ resistor at 20°C over a 1 MHz bandwidth?
10. a. A 2000 km system consists of 100 repeater sections. If the repeater gain at the highest 300 to 3400 Hz bandwidth channel frequency is 50 dB and the noise figure is 10 dB at this frequency, what is the total noise from the system in dBm?
    b. If the relative output level of the repeater is −10 dBr, what is the noise referred to the 0 TLP in dBm0's?
    c. If this white noise over the 4 kHz bandwidth is C-message weighted, what is the C-message-weighted noise at the 0TLP in dBm's and in picowatts?
11. Define activity factor, and give reasons why it is less than 100%.
12. To achieve maximum S/N ratios, can signal levels be continuously increased? Why or why not?
13. State the nature and origin of impulse noise in the telephone network.
14. a. State the chief causes of attenuation and envelope delay distortion in a cable pair.
    b. What name is given to the devices that partially compensate out the attenuation and envelope delay distortion?
15. The phase delay representing the transmission time of a single frequency signal component through a transmission line is given by

$$T = \frac{b}{\omega}$$

where b is total phase shift in radians
$\omega = 2\pi f$
T = phase delay in unit of time

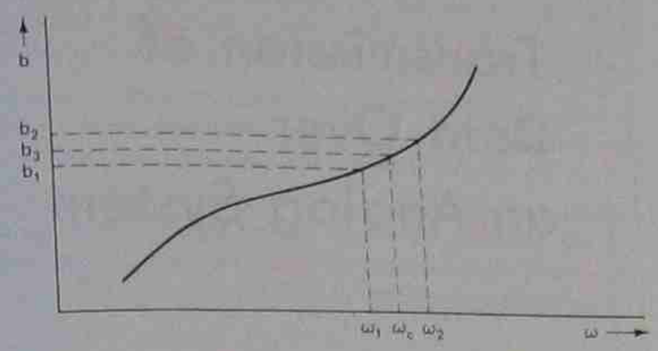
The envelope or group delay is given by:

$$T_g = \frac{db}{d\omega} \text{ (see Eq. (3-20))}$$

This represents the *slope* of the phase curve at the frequency in question.

a. Under what conditions are the phase delay and envelope delays identical?

b. Use the following graph to illustrate how T and T_g can be determined at ω_c.



c. Use Eq. (3-21) to explain why the modulation frequency in the group delay measurement must be kept small.
d. What is the difference between absolute and relative envelope delay?

16. Find the difference in phase delay in units of time of the two frequency components illustrated in Fig. 3-17. Refer to Eq. (3-19).
17. a. State the advantages and disadvantages of switched and private telephone lines.
    b. What is a conditioned line?
    c. What is the difference between C- and D-conditioning?
18. What is the purpose of a compandor?
    What is its principle of operation?

4

Transmission of Data Over an Analog System

4–1 INTRODUCTION

Telecommunication lines are increasingly required to carry digital signals, or signals in which only a finite number of separate states are permitted. An example of such a signal is the binary digital message in which

a. Only one of two *levels* can exist at any particular time, and
b. Only one of two *frequencies* can exist at any particular time, commonly known as Frequency Shift Keying (FSK).

This text will primarily use only binary originating, or baseband, messages when dc signaling and the modem are discussed. The binary symbol employed to represent one of two states is called a binary digit, or bit, and is usually symbolized by one of the following combinations:

1	0
HI	LO
MARK*	SPACE*

* these may be interchanged

One method of transmitting a binary pulse train on a local subscriber loop is to employ a dc signaling technique. In this case, a switch at the telephone set is opened and closed (representing 1's and 0's) to provide current pulses at the local exchange. These current pulses can be detected by a relay at the local exchange which determines whether a 1 or 0 was transmitted. Since dc or low frequencies cannot be directly transmitted between exchanges, other schemes must be used in those cases.

To understand some of the major difficulties experienced in dc signaling, consider the similar case of transmitting voltage pulses down a transmission line. This is a common method of communicating between computers which are in close proximity (100 m or less) of each other. Implementation of dc signaling tends to be simple and inexpensive. For somewhat longer distances, line drivers and possibly amplifiers need to be added. For distances of a few miles or greater, some form of modulation must be employed.

The distance limitation on dc signaling is due to distortion introduced by resistance, shunt capacitance, and series inductance associated with any real transmission line. The dielectric or shunt losses due to leakage through the insulation material are usually insignificant and therefore ignored. Line resistance is very much a function of frequency which causes the various frequency components making up a voltage pulse to experience differences in attenuation and propagation velocity. Resistance variation due to frequency is chiefly the result of skin effect, proximity effect, and radiation losses.

Skin Effect

Because the center of a conductor experiences more self-flux linkages than the outer region of the conductor (see Fig. 4-1), self-inductance increases upon approaching the center region of a conductor. Thus, the associated impedance is less near the conductor surface which causes more of the current to flow near the conductor surface than in the central region. This effect, known as skin effect, becomes more and more pronounced as the frequency is increased. Typical current distributions for an isolated conductor are shown in Fig. 4-2.

Because current flows through a smaller net area at higher frequencies, there is an increase in the effective resistance of the conductor. This in turn causes the signal to experience greater attenuation at the higher frequencies.

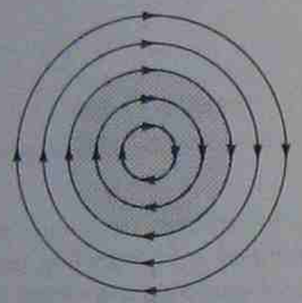


Fig. 4-1 Flux Around and In Current-Carrying Conductor

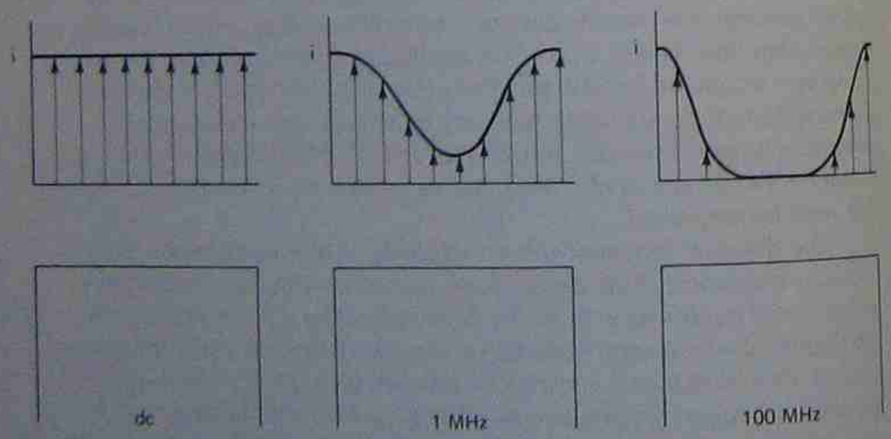


Fig. 4-2 Current Distribution Across a Conductor at Three Different Frequencies

Proximity Effect

Effective resistance is not only frequency dependent but is also dependent upon other conductors that are located in its proximity. Because of mutual repulsion or attraction of the charges flowing in nearby conductors, the current density undergoes a redistribution which tends to further increase the effective resistance of the conductor. This proximity effect is a function of conductor diameter, spacing, and frequency. It contributes significantly to the loss in twisted wire cable and parallel wire lines. Because of symmetry and shielding, it is of no consequence in coaxial lines.

Radiation

Radiation losses can contribute to an increased effective resistance, particularly at the higher frequencies. For closely spaced inductors, the loss is not significant unless the spacing is some appreciable portion of the electrical wavelength in size ($\lambda = C/f$). Coaxial lines experience no radiation loss because of shielding.

Because both the line attenuation and signal propagation velocity along the line are resistance dependent (see Sec. 2–6), both the line attenuation and signal velocity are frequency dependent. Typical plots are shown in Figs. 4-3 and 4-4 for a #22 twisted-wire pair. Both attenuation and signal velocity increase with frequency. The variation of signal velocity with frequency is termed dispersion.

Dispersion and attenuation applied to a transmission line both delay and distort the pulse, so that the detected pulse at the receiving end

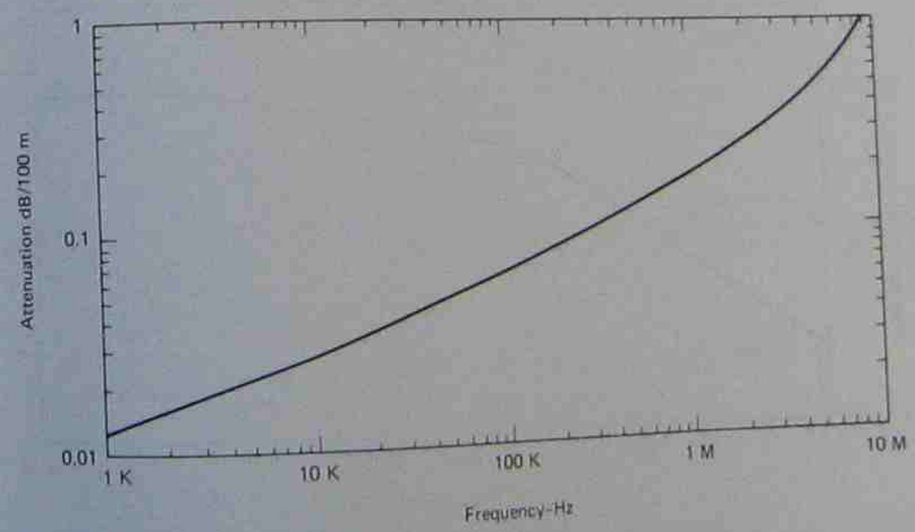


Fig. 4-3 Attenuation vs. Frequency on #22 AWG Line

is more difficult to interpret. The leading or lagging edges of a pulse containing high frequency components tend to be subject to extreme delays whereas the pulse top with low frequency content experiences only small delays. This causes the various parts of the pulse to arrive at the receiver at different times, producing distortion of the original signal. If a pulse train is present, as is normal, intersymbol interference, in which portions of neighboring pulses spill over into each pulse, results. Intersymbol interference (ISI) forces a reduction in the allowable permitted pulse rate for a given line length in order to maintain adequate distinction between adjacent pulses. As the line length is increased, dispersion increases, and the pulse rate must be further reduced. Delay equalizers must reduce this type of interference if high data rates are to be used where the pulses are more closely spaced.

Figure 4-5 shows an example of baseband data transmitted down a line. For each pulse transition applied to the line, the received pulse is transformed to a rounded-out and delayed transition as a result of line attenuation and dispersion. The line detector has a decision threshold, or slicing level, midway between the voltage extremes. When the received voltage exceeds this threshold, a one logic state is produced. A zero logic state is produced when the voltage falls below the threshold level.

The threshold can be considered the current drive necessary to change the state of a set of relay contacts. Exceeding this threshold current may cause the relay contacts to close, which would result in a sharp data output transition. As a first approximation, the pulses at the received

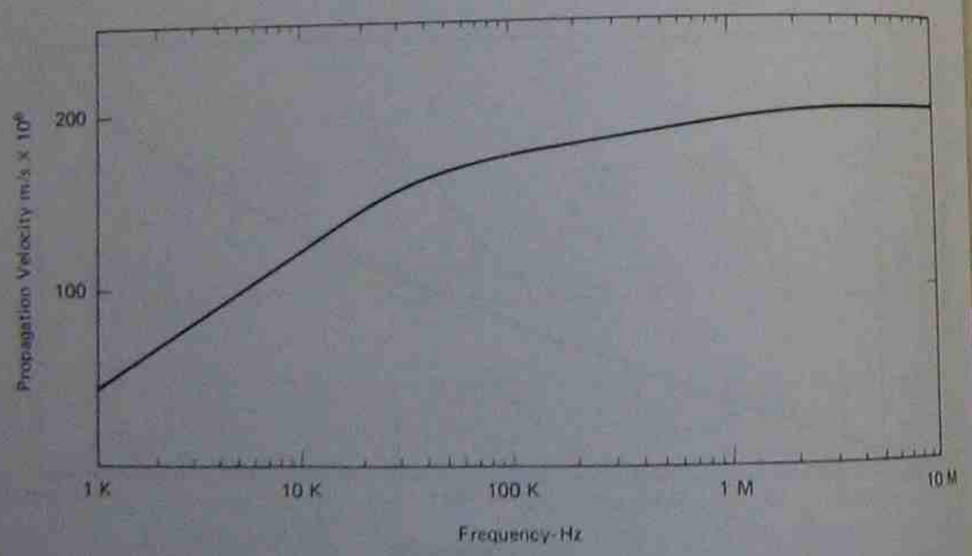


Fig. 4-4 Propagation Velocity vs. Frequency on #22 AWG Line

end can contain gradually increasing or decreasing transitions due to the resistive-capacitive nature of the line.

This is illustrated by the arrival transition sketch of Fig. 4-5. Observe the severe distortion of the received bit pattern because of the line dispersion.

If the pulse rate increases much more, even by adjusting the threshold level, the received bit pattern would still be in error. Varying the setting of the threshold level bias can cause time expansion or elongation of either the high state or low state, depending upon the direction of threshold movement. If the threshold level is lowered, the low-state times are increased at the expense of the higher state time. An incorrectly adjusted threshold level can severely degrade a system, particularly when several links are operated in tandem. This elongation of the high-state bits (or the low-state bits, if that is the case) is known as bias distortion.

The output bit stream is obtained by sampling the received data at the instant noted by the vertical arrow. This is merely a delayed version of the input clock pulse at the transmitting end.

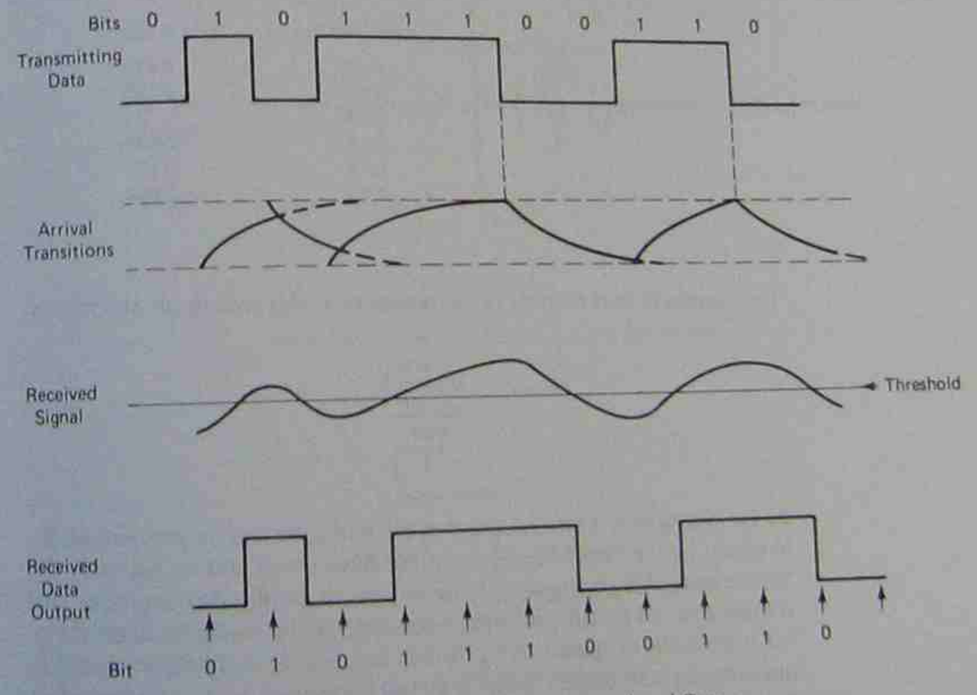


Fig. 4-5 Transmission of Baseband Data

4–2 BANDWIDTH LIMITING OF A PULSE TRAIN

To gain a better understanding of the detrimental effects of band limiting on a signal, consider a periodic pulse train of ¼ ms pulse duration and 1 ms period applied to a low-pass filter. Assume the filter is ideal, the skirt has an instant cut-off, and the time delay for the various frequencies is a constant. As an example, look at a 4000 b/s bit stream, where each digit represents a binary state of 1 or 0.

The Fourier series of the pulse stream, shown in Fig. 4-6(a) and explained in Appendix A, is given by:

$$v(t) = \frac{At_p}{T} + \frac{2At_p}{T} \sum_{n=1}^{\infty} \text{sinc} \frac{nt_p}{T} \cos \frac{n2\pi}{T} t \qquad (4\text{-}1)$$

where the sine terms are absent due to the even symmetry. The amplitudes for the various frequency components are:

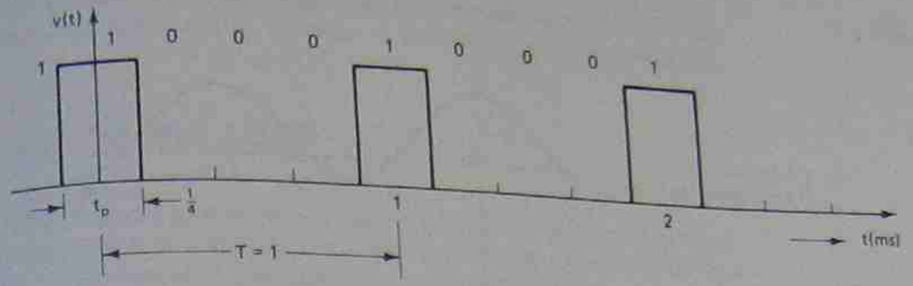
		Peak Amplitude
$DC\left(\frac{At_p}{T}\right)$		0.25
Harmonics $\left(\frac{1}{2}\frac{\frac{\sin n\pi}{4}}{\frac{n\pi}{4}}\right)$	Fundamental (n = 1)	0.450
	n = 2	0.318
	n = 3	0.150
	n = 4	0
	n = 5	−0.090
	n = 6	−0.106
	n = 7	−0.064

The waveform thus consists of harmonics of 1 kHz fading off at a rate of:

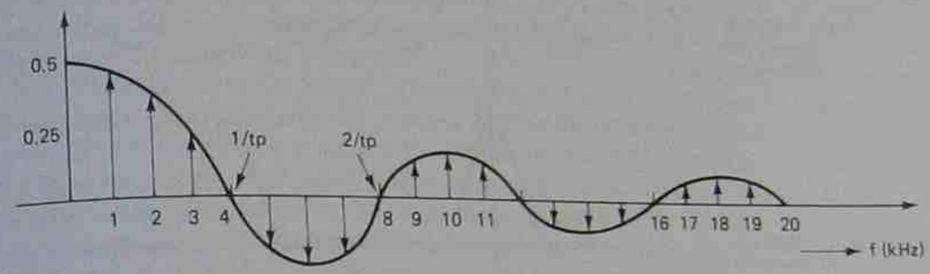
$$\frac{\sin \frac{n\pi t_p}{T}}{\frac{n\pi t_p}{T}}$$

By permitting only a finite number of the harmonics to be present, as determined by the cutoff frequency of the filter illustrated in Fig. 4-7, the transmitted pulse undergoes varying degrees of amplitude distortion. This is illustrated in Fig. 4-8 where the bandwidth of the low-pass filter of Fig. 4-7 is gradually increased. In Fig. 4-8(a), only the dc and fundamental frequencies are transmitted. In Fig. 4-8(f) all harmonics up to and including the sixth are passed unaffected.

By employing a nonperiodic bit stream of the same bit rate, the fre-

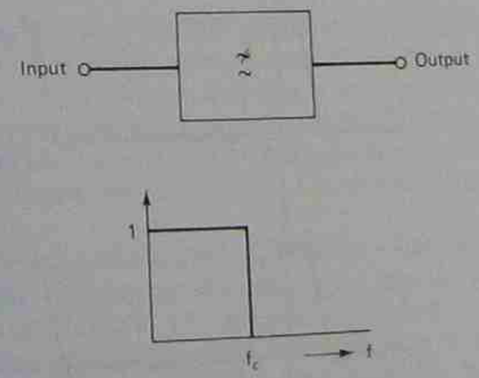


(a) A 4000 b/s Bit Stream



(b) Frequency Spectrum

Fig. 4-6 Frequency Spectrum of a 4000 b/s Bit Stream having a fundamental frequency of 1 kHz



Filter Response

Fig. 4-7 Ideal Low Pass Filter

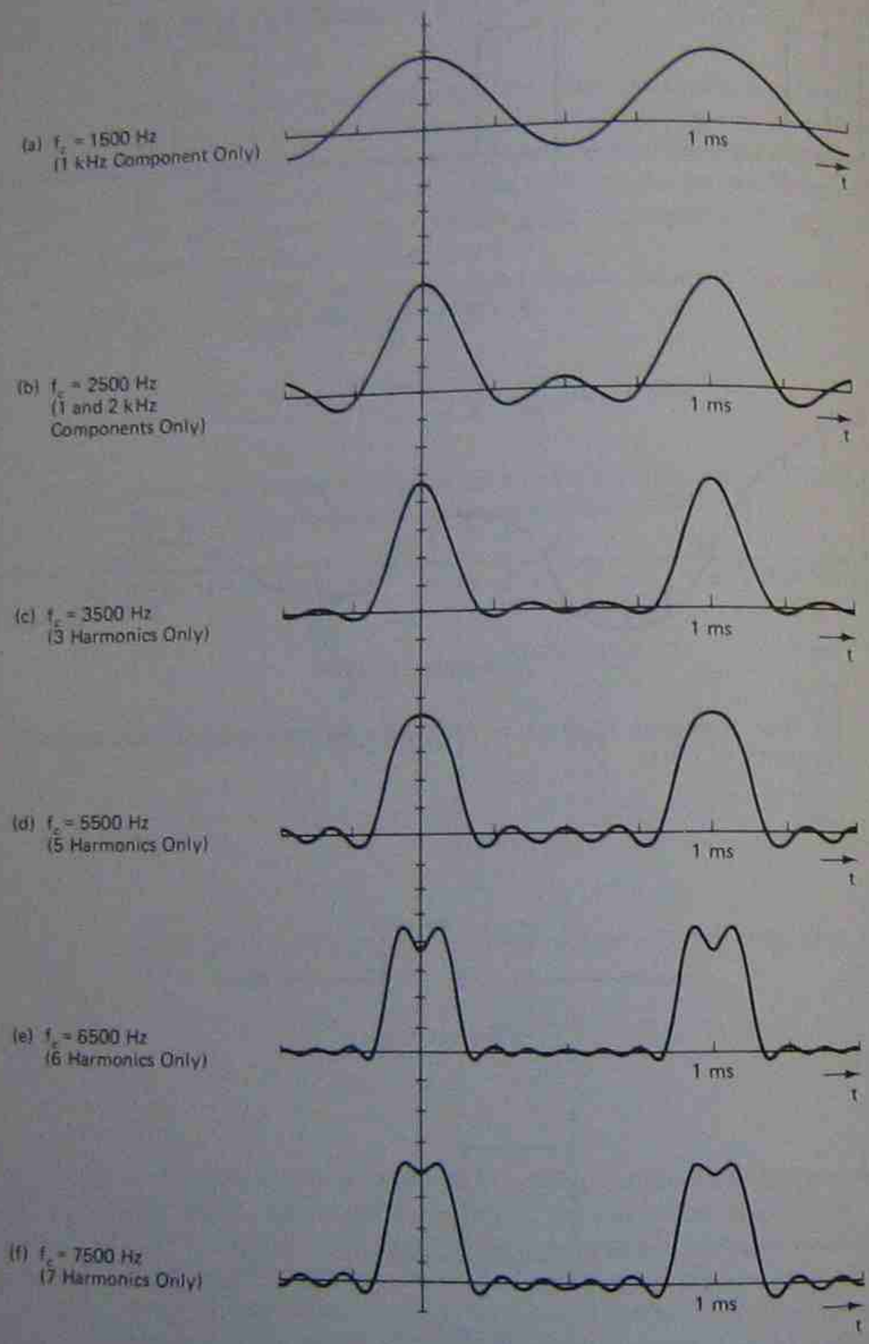


Fig. 4-8 The Effect of Filtering on a Periodic 4000 b/s Bit Stream Having a Fundamental Frequency of 1 kHz.

quency spectrum becomes spread out or continuous rather than discrete. The null points, however, continue to occur at integral intervals of n/t_p.

4–3 20 mA LOOP

A 20 mA loop circuit is usually used to connect computer or data terminal equipment to a teletypwriter. (See Fig. 1-20.) An example of this circuit is the interconnection between an Intel 8080 microcomputer and a teletypewriter (TTY) via an 8251 Universal Synchronous Asynchronous Receiver/Transmitter (USART). The USART accepts data from the microprocessor in parallel format and converts it into a serial data stream for transmission. It can also simultaneously receive a serial data stream and convert it into parallel data characters.

Pin Number 19 of the USART, shown in Fig. 4-9, is the Transmitter Data Connection (TxD) while Pin Number 3 is the Receiver Data Connection (RxD). With TTY strapping, a LO state on the TxD lead causes Q_2 and Q_3 to remain in cut-off preventing any current to flow to the TTY. On the other hand, a HI state on the TxD lead causes Q_2 and thereby Q_3 to go into saturation. Current can now flow from the +12 volt supply via Q_3 and through R_{13}. The TTY and its associated line flow through R_{20} to the negative 10 V supply line.

By selecting the smaller R_{13} value, longer line lengths can be used to the TTY. If a minimum of 20 mA is to be maintained and R_{13} is selected at 390, the line may contribute a maximum total resistance of 280 Ω:

$$\left(20 \times 10^{-3} = \frac{12 + 10}{390 + 430 + R_{line}}\right)$$

When receiving data from the TTY, no current can flow between Pins 2 and 27 of J_3 with the brush contacts open. The −10 volt supply causes Q_1 to remain OFF, resulting in a HI state at its output because of the 5.1 K pull-up resistor. The 7406 inverter causes the RxD pin to see a LO. If the TTY brushes are closed, the +12 V supply causes Q_1 to turn ON, lowering the input voltage to the 7406 inverter. This results in a HI state at the RxD terminal.

Figure 4-9 also shows the standard TTL (Transistor-Transistor Logic) levels strapped to the input/output terminals. In this case 0.0 to 0.8 V represents a logical LO, and 2 to 5 V represents a logical HI.

In general, line lengths must be kept somewhat short, or less than 20 m, in order to keep the pulse distortion at a reasonable level. For longer line lengths, line drivers must be employed. Figure 4-10 illustrates the techniques used for carrying data over various transmission distances.

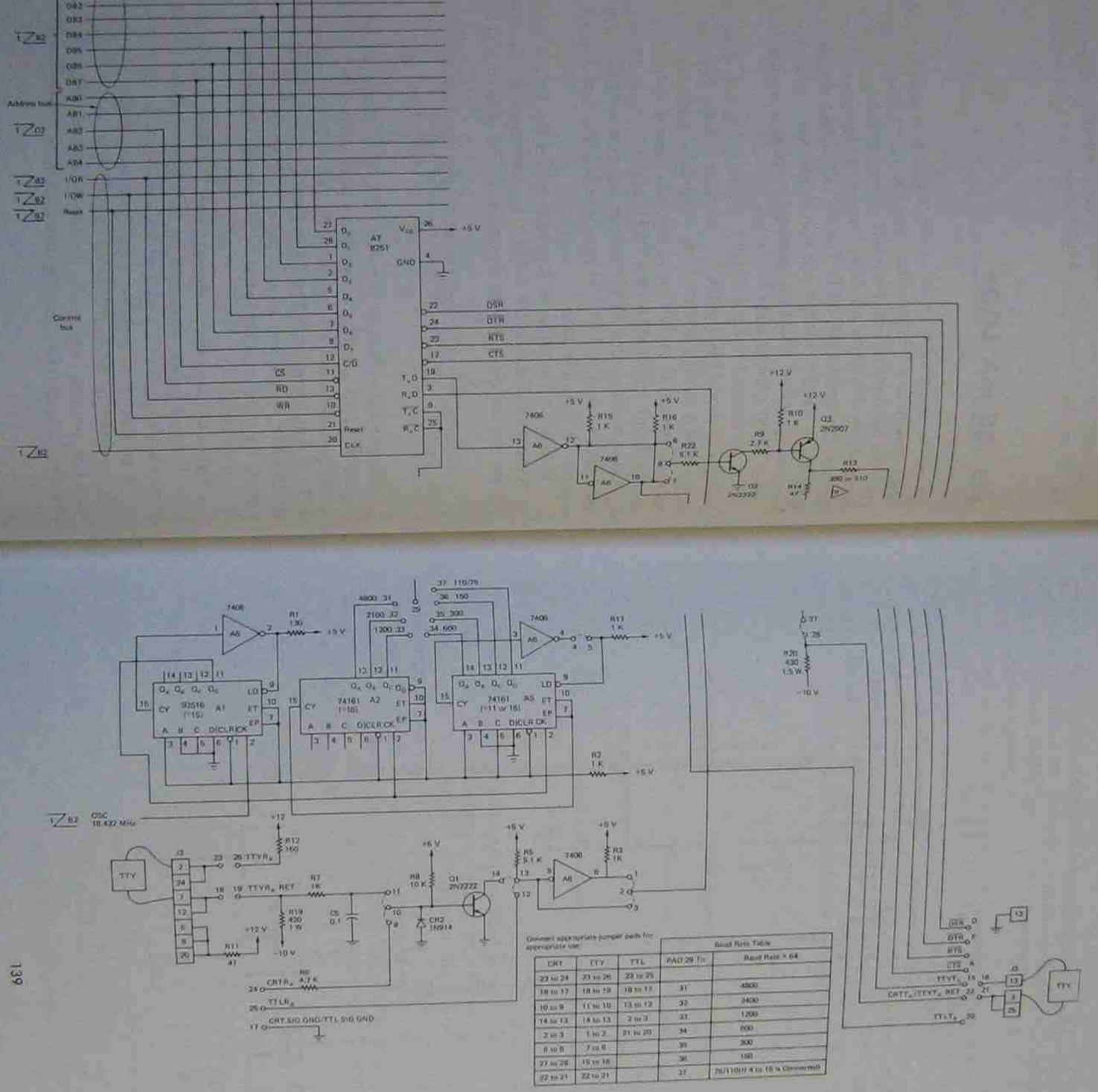


Fig. 4-9 8251 USART Employing 20 mA Current Loop (Courtesy of Intel Corporation)

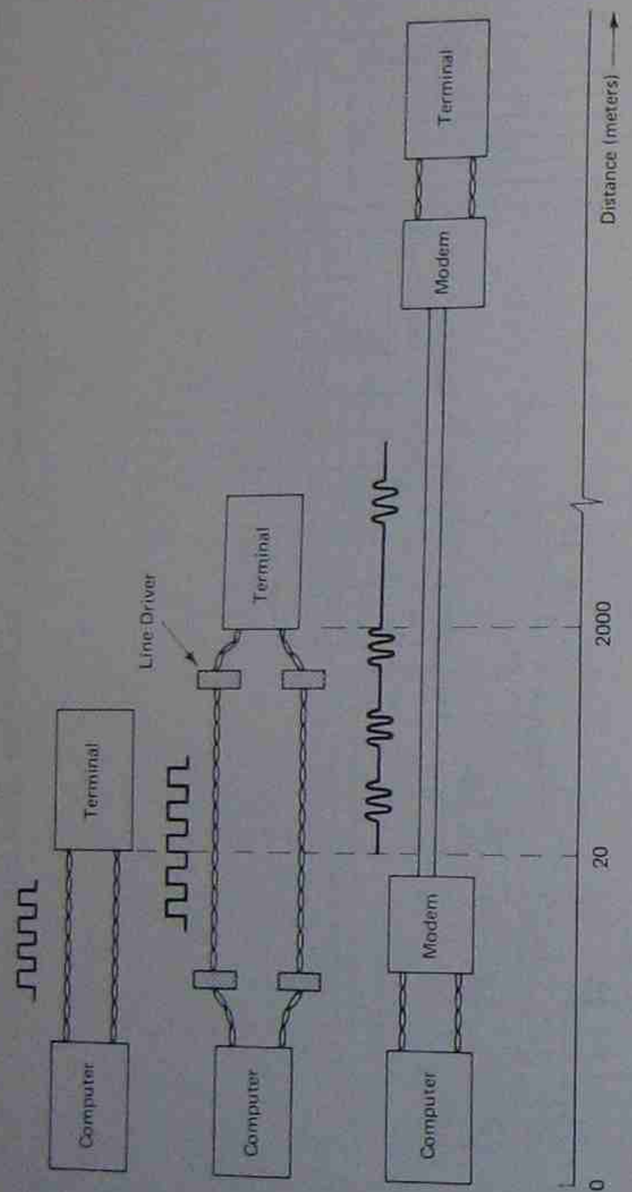


Fig. 4-10 Approximate Maximum Distance for Various Methods of Data Transmission

4–3.1 Line Drivers

The line driver system reduces the line capacitance problem by translating the signal into a low impedance output so that the time constant is lowered to an acceptable value. In this manner the pulse rise and fall times are greatly reduced.

Although unbalanced lines in which one line is kept at ground are still in common use, it is becoming increasingly popular to go to a balanced system. This is clearly illustrated by the recently developed EIA RS–422 standard recommending a balanced interface as opposed to the long established RS–232–C standard which recommends an unbalanced line. The former permits data rates of 20k b/s over a 200 ft distance, whereas the latter is limited to a 50 ft distance for the same bit rate. Examples of balanced and unbalanced lines are shown in Fig. 4-11.

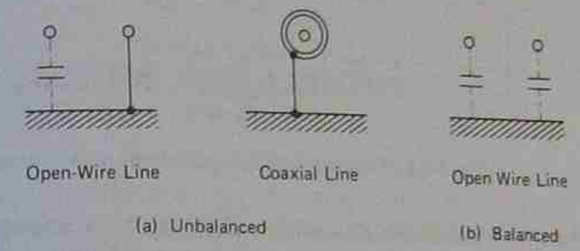


Fig. 4-11 Examples of Balanced and Unbalanced Lines

To more clearly understand the advantages of balanced lines, consider two typical applications: a single-ended operation using an unbalanced line and a differential operation using a balanced line. In a single-ended operation, as illustrated in Fig. 4-12, induced noise is added to the signal voltage as a result of magnetic or electric coupling of adjacent signal lines and as a result of any ground voltage difference due to the presence of 60 Hz. There is no way of removing such noise once it is present in the system.

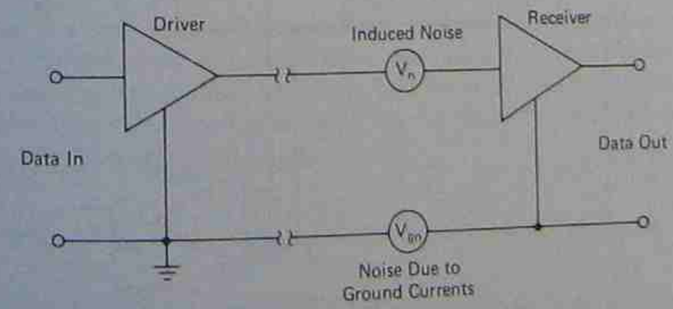


Fig. 4-12 Single-Ended Operation

The differential operation can ignore any ground voltages since they are common to both inputs of the differential receiver. (Refer to Fig. 4-13.) In addition, the differential operation is insensitive to any inductively or capacitively coupled noise which is commonly found in both conductors of the balanced line. Net noise resulting from any unbalance in coupling can be reduced by employing twisted lines which cancel out any transverse magnetic field by alternating the polarity of the induced voltage.

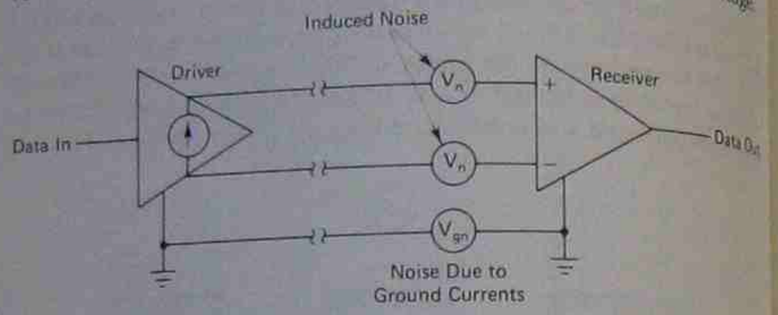


Fig. 4-13 Differential Circuit Operation

Capacitively coupled noise can be minimized by assuming longitudinal balance so that both lines are equally affected by noise. Shielding and keeping all line lengths short also aids in minimizing noise.

An illustration of a properly terminated balanced transmission line, in which each conductor has identical impedances with respect to ground, is shown in Fig. 4-14. This system employs a differential line driver and receiver. Each leg of the line is terminated with an $R_0/2$ re-

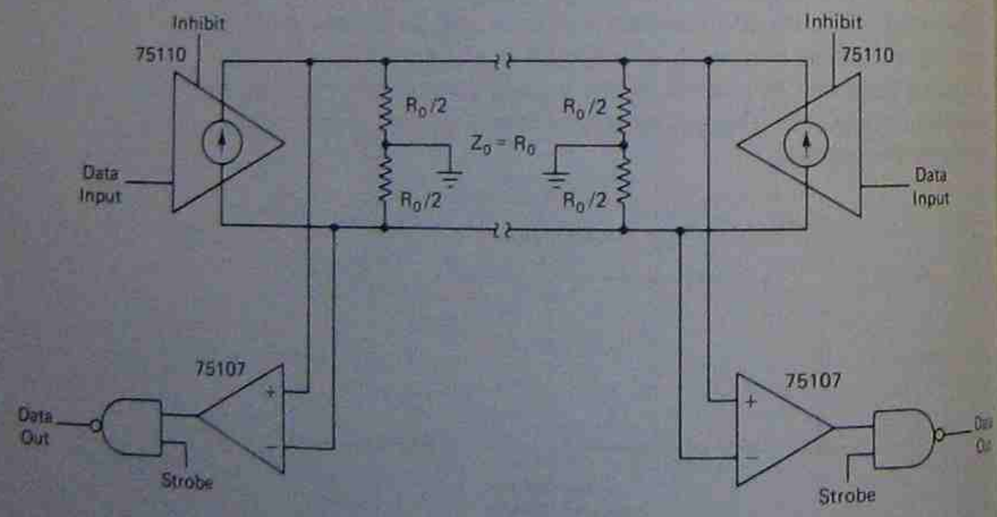


Fig. 4-14 Differential Transmission Employing Fairchild Devices

sistor, resulting in no reflections if a balanced line of characteristic impedance R_0 us used. Data is inserted on the balanced line by unbalancing the line voltage with driver current. The strobe feature will allow line-sharing or party-line operation, thus conserving twisted-pair line costs.

When the transmitting distances go beyond a few thousand meters, modems or data sets must be used to convert the digital or two-state signals to analog signals at the transmitting end and to perform the reverse function at the receiving end. The resulting analog line signal has less distortion than if the original baseband had been employed and therefore can be transmitted over much longer distances. Inserting appropriate line drivers can further increase the transmission distance.

4–4 CABLING PRECAUTIONS

When considering the use of already existing cables at higher frequencies, such as in the case of digital transmission, special precautionary procedures must be followed. Loading coils, which act as low-pass filters, must be removed. Building-out networks, which are added to make a cable appear electrically longer in order that standard loading plans can be used, should be removed since they severely attenuate the high frequency components in a signal. In addition, taped circuits and cross-split pairs should be removed. The latter is caused by improperly sorting out continuous pairs during splicing. Unused taped circuits are often left in place when a customer's connecting loop is cut to disconnect him from the line. These loose ends can cause excessive phase distortion because of the added capacitance. All are illustrated in Fig. 4-15.

4–5 BAUDS AND BITS

A baud is a term used in communication systems to indicate the rate at which changes occur in a signal over a given period of time. It is the unit of signaling speed, or modulation rate, and is found by taking the reciprocal of the shortest signaling element. For the Baudot code illustrated in Fig. 1-18, the modulation rate is:

$$\frac{1}{0.0135} = 74 \text{ baud}$$

A signal which can take on *n*-signaling conditions, such as a multi-frequency or multi-level signal having *n* unique frequencies or levels, can

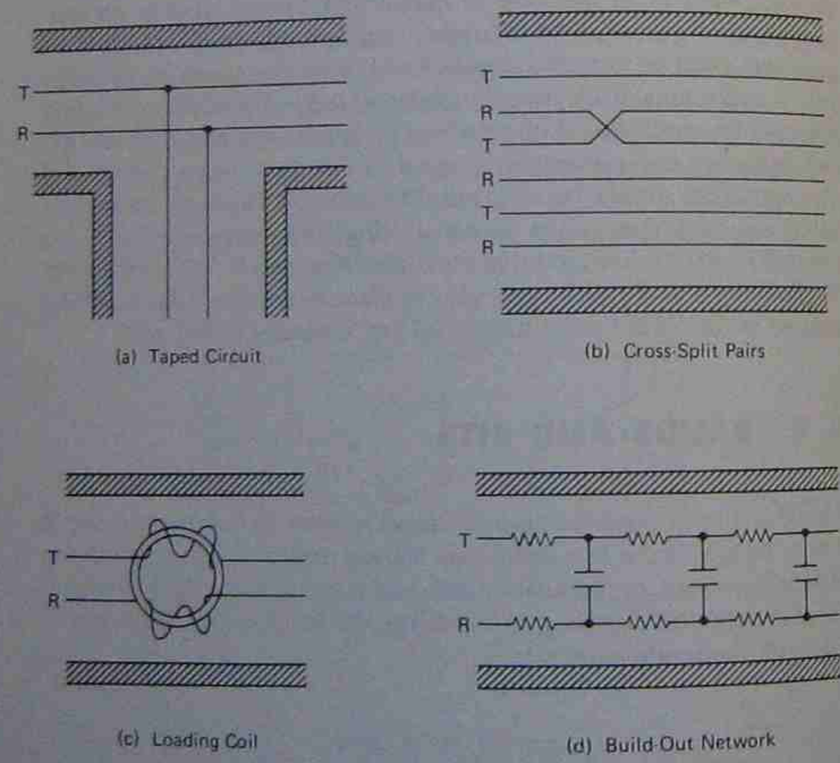
be represented in its various conditions by $\log_2 n$ bits. For instance, a four-level signal can be represented by $\log_2 4 = 2$ bits; i.e., 00, 01, 10 and 11.

To express the rate at which information can be transmitted, state the data signaling rate in bits per second (b/s). For serial transmission this is defined as:

$$\frac{1}{T} \log_2 n \text{(b/s)} \qquad (4\text{-}1)$$

where T = length of the unit signal element
n = the number of signaling conditions

The Baudot code illustrated in Fig. 1-18 has a data signaling rate of $1/.0135 \log_2 2 = 74$ b/s, which is the same as the baud rate. This parallelism occurs in any two-level system since each state can be represented by one bit; that is, there is one bit per baud. For a four-level system, there are two bits per baud; for eight levels, three bits per baud, and so forth.



(a) Taped Circuit

(b) Cross-Split Pairs

(c) Loading Coil

(d) Build-Out Network

Fig. 4-15 Circuits or Components That Need to Be Eliminated When Converting to Digital Transmission

In an eight-state signaling system with 10 mS unit signaling elements, the data signaling rate would be

$$\frac{1}{0.01} \log_2 8 = 300 \text{ b/s}$$

The signaling rate is $\frac{1}{0.01}$ or 100 baud.

To obtain the signaling rate for a parallel system, the signaling rate for each channel is calculated and the results summed.

Take the case of a modem that transmits a constant amplitude signal and takes on one of four different phases. The four phases can be represented by two bits as 00, 01, 10, and 11. In this case, each baud represents two bits of information. For a 1200 baud system, the information rate will be 2400 b/s.

The Data Transfer Rate is another term used to describe data transfer rates on an operating system. It is defined by the CCITT as "the average number of bits, characters or blocks per unit time passing between corresponding equipment in a data transmission system." Since systems normally introduce errors, the information bits accepted by a terminal are less than the number that are transmitted. Also, as in the case of the Baudot code, the start/stop elements do not contribute to the information received. Assuming error free transmission, 2500 bits would be received for 500 five-unit Baudot characters transmitted. This is an average of $2500/(500 \times 0.1) = 50$ b/s for the Baudot code of Fig. 1-18.

The average is less than the data signaling rate of 74 b/s because of the redundant start/stop elements. In a real system, the data transfer rate would be even less than this because of errors. Because data transfer rates depend upon the characteristics of a particular system, b/s represents signaling rates in this text.

4–6 MODEMS

In order to transmit data over existing analog common carrier equipment, modems are inserted at each end to convert information from digital to analog and back to digital. This is illustrated in Fig. 4-16.

The modem can be directly or indirectly coupled to the communications line. With indirect, or acoustic coupling, the data signals are converted into audible sounds which are readily picked up by the telephone handset microphone, or transmitter. The audible signals are then converted back to electrical signals and transmitted over the telephone system.

Because of the extra conversion involved, acoustical coupling tends to introduce more noise and distortion than direct coupling and hence is limited to information rates of less than 1200 b/s.

For permanent installations, direct hook-up should be chosen over acoustical coupling. When a non-Bell modem is used, a Data Access Arrangement (DAA) device or some other suitable protective device must be inserted to make a connection to the Bell-switched network. A DAA is not required on leased lines.

The data signal spectrum, which looks somewhat like the curve in Fig. 4-6, cannot be directly transmitted over an analog channel that has the characteristics shown in Fig. 2-6. The data contains significant low-frequency components that would be attenuated by the telephone channel. The modem performs the function of translating the digital signal spectrum to an appropriate spectrum for the analog system.

4–6.1 Turn Around and Message Acknowledgement

A typical modem might use two different frequencies to represent the MARK or SPACE input data. The frequencies are operated in either the half-duplex or full-duplex mode.

In half-duplex operation, the operation is bidirectional but only in one direction at any given time. When modems switch from transmit to receive and vice versa, some data transmission time is lost in the line turn-around process.

Many systems require the receiver to send an acknowledgement back to the transmitter whenever a block of data is detected. This transmission reversal significantly reduces net data throughput since time is expended to assure the establishment of a proper link; i.e. appropriate signal levels and synchronization.

As a compromise, a narrow reverse channel bandwidth is often used simultaneously with the main, or data-channel, transmission. The reverse or secondary channel does not require a large bandwidth since only a positive acknowledgement (ACK) or a negative acknowledgement (NAK) need be returned. Such a system is still considered a half-duplex system despite the fact that both forward and reverse channels coexist.

Bell System 202 modems employ a main data channel which denotes a MARK with a 1200 Hz tone and a SPACE with a 2200 Hz tone. Such frequency shift keying (FSK) causes the data channel to extend over

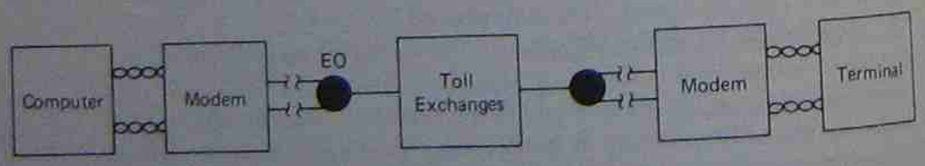


Fig. 4-16 Data Link Employing Modems

a bandwidth as shown in Fig. 4-17. The secondary channel, which is narrowly centered around 387 Hz, can be used to inform the transmitting station that the previous message block contained errors and should be retransmitted.

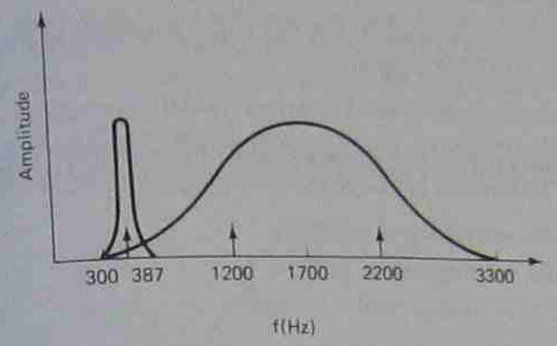


Fig. 4-17 Frequency Utilization of a Half-Duplex Bell System 202 Modem

Full-duplex operation is characterized by simultaneous bidirectional data transmission. Bell System 103/113 modems used on a two-wire line divide the channel bandwidth into two transmit and receive subchannels. When the modem is in the originate mode, it transmits on 1070/1270 Hz. When it is in the answer mode, it transmits on the 2025/2225 Hz frequencies. The modem goes into the originate mode when a call is originated through a modem. It goes into the answer mode if it initially answers a call. Figure 4-18 illustrates the spectrum utilization of a typical full-duplex modem using a two-wire line.

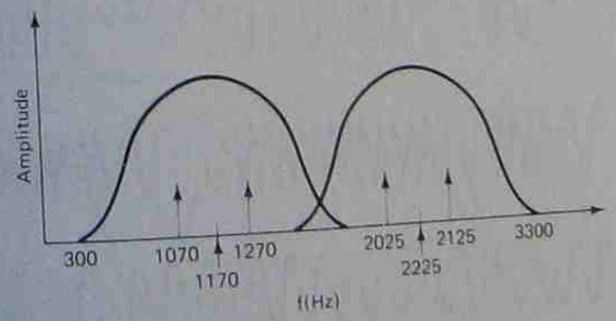


Fig. 4-18 Frequency Utilization of a Full-Duplex Bell System 103/113 Modem

4–6.2 Modem Modulation Techniques

Three basic types of modulation schemes used to convert digital data into suitable analog signals for the telephone lines are:

a. Amplitude Shift Keying (ASK)
b. Frequency Shift Keying (FSK)
c. Phase Shift Keying (PSK)

In addition, a combination of ASK and PSK is employed at the higher bit rates to produce what is called:

d. Quadrature Amplitude Modulation (QAM)

For digital data transmitted at low bit rates, the modulated signal usually carries the information in one of the following binary forms:

a. Two discrete voltage levels (ASK)
b. Two discrete frequencies (FSK)
c. Two discrete phases (PSK)

Each of these one-bit-per-state schemes is illustrated in Fig. 4-19.

Because FSK is relatively simple in its operation and thus inexpensive to implement, it is used extensively in low-speed modems for data rates below 1200 b/s. FSK is not efficient, however, in its use of spectrum space.

For the higher data rates, multiphase PSK synchronous data transmission is generally used. Typically, four-phase PSK is used at 2400 b/s

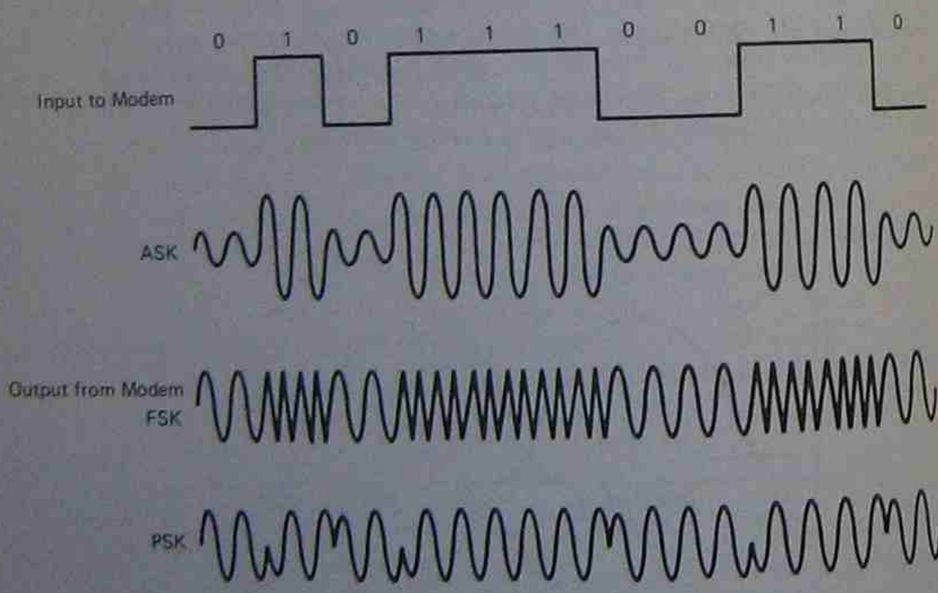
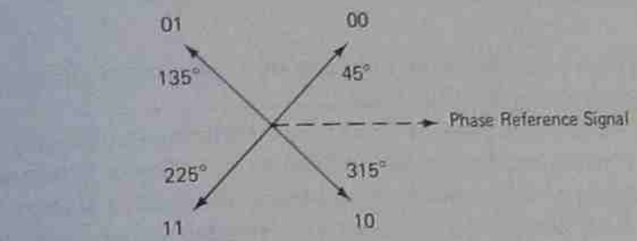


Fig. 4-19 One Bit Per Pulse Moduation Schemes Suitable for Telephone-System Transmission.

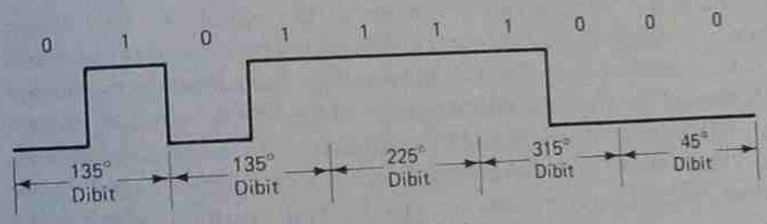
and eight-phase PSK is used at 4800 b/s. For the four-phase PSK system, the data stream is divided into blocks of two bits each, called dibits, with each dibit related to one of the four phases. A 1200 baud system would operate at an information rate of 2400 b/s.

A typical waveform of a four-phase system is shown in Fig. 4-20. In this example, the Gray code is used in which adjacent phases undergo only a one bit change in the two bits. This scheme reduces the effects of noise on a data signal. In the straight binary code, going from a 1(01) to a 2(10) represents a two-bit error whereas for the Gray code only a single bit error is produced.

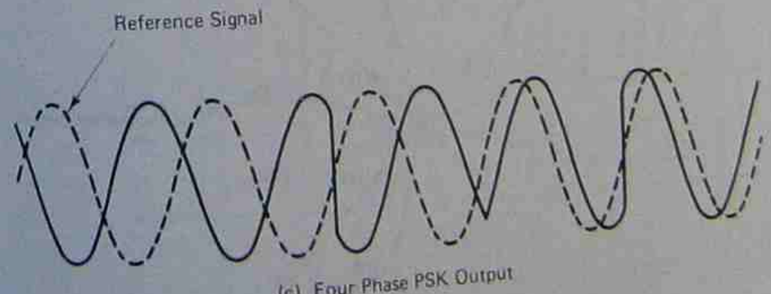
Figure 4-21 illustrates the phase diagram of an eight-phase, Gray-encoded PSK signal. Each phase is represented by three bits, called a



(a) Phaser Diagram



(b) Input to Modem



(c) Four Phase PSK Output

Fig. 4-20 Four-Phase PSK Signal

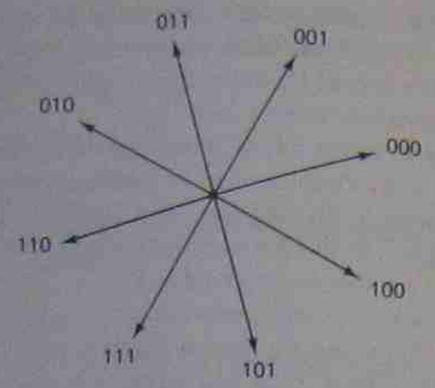


Fig. 4-21 Eight-Phase PSK Phase Diagram

tribit. Moving from phase to phase, only one bit undergoes a change—the chief characteristic of the Gray code.

For data transmission rates of 9600 b/s, a technique employing a combination of phase and amplitude modulation is finding acceptance. The phase diagram of this quadrature amplitude modulated (QAM) is shown in Fig. 4-22. Because 16 different Gray-encoded signal states are possible, $\log_2 16$ or 4 bits can be transmitted per signaling pulse. Thus, for a 2400 baud signal, a 4 × 2400 or 9600 b/s information rate is attained.

Table 4-1 offers a quick glimpse at the variety of Bell system modems presently in use. Note that for the higher bit rates, a separate pair of wires is used for each direction (4-wire line). It is common to equalize or flatten out the electrical characteristics of the line for bit rates exceeding 2400 b/s to achieve satisfactory performance. Generally, the attenuation characteristic and the propagation velocities experienced by the various frequency components in a signal are not constant, which results

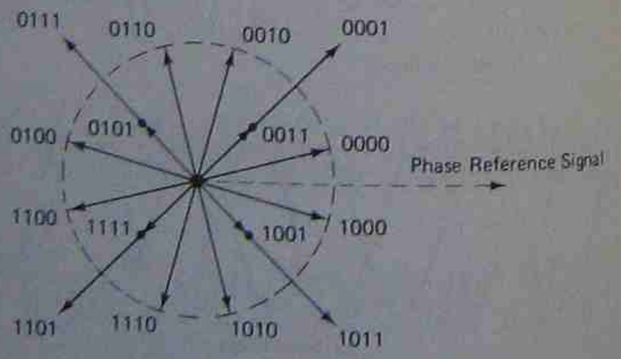


Fig. 4-22 QAM Phasor Diagram

Table 4-1 Bell System Modem Types*

	103J	113D	202S	212A	202T	208A	208B	209A	LADS	500A Data Service Unit
Data rates (bits/s)	Up to 300	Up to 300	Up to 1200	Up to 1800	Up to 1200 and 1200	4800	4800	9600	2400/4800/7200/9600/19,200	2400/4800/9600/56,000
Channel types Line interface Line conditioning	2-wire switched None	2-wire switched None	2-wire switched None	2/4-wire private C2 over 1200 bits/s	2-wire switched None	4-wire private None	2-wire switched None	4-wire private D1	4-wire[1] None	4-wire[2] N/A
Transmission mode[3]	Full duplex	Full duplex	Half duplex[4]	Half-duplex (2-wire) Full-duplex (4-wire)	Full duplex	Full duplex	Half duplex	Full duplex	Full duplex	Full duplex
Modulation scheme Diagnostics	FSK Remote & local loopback, Self-test	FSK Remote & local loopback, Self-test	FSK Remote & local loopback, Self-test	FSK Remote & local loopback, Self-test	FSK Remote & local loopback, Self-test	PSK 4-Phase Remote & local loopback, Self-test	PSK 8-phase Remote & local loopback, Self-test	QAM Remote & local loopback, Self-test	PSK 2-phase Remote & local loopback	None—digital regeneration Remote & local loopback
Equalization Line turnaround Synchronization	Fixed — Synchronous	Fixed — Synchronous	Fixed — Synchronous	None Asynchronous or Synchronous[5]	None Asynchronous	Automatic — Synchronous	Automatic — Synchronous	Automatic — Synchronous	Fixed None Synchronous	N/A None Synchronous
Business-machine interface	RS232C	RS232C	RS232C	RS232C	RS232C	RS232C	RS232C	RS232C	RS232	RS232C CCITT V.35[6]

1. Local Area Data Set (LADS) operates over Type-B3 Local Area Data channel only.
2. 500A Data Service Unit operates over private DATAPHONE (service mark of AT&T) Digital Service (DDS) channel only.
3. Transmission mode indicated is that of the transmission facility; full-duplex lines support half-duplex protocols only.
4. 202S and 202T are optionally available with a 5-bit/s, 387-Hz reverse channel.
5. Asynchronous operation at 0 to 300 bits/s or 1200 bits/s; synchronous operation at 1200 bits/s.
6. EIA RS232C interface provided at 2400, 4800, 9600 bits/s; CCITT V.35 interface provided at 56 kbits/s.

* Reprinted with permission from *Electronic Design*, vol. 27, no. 16, © Hayden Publishing Co., Inc., 1979.

in a rather distorted signal at the receiving end. These variations in frequency can be compensated out by employing line equalizers. For leased lines (permanent individualized or private links), the equalizers can be manually adjusted. Automatic equalizers are also available, but they consume up to 50 ms in the equalizing process. This initial training time is significant because it reduces the net data throughput. Techniques are rapidly being developed to reduce this training time.

The power spectrum of a typical voice band data signal is quite different from that of the human voice. The power spectral density of most data signals is designed to have a raised cosine spectrum centered at about 1800 Hz. This is typified by the signal shown in Fig. 4-17 and again in Fig. 4-23, which shows the power spectrum of a 1200 baud, four-phase PSK voice band data signal.

Most voice band channels are designed for speech signals (see Fig. 2-6) because of the different characteristics in spectral density of voice band data. The performance is less than optimal when voice band data is transmitted over the same system. In the future, coding systems may be implemented in the modems to take this into account.

4–6.3 Theory of Phase Shift Keying

PSK is a commonly used digital modulation technique which is very easy to handle analytically. This section will discuss a two-phase or binary phase shift keying system and then briefly discuss the implementation of a four-phase PSK system.

Binary phase shift keying (BPSK) is obtained through the system shown in Fig. 4-24. The data signal is translated into a non-return-to-zero (NRZ) code having no dc component. The resulting pulses of amplitudes $+A$ or $-A$ represent the binary 1's and 0's. The muliplier output is given by:

$$v_o' = \pm A \cos \omega_c t \tag{4-2}$$

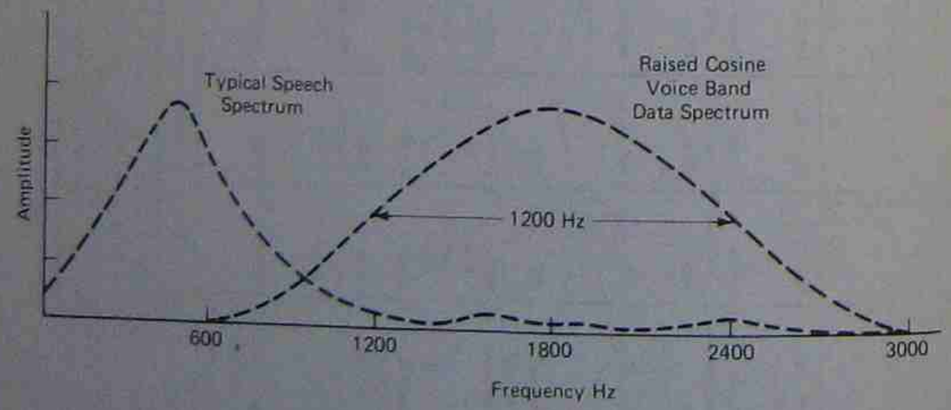


Fig. 4-23 Spectral Density of Speech and Voice Band Data: Four-Phase PSK 1200 baud 2400 b/s.

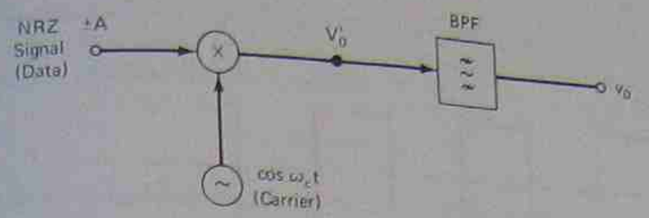


Fig. 4-24 The BPSK Modulator

where it is assumed that the gain of the multiplier is unity and where $f_c = \omega_c/2\pi$ is the unmodulated carrier frequency.

Since $-\cos \omega_c t = \cos (\omega_c t + 180°)$, the output voltage can be expressed by:

$$V_o' \rightarrow \begin{cases} A \cos \omega_c t \\ \text{or} \\ A \cos (\omega_c t + 180°) \end{cases}$$

This modulated signal has a constant maximum amplitude of A and can alternate only at 0° and 180° phase.

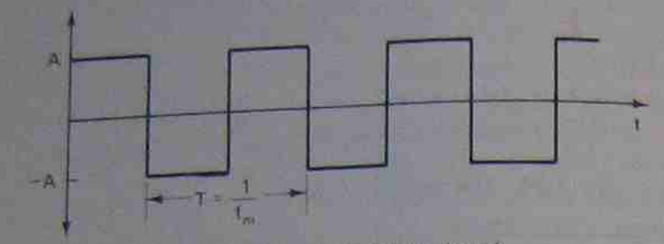
To obtain a general impression of the v_o' spectrum, consider the modulating NRZ signal to consist of a train of square wave pulses without a dc bias. If the pulse rate is f_m, the spectrum of this pulse train would have the typical $(\sin x)/x$ amplitude response for the odd harmonics of f_m as shown in Fig. 4-25.

When one sinusoid is multiplied by another sinusoid, the following sum and difference frequencies are noted:

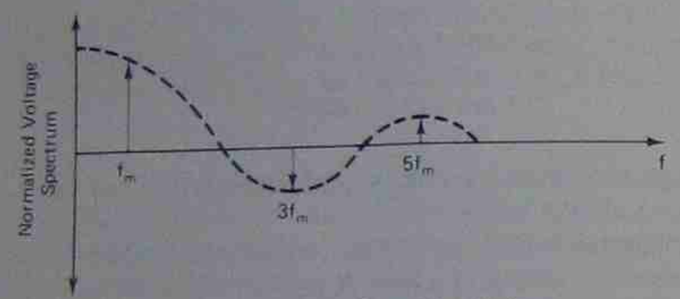
$$\cos \omega_1 t \cos \omega_2 t = \frac{1}{2} \cos (\omega_1 + \omega_2)t + \frac{1}{2} \cos (\omega_1 - \omega_2)t \tag{4-3}$$

Thus when the square wave, which contains only odd harmonics of f_m, is multiplied by the sinusoid of frequency f_c, as dictated by Eq. (4-2), the harmonics of f_m are obtained as sidebands around the frequency f_c where $f_c \gg f_m$. This is shown in Fig. 4-26. Since no dc is present in the NRZ signal, the carrier at f_c is suppressed.

The spectrum indicates that the bandwidth taken up by the modulated signal is twice the baseband or modulation bandwidth. The bandpass filter is inserted to limit the transmitted spectrum. A possible negative effect of adding such a filter might be a non-uniform time delay of the various frequency components in the signal, particularly for those lying near the filter skirts if the filter is not correctly designed. This results in the system being more susceptible to noise. The block diagram of a corresponding BPSK demodulator or detector is shown in Fig. 4-27.



(a) Time Domain Representation of the NRZ Signal



(b) Spectrum of the Square Wave Having a Pulse Rate of f_m

Fig. 4-25 Frequency and Time Domain Representations of a NRZ Square Wave

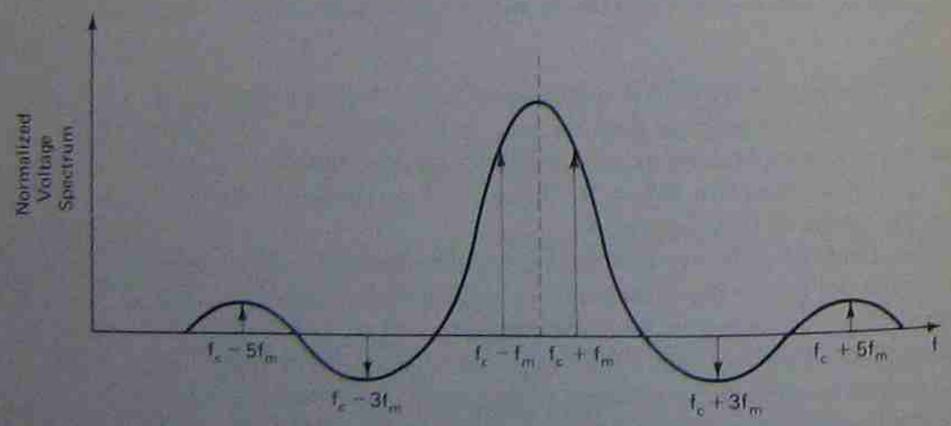


Fig. 4-26 Spectrum of a Square Wave Modulated Sinusoidal Signal

Ignoring all filtering effects of the filter and transmission medium aside from the inherent delay of θ radians between the transmitter and receiver, the received signal, v_s, will be:

$$v_s = \pm A \cos(\omega_c t + \theta) \tag{4-4}$$

In order to demodulate this signal, the carrier at the correct phase must be extracted. This is accomplished in the carrier recovery section by squaring the received signal and then filtering out the desired second harmonic with a bandpass filter (or a phase-lock loop). The squared signal can be expressed as:

$$v_s^2 = A^2 \cos^2(\omega_c t + \theta) = \frac{A^2}{2} + \frac{A^2}{2} \cos 2(\omega_c t + \theta) \tag{4-5}$$

where the latter term represents a frequency component at twice the carrier frequency. The divide-by-two circuit yields the desired $\cos(\omega_c t + \theta)$ carrier.

Mixing or multiplying the extracted carrier with the received modulated signal results in:

$$\pm A \cos(\omega_c t + \theta) \cos(\omega_c t + \theta) = \underbrace{\pm \frac{A}{2}}_{\text{desired signal}} \underbrace{\pm \frac{A}{2} \cos 2(\omega_c + \theta)}_{\text{removed by LP filter}} \tag{4-6}$$

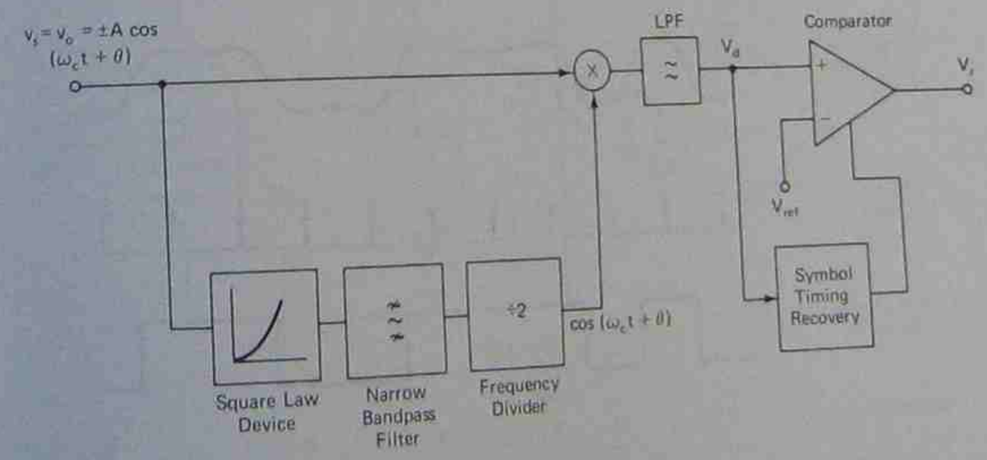


Fig. 4-27 A BPSK Demodulator

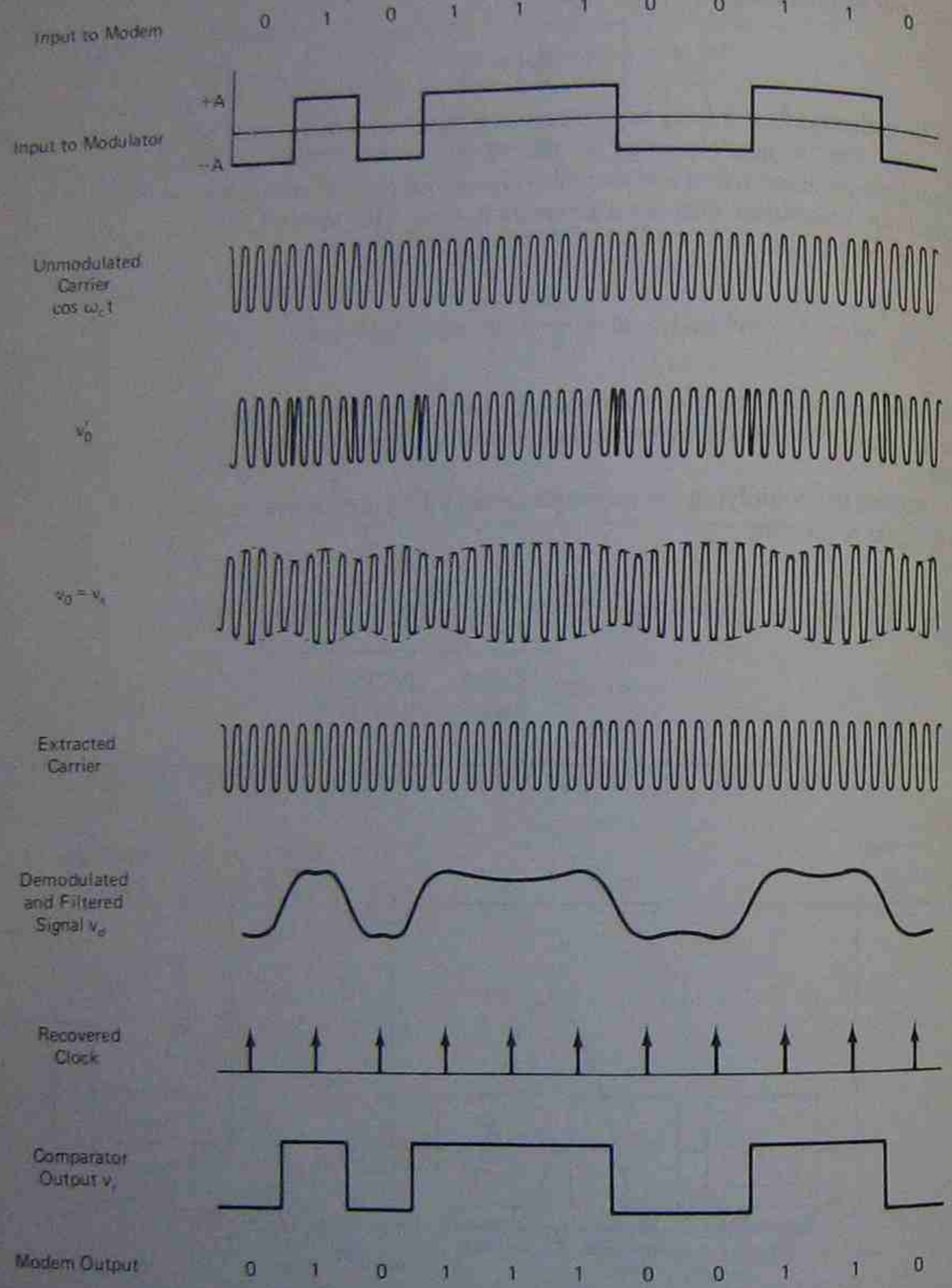


Fig. 4-28 Waveforms of the BPSK Modem

The desired signal, which contains the original information, is extracted by the low-pass filter. The comparator circuit, which is clocked at the symbol rate, finally determines whether a 1 or a 0 logic state is detected. Typical waveforms at the indicated locations of Figs. 4-24 and 4-27 are illustrated in Fig. 4-28.

To achieve higher bit rates, four- or eight-phase modulation is used in which each interval or symbol carries two or three bits of information. At the transmitter, a serial-to-parallel conversion is performed to group the incoming data stream into groups of two or three bits, depending upon whether four- or eight-phase modulation is used. The receiver performs the complementary parallel-to-serial conversion.

Figure 4-29 shows the block diagram of a four-phase modulator. The serial-to-parallel encoder pairs up the binary digits, called dibits, which results in a transmission speed of the A and B channels that is half that of the incoming serial data stream. Figure 4-30 shows the channel waveform corresponding to the signal earlier shown in Fig. 4-20.

The lower modulator in Fig. 4-29 is fed with a 90° carrier ($\cos \omega_c t + \pi/2$) and the upper modulator with a 0° carrier ($\cos \omega_c t$). When the lower modulator obtains an input A digit of 0, the output consists of a 90° carrier. When the input A digit is a 1, the output consists of a −90° carrier. Similarly, the output from the upper modulator is at 0° or 180°, depending upon whether the input B digit is an 0 or a 1. For instance, if the binary values of A and B represent the voltage signals listed in Table 4-2, the corresponding output voltage from the modulator can be

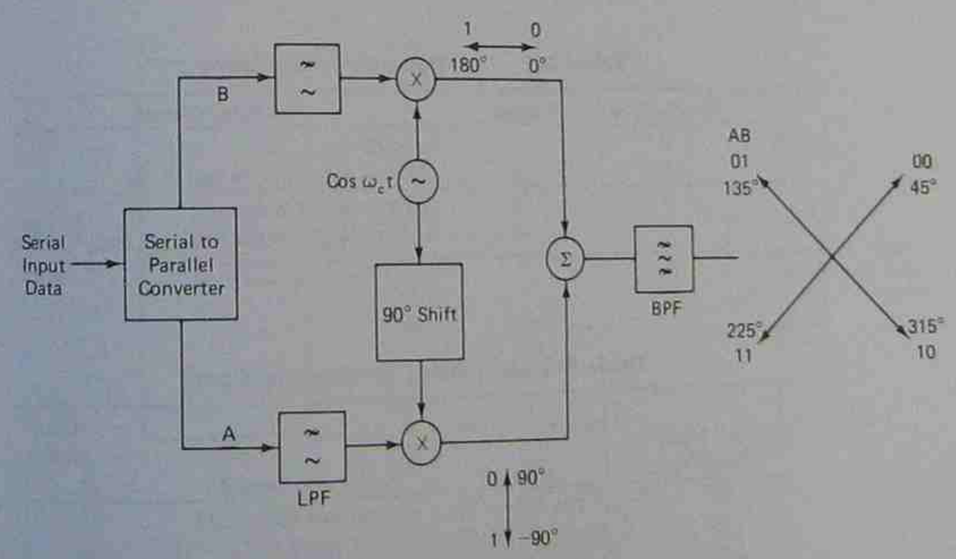


Fig. 4-29 Four-Phase Modulator

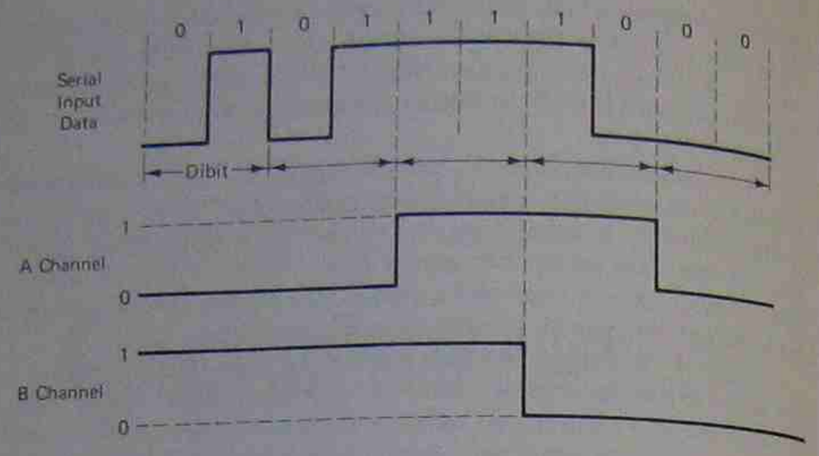


Fig. 4-30 Channel Waveforms from a Four-Phase Serial-to-Parallel Encoder

represented by the phasor shown in Fig. 4-29. By summing the two output signals from the modulators, the output phasor diagram is obtained. (See Fig. 4-20.)

The four-phase signal can be readily detected if a fixed reference phase signal is available at the receiver. Figure 4-31 shows such a receiver.

Table 4-2 Four-Phase Modulator Signal

	Binary Value	Voltage	Modulator Output
B	0	V	$V \cos \omega_c t$
	1	$-V$	$-V \cos \omega_c t = V \cos (\omega_c t + \pi)$
A	0	V	$V \cos (\omega_c t + \pi/2)$
	1	$-V$	$-V \cos (\omega_c t + \pi/2) = V \cos (\omega_c t + \pi/2)$

Table 4-3 Four-Phase Demodulator Signal

Received Signal Phase	A Output	B Output	Data Output A	Data Output B
45°	½ cos 45° = .35	½ cos 45° = .35	0	0
135°	½ cos 45° = .35	½ cos 135° = −.35	0	1
225°	½ cos 135° = −.35	½ cos 225° = −.35	1	1
315°	½ cos 225° = −.35	½ cos 315° = .35	1	0

Using Eq. (4-3) with a representative received signal of cos $(\omega_c t + \theta)$, the output from the multipliers or demodulators will be:

$$\cos (\omega_c t + \theta) \cos (\omega_c t + \phi) = \frac{1}{2} \cos (\theta - \phi) + \frac{1}{2} \cos (2\omega_c t + \theta + \phi) \qquad (4\text{-}7)$$

where ϕ takes on a value of 0 or 90°, depending upon which modulator is being considered.

The low-pass filter will remove the second harmonic components of $2\omega_c$, leaving the term of the form $1/2 \cos (\theta - \phi)$. Since the incoming signal phase must take on one of the θ values of ±45° or ±135°, when this signal is multiplied by the reference oscillator of phase $\phi = 0°$ or 90°, the signal obtained at the LPF output is one tabulated in Table 4-3. The comparator circuits readily detect the dc signal level from the LP filters to note the digital assignment. With the comparators shown, the comparator output goes LO when the input exceeds 0 volts, indicating a 0. The parallel-to-serial converter performs the process of converting the two-bit symbol back to two bits of serial data.

Although it is quite feasible to obtain a fixed phase recovered carrier, it is difficult to obtain the required *absolute* phase. The receiver must be given some indication of the proper phase reference since it is relatively easy for the carrier recovery phase to rest at 90° or 180° from

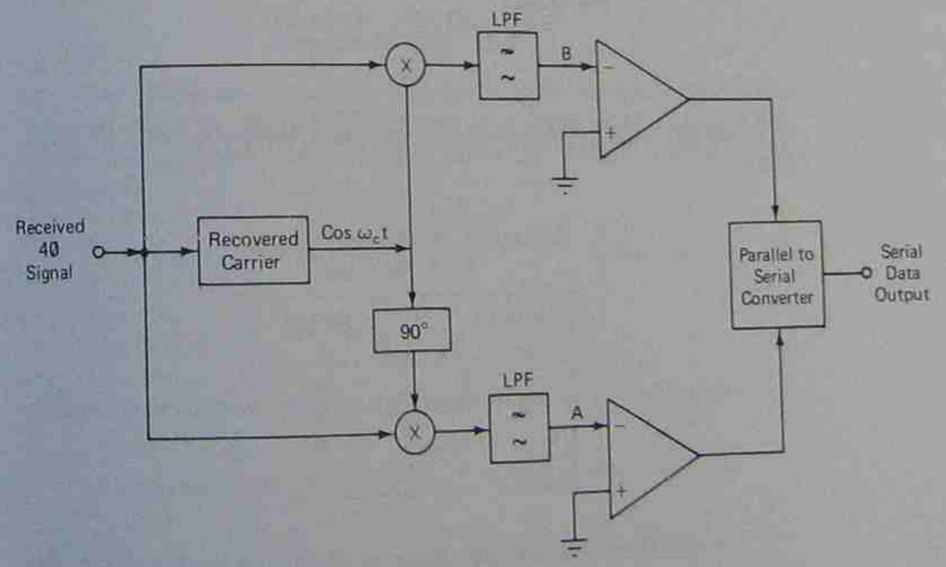


Fig. 4-31 Four-Phase Demodulator

the desired phase reference. Because of this difficulty, differential phase modulation and detection is preferred. In this modulation scheme, the information is encoded in terms of phase *changes* between adjacent symbols, rather than as an absolute phase for each symbol.

In the two-phase, coherent PSK system previously examined, a 1 corresponded to a 0° phase signal and a 0 corresponded to a 180° phase signal. In the modified differential phase modulation system, a 1 may correspond to a 0° *change* between successive signals and a 0 correspond to a 180° *change* between successive signals. At the receiver, each succeeding signal is compared with the phase of the previously received signal. Since noise interferes with *both* the current signal and the reference signal, the signal-to-noise ratio must be slightly greater in order to achieve similar error rates.

The encoding of a differential phase shift keyed (DPSK) signal is shown in the block diagram of Fig. 4-32.

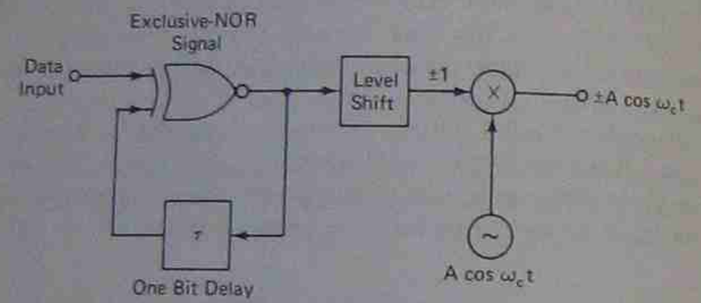


Fig. 4-32 DPSK Modulator

The exclusive-NOR gate has a truth table, which is given in Table 4-4.

Table 4-4 Exclusive-NOR Truth Table

Inputs		Output
A	B	C
0	0	1
0	1	0
1	0	0
1	1	1

The encoding of a data bit stream is illustrated in Table 4-5. To begin the encoded signal, a reference binary digit of 1 is chosen. Because of the delay in the feedback path, the present message bit (data input) is compared with the previously encoded signal bit to arrive at the following digit for the encoded signal. In the example given, the first digit in the message, namely a 0, is compared with the initialization bit of 1, resulting in a 0 appearing as the next encoded bit. This latter 0 bit is the reference for the next digit to be encoded. Comparing this 0 with the next message digit of 1 results in another 0 as the next encoded digit. The resulting encoded message is then level shifted to a + or −1, to be multiplied by the unmodulated carrier $A \cos \omega_c t$, which results in a phase-shifted carrier of 0 or π radian.

Table 4-5 DPSK Example

Data Input		0	1	0	1	1	1	0	0	1	1	0
Encoded Message	1	0	0	1	1	1	1	0	1	1	1	0
reference digit →												
Transmitted Phase	0	π	π	0	0	0	0	π	0	0	0	π

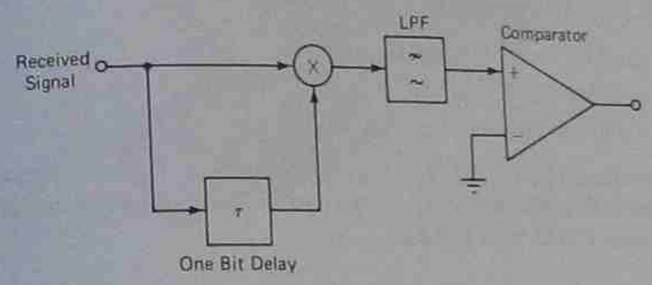


Fig. 4-33 DPSK Demodulator

The block diagram of Fig. 4-33 can be used for demodulation. The received signal is compared bit by bit, or, in other words, adjacent bits are compared. If they are of the same phase, the multiplier output will be positive, i.e.:

$$(A \cos \omega_c t)^2 = [A \cos (\omega_c t + \pi)]^2 = \underbrace{\frac{A^2}{2}}_{\text{desired component}} + \frac{A^2}{2} \cos 2\omega_c t$$

If they are of different phases, the multiplier output will be negative, i.e.:

$$(A \cos \omega_c t)\, A \cos (\omega_c t + \pi) = \frac{A^2}{2} \cos \pi + \frac{A^2}{2} \cos (2\omega_c t + \pi)$$

$$= \underbrace{\frac{-A^2}{2}}_{\text{desired component}} + \frac{A^2}{2} \cos (2\omega_c t + \pi)$$

The dc term is allowed to go through the low-pass filter so it can be applied to a comparator. The comparator translates the dc levels of $+A^2/2$ and $-A^2/2$ to 1's and 0's. When the threshold voltage of 0 volts is exceeded, a 1 results. If the input voltage is below the 0 threshold voltage, a 0 results. The detected message of Table 4-5 is illustrated in Table 4-6.

4–6.4 Interface Standards

Some common interfacing standard must be adopted in order to connect Data Terminal Equipment (DTE) such as computers, teletypewriters, business machines, and printers to Data Communication Equipment (DCE), or, more specifically, Modem or Data Sets. Equipment manufacturers need an agreed upon standard in order to design terminal and communication equipment with minimal interface problems.

Early in the history of data communication network planning, the Electronic Industries Association (EIA) accepted a set of standards known as EIA RS-232-C. This set of standards ensures:

a. Compatible voltage and signal levels,

b. Common pin wiring configurations, and

c. A minimal amount of control information

between the DTE and DCE, as illustrated in Fig. 4-34. These standards are adhered to throughout North America and conform to the internationally recognized CCITT V.24 standard.

Table 4-6 Detection of DPSK Signal Given in Table 4-5

Received Message	0	π	π	0	0	0	0	π	0	0	0	π
Comparator Input		$\frac{-A^2}{2}$	$\frac{A^2}{2}$	$\frac{-A^2}{2}$	$\frac{A^2}{2}$	$\frac{A^2}{2}$	$\frac{A^2}{2}$	$\frac{-A^2}{2}$	$\frac{-A^2}{2}$	$\frac{A^2}{2}$	$\frac{A^2}{2}$	$\frac{-A^2}{2}$
Output Bit Stream		0	1	0	1	1	1	0	0	1	1	0

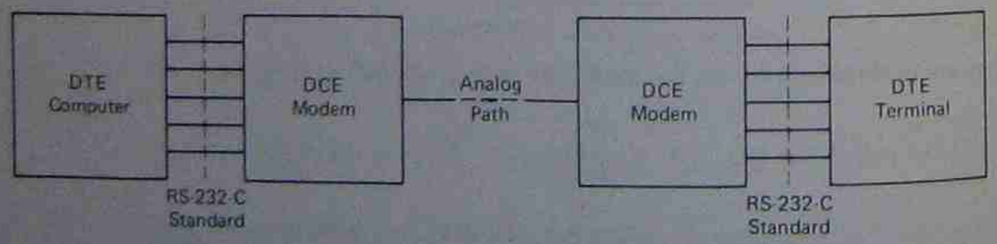


Fig. 4-34 Data Transmission System

The RS-232-C standard not only conforms to the V.24 standard but it also specifies the mechanical connector type and pin assignments for each signal. The 25 pin connector specified by the EIA standard is shown in Fig. 4-35.

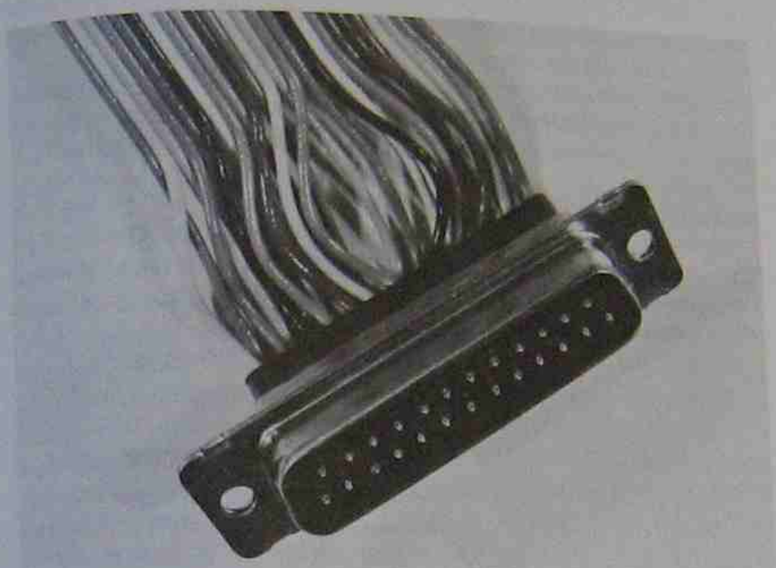


Fig. 4-35 RS-232-C 25 Pin Connector

In addition to the protective and signal grounds, all circuits carry bi-polar voltage signals. Voltages at the connector pins with respect to the signal ground (AB) may not exceed ± 25 V. Voltages between + 3 V lie in the transition region and are not defined. A mark is indicated by a voltage in the −3 to −25 volt range, while a space is indicated by a voltage between 3 to 25 V.

The interchange circuit can be used for control, data, or timing, and its driver can be located in the DTE or DCE. Its equivalent circuit, however, must always meet the criteria illustrated in Fig. 4-36.

$$|V_o| < 25 \text{ V}$$

$$3000\ \Omega \leq R_L \leq 7000\ \Omega$$

$$C_L < 2500 \text{ pF}$$

$$\text{voltage slew rate: } \frac{dv}{dt} < \frac{30 \text{ V}}{\mu s}$$

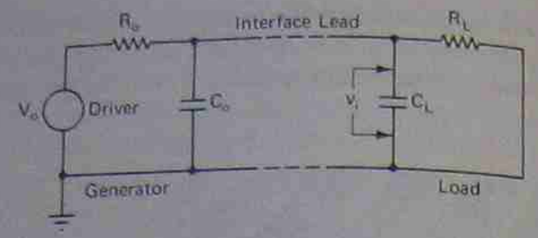


Fig. 4-36 Interchange Equivalent Circuit

For the control circuit, the time in the transition region must be less than 1 ms. For data circuits, the time in the transition region must be less than 1 ms or 4% of the nominal signal element duration, whichever is the lesser. The generator circuit must be able to withstand an open circuit, or a shorting condition, to any other circuit at the interface.

Figure 4-37 illustrates the RS-232-C load signal regions.

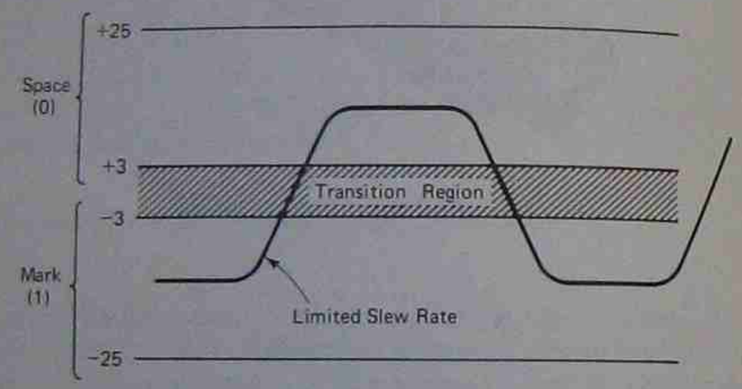


Fig. 4-37 RS-232-C Load Signal

The data and control lead terms can be summarized as follows:

Data Circuit Interface Terms (BA and BB):

Binary state	1	0
Signal Condition	MARK	SPACE
Voltage	Negative	Positive
Paper Tape	Hole	No hole

Control Lead terms:

Voltage	Negative	Positive
Control Function	OFF	ON

This standard is applicable to signaling rates up to 20,000 b/s. The female side of the 25 pin connector is associated with the DCE. Since cables must not present more than 2500 pF at the interface point, the distances between the DCE and DTE are usually kept to below 100 m. Short cables of less than 50 ft (15 m) are recommended. Table 4-7 lists the most frequently used interchange circuits and their connector pin assignments.

Just because a piece of equipment adheres to the RS-232-C specifications does not mean that it will operate with other equipment following the same EIA standards. The EIA standard, for instance, does not control clock timing, idling characteristics, or control characteristics. These may be specified by the user or manufacturer.

To come to some understanding of the RS-232-C specifications, consider the general example of a computer tied to a terminal via a pair of modems operated in the half-duplex mode. Ignore, for the moment, the sequence of events that takes place when the channels are initially established. This will be considered in much more detail on a full-duplex-operated system.

To begin transmitting data on the computer shown in Fig. 4-38 the request-to-send line (RTS) must be turned on to inform the modem to begin transmitting a carrier. After a short delay, the computer's modem gives the computer a clear-to-send signal (CTS) and the computer begins transmitting data. The delay is necessary to assure that the terminal modem is ready to accept the data.

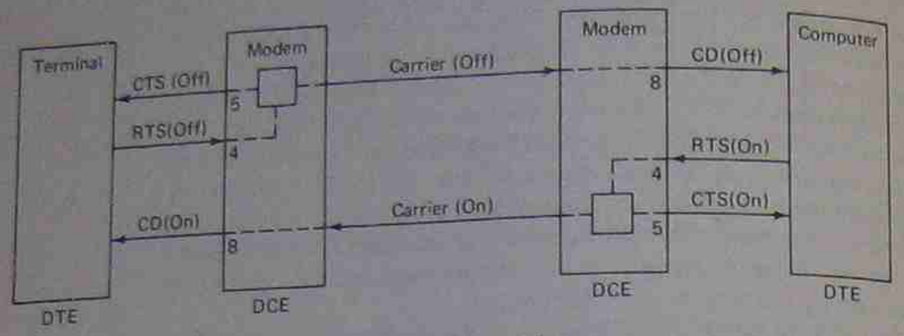
If the computer is ready to accept data, it turns off its RTS lead, and the carrier from the computer's modem dies. The modem at the terminal end detects this and turns its carrier detect (CD) off. If the terminal now desires to send data, it turns on its RTS lead which tells the modem to start sending a carrier. Again, after a timed wait, the terminal's CTS lead is turned on, which informs the terminal that it can begin transmission.

4–6.5 T103A Data Set

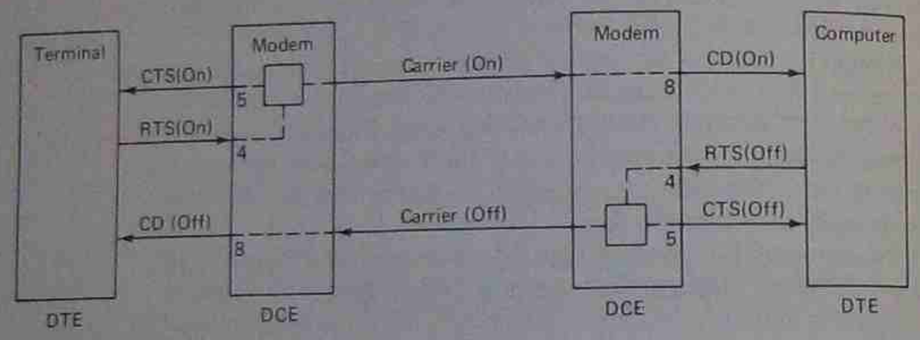
The application of the RS-232-C specifications can be seen in more detail by examining the commonly used T103A data set, or modem, which achieves bit rates of 300 b/s when it is used in full-duplex operations over the direct distance dialing network (DDD). This examination will specifically concentrate on the RIXON T103A Data Set, the simplified block diagram of which is shown in Fig. 4-41.

The system block diagram of Fig. 4-39, which uses the RIXON data set, illustrates the interconnections of the DTE, the data set, and the control unit as well as the tie-in to the DDD network.

Full-duplex operation is achieved by operating the data sets over two frequency bands, f_1 and f_2 (see Fig. 4-18). Within each band, there is a MARK and a SPACE frequency. One data set operates as the originating station, while the other acts as the answering station. Both data sets



(a) Computer in Transmit State



(b) Terminal in Transmit State

Fig. 4-38 Application of RS-232-C Specifications

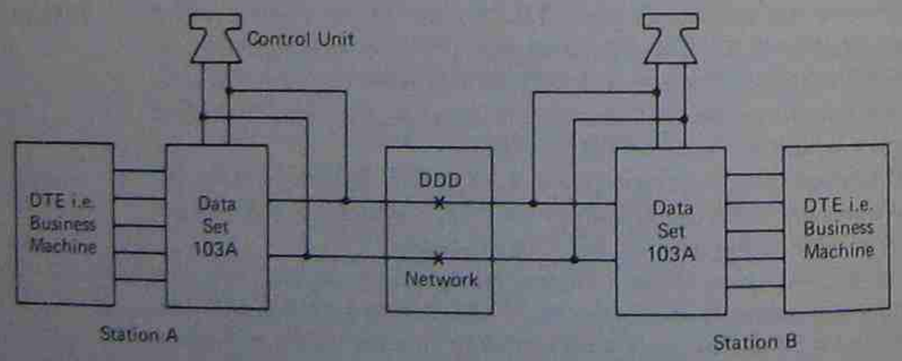


Fig. 4-39 Data Phone System

Table 4-7 Frequently used RS-232-C Signal

	Interchange Circuit	Pin Number	Name	Description	Nmemonic	From DCE	To DCE
for data	AA	1	Protective Ground	A connection to the machine or equipment chassis			
	AB	7	Signal Ground	Common ground and return path for all signals			
	BA	2	Transmitted Data	Data path for signals generated by DTE	TDATA		X
	BB	3	Received Data	Data path for signals generated by the remote DTE and received via the DCE	RDATA	X	
for control	CA	4	Request to Send	Activated by the DTE to inform the DCE to prepare to receive data from the DTE and to enter the transmit mode	RTS		X
	CB	5	Clear to Send	Activated by the DCE to inform the DTE that it is ready to receive and to retransmit data from the DTE	CTS	X	
	CC	6	Data Set Ready	DCE informs the DTE that the DCE is operational	DSR	X	
	CF	8	Received Line Signal Detector Carrier Detect	Activated by the DCE informing the DTE that the remote DCE is sending a carrier	CD	X	
	CD	20	Data Terminal Ready	Activated by the DTE informing the DCE that the DTE is operational	DTR		X
	CE	22	Ring Indicator	Signal from DCE informing the DTE that a ringing signal has been received over line	RI	X	
	DA	24	Transmit Signal Element Timing	Clock pulses provided by DTE			X

receive and transmit, but the station which initiates a call becomes the originating station. The originating station transmits in the f_1 band and receives in the f_2 band. The answering station transmits in the f_2 band and receives in the f_1 band. Table 4-8 shows the frequency allocations.

Depending upon which key is depressed on the control unit, the customer can go into one of the several available modes of operation. Figure 4-40 shows the control unit with its keys, and Table 4-9 gives the options available to each key position. To initiate a call, the customer dials the location he wishes to transmit data to or from via the DDD network. At the receive end, the attendant may answer the call and go into the data mode by pushing the data key on his telephone set. If the telephone set is placed on automatic answer, the data set will receive the call and control the line and business machine.

In general, the data set accepts serial binary data from the DTE and converts the binary data into a FSK signal. The FSK signal, in turn, is transmitted over the voice band telephone system. At the receiving end, the FSK signal is demodulated into a binary data stream and sent to the adjoining DTE. Either station may be the originating or answering station. The calling station becomes the originating station. If no carrier is present, the data set signals a MARK condition to the local DTE.

Before discussing the processing of a call, it is important to examine the significant blocks of the T103A data set that are shown in Fig. 4-41.



Fig. 4-40 567PB-61 Telephone/Control Set (Courtesy of Rixon Inc.)

Table 4-8 T103A Data Set Frequency Allocation

Data Set Mode	Transmit Band	Designation	Frequency
Originating	f_1	f_{1M} (MARK)	1270
		f_{1S} (SPACE)	1070
Answering	f_2	f_{2M} (MARK)	2225
		f_{2S} (SPACE)	2025

The triangular shaped termination points correspond to the pin numbers specified in Table 4-7 and are interconnected to the local DTE. Note that the protective and signal grounds are connected together (Pins 1 and 7). These provide the reference point for all signal voltages. The following discussion will assume that the Data Terminal Ready lead (CD) is ON, which causes the RR relay to operate.

Data Set Block Diagram

1. **Transmitter:** When transmitting data, the multivibrator (MV) operates at one frequency for a MARK input on the data lead and at another frequency for a SPACE input on the data lead. The time constant required to bring the OFF transistor to an ON state in the multivibrator circuit is altered by selecting a different charging resistance. Figure 4-42 gives a simplified version of the modulator. In order to remove undesirable harmonics and reduce sideband levels, bandpass filter Z_1 bandlimits the signal from the modulator.
2. **Receiver:** The Tip and Ring (T-R) leads from the telephone line are connected across the ringing detector (a voltage doubler circuit)

Table 4-9 Key Designations 567PB-61 Telephone Control Set (Courtesy of Rixon Inc.)

Index Number	Key Designation	Function
1	DATA	Nonlocking key used to shift from talk to data mode (also takes data set out of test mode if connected in error).
2	TALK	Locking key used to obtain normal telephone operation.
3	TEST 1	Nonlocking key used for remote testing of answer mode.
4	TEST 2	Nonlocking key used for remote testing of originate mode.
5	LOCAL	Locking key used to allow the telephone or business machine to operate to half duplex to make tapes or check their accuracy.
6	AUTO	Locking key used to provide automatic answer of incoming calls.

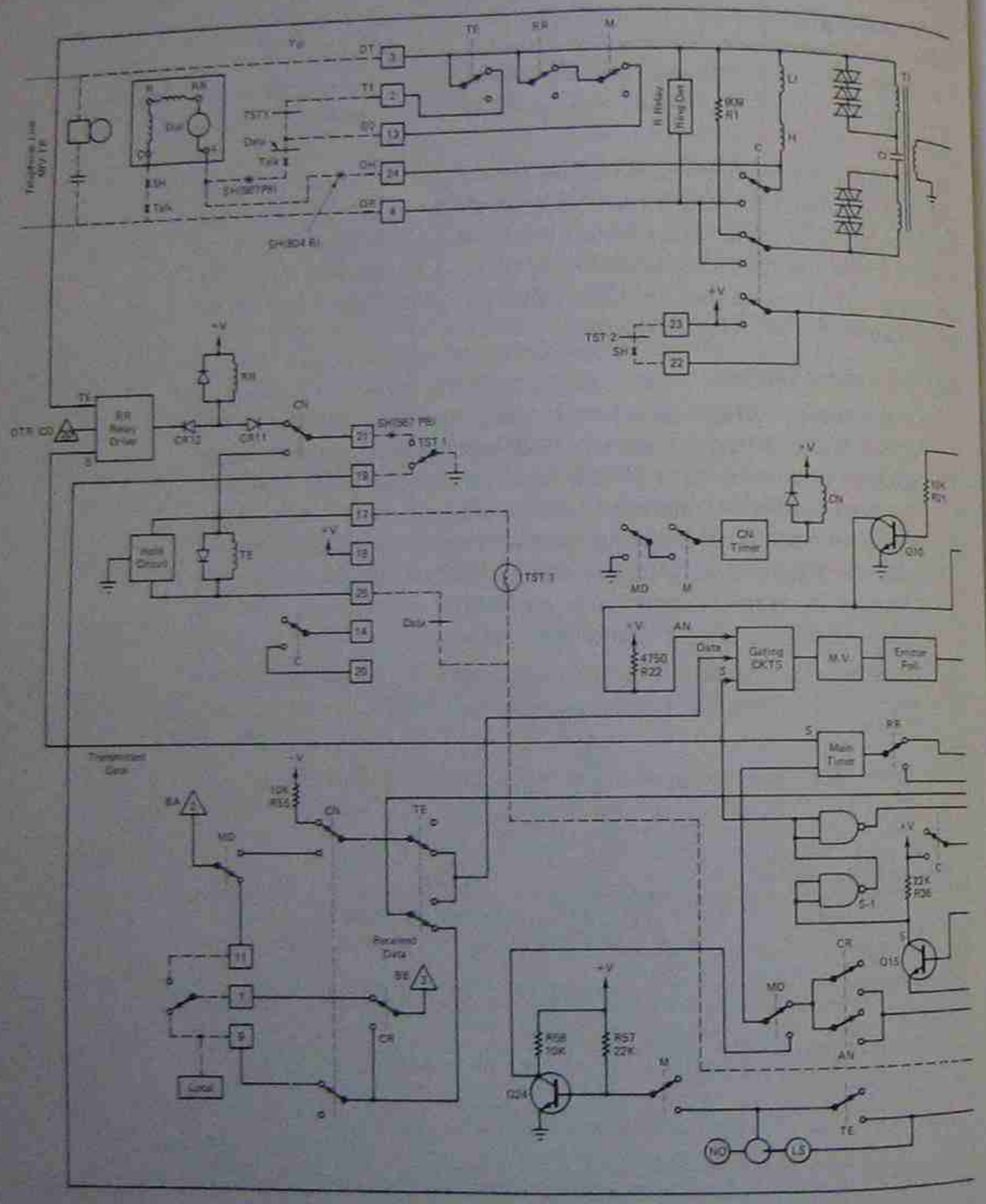


Fig. 4-41 T103A Data Set Simplified Block Diagram (Courtesy of Rixon Inc.)

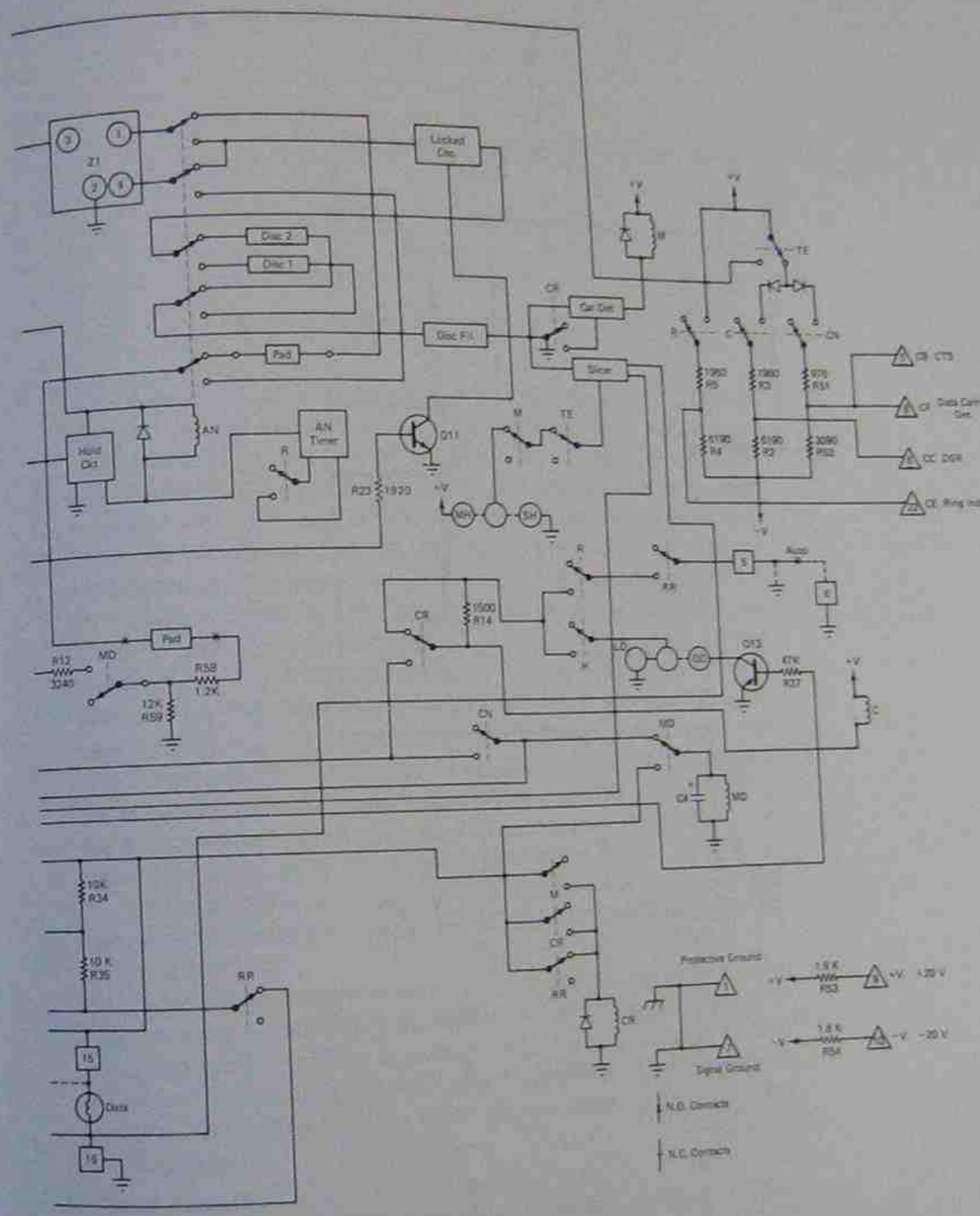


Fig. 4-41 Concluded

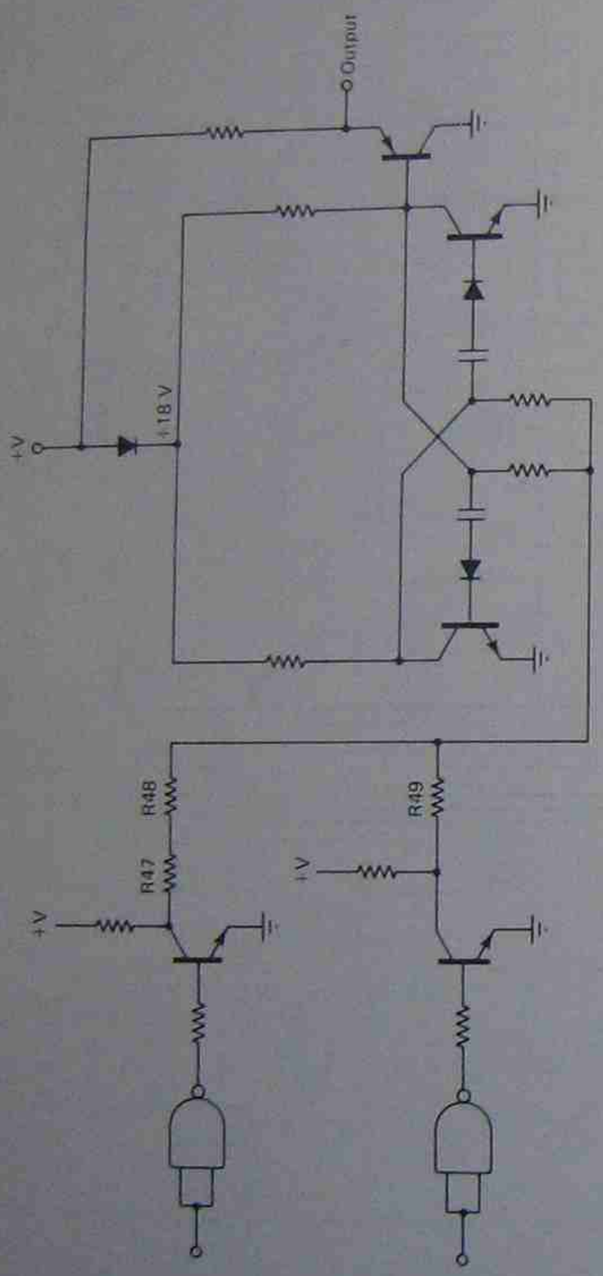


Fig. 4-42 Data Set Modulator

which, when activated, causes the R relay to operate. Figure 4-43 illustrates the ringing relay circuit.

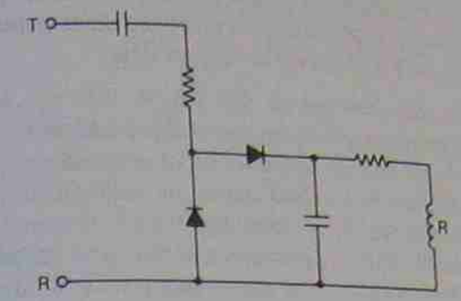


Fig. 4-43 Ringing Detector Circuit

With the R contacts switched, ring indicator Pin 22 (CE) goes high and the line transformer, T_1, is connected to the T-R leads provided that the CD lead (Pin 20) is ON. (Relay C is operated via contacts R and RR.) The secondary of the line transformer is connected to a filter which, in turn, directs the signal to the locked oscillator. When no signal is present, the oscillator free-wheels at a frequency centered midway between the MARK and SPACE frequency: 1170 Hz for f_1 and 2125 for f_2. The discriminator is tuned negatively for 1270 Hz and positively for 1070 Hz when in the f_1 range. It is tuned negatively for 2225 Hz and positively for 2025 Hz when in the f_2 range. Thus, the discriminator output will be at 0 when the oscillator is free-wheeling, i.e., running midway between the MARK and SPACE frequencies. When a signal of sufficient magnitude appears, the oscillator locks to the frequency of the applied signal, causing the discriminator output to go positive or negative. The f_2 band discriminator is shown in Fig. 4-44.

If the discriminator output is significantly large and stable, the carrier detector circuit activates the M relay. High-level noise tends to be rejected.

The slicer circuit is a zero crossing detector that reshapes the discriminator output into a digital data wave. Depending upon the strapping used, the slicer is either clamped in the marking (MARK HOLD) or the spacing (SPACE HOLD) condition in the absence of carriers. The output from the slicer appears at the received data BB leads (Pin 3).

3. **Timer Circuits:**
   Main timer: The main timer generates a number of delays while a call is being set up. These delays include a guard interval of about 150 ms, a 1.5 s delay for disabling the echo suppressors, and disconnect time intervals.

AN timer: AN is a memory device which locks up after ringing has occurred.
CN timer: The CN timer delays informing the local DTE that it may transmit data in order to ensure that the distant station has detected the carrier and is prepared to receive data.

Now that the workings of the T103A data set have been described, call processing will be more easily understood.

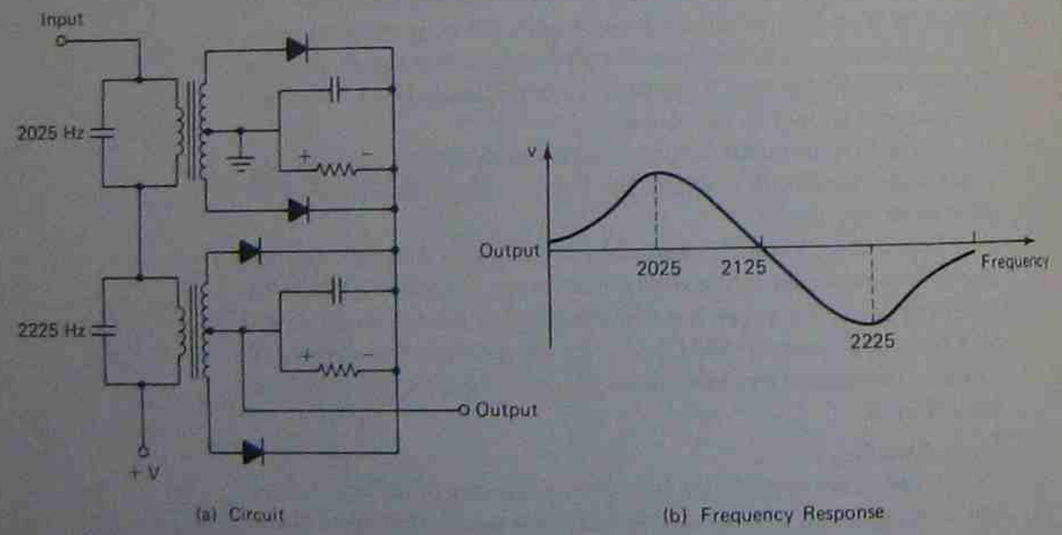
The discussion on processing a call, or of establishing a connection from a calling station to a called station, is outlined in tabular form in Table 4-10. It is assumed in the table that the CD terminal is ON at both ends; that is, that RR relay is activated, and that time advances with each step. Notes written alongside the steps explain what is happening within the modem (Fig. 4-41) when the AUTOMATIC ANSWER is in operation.

While following the channel establishment sequence, it may be helpful to observe the graphical counterpart illustrated in Fig. 4-45. Figure 4-46 graphically illustrates the disconnect sequence when the called station is in AUTO mode.

In the local mode, the BA (transmitted data) and BB (received data) leads are interconnected via MD, LOCAL and CR.

4–6.6 Line Drivers

To extend the EIA interface beyond 50 ft, a line driver pair can be inserted, as shown in Fig. 4-47. The line driver system translates the RS-232-C signals into low-impedance outputs suitable for overcoming



(a) Circuit (b) Frequency Response

Fig. 4-44 Discriminator Circuit and Action

Table 4-10 Processing of a Call via a T103A Data Set

CD Lead is ON at Both Ends

Calling Station (Originate)	Called Station (Answer)	(Notes)
OFF Hook Condition Presented to Central Office. Lifting the Handset causes the Switchook (SH) to Operate. Tipside of the Telephone Line (DT) Connected to Terminal F of Telephone Set Via TE, TST1, DATA, TALK and SH Contact Points. Ring Side (DR) Connected to Terminal C of Telephone Set Via TALK and SH Contacts.		
Attendant Operates TALK Key and Lifts Handset		
Dial Tone is Heard		
Normal Dialing on DDD Network		
Audible Ringing	Attendant / Automatic Answer	
A Short Exists Across H Winding via TE TST1, DATA, TALK, SH	Ringing Voltage Applied to Data Set	"R" Relay Operates — Ring Indicates (CE Lead) Goes ON
If Busy Tone is Received Call Does Not Proceed	Attendant Hears Ringing and Operated Talk Key	"C" Relay Operates — Grounding of C Coil thru CR, R, RR and Auto
	A	B

Table 4-10 (Continued)

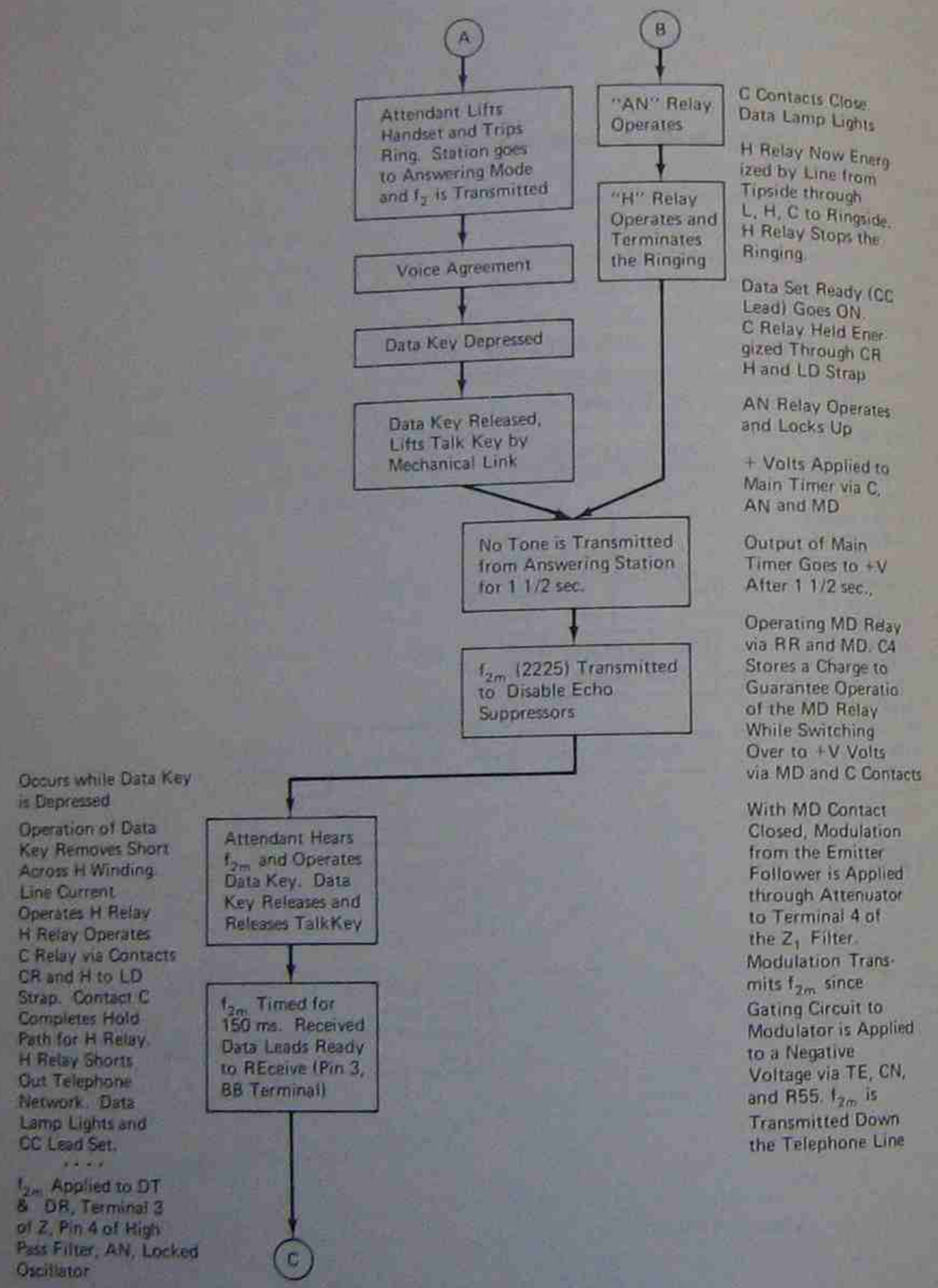


Table 4-10 (Continued)

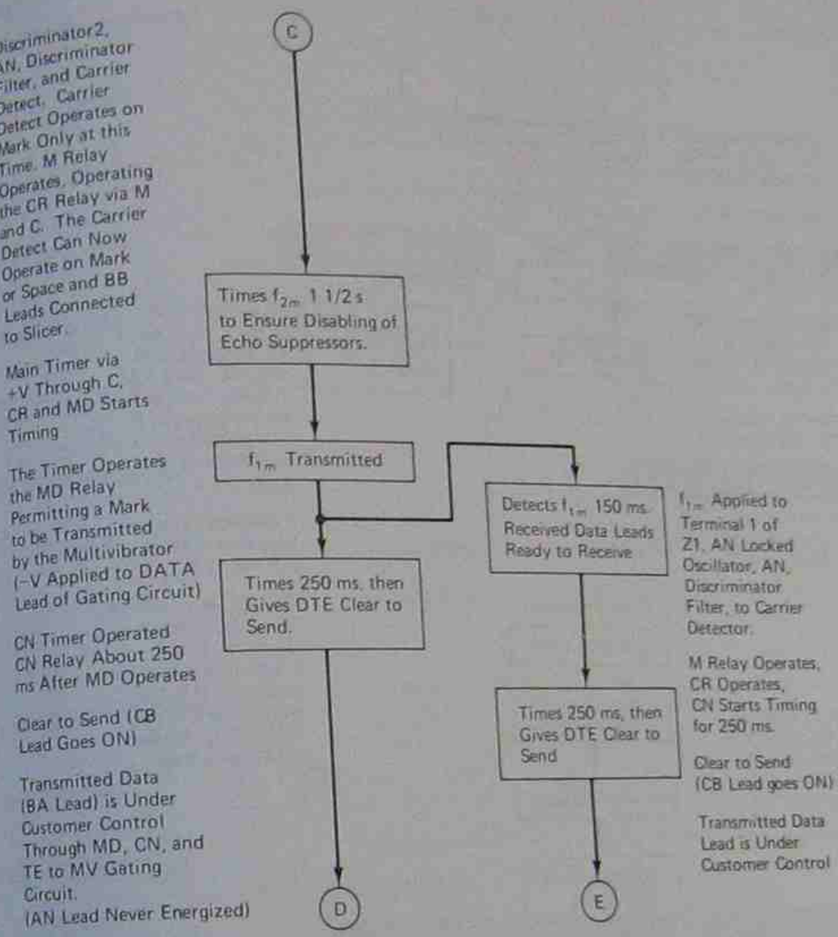
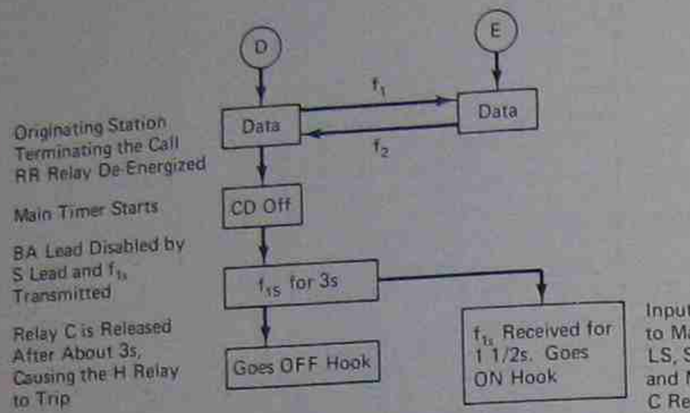


Table 4-10 (Concluded)



Originating Station Terminating the Call RR Relay De-Energized

Main Timer Starts

BA Lead Disabled by S Lead and f_{1s} Transmitted

Relay C is Released After About 3s, Causing the H Relay to Trip

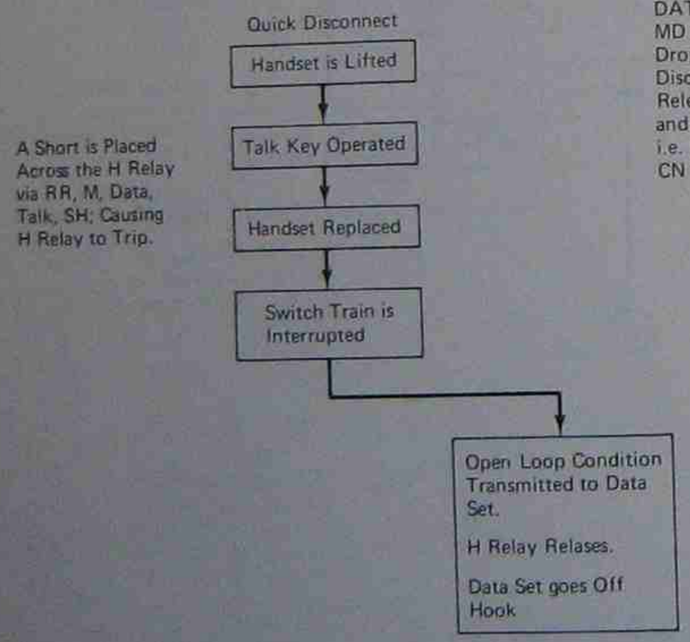
Data Set Ready on CC Lead Turns OFF

C Relases CR and MD, Causing CN to Release CF and CB Leads (Data Carrier Detect and Clear to Send) Go OFF

Input from Slicer to Main Timer via LS, Strap, M, Q24 and MD. After 1.5s, C Relay Releases. CC Lead goes OFF. (+V Volts Applied to Both Ends of Relay Coil.)

C Contacts Open, Dropping H Relay and AN Relay

DATA Lamp Turns OFF, MD and CR Relays Dropped. Line is Disconnected by Release of H Relay and Carrier is Lost i.e. M Relay Releases. CN Timer is Reset.



A Short is Placed Across the H Relay via RR, M, Data, Talk, SH; Causing H Relay to Trip.

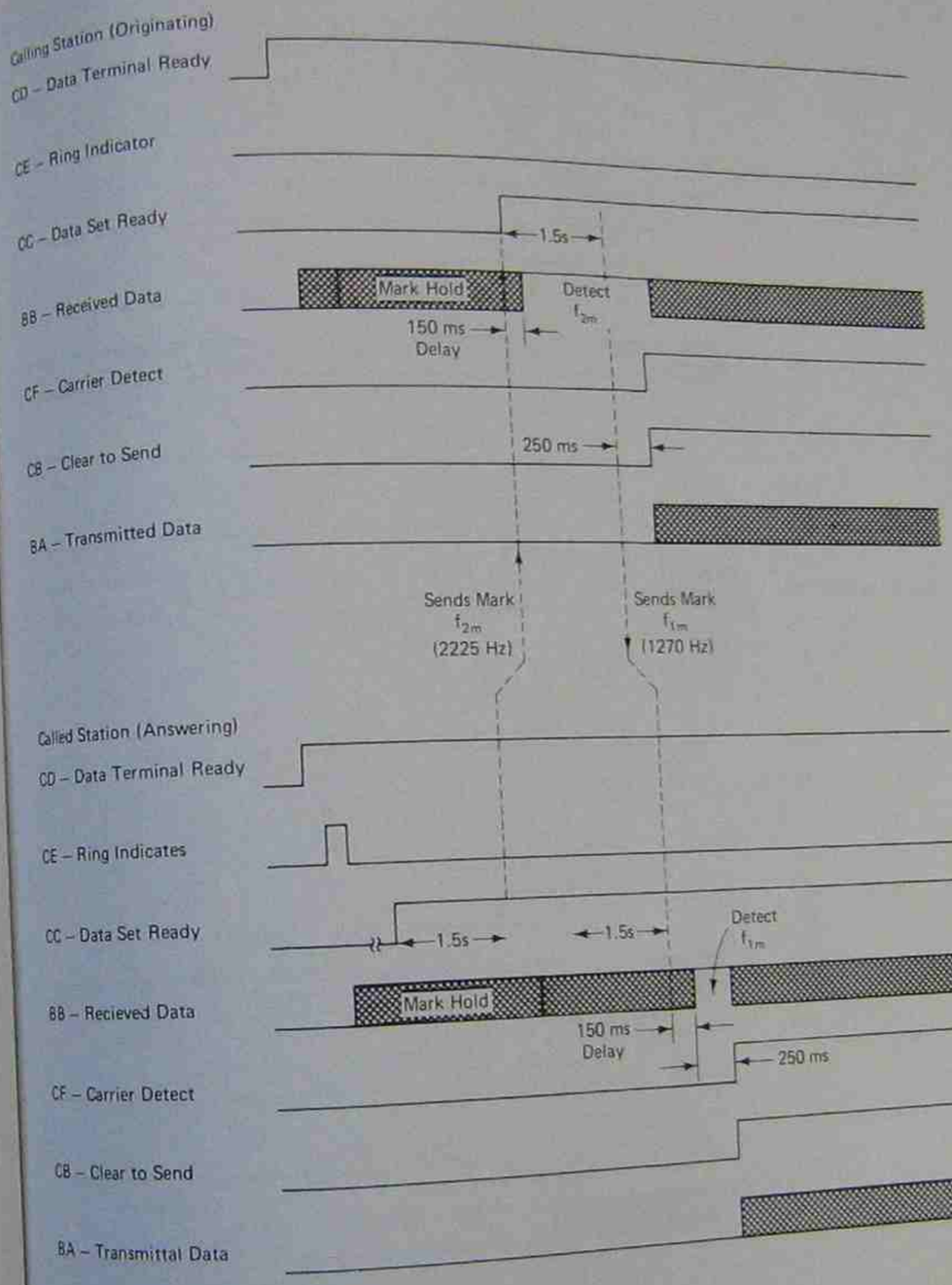


Fig. 4-45 Channel Establishment Sequence of a T103A Data Set

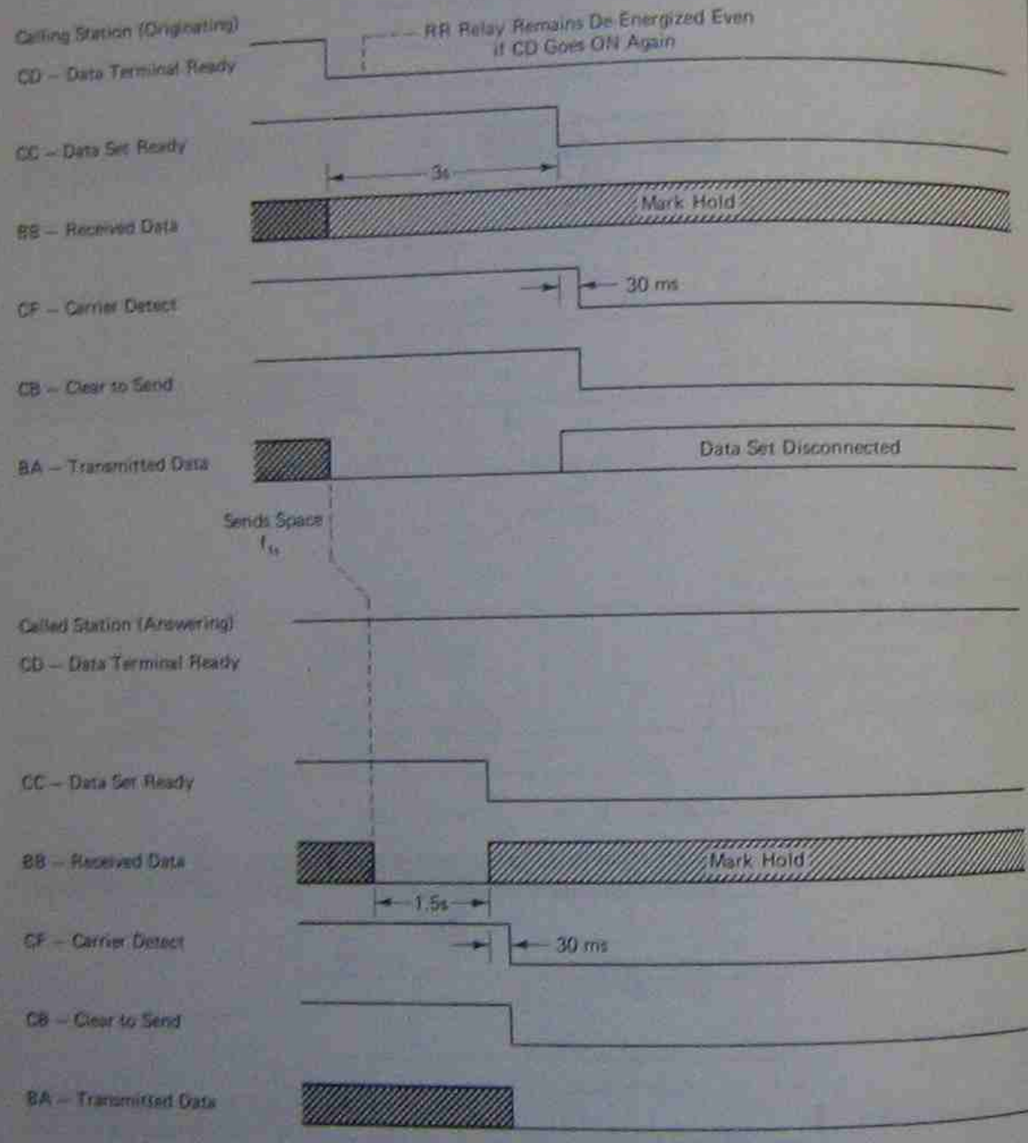


Fig. 4-46 Channel Disconnect Sequence of a T103A Data Set

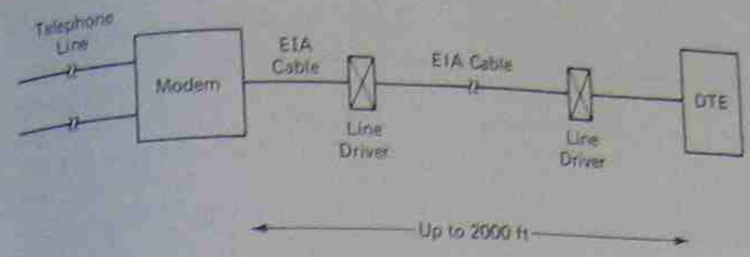


Fig. 4-47 Extension of EIA Interface

the impedance of a long cable. By using proper twisted-pair cable, modem-unit *interfaces* can be extended by up to 2000 ft at 9600 b/s.

4–6.7 Modem Chips

Many of the modem functions are now available in chip form, including the low-speed Motorola MC6860 and MC14412 modems. The MC6860 device shown in Fig. 4-48 includes a modulator, a demodulator, and supervisory control. The modulator produces an eight-step synthesized sinewave which is cleaned up by a bandpass transmit filter. If the modem is in the originate mode, the transmit filter must pass frequencies of 1070–1270 Hz. If it is in the answer mode, it must pass frequencies of 2025–2225 Hz.

The filters should have steep roll-offs when going into the stopband region and should provide linear phase or group delay in the passband. An ideal filter is shown in Fig. 4-49.

If the attenuation and group delay characteristics are not flat in the passband, the various frequencies which make up the signal experience different attenuations and delays, resulting in distortion. Refer back to Sec. 3–2.5 and 3–2.6 for a more complete explanation of this type of distortion.

The duplexer in the modem schematic is a circuit which causes the transmitted signal to follow a path different from that followed by the received signal, even though both may be simultaneously present on a portion of the path. In the modem under consideration, the signal from the transmit filter reaches the data coupler but not the receive filter. A strong transmit signal is not desirable in the local receiver since the receiver can become overloaded, thus causing large intermodulation products. The received signal from the data coupler can reach both the transmit and receive filters.

Telephone companies have historically required that customer-owned data sets be connected to the telephone network through a data coupler furnished by the telephone company. These data couplers prevent hazardous voltages from occurring on the network by providing data

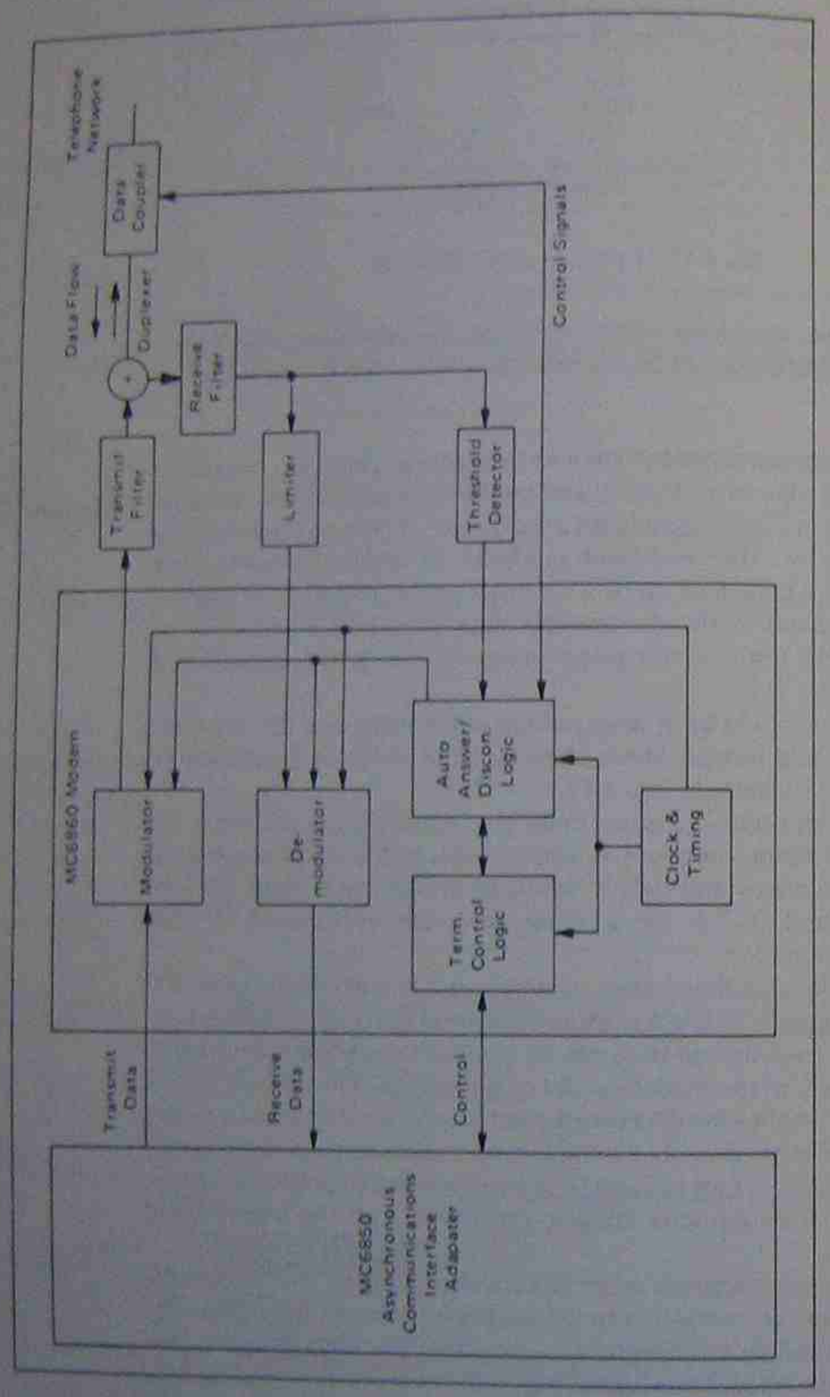


Fig. 4-48 Low-Speed FSK Modem (Courtesy of Motorola Inc.)

access arrangements (DAA) and sufficient electrical isolation between the customer's device and the telephone network. In addition, the couplers provide suitable signal and control voltage levels which will protect the equipment and will also prevent disturbances from occurring on neighboring telephone circuits. Although requirements for renting a data coupler are changing, the customer must still meet strict coupling regulations. The data coupler in the modem illustrated in Fig. 4-48 meets the requirements of the Bell system CBS and CBT units.

4–6.8 FSK Frequency Response

The FSK signal shown in Fig. 4-50(a) illustrates the bandwidth required for the transmission of FSK signals. By using the switching function outlined in Appendix A, the FSK waveform can be constructed by multiplying an f_2 sinusoid by $S(t)$ and adding the result to an f_1 sinusoid multiplied by $1 - S(t)$. This is illustrated in Fig. 4-50(b) and (c).

Before obtaining the frequency content of the FSK signal, the frequency content of the waveforms, shown in Fig. 4-50(b) and (c), must be found. Since the switching function is a square wave with a period of 2τ, it contains all the odd harmonics of $1/2\tau$ Hz which roll-off at the sin f/f rate (see Appendix A). Since multiplying two sinusoids results in the sum and difference frequencies (see Eq. (4-3), when the sinusoid $\cos 2\pi f_2 t$ is multiplied by the switching function containing odd harmonic sinusoids, the following sum and difference frequencies are obtained:

$$f_2 + \frac{1}{2\tau},\ f_2 - \frac{1}{2\tau},\ f_2 + \frac{3}{2\tau},\ f_2 - \frac{3}{2\tau} \cdots$$

Similarly, when $\cos 2\pi f_1 t$ is multiplied by $1 - S(t)$, the following sum and difference frequencies are obtained:

$$f_1 + \frac{1}{2\tau},\ f_1 - \frac{1}{2\tau},\ f_1 + \frac{3}{2\tau},\ f_1 - \frac{3}{2\tau} \cdots$$

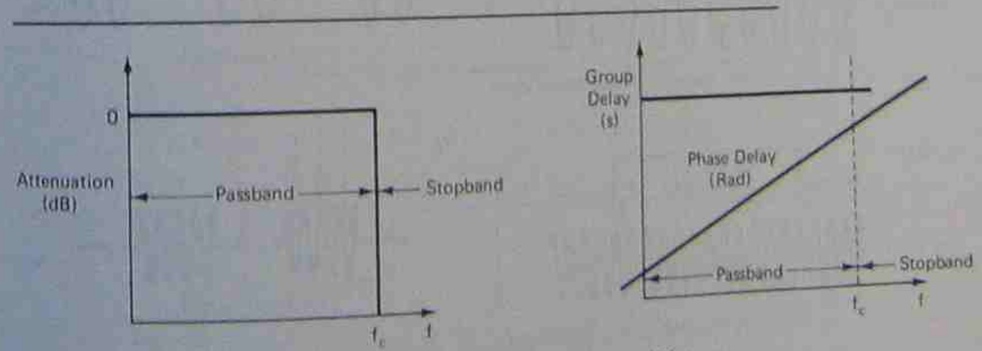


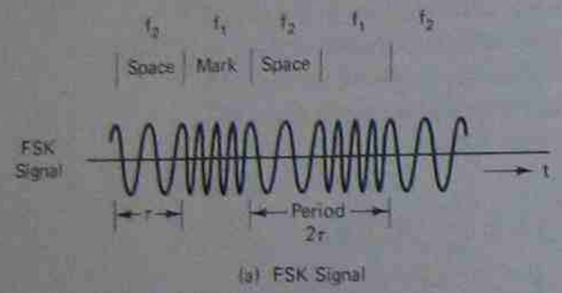
Fig. 4-49 Ideal Filter Characteristics

The negative sign on $S(t)$ indicates that the resultant signal undergoes an extra 180° phase shift.

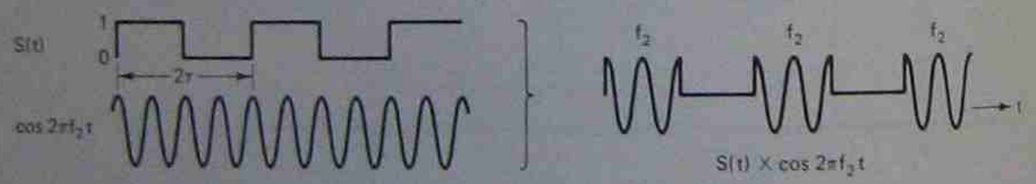
Figure 4-51 illustrates the resulting frequency spectrum of the FSK signal. The bandwidth required to pass an FSK signal usually includes the components that reside within the first sin f/f zero. For the signal under consideration, this would be $f_1 - f_2 + 2/\tau$ Hz.

4–7 UNIVERSAL SYNCHRONOUS/ ASYNCHRONOUS RECEIVER/ TRANSMITTER (USART)

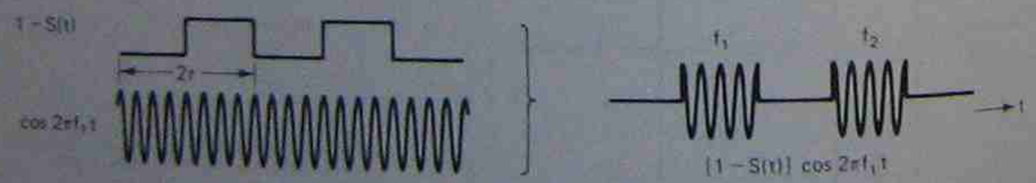
Serial data transmission is generally used to transmit data over relatively long distances. The device which converts the parallel bits from a central processing unit (CPU) or some data terminal into a continuous serial data



(a) FSK Signal



(b) Generation of f_2 Space Signal



(c) Generation of f_1 Mark Signal

Fig. 4-50 Generation of FSK Signal

stream, or vice-versa, is known as a Universal Synchronous/ Asynchronous Receiver/Transmitter (USART). Either synchronous or asynchronous operation is selected. If the device permits only asynchronous operation, it is known as a UART. This section will briefly consider two such devices: a COM2502/H UART and the Intel 8251A USART.

COM 2502/H UART

This MDS/LSI monolithic circuit is a 40 pin device that transmits and receives asynchronous data in either the half-duplex or full-duplex mode. The transmitter accepts parallel data and converts it into a serial asynchronous output. The receiver and transmitter can operate simultaneously and at different baud rates. A transmit/receive clock must be provided at 16 times the baud rate.

The UART is programmable in the sense that the data word length, parity mode, and number of stop bits can be selected by the control lines. There may be 5, 6, 7, or 8 data bits; even/odd or no parity; and 1, 1.5, or 2 stop bits. Figure 4-52 illustrates the format of a typical serial character.

Both the receiver and transmitter have double character buffering so that one character time interval is always available to exchange a character with the external device. Figure 4-53 gives the pin configuration of the UART, and Table 4-11 describes each pin function.

Description of Transmitter Operation

A simplified block diagram of the UART Transmitter is shown in Fig. 4-54. When the power is turned ON and the master reset (MR) is pulsed, the Transmitter Buffer Empty (TBMT) line will go high indicating that the transmitter holding register is empty and that the control bit may be set. A high-level signal on the control strobe enters the control bits NDB1, NDB2, NSB, POE and NPB.

The data is loaded into the transmitter holding register by applying a low-level input to the data strobe. If the transmitter shift register is empty,

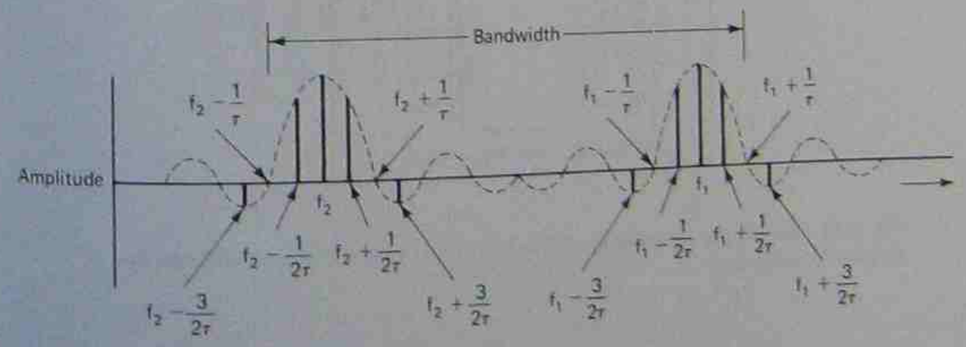


Fig. 4-51 Frequency Spectrum of FSK Signal

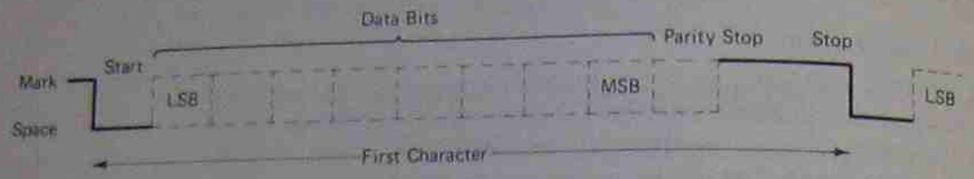


Fig. 4-52 Format of Typical Serial Asynchronous Character

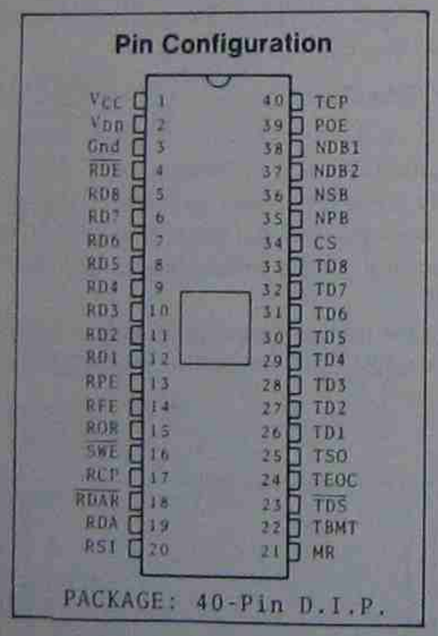


Fig. 4-53 Pin Configuration of the COM 2502/H UART. (Courtesy SMC Microsystems Corporation.)

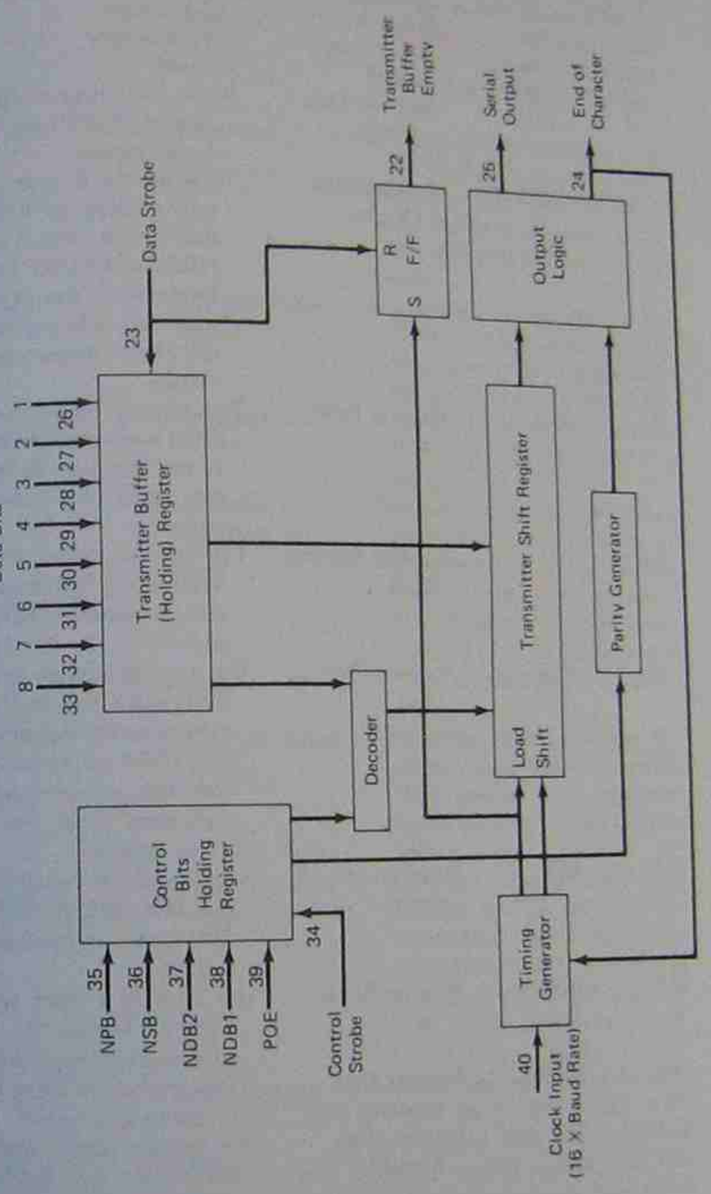


Fig. 4-54 Simplified UART Transmitter

Table 4-11 Description of Pin Functions (Courtesy SMC Microsystems Corporation)

Pin No.	Symbol	Name	Function
1	V_{CC}	Power Supply	+5 V Supply
2	V_{DD}	Power Supply	−12 V Supply
3	GND	Ground	Ground
4	$\overline{RDE}$	Received Data Enable	A low-level input enables the outputs (RD8-RD1) of the receiver buffer register.
5–12	RD8-RD1	Receiver Data Outputs	These are the 8 tri-state data outputs enabled by $\overline{RDE}$. Unused data output lines, as selected by NDB1 and NDB2, have a low-level output, and received characters are right justified, i.e. the LSB always appears on the RD1 output.
13	RPE	Receiver Parity Error	This tri-state output (enabled by $\overline{SWE}$) is at a high-level if the received character parity bit does not agree with the selected parity.
14	RFE	Receiver Framing Error	This tri-state output (enabled by $\overline{SWE}$) is at a high-level if the received character has no valid stop bit.
15	ROR	Receiver Over Run	This tri-state output (enabled by $\overline{SWE}$) is at a high-level if the previously received character is not read (RDA output not reset) before the present character is transferred into the receiver buffer register.
16	$\overline{SWE}$	Status Word Enable	A low-level input enables the outputs (RPE, RFE, ROR, RDA, and TBMT) of the status word buffer register.
17	RCP	Receiver Clock	This input is a clock whose frequency is 16 times (16×) the desired receiver baud rate.
18	$\overline{RDAR}$	Receiver Data Available Reset	A low-level input resets the RDA output to a low-level.
19	RDA	Receiver Data Available	This tri-state output (enabled by $\overline{SWE}$) is at a high-level when an entire character has been received and transferred into the receiver buffer register.

Table 4-11 (Continued)

Pin No.	Symbol	Name	Function
20	RSI	Receiver Serial Input	This input accepts the serial bit input stream. A high-level (mark) to low-level (space) transition is required to initiate data reception.
21	MR	Master Reset	This input should be pulsed to a high-level after power turn-on. This sets TSO, TEOC, and TBMT to a high-level and resets RDA, RPE, RFE and ROR to a low-level.
22	TBMT	Transmitter Buffer Empty	This tri-state output (enabled by $\overline{SWE}$) is at a high-level when the transmitter buffer register may be loaded with new data.
23	$\overline{TDS}$	Transmitter Data Strobe	A low-level input strobe enters the data bits into the transmitter buffer register.
24	TEOC	Transmitter End of Character	This output appears as a high-level each time a full character is transmitted. It remains at this level until the start of transmission of the next character or for one-half of a TCP period in the case of continuous transmission.
25	TSO	Transmitter Serial Output	This output serially provides the entire transmitted character. TSO remains at a high-level when no data is being transmitted.
26–33	TD1-TD8	Transmitter Data Inputs	There are 8 data input lines (strobed by $\overline{TDS}$) available. Unused data input lines, as selected by NDB1 and NDB2, may be in either logic state. The LSB should always be placed on TD1.
34	CS	Control Strobe	A high-level input enters the control bits (NDB1, NDB2, NSB, POE and NPB) into the control bits holding register. This line may be strobed or hard wired to a high-level.

Table 4-11 (Concluded)

Pin No.	Symbol	Name	Function
35	NPB	No Parity Bit	A high-level input eliminates the parity bit from being transmitted; the stop bit(s) immediately follow the last data bit. In addition, the receiver requires the stop bit(s) to follow immediately after the last data bit. Also, the RPE output is forced to a low-level. See pin 39, POE.
36	NSB	Number of Stop Bits	This input selects the number of stop bits. A low-level input selects 1 stop bit; a high-level input selects 2 stop bits. Selection of two stop bits when programming a 5 data bit word generates 1.5 stop bits from the COM2017/H.
37–38	NDB2, NDB1	Number of Data Bits/Character	These 2 inputs are internally decoded to select either 5, 6, 7, or 8 data bits/character as per the following truth table: NDB2 NDB1 data bits/character L L 5 L H 6 H L 7 H H 8
39	POE	Odd/Even Parity Select	The logic level on this input, in conjunction with the NPB input, determines the parity mode for both the receiver and transmitter, as per the following truth table: NPB POE MODE L L odd parity L H even parity H X no parity X = don't care
40	TCP	Transmitter Clock	This input is a clock whose frequency is 16 times (16×) the desired transmitter baud rate.

the data is automatically transferred to the shift register. TEOC goes low and TBMT goes high, indicating that the transmitter buffer register has been loaded into the shift register and that the buffer register is available to receive new data. The stop and parity bit are then added to the data, and serial transmission is started. The start bit appears first, followed by the data bits, the parity bit (if selected), and the stop bit(s). TEOC goes high after the last stop bit has been on for one time bit interval. If the transmitter buffer register is loaded (TBMT, low), a new character is transfered to the shift register to be transmitted. If TBMT is high, transmission is stopped until a new character is loaded into the transmitter holding register. New control bits may also be entered prior to the next data transmission.

Description of Receiver Operation

A simplified block diagram of the UART receiver is shown in Fig. 4-55. The receiver is under the same type of control as the transmitter.

A parity check circuit checks for even or odd parity if parity was selected at the transmitter stage. If parity does not check, the parity error signal (RPE) goes high. The logic circuit also checks the first stop bit of each character. If the stop bit is not a logic one as it should be (see Fig. 1-18) the framing error signal (RFE) goes high.

With the power turned ON and the master reset (MR) pulsed, the receiver data available signal (RDA) goes low, indicating that the receiver buffer is empty. The control bits are then entered with the control strobe.

Serial asynchronous data is sent to the serial input line (RSI). The start bit detects circuit searches for a high to low transition (MARK to SPACE) on the input line. If this transition is found, a counter is reset and allowed to count until the center of the start bit is reached. If the input is still low (SPACE), the signal is assumed to be a valid start bit and the sampling continues at the center of all consequent data and stop bits. The character is thus assembled in the receiver shift register. The parity error and framing error signals are put into the status word buffer register. After the stop bit is detected, the received character is transferred to the holding register and the data received line (RDA) is set high, which indicates to the external device that the output data in the buffer register may now be sampled. After the external device samples the output, it should strobe the Receiver Data Available Reset ($\overline{\text{RDAR}}$) with a low, thus resetting the RDA line to a low. If the RDA line is not reset before the next character is moved into the receiver holding register, the overrun line will be set high, indicating that a character is lost.

If the UART is to be interfaced with a modem, the transmitter serial output (TSO) and receiver serial input (RSI) must be converted to RS-232 levels.

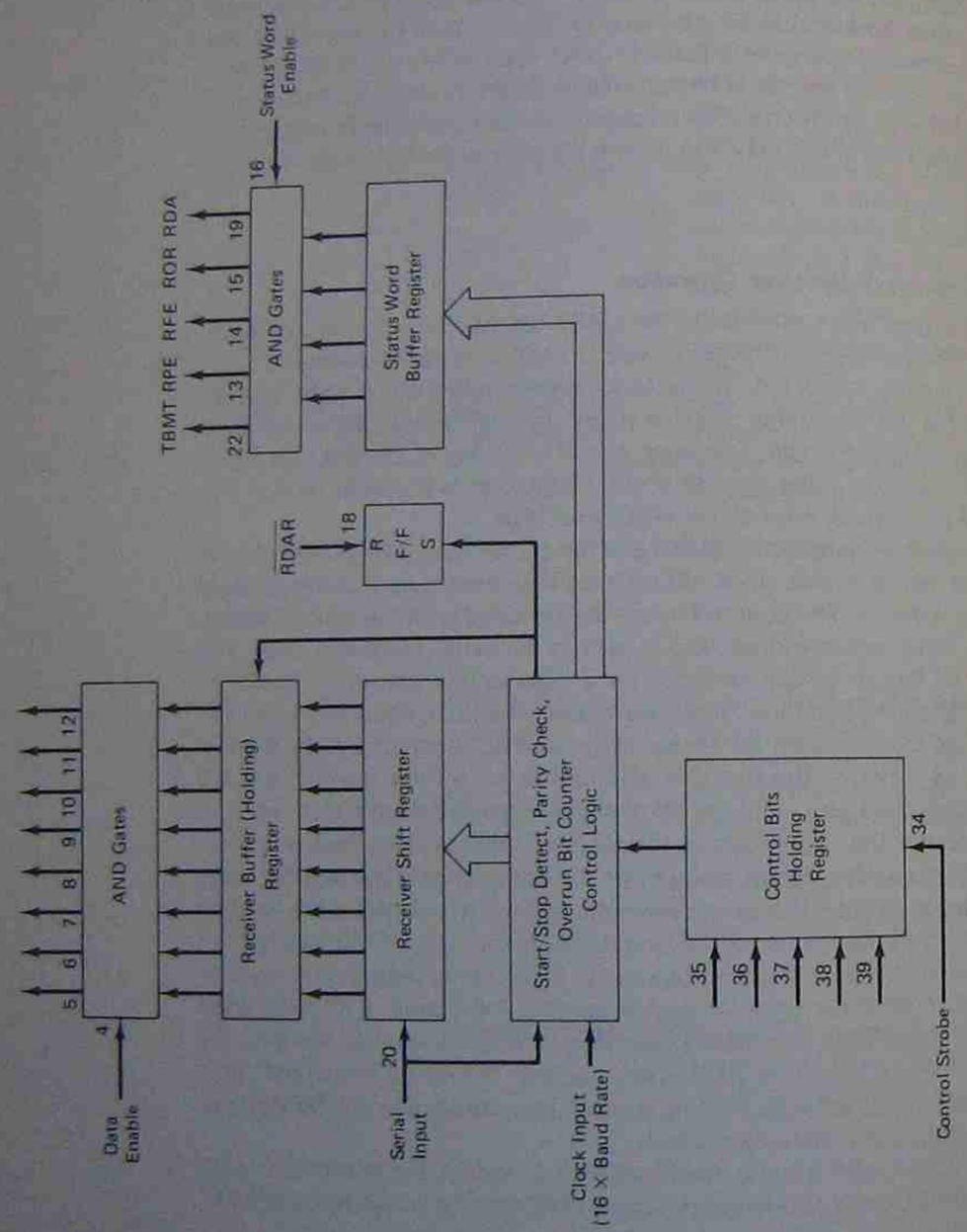


Fig. 4-55 Simplified UART Receiver

INTEL 8251 USART

The Intel 8251 is specifically designed to operate with the 8080 microcomputer system. It is used as a peripheral device and is programmed by the microprocessor (μPU) to operate using virtually any serial data transmission technique presently in use. The USART accepts data characters from the μPU in parallel format and then converts them into a continuous serial data stream for transmission. It can simultaneously receive serial data streams and convert them into parallel data characters for the μPU. The USART will signal the μPU whenever it can accept a new character for transmission or whenever it has received a character for the μPU.

The 8251 contains seven major blocks as illustrated in Fig. 4-56. Two of these blocks serve as the interface to the microprocessor. Two blocks provide serial, transmitter data, and control signals, and two more blocks provide serial receiver data and control. The final block provides signals for modem control.

BLOCK DIAGRAM

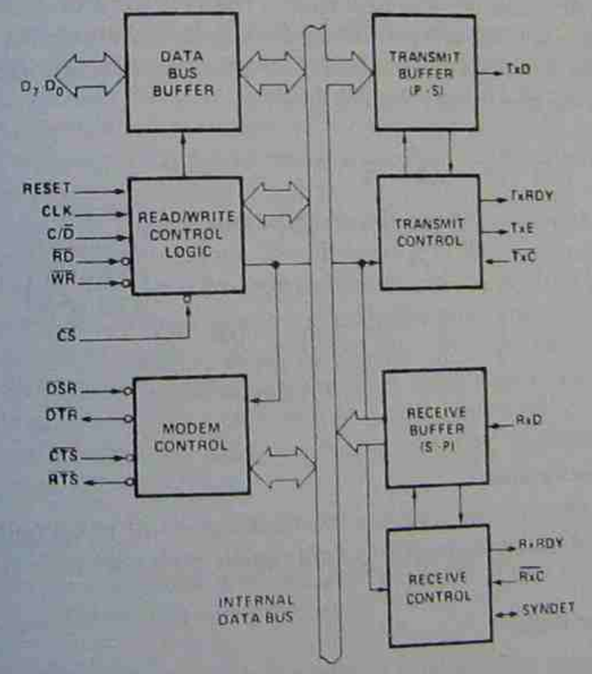


Fig. 4-56 Block Diagram of the Intel 8251 USART (Courtesy of Intel Corp.)

Data Bus Buffer

The Data Bus Buffer is a three-state, bi-directional, eight-bit buffer that is used to interface the 8251 to the 8080 System Data Bus (D_7-D_0). Data is transmitted or received by the buffer upon execution of input or output instructions of the 8080 μPU. Control words, command words, and status information are also transferred through the data bus buffer.

Read/Write Control

The Read/Write control logic accepts input from the 8080 control/address bus (see Fig. 4-9) and generates control signals for overall device operation. It contains the control word register and command word register that store the various control formats for device functional definition. The reset input forces the 8251 into an "idle" mode (TxE output HI). The 8251 will remain in the idle mode until a new set of control words is written to program its functional definition. The CLK input is used to generate internal device timing.

The $\overline{WR}$ (write) and $\overline{RD}$ (read) inputs are used to control the transfer or data between the μPU and the 8251 data buffer. A LO on the $\overline{WR}$ input allows the μPU to output data or control words to the 8251; that is, it writes out to the 8251. A LO on $\overline{RD}$ input allows the μPU to read from the 8251 data or status information. The C/$\overline{D}$ (control/data) input in conjunction with the $\overline{WR}$ and $\overline{RD}$ inputs determine whether the information on the data bus is data character, control word, or status information. The $\overline{CS}$ (chip select) input enables the 8251.

C/$\overline{D}$	$\overline{RD}$	$\overline{WR}$	$\overline{CS}$	
0	0	1	0	8251 → Data Bus (Receive Character)
0	1	0	0	Data Bus → 8251 (Transmit Character)
1	0	1	0	Status → Data Bus
1	1	0	0	Data Bus → Control
X	X	X	1	Data Bus → 3-State

Transmitter Control

The transmitter control manages the transmission of serial data by accepting and issuing transmitter control signals both internally and externally.

Transmitter Buffer

The transmitter buffer accepts parallel data from the data bus buffer, converts it to a serial bit stream, inserts the appropriate characters or bits (depending on the serial format), and outputs a composite serial stream of data on the TxD output pin. The TxRDY (Transmitter ready) output signals

the μPU that the transmitter is ready to accept a data character. It can be used to interrupt the system or for polled operation. The checking status μPU can check TxRDY using a staus read operation. TxRDY is automatically reset when a character is loaded from the μPU.

The TxE (transmitter empty) output goes high when the 8251 has no character to transmit. In the synchronous mode, a high on this output indicates that sync characters are being transmitted automatically as fillers.

The $\overline{TxC}$ (transmitter clock) controls the rate at which the character is to be transmitted. In the synchronous transmission mode the frequency of $\overline{TxC}$ is the actual baud rate (1x). In the asynchronous transmission mode, the frequency of $\overline{TxC}$ is a multiple of the actual baud rate. A portion of the mode instruction selects the value of the multiplier, which can be 1×, 16×, or 64× the baud rate.

Receiver Buffer

The receiver accepts serial data, converts it to parallel format, checks for bits or characters that are unique to the communication technique, and sends an assembled character to the μPU.

Receiver Control

The receiver control manages the reception of serial data characters by accepting and issuing receiver control signals both internally and externally. The RxRDY (receiver ready) output indicates the receiver buffer contains a character that is ready to be put into the μPU. RxRDY can be used as an interrupt to the system or for polled operation. The μPU can check RxRDY using a status read operation. RxRDY is automatically reset when the character is read by the μPU.

The receiver clock controls the rate at which the character is to be received in synchronous mode. The frequency, or $\overline{RxC}$, is equal to the actual baud rate (1×). In asynchronous mode, the frequency of $\overline{RxC}$ is a multiple of the actual baud rate. A portion of the mode instruction selects the value of the multipliers, which can be 1×, 16×, 64× the baud rate.

The SYNDET (sync detect) signal is used in synchronous mode only. When used as an output, or internal sync mode, the SYNDET pin will go high to indicate that the 8251 has located the sync character in the receive mode.

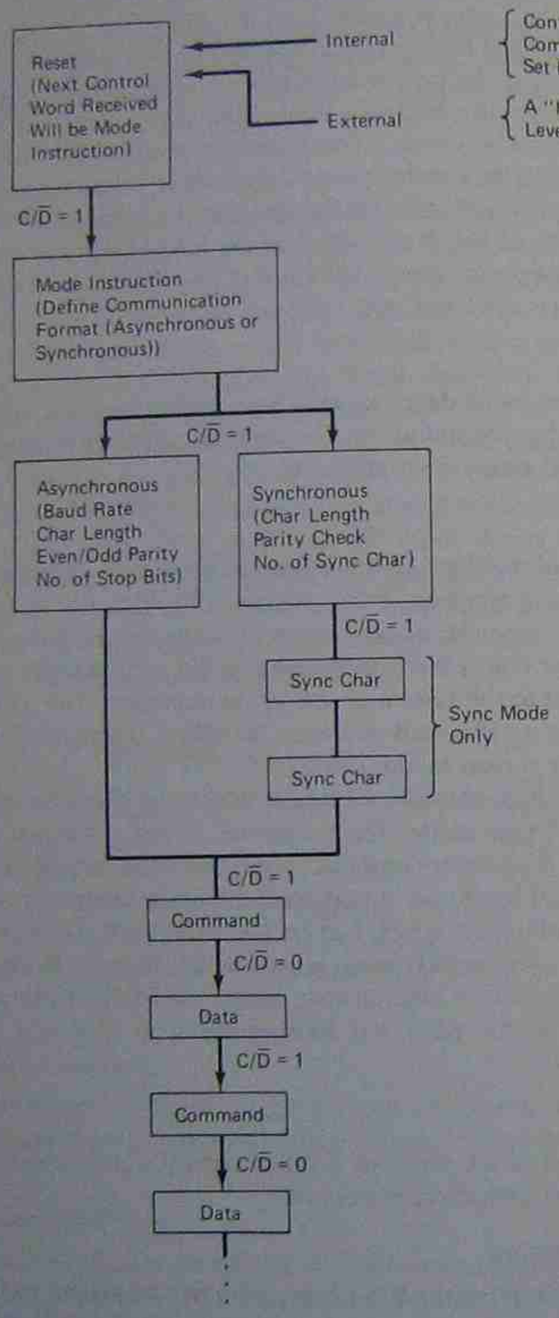
Modem Control

The modem control block provides a set of control inputs and outputs generally associated with modem control.

Operational Description

The functional definition of the 8251 is programmed by system software. Table 4-12 illustrates the sequencing for programming the 8251. After a RESET operation, which can be externally operated or internally operated

Table 4-12 Programming the 8251 USART

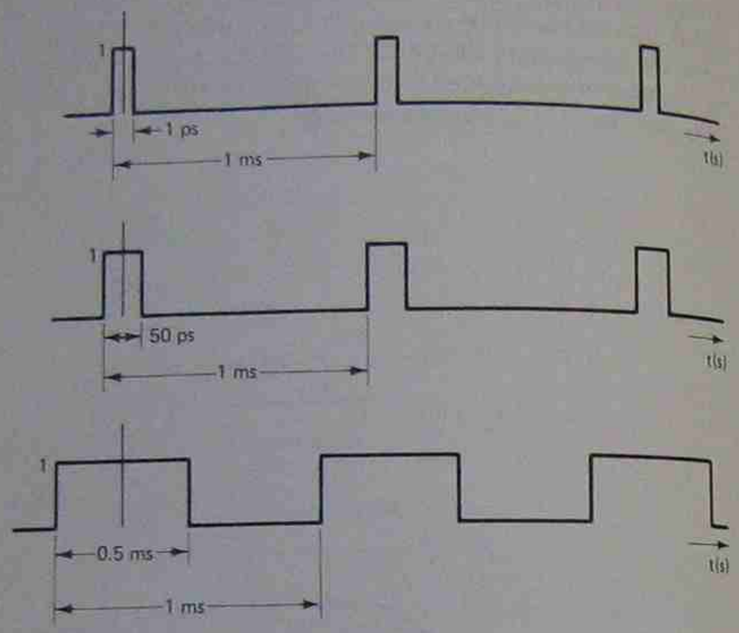


by setting a bit in the COMMAND instruction, the MODE instruction must be loaded by the μPU. The MODE instruction determines whether the data communication is to be asynchronous or synchronous, and also determines the baud rate, character length, and so forth. COMMAND instructions must follow the MODE instructions or Sync characters. COMMAND instructions can be written into 8251 in the data block at any time during the operation of the 8251. The COMMAND instruction controls the operation of the selected format. It provides for Enable Transmit/Receive, Error Reset, and Modem controls.

PROBLEMS

1. a. State the nature and causes of the skin and proximity effect.
   b. Give the reasons why skin and proximity effect cause pulse distortion.
   c. What is radiation resistance, and under what electrical conditions must it be taken into account?
   d. Define the electronic term dispersion.
2. What is intersymbol interference, and what effect does it have on maximum pulse rates over a transmission medium?
3. Define the following terms as they relate to electronic circuits:
   a. threshold or slicing level
   b. MARK and SPACE
   c. bias distortion

4. a. Sketch the frequency spectra of the following pulse trains giving the peak amplitude and the first three zero crossings for each signal. Include the sinc envelope in dotted form.



   b. What effect does the pulse duration have on the number of significant frequency components in its frequency spectrum? If only the frequency components up to the first zero crossing in the frequency spectrum are considered significant, how does the required transmission bandwidth relate to the pulse duration?

5. a. How far from the USART of Fig. 4-9 can a teleprinter be located if #26 AWC copper wire is used for the interconnecting line and R13 is selected at 390 Ω? Assume a minimum loop current of 18 mA.
   b. What alterations can be made to the circuit in (a) to increase the distance to the teleprinter?

6. What are the functions of the following devices?
   a. line drivers
   b. modems
   c. acoustic couplers
   d. UART

7. a. Why is induced noise less of a problem in a differential circuit/transmission line operation than in a single-ended operation?
   b. How is noise coupled from one cable pair to another cable pair? How can such induced noise be minimized?

8. When using existing cable for digital transmission, what cabling precautions should be taken?

9. a. Define the following terms and illustrate with an example:
      a) baud rate
      b) data signaling rate
      c) data transfer rate
   b. Each element of the QAM signal shown in Fig. 4-22 has a time duration of 1 μs. If such a signal experiences an error rate of 1×10^{-2}, obtain its:
      a) baud rate
      b) data signaling rate
      c) data transfer rate

10. a. Describe the difference between full-duplex and half-duplex modem operation. Relate these to the Bell System 103 and 202 modems.
    b. What is the function of the reverse channel in the Bell System 202 modem?
    c. Describe the difference in operation when a modem is in the originate and the answer mode. What determines the mode a modem undertakes?

11. With the aid of graphical illustrations, show the three primary types of modem modulation techniques.

12. a. What is a dibit? Using the waveforms shown in Fig. 4-30, what is the phase of the output signal when the dibit is an 11? A 10?
    b. What is QAM?
    c. Describe the nature of the Gray code and give its chief advantage.

13. a. How can the carrier be extracted in a BPSK demodulator?
    b. What is the purpose of the LPF in the demodulator of Fig. 4-27?

14. Explain why DPSK is frequently used rather than 4Ø PSK on digital radio systems.

15. Give the truth table of the exclusive NOR gate.

16. a. What is the purpose of the level shift block in Fig. 4-32?
    b. If the reference digit is a 0, obtain the encoded message for the same data as given in Table 4-5.

17. A signal element A cos $\omega_c t$ (representing a 1) succeeded by a signal element A cos $(\omega_c t + \pi)$ (representing a 0) is applied to the DPSK demodulator of Fig. 4-33. Give the analytical expression for the

signal during the second signal element at the following locations:
a. prior to the filter
b. after the LPF
c. at the comparator output
Assume all gains to be unity.

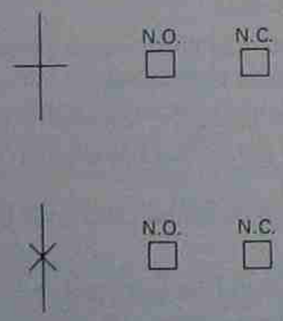
18. a. To what type of equipment does the RS-232-C standard apply?
b. What does the RS-232-C standard assure to the installer or user?
c. State the allowed MARK and SPACE voltage ranges. For what voltage range is the control voltage ON?
d. Why does the RS-232-C standard usually result in less than a 50 ft distance between the DCE and DTE?
e. With which equipment (DTE or DCE) is the female connector associated?

19. Describe the nature and direction of the signal associated with the following interchange circuits: BA, CA, CB, and CE.

20. What is the function of the control unit associated with the T103 A Data Set.

Questions 21 to 26 refer to the T103 A *Data Set block diagram* (Fig. 4-41).

21. What is the state of the following contacts?



22. What relays must be activated to obtain the following conditions?

	Relay(s)
a. DATA LAMP ON	
b. DATA SET READY	
c. RING INDICATOR	
d. TRANSMITTED DATA LEAD CONNECTED	

23. a. What is the condition of the locked oscillator when no signal is present?
b. What type of signal operates the R relay?

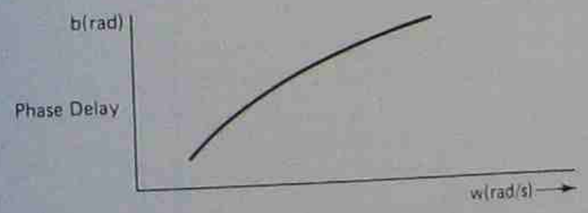
c. If the CR relay is not activated, under what two conditions will the C relay be activated?
d. What is the purpose of the capacitor across the MD relay?

24. a. Why are there two discriminators circuits, DISC 1 and DISC 2, in the DATA SET?
b. What effect does the DATA gating circuit have on the multivibrator block?
c. What activates the AN relay? Is the AN relay activated in both the originating and answering modems?
d. The AN relay affects the gating circuit to the multivibrator, the choice of discriminator, and free running locked oscillator frequency. Give the center frequency of each block, with AN relay activated and deactivated.

25. a. Under what relay condition(s) is the transmitted Data lead (Pin 2) connected to the M.V. gating circuit?
b. Under what relay condition(s) is the Received Data lead (Pin 3) connected to the slicer output?
c. What is the function of the slicer block?

26. a. What is the □o circuit across the T-R input lines? Why must a capacitor be placed in series with this circuit?
b. What is the purpose of the varistors across the transformer windings?
c. Under what condition does the DATA lamp turn ON?

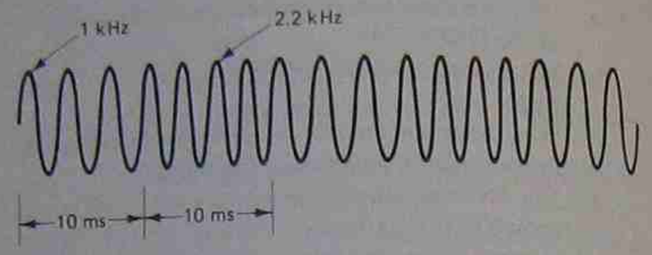
27. a. With the use of the phase curve shown, define in words and by an equation the group or envelope delay.



b. Sketch a desired phase delay characteristic on the figure of (a).
c. Sketch an ideal attenuation vs. frequency characteristic of a telephone network.

28. State the functions of the following:
a. duplexer
b. DAA network

29. a. Sinusoidal multiplication results in sum and difference frequencies. Sketch the frequency spectrum of the following 1 kHz pulsed-carrier signal.



b. If this waveform is considered an overmodulated AM signal, what effect does overmodulation have on the frequency spectrum of an AM signal?

30. Sketch the frequency spectrum of the following FSK signal.



5

Analog Pulse Modulation

5–1 INTRODUCTION

This chapter will consider the transmission of information via analog pulses, whereby each sample of the analog signal (message) can take on a continuous range of values. In digital transmission, which will be discussed in the next chapter, each sample can take on only a limited number of values.

An analog signal that is sampled often enough can be exactly specified, even if it is continuously varying with time, as long as it is band limited. The analog signal can be fully reconstructed from the sampled pulses if no noise or distortion has been added during transmission.

In analog pulse modulation, the samples directly modulate a periodic pulse train with one pulse for each sample. This discussion will focus on pulse amplitude modulation (PAM), which is probably the most common form of analog pulse modulation. A brief discussion of pulse duration and pulse position modulation schemes will follow. Although all these modulation schemes are widely used, the most favored is pulse code modulation (PCM). This will be discussed in Chap. 6.

5–2 TIME SAMPLING

Analog signals representing voice or temperature are continuous functions with time. Time quantization is used to transmit such signals in a pulsed format, whereby the signal is defined or measured at only discrete time instants. If the sampling rate is sufficiently high, information should not be lost by this technique. The signal can be reproduced from these samples through various means.

The samples ideally consist of pulses of zero width having an amplitude equal to the signal amplitude at the instant of sampling. Since ideal sampling is not a physically realizable process and the resultant pulses do not contain any energy, the following types of sampling will be considered: (a) natural sampling and (b) sample-and-hold.

Natural sampling consists of samples which are of short, but not infinitesimal duration, that have amplitudes which briefly follow the signal. Figure 5-1(b) shows such a waveform. Sample-and-hold is undoubtedly the most common technique since it retains the sampled value for a longer period, as shown in Fig. 5-1(c). When applied to the encoder, sample-and-hold allows time for a sequence of operations rather than a single very rapid operation. The sample-and-hold circuit imposes a non-return-to-zero (NRZ) sampling in the input signal shown.

An example of an electronic switch is the MOS-FET circuit shown in Fig. 5-2. The P-channel acts as a low impedance, or closed, switch when

its gate voltage goes negative. Thus, when the input to the driver goes low, G_1 goes negative, and the switch closes. This allows $m(t)$ to appear at the output. When the driver input goes high, G_1 goes high and the resistance across the drain and source is in the tens of megohms.

Sample-and-hold circuits depend upon the rapid charging of a capacitor to the potential of the input sample amplitude, and then the opening of the switch to retain the charge for the following circuits. The switching in Fig. 5-3 is done by the N-channel FET which presents a fairly low time constant in conjunction with the capacitor when the input sampling pulse goes high. The FET gate opens when the pulse goes to zero and the capacitance charge is maintained except for the leakage of the connecting high impedance circuits.

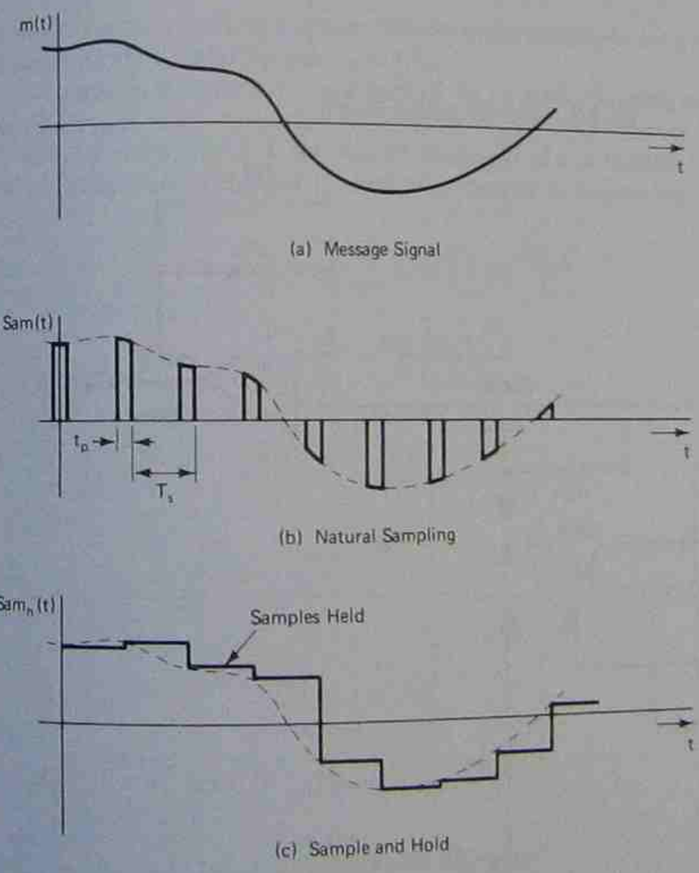


Fig. 5-1 Sampling Techniques

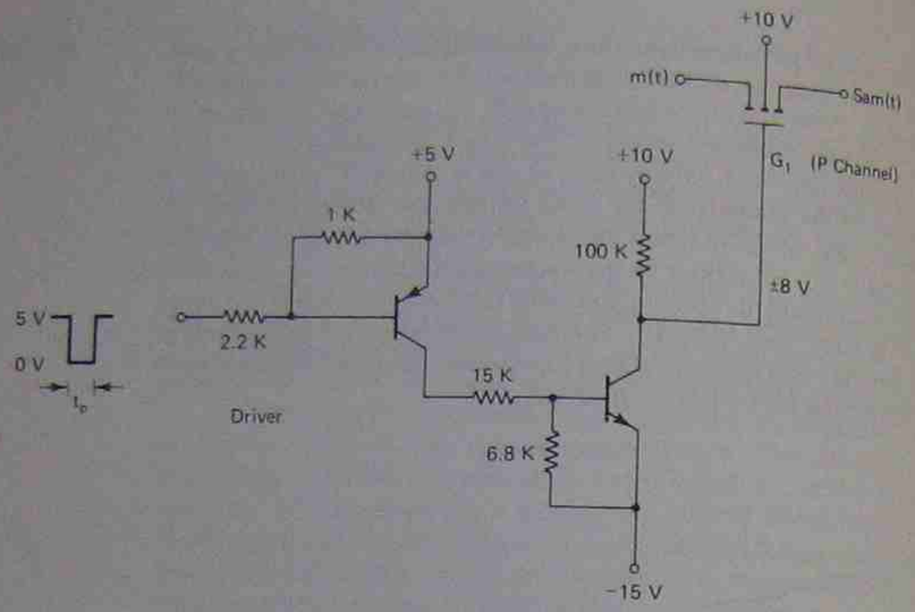


Fig. 5-2 MOS-FET Analog Switch

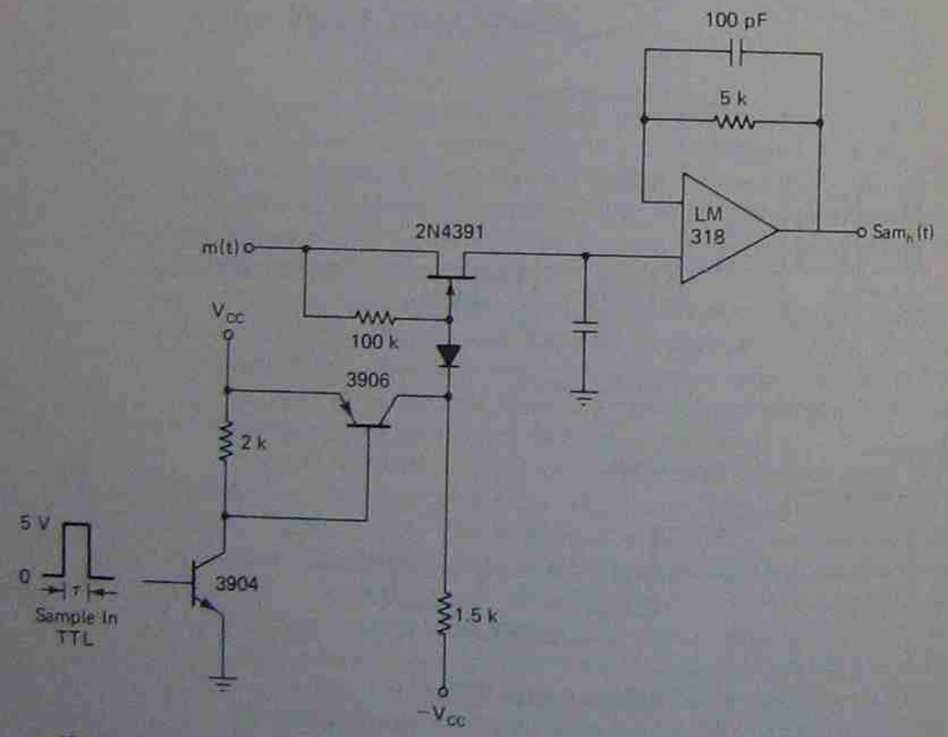


Fig. 5-3 Sample-and-Hold Schematic for a Maximum Sampling Rate of About 100 kHz

For high speed sampling, switching diodes in a bridge format can be used. The pulse transformer in Fig. 5-4 isolates the sampling pulse from the signal path, which results in the bridge floating in the signal path.

5–3 THE SAMPLING THEOREM: NATURAL SAMPLING

Consider a signal $m(t)$ which is band limited and has a highest frequency component of f_{max}. Let this signal be naturally sampled at a pulse rate, f_s, and with sampling times, t_p. From Fig. 5-5, observe that the sampled waveform can be expressed as:

$$Sam(t) = m(t) \times S(t) \tag{5-1}$$

where $S(t)$ is the sampling function consisting of a rectangular wave train of pulse width, t_p, and period, $T_s = 1/f_s$.

To obtain the frequency spectrum of the sampled waveform, multiply the Fourier series expansion of $S(t)$ by the frequency spectrum of $m(t)$. As derived in Appendix A, the Fourier expansion of a rectangular pulse train of Amplitude 1, period T_s, and pulse duration t_p is given by

$$S(t) = \frac{t_p}{T_s} + \frac{2t_p}{T_s} \sum_{n=1}^{\infty} \text{sinc}\, \frac{nt_p}{T_s} \cos \frac{2\pi nt}{T_s} \tag{5-2}$$

where $\text{sinc}\, \dfrac{nt_p}{T_s} = \dfrac{\sin \pi t_p\, n/T_s}{\pi\, t_p\, n/T_s}$

and T_s is the sampling time.

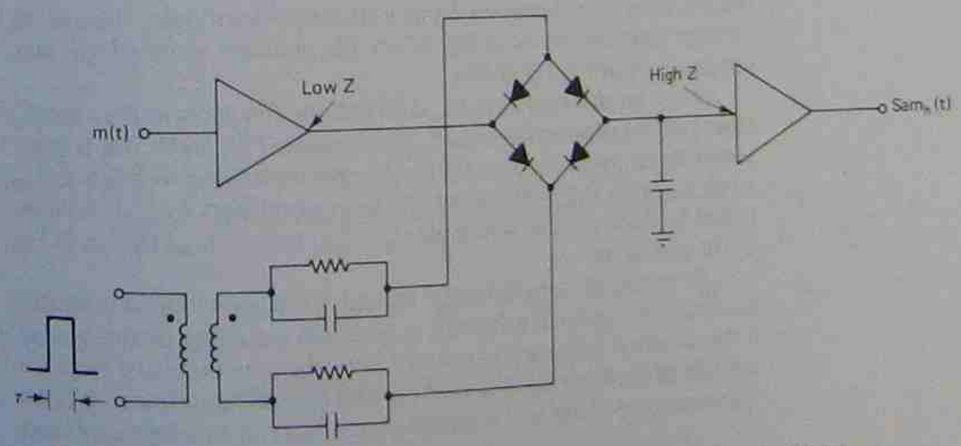


Fig. 5-4 Sample and Hold Circuit for Very Narrow Sampling Times ($\tau < 1$ ns)

This function consists of an infinite number of harmonics of $f_s = 1/T_s$, which have a decaying amplitude given by sinc nt_p/T_s. Since the numerator of the sinc function has a maximum amplitude of 1, the amplitude decays as $1/n$. The sinc function goes to zero whenever its numerator is an integral number of π radians since sin $n\pi = 0$.

It may appear that the sinc function goes to zero for $n = 0$ since sin $0 = 0$; however, the denominator also goes to 0 when $n = 0$, and division by 0 is not allowed. In actual fact, the sinc function has a maximum value of 1 when its argument is zero. This can be checked by substituting an extremely small value for x into sin x/x and calculating the result.

Thus, $S(t)$ goes through zero whenever

$$\pi t_p \frac{n}{T_s} = m\pi \text{ or when } \frac{n}{T_s} = \frac{m}{t_p}$$

The spectrum of the switching function is shown in Fig. 5-6

The sampled baseband signal $m(t)$ $S(t)$ is thus given by the expression

$$Sam(t) = \frac{t_p}{T_s} m(t) + \frac{2t_p}{T_s}\left[\text{sinc}\,\frac{t_p}{T_s}\, m(t)\cos\frac{2\pi t}{T_s} + \text{sinc}\,\frac{2t_p}{T_s}\, m(t)\cos\frac{2\pi 2t}{T_s} + \cdots\right] \quad (5\text{-}3)$$

The spectrum of the sampled function is now given by the multiplication of the frequency components of the message, $m(t)$, by the spectrum shown in Fig. 5-6. Recall that when two sinusoidal functions are multiplied, such as $\cos \omega_1 t \times \cos \omega_2 t$, the sum and difference components, $\frac{1}{2}\cos(\omega_1 + \omega_2)t + \frac{1}{2}\cos(\omega_1 - \omega_2)t$, are obtained. Each message frequency component when multiplied by one of the harmonics of $S(t)$ will result in a lower and upper sideband around each harmonic. Thus, for the message spectrum shown in Fig. 5-7(a), the naturally sampled spectrum appears as shown in Fig. 5-7(b).

It can be observed from Eq. (5-3) that the first term in the series is, aside from the constant factor t_p/T_s, the message $m(t)$ itself. This is represented in the frequency domain by the spectrum lying in the 0 to f_{max} range shown in Fig. 5-7(b). Similarly, the second term in (5-3) is represented by the first set of sidebands ($f_s - f_{max}$ to $f_s + f_{max}$) shown in Fig. 5-7(b), and so on.

By passing the sampled signal through a low-pass filter, it is possible to extract the signal $m(t)$ if the various spectrum groupings do not overlap. If the sampling frequency is reduced until the spectrum just begins to overlap, as shown in Fig. 5-8, $m(t)$ can still be separated from the rest of the spectrum. As long as f_s is greater than 2 f_{max}, no interference occurs.

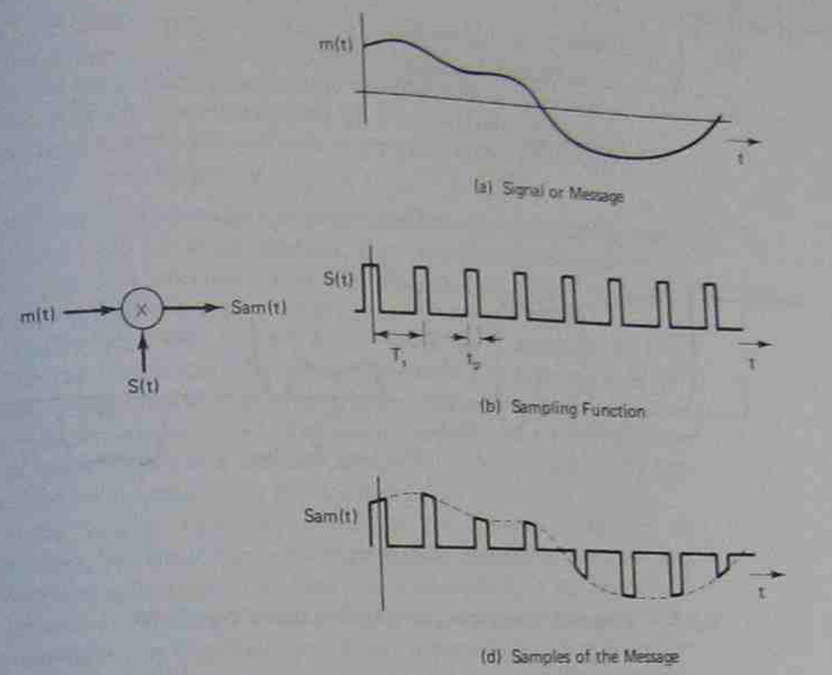


Fig. 5-5 Generation of a Naturally Sampled-Signal

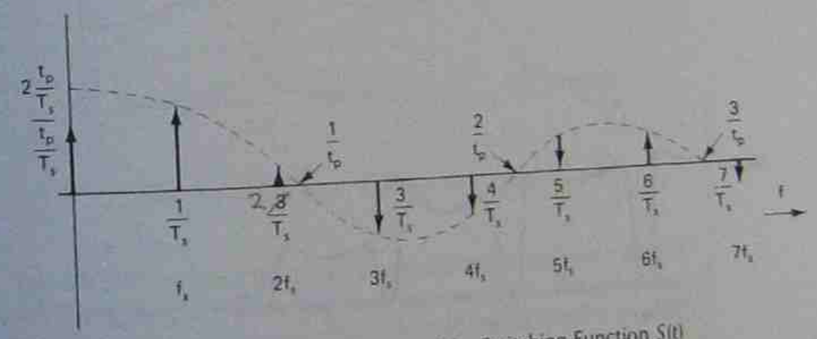
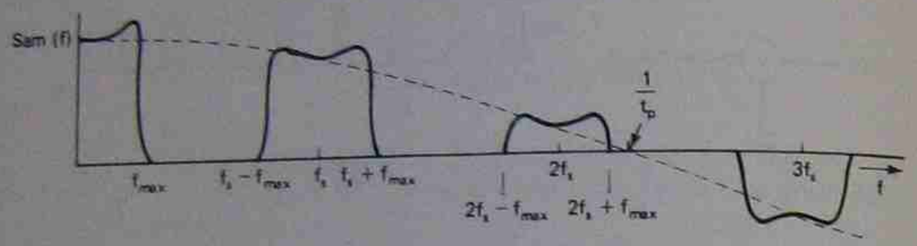


Fig. 5-6 Line Spectrum of the Switching Function $S(t)$



(a) Message Spectrum



(b) Sampled Spectrum

Fig. 5-7 Frequency Spectrum of a Natural Sampled Waveform

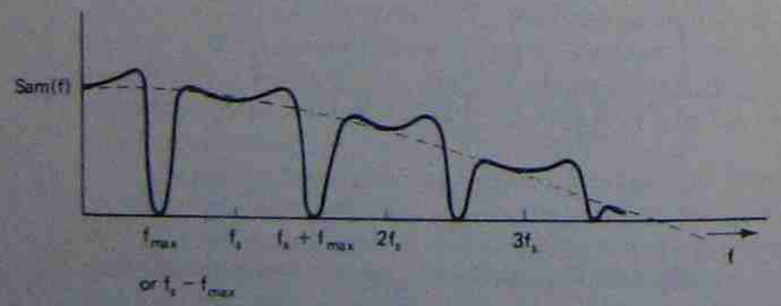


Fig. 5-8 Frequency Spectrum when $f_s = 2\ f_{max}$

This is a simple proof of the Sampling Theorem. It can be stated as follows:

> A band limited signal of maximum frequency f_{max}, when sampled at regular intervals at a rate of at least $2\ f_{max}$, can be completely reconstructed from these samples with no distortion. That is, a signal waveform can be properly represented if at least two samples are taken for every cycle of the highest significant signal frequency.

CCITT has set a standard of 8000 samples per second for the standard 300 to 3400 Hz voice channel. This complies with the Sampling Theorem, and all information of the original audio signal is recoverable. Thus, a sample is taken every $^1/_{8000}$ s or every 125 μs.

Actual analog signals, such as voice, are not so sharply band limited. They roll off gradually, resulting in unnecessarily high sampling frequencies. Filters can be employed to reduce the higher frequencies, but due to their finite roll-off, can never eliminate them. Also, since the spectrum of noise that creeps into any system is broad, noise from the first lower sideband overlaps into the baseband.

If the original signal has frequency components above $f_s/2$, as shown in Fig. 5-9, the lower sideband at f_s will overlap with the baseband and distortion will occur. This effect is termed aliasing, or frequency foldover distortion. Aliasing can be reduced by improving the filter slope characteristics or by increasing the sampling frequency. Going to a higher pole filter for faster roll-off raises costs. Increasing the sampling frequency increases the bandwidth requirement of the transmission medium.

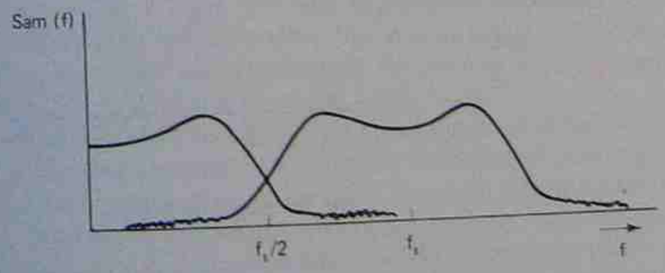


Fig. 5-9 Aliasing or Foldover Distortion

The effect of an inadequate sampling rate on a sinusoidal signal is illustrated in Fig. 5-10. An alias frequency appears in the recovered signal. This low frequency alias signal, shown by the dotted line, is a result of sampling at less than twice per cycle.

The effect of sampling frequency and rate of filter roll-off on distortion is discussed in the following somewhat simplified case.

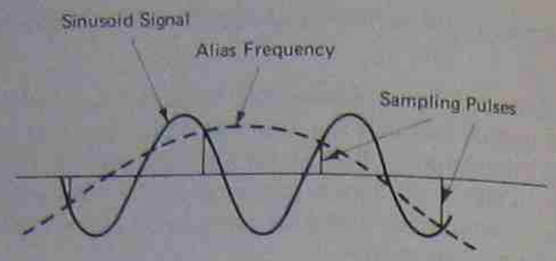


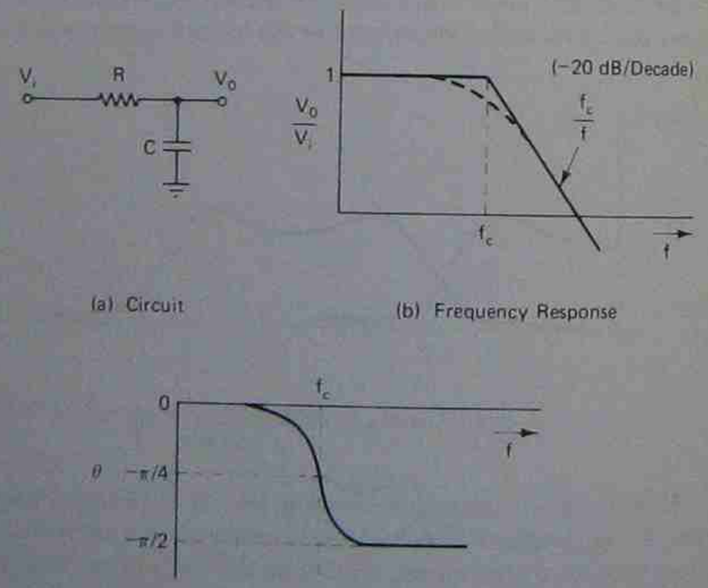
Fig. 5-10 Alias Frequency as a Result of Inadequate Sampling Rate

For a single pole filter, as shown in Fig. 5-11, the ratio of output to input voltage can be expressed as

$$\frac{V_o}{V_i} = \frac{-j/\omega C}{R - j/\omega C} = \frac{1}{1 + j\omega RC}$$

$$\text{thus} \quad \left|\frac{V_o}{V_i}\right| = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_c}\right)^2}} \tag{5-4}$$

where $\omega_c = \frac{1}{RC}$ represents the radial cut-off frequency.



(a) Circuit (b) Frequency Response

(c) Phase Response

Fig. 5-11 Transfer Function Amplitude and Phase Plot of a Single Pole Filter

If n such filters are cascaded, each isolated from another, the response becomes

$$\left|\frac{V_o}{V_i}\right| = \left(\frac{1}{\sqrt{1 + (\omega/\omega_c)^2}}\right)^n \tag{5-5}$$

which, for frequencies much above cut-off, approaches the asymptote

$$\left|\frac{V_o}{V_i}\right| = \left(\frac{1}{\omega/\omega_c}\right)^n = \left(\frac{\omega_c}{\omega}\right)^n = \left(\frac{f_c}{f}\right)^n \tag{5-6}$$

In this theoretical development, assume the sharp corner square frequency response holds true, as illustrated by the solid line of Fig. 5-11(b). It is sufficiently accurate for the purpose of this case.

When considering the ratio of output to input power, the voltage response is squared. Therefore, for $f \gg f_c$

$$\frac{P_o}{P_i} = \left(\frac{f_c}{f}\right)^{2n} \tag{5-7}$$

In logarithmic terms, the roll-off occurs at a rate of

$$10 \log \left(\frac{f_c}{10\, f_c}\right)^{2n} = -20\, n \text{ dB/decade}$$

Consider an audio signal with a flat frequency response of unity amplitude that is applied to an n-pole filter and then sampled at a rate of f_s, as shown in Fig. 5-12(a). The filter will determine the signal baseband bandwidth applied to the sampler. The resulting frequency response after sampling, *Sam* (*t*), closely resembles the previous sampling spectrum shown in Fig. 5-7. This is illustrated in Fig. 5-12(b).

An examination of Fig. 5-12(b) shows that some of the lower sideband of the first sampled harmonic falls into the band of the baseband signal. This appears as an undesirable signal and distorts the baseband signal. This type of distortion is known as aliasing, or foldover distortion. The aliasing distortion shown by the shaded area X is due to the finite rate of attenuation of the low-pass filter. Since the shaded area Y, which is equivalent to the shaded area X, is the easiest of the two to express mathematically, it will be used to obtain the distorting noise power.

The bandlimited baseband signal amplitude response can be expressed mathematically as:

$$Sam(f) = 1 \quad \text{for} \quad 0 < f < f_c \tag{5-8(a)}$$

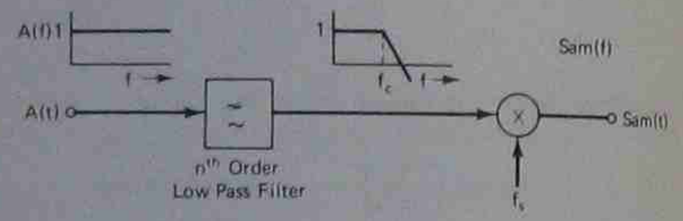
$$= \left(\frac{f_c}{f}\right)^n \quad \text{for} \quad f > f_c \tag{5-8(b)}$$

where n is an integer representing the number of poles of the filter, and f_c represents the filter 3 dB cut-off frequency. For convenience, let the sampling frequency be represented by some constant that is k times the minimum allowable sampling rate of $2f_c$, where the significant baseband signal power is assumed to lie in the frequency range up to f_c.

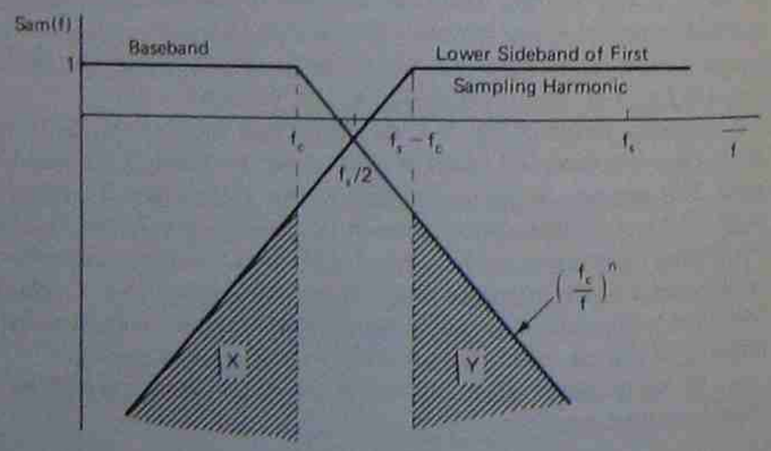
$$f_s = k\ 2f_c \tag{5-9}$$

The signal power, S, can thus be obtained by integrating the power spectral response (the square of the $Sam(f)$ response) from 0 to f_c.

$$S = \int_0^{f_c} 1^2\ df = f_c \tag{5-10}$$



(a) Natural Sampling Scheme Complete with Frequency Plots



(b) Frequency Response of Sampled Signal

Fig. 5-12 Aliasing with n-Pole Filter

The corresponding distortion power, represented by the shaded area X or Y, is given by integrating the square of Eq. (5-8(b)) from

$$f_s - f_c \quad \text{to} \quad f_s$$

$$N = \int_{f_s - f_c}^{f_s} \left(\frac{f_c}{f}\right)^{2n} df$$

$$= \frac{1}{-2n + 1}\left[\frac{f_c}{(2k)^{2n-1}} - \frac{f_c}{(2k-1)^{2n-1}}\right] \tag{5-11}$$

Hence the signal power-to-distortion power ratio is given by:

$$\frac{S}{N} = \frac{1}{\frac{1}{-2n+1}\left[\frac{1}{(2k)^{2n-1}} - \frac{1}{(2k-1)^{2n-1}}\right]}$$

which, in decibels, is

$$\left(\frac{S}{N}\right)_{dB} = -10 \log\left\{\frac{1}{-2n+1}\left[\frac{1}{(2k)^{(2n-1)}} - \frac{1}{(2k-1)^{(2n-1)}}\right]\right\} \tag{5-12}$$

Table 5-1 and Fig. 5-13 show how the filter roll-off affects the aliasing distortion. The curves show that the greater the number of poles in the filter, or the greater the filter roll-off, the less will be the distortion. The sampling frequency also has a significant effect on distortion in that an increased sampling frequency inserts a larger guard band between the baseband and the sideband, thus improving the S/N ratio.

5–4 SAMPLE AND HOLD: FLAT-TOP SAMPLING

More practical sampling systems use the sample-and-hold circuit. This circuit samples the input at precise intervals and holds the input amplitude constant for the duration of the sampling time even though the measured signal is varying. This holding time can extend in time to the next sampling instant, but may be of a shorter duration, as shown in Fig. 5-14(b).

Convolution must be used to analyze the sampled waveform. Since this technique may be unfamiliar to many readers, this discussion will merely state its results. The frequency spectrum of the sampled signal consists of a baseband around the harmonics of the sampling frequency very much like that shown in Fig. 5-7. However, there is one additional

Table 5-1

N	k	S/N	N	k	S/N
1.00	1.00	3.01	5.00	2.50	64.35
2.00	1.00	5.35	6.00	2.50	77.03
3.00	1.00	7.13	7.00	2.50	89.65
4.00	1.00	8.49	8.00	2.50	102.23
5.00	1.00	9.55	1.00	3.00	14.77
6.00	1.00	10.42	2.00	3.00	29.49
7.00	1.00	11.14	3.00	3.00	44.17
8.00	1.00	11.76	4.00	3.00	58.80
1.00	1.50	7.78	5.00	3.00	73.39
2.00	1.50	15.33	6.00	3.00	87.93
3.00	1.50	22.65	7.00	3.00	102.43
4.00	1.50	29.79	8.00	3.00	116.90
5.00	1.50	36.75	1.00	3.50	16.23
6.00	1.50	43.58	2.00	3.50	32.43
7.00	1.50	50.30	3.00	3.50	48.59
8.00	1.50	56.93	4.00	3.50	64.73
1.00	2.00	10.79	5.00	3.50	80.82
2.00	2.00	21.46	6.00	3.50	96.89
3.00	2.00	32.02	7.00	3.50	112.93
4.00	2.00	42.47	8.00	3.50	128.94
5.00	2.00	52.82	1.00	4.00	17.48
6.00	2.00	63.08	2.00	4.00	34.94
7.00	2.00	73.27	3.00	4.00	52.37
8.00	2.00	83.39	4.00	4.00	69.77
1.00	2.50	13.01	5.00	4.00	87.15
2.00	2.50	25.95	6.00	4.00	104.51
3.00	2.50	38.82	7.00	4.00	121.84
4.00	2.50	51.62	8.00	4.00	139.16

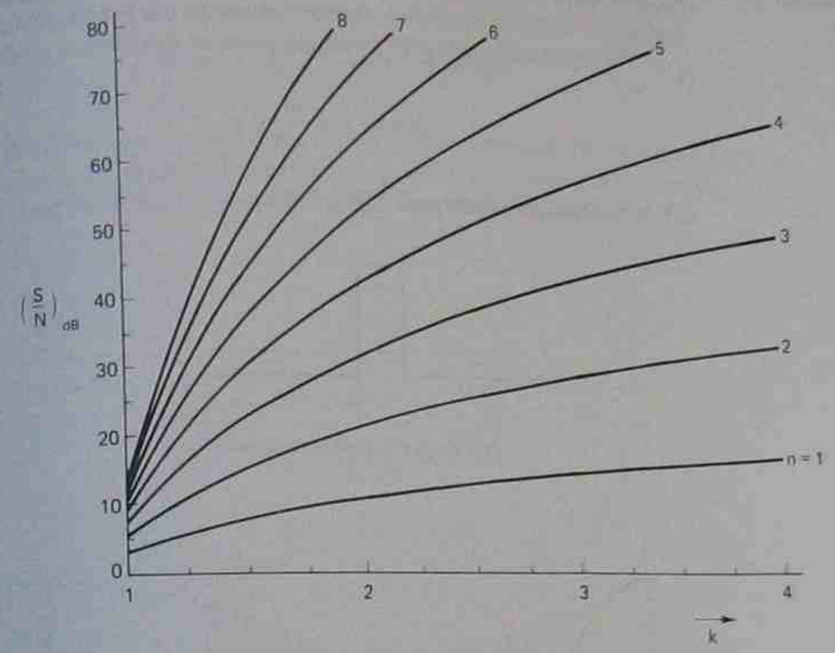


Fig. 5-13 S/N Ratio Due to Aliasing

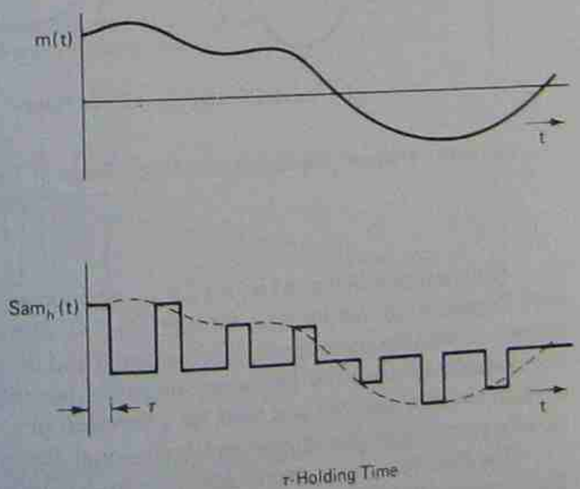


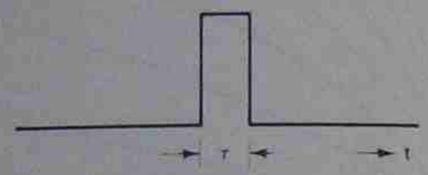
Fig. 5-14 Sample-and-Hold Waveform

contribution to distortion, called aperture distortion. Aperture distortion depends upon the holding time, and follows the frequency spectrum envelope of a single pulse of time duration equal to the holding time, τ.

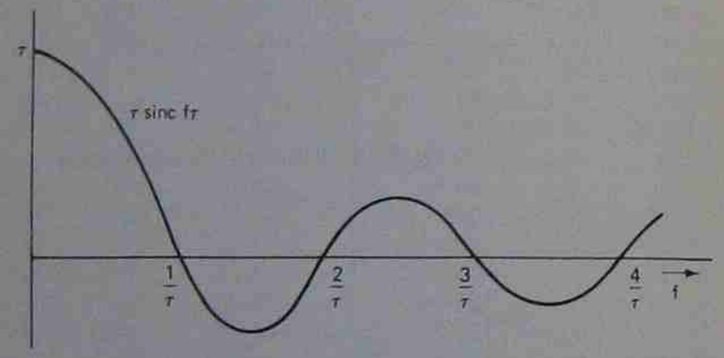
The frequency spectrum of a single pulse of duration is given by:

$$\tau \text{ sinc } f\tau = \tau \frac{\sin \pi f\tau}{\pi f\tau} \tag{5-13}$$

which is graphically illustrated in Fig. 5-15(b).



(a) Single Pulse of Duration τ



(b) Frequency Spectrum of a Single Pulse

Fig. 5-15 Frequency Spectrum of a Single Pulse of Time Duration τ

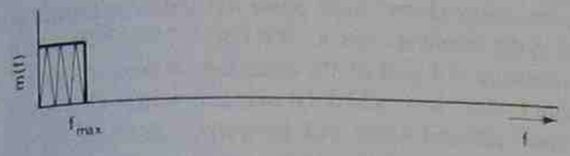
The spectrum shown in Fig. 5-15 is a continuous one which decays with frequency, as can be seen from Eq. (5-13). Furthermore, as τ decreases, so does the energy content of the signal and the various frequency components. Given a flat frequency response for $m(t)$ (shown in Fig. 5-16(a)) which is sampled and held for a duration of τ, results in the frequency spectrum for the sample-and-hold signal shown in Fig. 5-16(b).

The distortion can be more clearly illustrated by expanding the low frequency region that contains the frequencies transmitted after filtering.

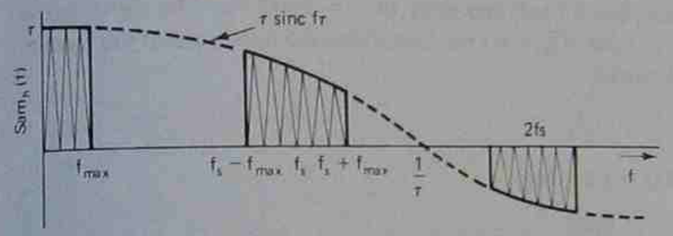
Thus, from Fig. 5-16(c), the percentage error due to aperture distortion can be calculated by:

$$\% \text{ error} = \left(1 - \frac{\sin \pi f\tau}{\pi f\tau}\right) \times 100 \tag{5-14}$$

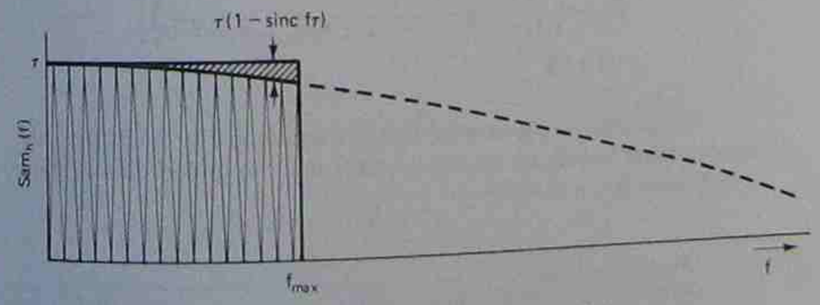
It can be readily seen from Fig. 5-16(b) that the larger the sampling aperture, τ, the faster the envelope drops off in amplitude and the more attenuated the higher frequency components of the signal become. Ob-



(a) Message Spectrum



(b) Frequency Spectrum of Sample-and-Hold Signal with a Holding Time of τ



(c) Aperture Distortion Indicated by the Shaded Area

Fig. 5-16 Sample-and-Hold Circuit Frequency Response

viously, this creates an error in the recovered signal. Figure 5-17 is a useful design chart which shows the percentage error introduced versus the sampling aperture time multiplied by the signal frequency. The rate of sampling does not affect aperture error.

Although it might appear wise, at first glance, to keep the aperture time short in order to minimize error, this results in reduced signal energy content which makes the system more susceptible to noise. As a result, the aperture time is usually maximized, as illustrated earlier in Fig. 5-1(c).

Fortunately, this type of distortion is linear. It can be taken out by inserting an equalizer in cascade with the input or output filter whose transfer function is the inverse of the single pulse frequency response of time duration equal to the sampling time, τ. (See Fig. 5-15(b).) That is, if the equalizer has a response of $1/\text{sinc } f\tau$, the equalizer action, in combination with the aperture effect, will yield a flat overall transfer characteristic between the input baseband signal and the output received signal. Some error always remains because the $u/\sin u$ transfer characteristic cannot be exactly synthesized.

Example 5-1 In PCM telephony, a flat signal is first filtered with a 5-pole filter at −100 dB/decade, and a break frequency of $f_c = 3400$ Hz. The signal is then sampled at a rate of $f_s = 8$ kHz. Determine the S/N ratio of the received baseband signal.

Solution

$$f_s = 2\,k\,f_c$$

$$8000 = 2\,k\,(3400)$$

$$k = \frac{8000}{6800} = 1.176$$

$$\text{S/N} = -10 \log \left\{ \frac{1}{-2 \times 5 + 1} \left(\frac{1}{(2 \times 1.176)^{(2\times5-1)}} - \frac{1}{(2 \times 1.176 - 1)^{(2\times5-1)}} \right) \right\}$$

$$= 21.4 \text{ dB}$$

The spectrum of the sampled signal is shown in Fig. 5-18. The sidebands around 16 kHz are not shown. Note that the folded data is about 13 dB down at $f_c = 3.4$ kHz.

Example 5-2 In the telephone system, the sampling frequency is standardized at 8 kHz. If the error due to sampling aperture cannot be more than ½% at 3 kHz, what aperture time is permitted?

Solution Using Fig. 5-17, for 0.5% error, $f\tau = 0.055$. Therefore the maximum aperature time is:

$$\tau = \frac{0.055}{f} = \frac{0.055}{3000} = 18.3\ \mu\text{s}$$

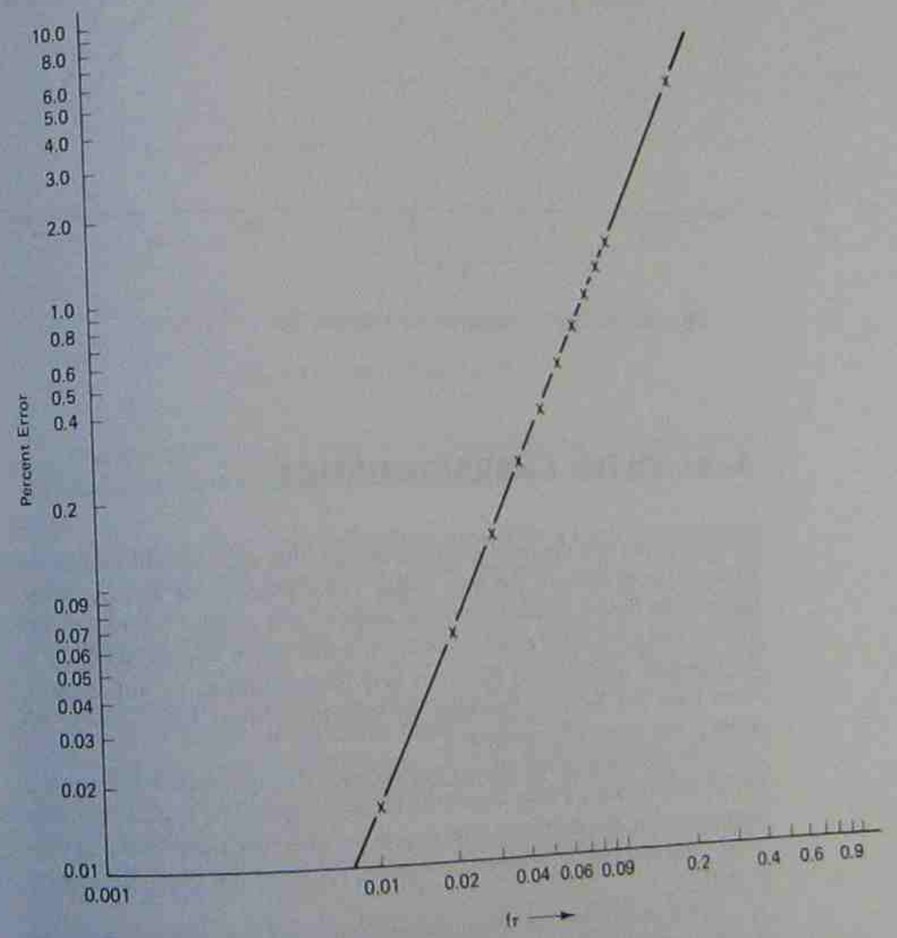


Fig. 5-17 Aperture Error

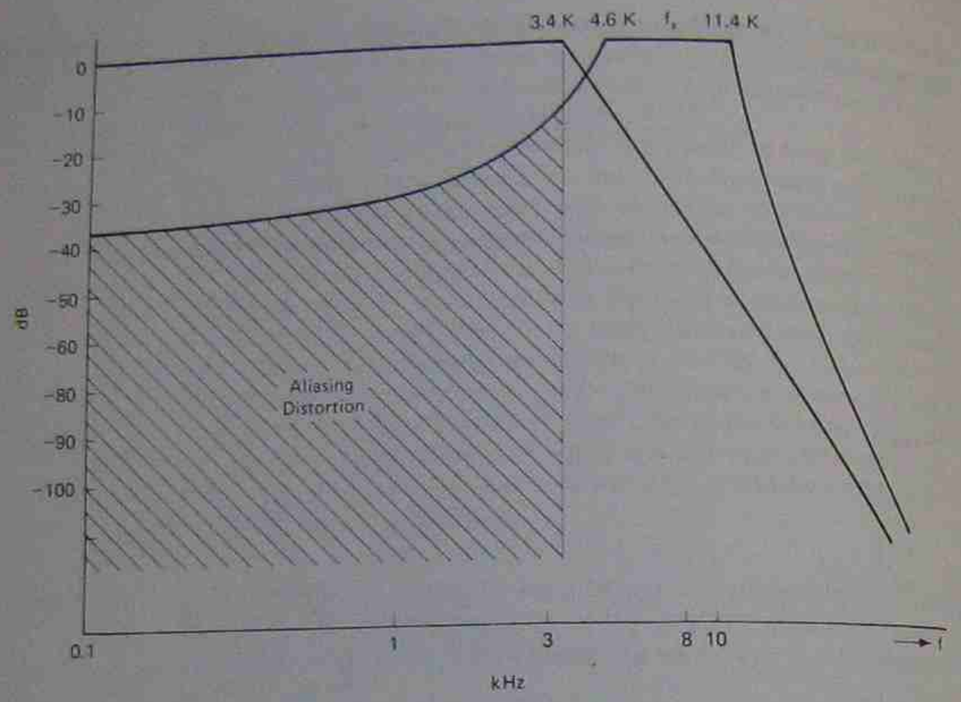


Fig. 5-18 Frequency Spectrum for Example 5-1

5–5 FILTER CHARACTERISTICS

Without going into the details of filter design, this section will outline the basic characteristics of the four major classifications of analog filters: namely, the Bessel, Butterworth, Chebyshev, and elliptic filters. The frequency and time delay characteristics of these filters will be sketched in the low-pass mode since the front end pre-sampling, or anti-aliasing, filters have this feature. Bandpass or high-pass filters have similar frequency and phase characteristics.

The passband of the filter is the frequency range in which a signal is transmitted through a filter with little insertion loss. The high attenuation region is called the stopband. The region between the passband and stopband, with somewhat flexible boundaries, is called the transition region, or skirt.

In the ideal filter, the amplitude response and group delay should be constant with frequency in the passband regions. The group, or envelope,

delay is related to the phase response by the expression

$$T_g = \frac{db}{d\omega} \quad \text{where } b \text{ is the phase response}$$

Refer back to Sec. 3–2.6 for a more detailed discussion of group delay. Group delay in the stopband is of little interest since this region rejects the undesired signal and noise components.

A typical feedback low-pass Bessel, Butterworth, or Chebyshev filter element is shown in Fig. 5-19. The choice of resistance and capacitance values determines the frequency and phase characteristics of the filter. By cascading several such networks, higher-order filters may be formed.

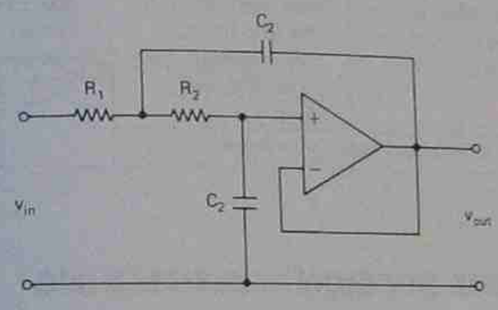


Fig. 5-19 Low-Pass Filter Element

The Bessel filter has a very linear phase response but a fairly gentle skirt slope, as illustrated in Fig. 5-20. It is well-suited for pulse applications.

The Chebyschev filter has a sharp roll-off but a significant passband ripple. The faster the roll-off, the greater the peak-to-peak ripple in the passband. The phase response is highly non-linear in the skirt region. Such unequal delay of data frequencies in the passband causes severe pulse distortion and thus increased errors at modem demodulators. This can be somewhat overcome by increasing the bandwidth of the filter so that the more-or-less linear phase region is extended.

The Butterworth filter has characteristics somewhere between those of the former two. It has a moderate roll-off of the skirt and a slightly non-linear phase response.

The elliptic filter has the sharpest roll-off of all the filters in the transition region but has ripple in both the passband and stopband regions. It can be designed to have very high attenuations for certain frequencies in the stopband, which lessens the attenuation for other frequencies in the stopband.

Table 5-2 presents a summary of the filters response characteristics discussed.

Table 5-2 Filter Response Summary

Bessel	Butterworth	Chebyschev	Elliptic
Maximally flat time delay in the *dc* vicinity. Monotonically decreasing in amplitude with frequency.	Maximally flat amplitude response in the *dc* vicinity.	Equiripple amplitude in the passband. Fastest roll-off for the same order of filter when compared to Bessel or Butterworth filters.	Equiripple amplitude in both passband and stop band. Sharper transition band and higher attenuation in the lower portion of the stop band.
"Best" phase response; "poorest" amplitude response.	Compromise between amplitude and phase when compared to Bessel and Chebyschev filters.	"Best" amplitude response. (More attenuation in the stop band as ripple is increased.) "Poorest" phase response.	High frequency attenuation lower than that of the other filters. Sensitive to component values.
Attenuation continues to increase as frequency increases			

5–6 TIME DIVISION MULTIPLEXING

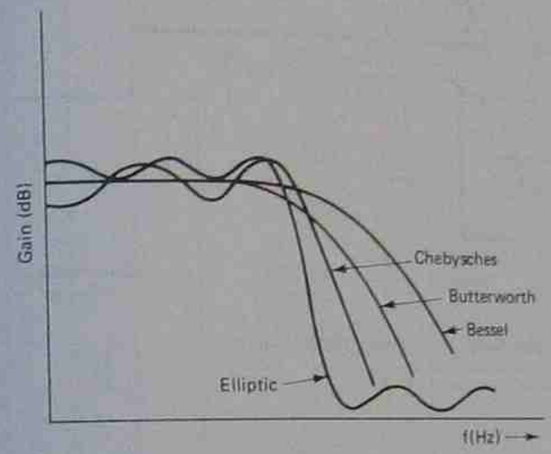
Any pulse modulation scheme involves translating the audio, or modulating, signal into a series of encoded pulses, sending these pulses over a transmission medium, and reverting the pulses back to an analog signal. Regardless of the encoding used, a PAM signal is always obtained initially.

The audio signal is first sampled at a sufficiently high rate in order to minimize distortion. By keeping the samples short in comparison to the time intervals between them, many different channels can be interleaved in one sampling period. This process, as performed by a commutator, is illustrated in Fig. 5-21. It is commonly called Time Division Multiplexing (TDM). TDM allows each channel the full bandwidth of the transmission medium whenever its signal is transmitted, although each channel is not continuously on the system. This is in contrast to Frequency Division Multiplexing (FDM) in which each channel is continuously on the system but allowed only a limited portion of the system bandwidth. This is clearly illustrated in Fig. 5-22.

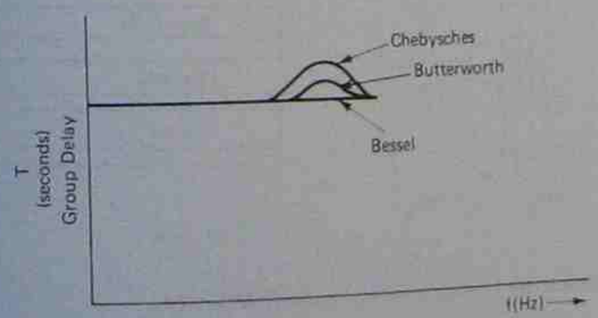
After a PAM signal is obtained, it can be transmitted directly or quantized and encoded to provide improved noise immunity. The latter is more commonly done. This will be further discussed in Chap. 6 when considering Delta and Pulse Code Modulations systems.

An example of a four-channel electronic sampling system is shown in Fig. 5-23. This system employs two flip-flops and MOS P-channel enhancement transistors as switches. When the gate to the MOS P-transistors becomes negative, the P-channels act as a low impedance circuit, with resistance at about 200 Ω. Normally, the resistance across the drain and source is in the tens of megohms.

Flip-flops A and B convert the serial clock input to binary parallel output. The four NAND gates decode this to sequentially drive the switches. When all the inputs to a NAND gate go HI, the output goes LO, thus causing the attached transistor to act as a low resistance path.



(a) Amplitude Response



(b) Group Delay Response

Fig. 5-20 Filter Characteristics

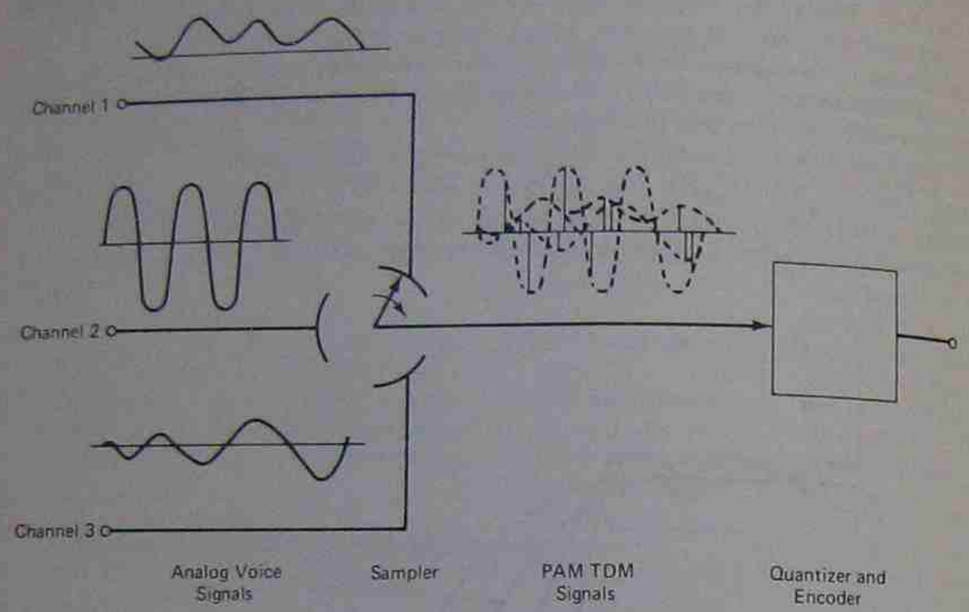


Fig. 5-21 Simplified TDM System

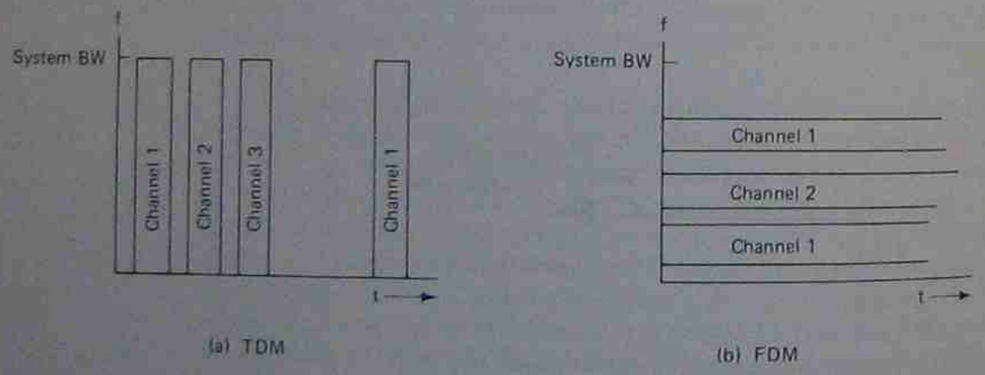


Fig. 5-22 Bandwidth Utilization of TDM and FDM Systems

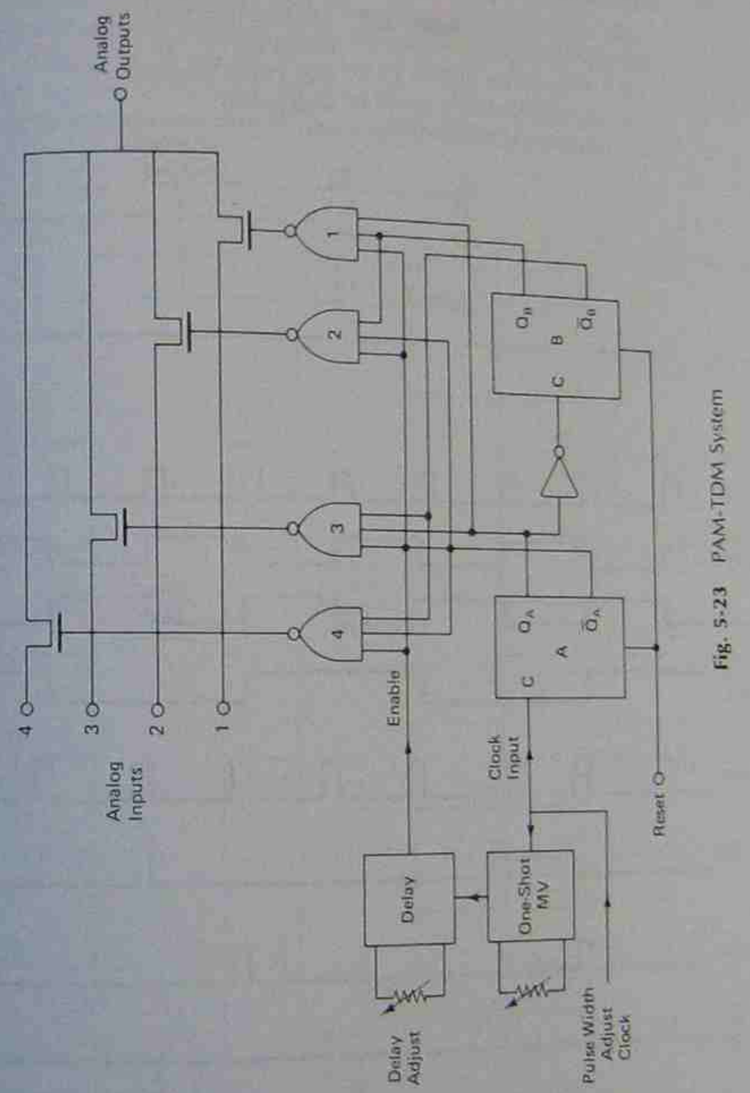


Fig. 5-23 PAM-TDM System

Since the two flip-flops provide two digits, the number of commutator switches that can be driven is 2^2 or 4. An eight-channel multiplexer would need three flip-flops; that is, $2^3 = 8$. Figure 5-24 shows the waveforms of the four-channel commutator. Table 5-3 gives the flip-flip conditions for the channel selection.

Table 5-3 Channel Selection Conditions for Fig. 5-23 Circuit (Enable must be HI)

Q_A	Q_B	Channel Selected
LO	LO	4
LO	HI	2
HI	LO	3
HI	HI	1

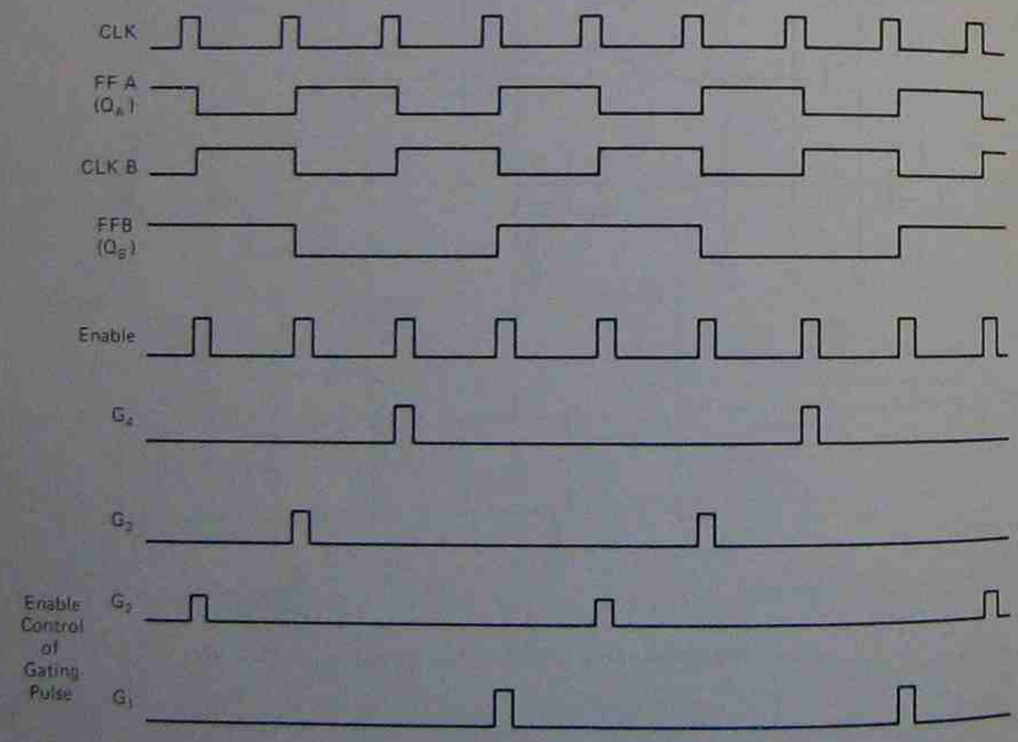


Fig. 5-24 Waveforms for the Four-Channel Commutator of Fig. 5-23

In order to establish a variable sampling aperture, the clock pulse is fed to a monostable multivibrator with pulse width adjusting control. This will establish the aperture time. A delay circuit then moves the sampling pulses such that the sampling pulse occurs some time away from the gates switching edges.

A typical commercial eight-channel multiplexer is the RCA COS/MOS CD4051A integrated circuit, the functional diagram of which is shown in Fig. 5-25. It has the three binary control inputs of A (LSB), B,C(MSB), and an inhibit input. The three binary inputs select which one of the eight channels is turned ON to connect the input to the output. For

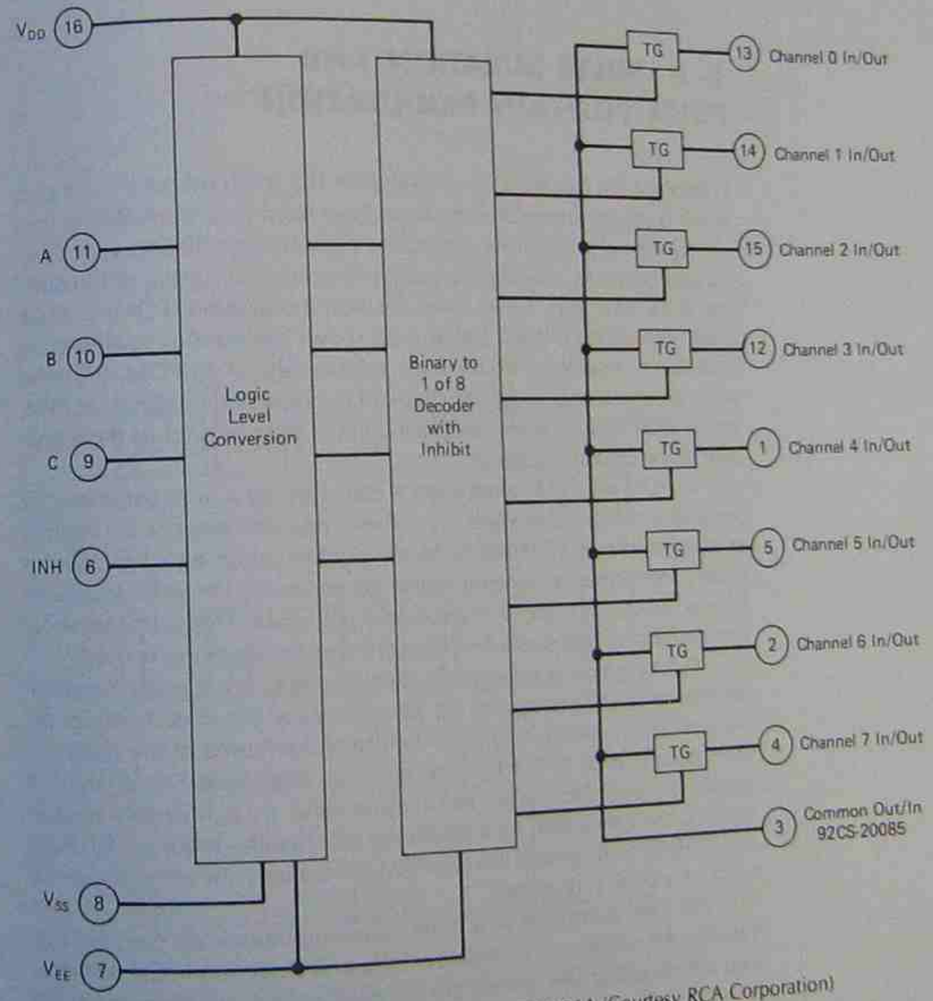


Fig. 5-25 Functional Diagram of CD4051A (Courtesy RCA Corporation)

instance, with A = 1, B = 1, C = 0, and inhibit = 0, Terminal 3 is connected to output Terminal 12 (Channel 3). Logic 0 = V_{SS}, and 1 = V_{DD}. The terminals marked CHANNEL IN/OUT and COMMON OUT/IN can be used as input or output terminals. When used on a multiplexer, the CHANNEL IN/OUT terminals are the inputs and the COMMON OUT/IN terminal is the output. When used as a demultiplexer, these roles are reversed.

When $V_{DD} - V_{EE} = 15$ V (as when $V_{DD} = 5$ V, $V_{SS} = 0$, V and $V_{EE} = -10$ V), the channel ON resistance is typically 50 Ω. The OFF resistance-input leakage is typically 10 pA for the same supply voltage condition.

5–7 PULSE DURATION AND PULSE POSITION MODULATION

In the PAM, the baseband signal modulates the amplitude of a pulse train which is spaced at regular intervals and has fixed time slots. Rather than varying the pulse amplitude, alternative modulation schemes vary either the pulse intervals, called pulse position modulation (PPM), or the duration of the time slots, called pulse duration modulation (PDM) or pulse width modulation (PWM). Figure 5-26 shows the various analog pulse modulations waveforms for the same message signal. In PDM, the pulse width is proportional to the amplitude of the modulating signal. In PPM, the pulse delay from some reference point is proportional to the amplitude of the modulating signal.

In PDM and PPM, information is conveyed by a time parameter, or the location of the pulse edges. Thus, these modulation types are referred to as pulse time. In PAM and PDM, the sample values equal to zero are usually represented by nonzero amplitude or duration in order to prevent missing pulses and to preserve a constant pulse rate. This is important for synchronization purposes when time division multiplexing is used.

Pulse position or pulse width measurements are usually based on the leading or trailing edge of the pulse. Usually the uncertainty of the pulse position is stated in terms of the rise and fall time of the pulse, (or time required to go from 10 to 90% of its full amplitude). An analysis of typical low-pass filters shows that the rise time, t_r, is inversely proportional to the bandwidth. As a result, the bandwidths required for PDM and PPM are much greater than for PCM, where only the presence or absence of a pulse is of interest.

For PAM systems, in which the pulse amplitudes are continuously variable, the problems of cumulative noise and distortion in the transmission system can greatly disturb the sampled waveforms. Because of the

inherent disadvantages mentioned for PDM, PPM, and PAM, PCM has become the most common digital technique in cable systems. The latter does not require excessive bandwidths, and its amplitude is of little importance except to note whether or not a pulse is present.

To produce PDM, which is often called pulse width modulation (PWM), a circuit employing a 555 timer can be used, as illustrated in Fig. 5-27. The internal block diagram of the 555 chip and the resulting waveforms are shown in Fig. 5-27(b) and (c). When Pin 2 is high (V_{cc}), Comparator 2 sets the output (Q_1) of the FF HI. This, in turn, will turn Q_T ON and move the buffer output to LO. When a clock pulse that is going nega-

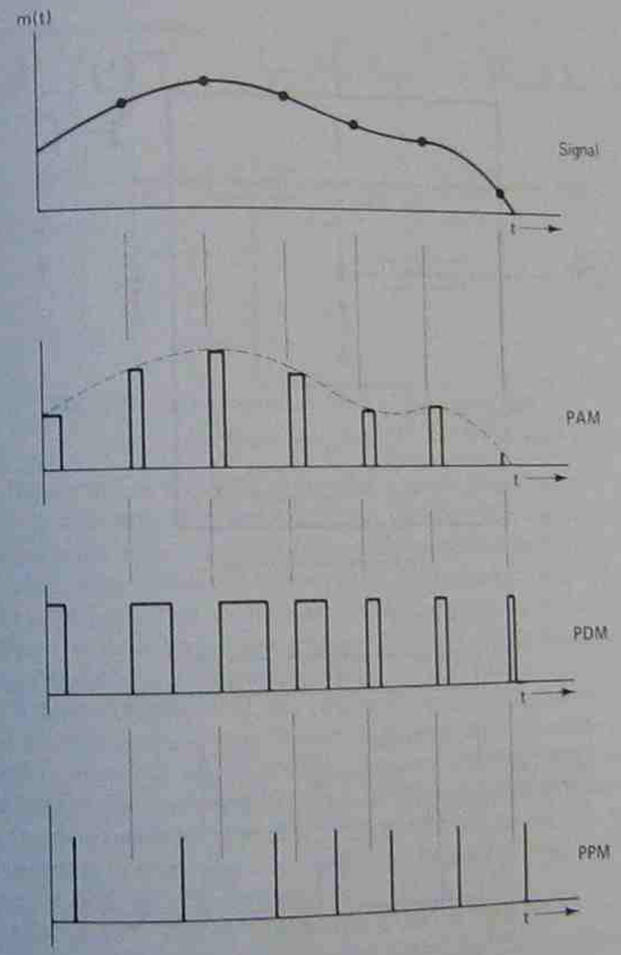


Fig. 5-26 Analog Pulse Modulation Schemes

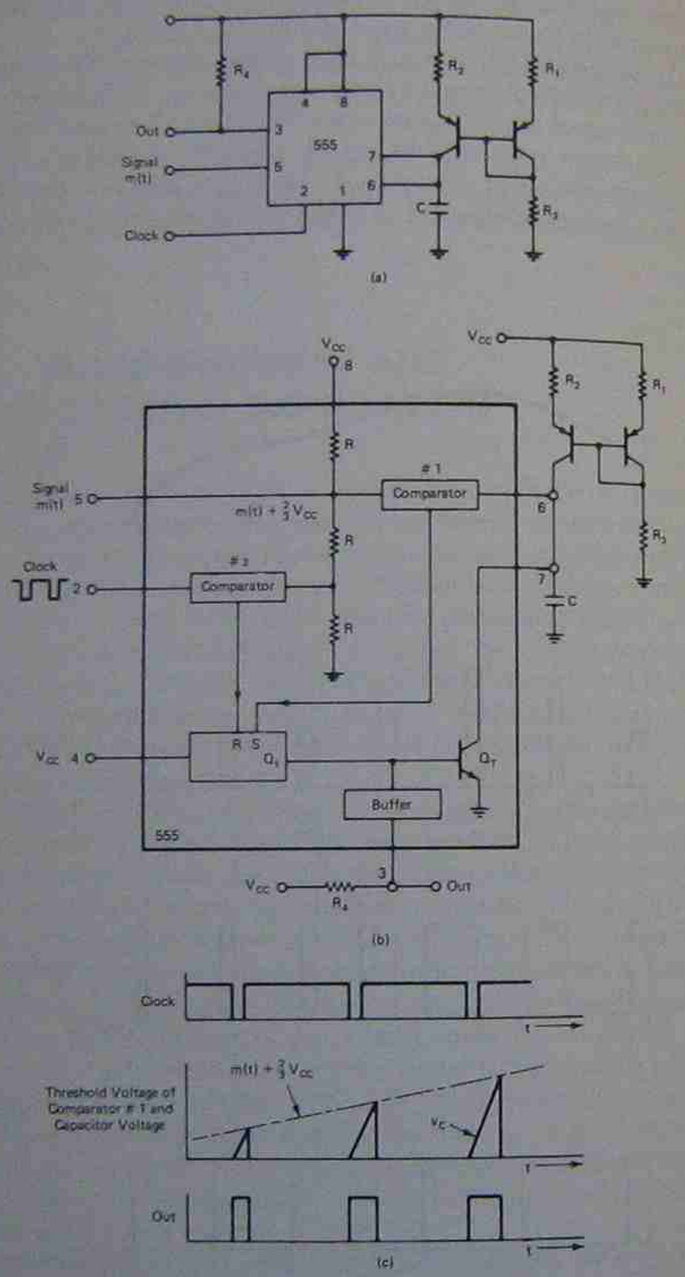


Fig. 5-27 Generation of PDM Employing a 555 Timer

tive is applied to Pin 2, the output of the FF resets to LO. Q_T is cut off and C begins to charge linearly, as illustrated in Fig. 5-27(c). The buffer output remains HI until the capacitor voltage, which appears at Pin 6, the input to Comparator 1, reaches the value of the signal voltage which is applied to the Pin 5 input of Comparator 1. This trips the output of Comparator 1 and sets the FF output HI. Buffer output goes low and the capacitor rapidly discharges through Q_T. This cycle repeats upon the application of the next pulse.

One method of transmitting PDM more efficiently is to transmit only the pulse signs. Figure 5-28 shows the arrangement and the waveforms. At the receiving end, the appropriate pulse can be stretched to recover the original pulse.

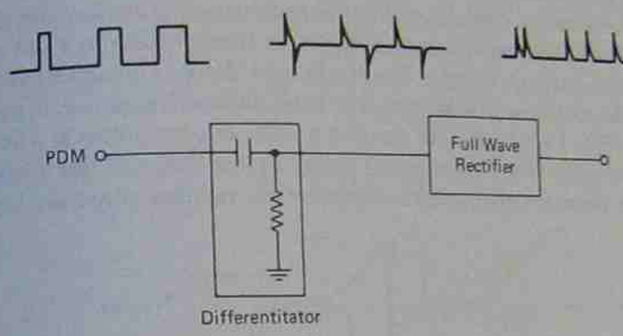


Fig. 5-28 Transmission of PDM Using Pulse Edges Only

The circuit of Fig. 5-29 shows the PDM recovery circuit.

It is assumed that the maximum duration of PDM is less than one-third of the signal period, T_s. The one shot will set the flip-flop to zero only when two successive PDM pulses pass through with separations greater than $T_s/3$. This assumes that Pulse 2 triggers the flip-flop to zero and that the next pulse that triggers the flip-flop corresponds to the leading edge of the transmitted pulse.

In demodulating PDM, the best results are obtained when it is converted to PAM and low-pass filtered. A general block diagram of a PDM or PAM is shown in Fig. 5-30. The constant current generator assures a linear voltage rise on the capacator. Note that the discharging pulses have to be located beyond the widest pulse width. The easiest way to obtain such pulses is to produce narrow pulses by triggering a multivibrator and delaying it with another monostable.

PPM is very closely allied with PDM. In fact, it can be derived directly from PDM, as is shown in Fig. 5-31. The PDM is differentiated, then rectified and shaped. Remember that in PDM it is the position of the

pulse edges that carry the information and not the pulse itself. Therefore, PPM carries exactly the same information as long as the position of the clock pulses (leading edge) is well defined in the received signal.

It is easy to see that PPM is superior to PDM for message transmission since the wide pulses of PDM require more energy than PPM when transmitted. Specifically, PPM is suited for communication in the presence of noise. Very high peak narrow pulses can be transmitted, and the pulse position can be determined even when the noise level is high. Note, however, that transmitting very narrow pulses requires a large bandwidth. When light is used as the media for transmitting analog signals, PPM or PCM are the most suitable types of modulation because the maximum power output in the modulated light source, such as LED or LASER, is achieved when it is pulsed at a very low duty cycle.

Mathematical analysis of PPM is much more cumbersome than PDM. The easiest way to demodulate is to convert PPM to PDM and follow with low-pass filtering. The PDM could also be converted to PAM and followed by low-pass filtering. The latter method is superior. In transmitting PPM, it is necessary to transmit a series of sync pulses at a much lower repetition rate than sampling pulses so that they will not interfere with the original signal and/or minimize the number of pulses trans-

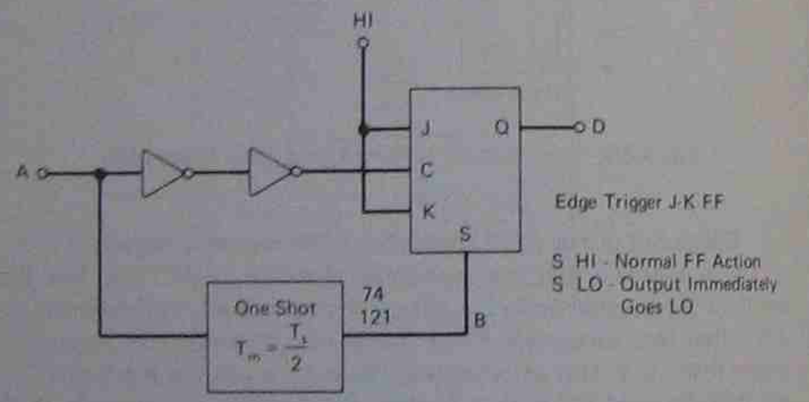


Fig. 5-29 PDM Recover Circuit

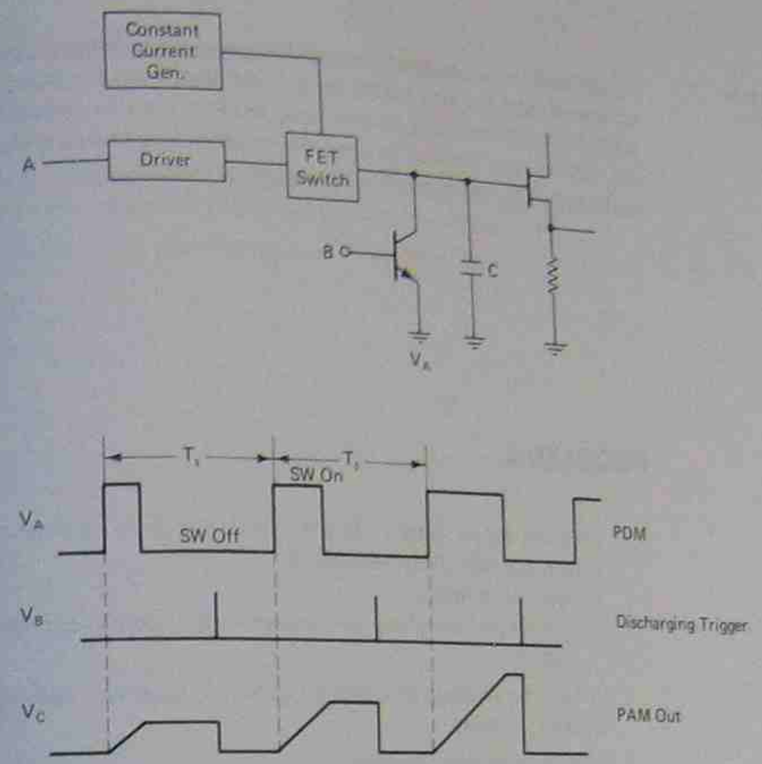


Fig. 5-30 Conversion of PDM to PAM

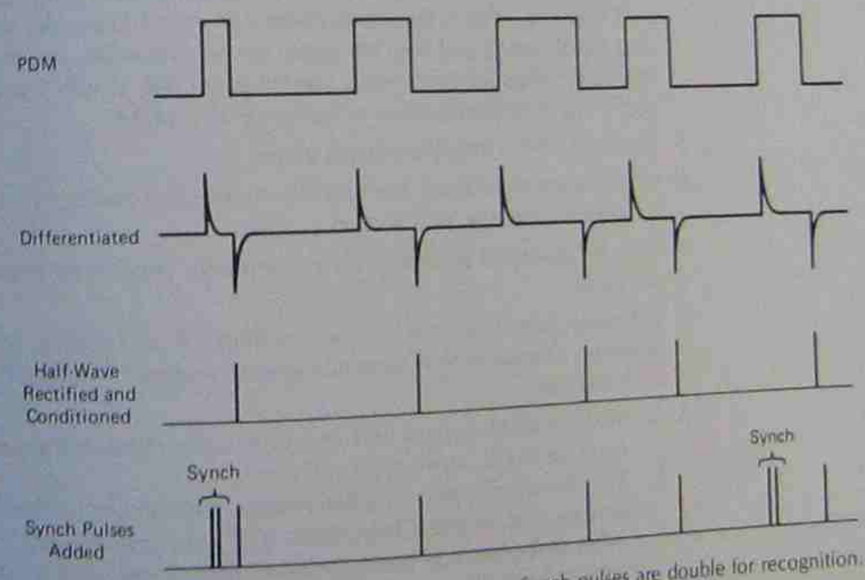


Fig. 5-31 Generation of PPM from PDM Note: Synch pulses are double for recognition.

mitted in order to conserve transmission power. Sampling pulses can be recovered from the sync pulses by a phase lock loop frequency synthesizer, or by synchronizing an oscillator. In PPM, a series of double pulses are used for synchronization since the synch pulses must be narrow and at the time distinguishable from the signal pulses.

PROBLEMS

1. With the aid of graphs, describe both the time domain and frequency domain characteristics of:
   a. natural sampling
   b. flat-topped sampling (also known as the sample-and-hold technique)
2. What is the aliasing, or foldover distortion? Give two techniques for reducing aliasing.
3. State the Sampling Theorem.
4. In a PCM telephony circuit, the voice signal is filtered with a 4-pole filter with a break frequency of 3300 Hz. If the signal is sampled at an 8 kHz rate, what is the approximate S/N ratio? Draw the sampled signal spectra and note the signal power and noise power by appropriate shading. Since voice spectra is not flat, would you expect a higher S/N ratio on an actual system? Explain.
5. Derive Eq. (*5-11*) from the integral given.
6. What is aperture time and aperture distortion? Give two techniques for reducing this type of distortion.
7. Why is flat-topped sampling more commonly used than natural sampling?
8. If the error due to aperture is to be less than 1% at 3.3 kHz, what maximum aperture time is permitted when the signal is sampled at an 8 kHz rate?
9. a. With the aid of a typical filter response curve, indicate the passband, stopband, and transition regions.
   b. With the aid of a gain plot, describe the differences in characteristics of an elliptic, Chebyschev, and Butterworth filter.
   c. What affect does non-linear envelope delay have on a rectangular pulse train?

10. With the aid of diagrams, define the following:
    a. TDM
    b. PDM
    c. PPM
11. Design a two-channel sampling system which employs N-channel FET's rather than the P-channel FET's shown in Fig. 5-23.
12. List some of the disadvantages of PDM, PPM, and PAM.

6 Digital Modulation

6–1 INTRODUCTION

Up to this point in the text, either the amplitude or the timing of the transmitted pulse modulated waveform was made to vary continuously with the amplitude of the modulating signal. These types of modulation systems are analog in nature. This chapter will consider two of the most common types of systems in which the encoded signal consists of binary digits. The two systems are Delta Modulation (DM) and Pulse Code Modulation (PCM).

6–2 SIMPLE DELTA MODULATION

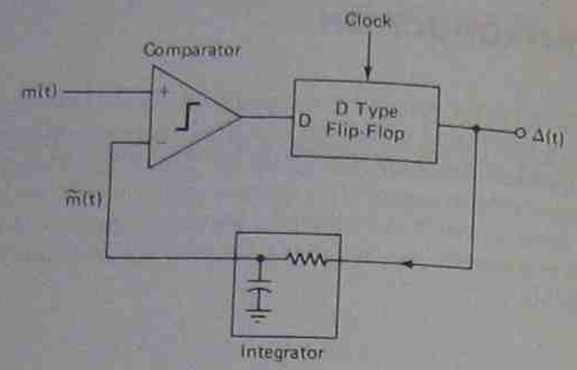
The less frequently used of the two systems is Delta Modulation (DM). Currently, it is being applied to satellite communications, television, and subscriber telephone loops. It is less complex and therefore less costly than PCM, more tolerant to transmission errors, and does not need the synchronization requirements of PCM. On the other hand, it is sensitive to slope overload (which will be described in more detail later) and is unsuitable for time sharing as an encoder/decoder among multiple channels.

The delta modulator transmits binary output pulses whose polarity depends upon the difference between the modulating signal and the feedback signal corresponding to the history of the signals previously sent. In the simple Delta Modulator of Fig. 6-1, the integrated feedback $\tilde{m}(t)$ is compared to the input signal $m(t)$.

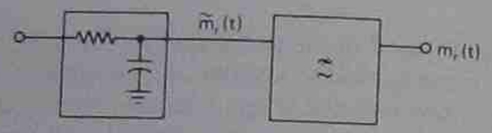
If $m(t) > \tilde{m}(t)$, a positive pulse forms the output, $\Delta(t)$, and if $m(t) < \tilde{m}(t)$, a negative pulse forms the transmitted signal. This output pulse train also forms the approximation $\tilde{m}(t)$ after being integrated.

The following waveform sketches will clarify the operation of this modulator. As illustrated in Fig. 6-2, a series of trangular waveforms results for $\tilde{m}(t)$ when there is no signal applied to the Delta Modulator. This is because integration of a constant gives a ramp. (See left-hand side of Fig. 6-2.) In this idling condition, the transmitted signal consists of alternate positive and negative pulses. The difference between the original signal, $m(t)$, and the reconstructed signal, $\tilde{m}(t)$, results in an error signal, which is often called granular or quantization noise. This noise can be decreased by either decreasing the magnitude of the step size "a" or by increasing the sampling frequency.

If $m(t)$ exceeds the feedback signal, $\tilde{m}(t)$, the transmitted wave train remains positive. When $\tilde{m}(t)$ begins to overshoot $m(t)$, the encoder output begins transmitting a negative pulse. If the modulating signal changes more rapidly than the encoder can follow, slope overload occurs. This



(a) Encoder



(b) Decoder

Fig. 6-1 Delta Modulator

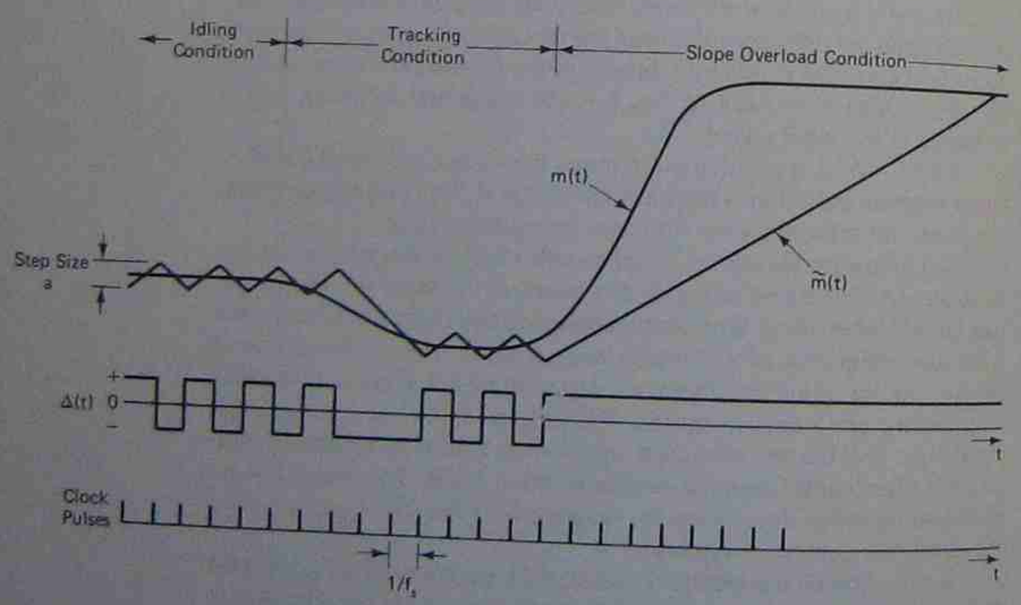


Fig. 6-2 Delta Modulation Waveforms Illustrating Idling, Tracking, and Slope Overload

slope overload, shown towards the right of Fig. 6-2, causes waveform distortion and is one of the severe limitations of Delta Modulation. The integrator is unable to closely track large amplitude, high frequency signals.

When in slope overload, the difference between $\tilde{m}(t)$ and $m(t)$ is called slope overload noise, a noise which greatly exceeds quantization noise. As a function of input signal power, the S/N ratio initially increases with signal power until slope overload is reached, as illustrated in Fig. 6-3. This increase is due to the signal power term increasing while the N-term, chiefly quantization noise, remains fixed. Upon a further increase in signal power, slope overload occurs and the S/N ratio suffers.

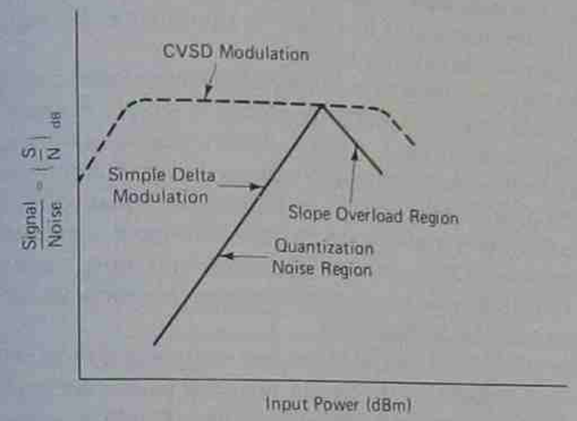


Fig. 6-3 Signal-to-Noise for Simple and Adaptive Delta Modulation

Given a modulating signal of $m(t) = A \sin 2\pi f_m t$, the maximum rate of rise would be $dm(t)/dt|_{max} = 2\pi f_m A$. For a step size "a", as shown in Fig. 6-2, and for a sampling rate of f_s, the rate of rise of the reconstructed signal $\tilde{m}(t)$ is af_s. Equating these to determine the maximum amplitude-modulating frequency product allowable produces

$$2\pi f_m A = af_s$$

$$\text{or} \qquad Af_m = af_s/2\pi \tag{6-1}$$

Thus, given a maximum of f_m, a maximum amplitude can be determined. As noted in Eq. (6-1), slope overload can be delayed by increasing the sampling frequency f_s or increasing the step size "a". This, of course, either forces an increase in the bandwidth or increases the quantization noise.

6–3 COMPANDED DELTA MODULATION

Studying the S/N response of the simple Delta Modulation Scheme shown in Fig. 6-3 shows that the S/N ratio varies considerably with the signal power. At low signal power, the S/N ratio is much worse than at high signal powers as long as it remains within the quantization noise region.

In order to maintain a more constant S/N ratio, the step size is made adaptable to the input signal amplitude. For small input signals, $m(t)$, the step size is kept small but increased as the input signal is increased. This makes the noise power vary with signal power, maintaining a constant S/N ratio over a fairly large range of input power levels. In addition, slope overload is less likely to occur since the step sizes are the largest when nearing this region.

For telephony, the type of adaptive delta modulation used is Continuous Variable Slope Delta Modulation (CVSD). A block diagram of such a modulator is shown in Fig. 6-4. The squaring circuit causes the gain of the feedback amplifier to increase with increase in signal amplitude, regardless of the polarity of the voltage. In addition, by squaring the quantization noise, power is made to vary linearly with input signal variations.

Single-chip CVSD encoders/decoders are now available. The Harris HC-55516/55532 and the Consumer Microcircuit of America FX209 are typical examples.

One chief advantage CVSD has over companded PCM is that it can handle voice at about half the bit rate. For good voice intelligibility, companded PCM requires about 64k bit/s whereas CVSD requires only 32k b/s for similar quality.

On the negative side, unlike PCM, CVSD is not suitable for time-sharing among multiple channels. Also, at 32k b/s rate, CVSD cannot handle tone and phase encoded modem transmission as well.

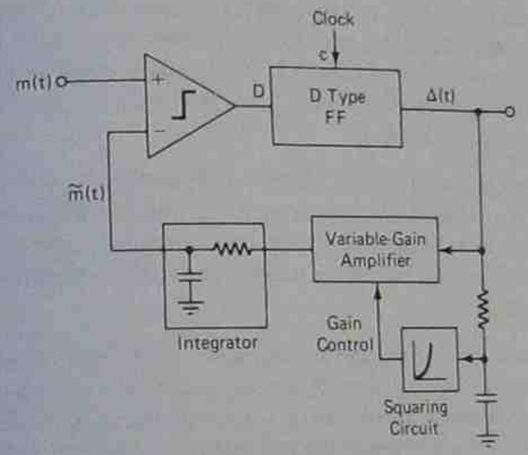
6–4 PULSE CODE MODULATION

The most common form of digital transmission is Pulse Code Modulation (PCM). PCM performs the three functions of

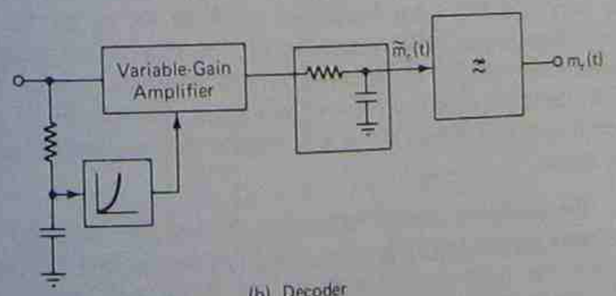
1. Sampling the analog signal,
2. Quantizing the sampled amplitudes, and
3. Encoding the quantized sample into a digital signal.

The principle area of PCM application has been in linking relatively close terminals or nearby local exchanges by wire or cable. In time, PCM will be employed in very large networks spanning great distances.

Consider a simplified version of a 24-channel PCM system that is generally used throughout North America. Since it often displaces the old FDM on VF cable, it is necessary to take certain precautions when the transfer is made. Voice usage on a cable requires consistent characteristics out to about 4 kHz whereas PCM requires consistent performance out to a much higher frequency of 2.5 MHz. As a result, the alterations or precautions described in Sec. 4–4 must be taken in order to prevent the cable from behaving as a low-pass filter with a cut-off frequency at around 4 kHz. That is, all loading coils, built-out networks, and taped circuits must be removed. In addition, there should be no split pairs (where a tip or ring of one cable is spliced to a tip or ring of another pair), high resistance splices, grounds, or moisture in the cable. These can all cause severe signal impairments at the higher PCM frequencies.



(a) Encoder



(b) Decoder

Fig. 6-4 Variable Slope Delta Modulator

In order to adequately understand the concept of sampling and to obtain the bandwidth required by a PCM signal, review the Sampling Theorem which states: For a signal bandwidth limited to B, all the information in this signal will be retained by the sample if the signal is sampled at regular intervals of time and at a rate higher than 2B. For a 4-kHz audio signal, for example, the sampling rate must be at least $2 \times 4000 = 8000$ samples/s, or one sample must be taken every 125 μs ($\frac{1}{8000}$ s) for no information to be lost. The 125 μs interval between samples is taken up by 24 channels, signaling pulses, and sync pulses. Because the samples are short compared to the time intervals between them, many different channels can be interleaved and transmitted in one sampling period. This process is known as time division multiplexing (TDM).

Figure 6-5 shows a basic block diagram of a PCM speech terminal. When transmitting, the audio signal is first bandlimited to minimize aliasing or foldover distortion and then sampled to form a PAM signal. At this point, the signals from all channels go to common equipment where the coding and additions of sync are performed. Since each individual channel requires very little equipment and the more exotic switching and coding is done in common circuitry, PCM terminals can be inexpensive for a large number of channels.

With the declining cost of Analog-to-Digitial (A/D) converters, the future trend is to have an encoder on each channel and then combine the 24 channels with a digital multiplexer. In this manner, processing, such as reducing redundancy before it is multiplexed, can be performed on the sampled signal.

After sampling, the voltage amplitude is "quantized" or is forced to take on certain discrete voltages. This new signal forms an approximation to the old one. Because the two forms are not identical, an inherent error occurs. The difference between the old and new signals is called quantizing error, and is referred to as quantizing noise when heard or measured. The error signal has frequency components extending from above the voice band. Because of aliasing, these higher frequency components fold back into the voice band and result in an error signal at the receive-filter output. In PCM transmission, this quantizing noise is one of the dominant sources of impairment, and it remains significant even when other noise contributions are minimized. The transfer function of a linear quantizer is shown in Fig. 6-6 along with the quantizing of a sampled signal.

The quantizing noise is reduced as the spacing between the quantized steps are reduced or, in other words, as the number of levels are increased. For voice communication, 256 levels are commonly used, which can be represented by an eight-digit code. In general, for an n-digit

code yielding 2^n equally spaced steps, the peak S/N ratio in decibels is given by:

$$\frac{S}{N} = 10.8 + 6n \text{ dB} \qquad (6\text{-}2)$$

which for 8 digits gives a $\frac{S}{N}$ ratio of 58.8 dB

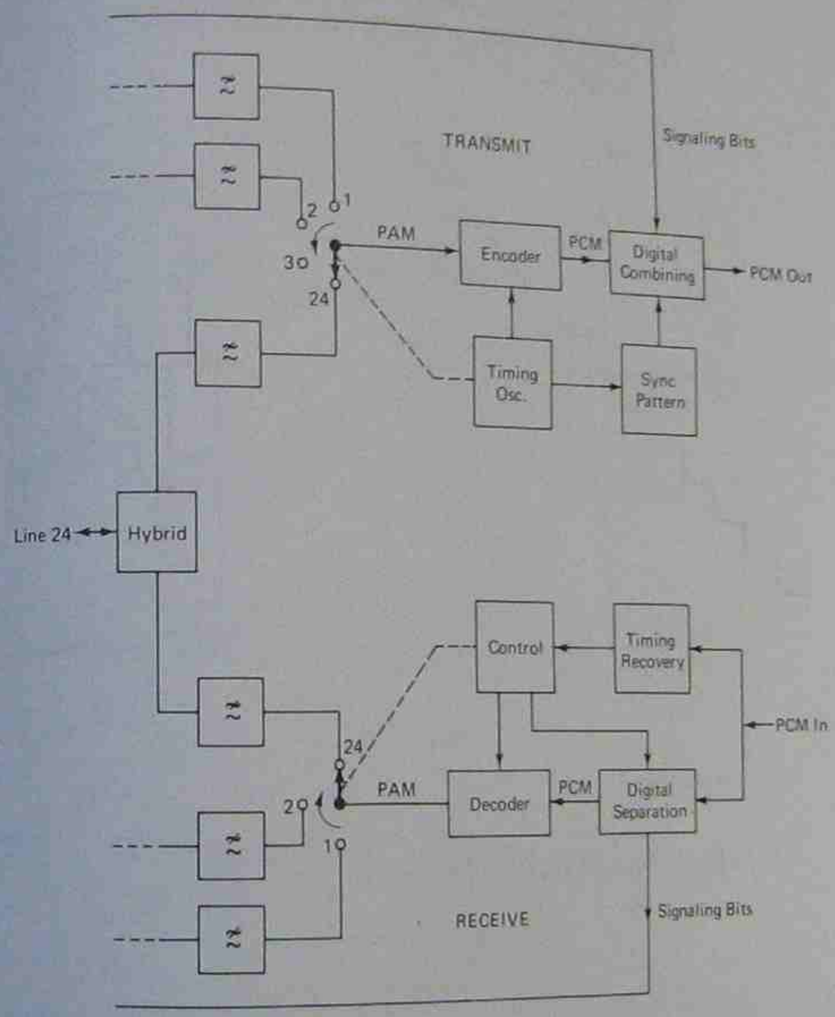


Fig. 6-5 Block Diagram of PCM Terminal

Equation (6-2) assumes that the maximum amplitude signal is obtained just before clipping occurs. In practice, low level signals are more probable than high level ones. Hence, the real S/N ratio can be much smaller than that shown in Eq. (6-2) since the quantizer noise is essentially independent of signal level for uniform step sizes. To obtain a constant S/N ratio for all signal levels, non-linear or piece-wise linear encoders are employed. This will be discussed in Sec. 6–7.

Once the level of quantization distortion is selected, immunity from channel noise and distortion can be obtained. As long as signal distortion is kept within certain bounds, identification of the proper signal level can be maintained. After amplitude quantization, the sample is converted

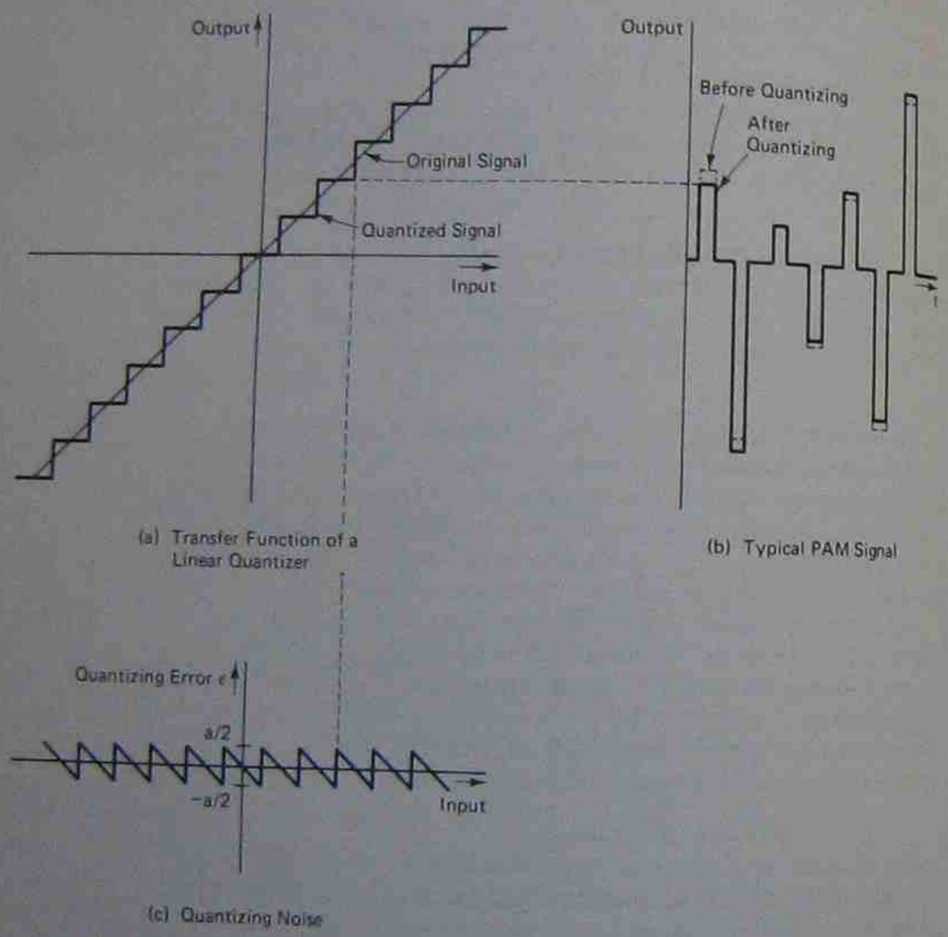


Fig. 6-6 Quantizing Process

into a code which consists of either seven or eight binary bits for the most common PCM system.

The signal is now discrete in both amplitude and time. If eight bits are transmitted for each sample as shown in Fig. 6-7, a fresh train of elements equivalent to the old ones can be regenerated with very little impairment. Although the signal is subject to noise, distortion, and bandwidth limitation, if these flaws are small enough, the regenerated signal will correspond to the original one. Thus, with suitable spacing, a series of regenerative repeaters will pass along a discrete signal without any accumulation or impairment along the route.

At the sampling instant, the regenerator compares the signal with the threshold decision level, or the level midway between the two ex-

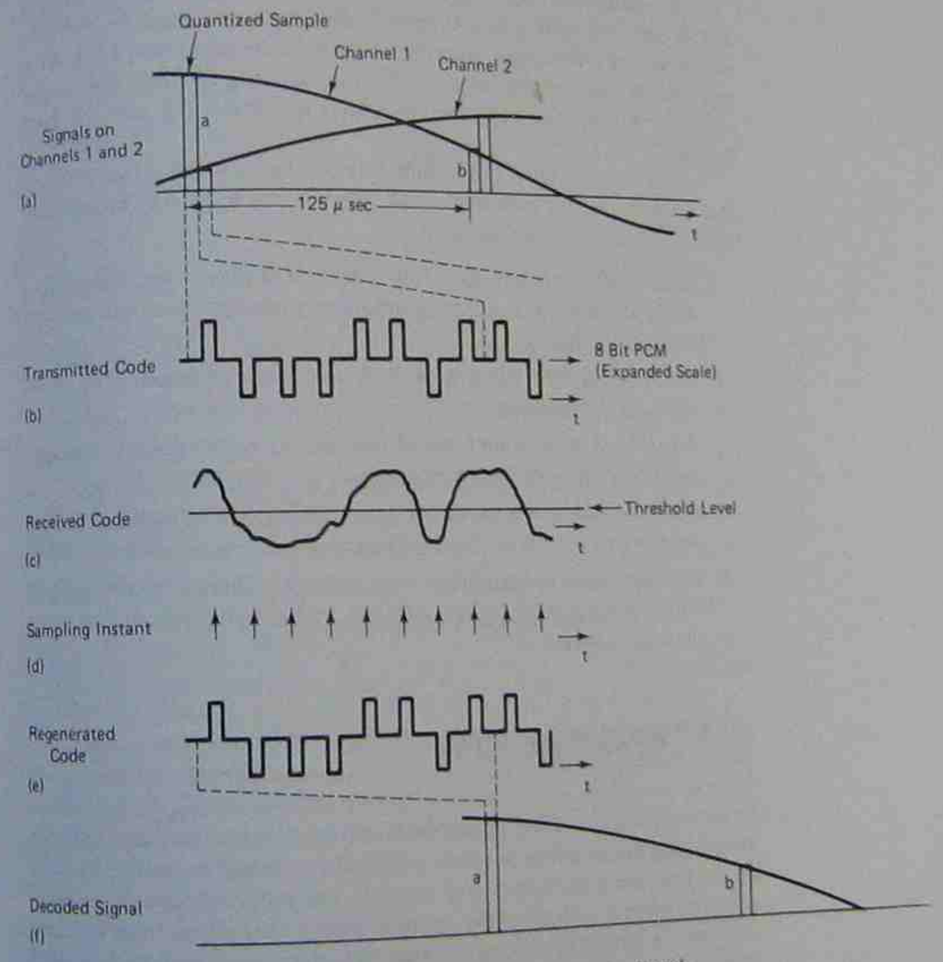


Fig. 6-7 Regeneration of Discrete Signals

pected signal levels. If positive, a positive pulse is generated; if negative, a negative pulse is generated. As can be seen from Fig. 6-7(c), fairly large impairments can be encountered before the signal is lost. This is one of the chief advantages that digital transmission has over analog transmission.

PCM has great immunity to interference and noise. This noise immunity is not gained without compromise, but at the expense of increased bandwidth. To transmit the pulse stream of 8 bits/sample shown in Fig. 6-7(e), $8 \times 2 \times 4000 = 64$k b/s must be transmitted for a 4 kHz bandwidth limited signal which represents a minimum baseband bandwidth of 32 kHz/s. Thus, practical PCM systems require 32k/4k, or 8 times the bandwidth of their analog counterparts.

In addition to speech signal information, signaling information such as ON and OFF hook and dialing must also be transmitted. In addition, synchronizing information (framing code) must be generated so that the receiver can sort out the code words and signaling bits into the proper channels. Alarm code can also be added to inform the user of local trouble.

In addition to PCM's immunity to noise, which gives transmission characteristics almost independent of distance, the following factors also favor this mode of modulation:

a. It can easily carry a mixture of traffic such as telephony, telegraphy, data, and encoded video information, provided the medium has sufficient capacity.
b. It can increase the capacity of single telephone channels over cable pairs by multiplexing.
c. It can lend itself to such novel facilities as cryptography, storage, and other forms of digital processing.
d. It is more suitable for the newer types of transmission media such as light beams in optical fibers and multiple access satellites.
e. Its signal characteristics allow easy access to electronic switching in which groups of digits are selected to be switched in turn onto various highways.

6–5 APERTURE TIME

When sampling an analog signal, the signal usually is changing in amplitude during the sampling process. In order to maintain a maximum limit on the amplitude uncertainty, or error, in the measurement of a signal which is varying with time, the aperture time or sampling time window must be of a limited time duration. This maximum aperture time is related to the rate of signal amplitude change.

For an A/D converter, the aperture time is the conversion time. For a sample-and-hold circuit, it is the signal averaging time during the sample-to-hold transition. A sample-and-hold circuit placed before an A/D converter effectively reduces the aperture times since the sample-and-hold can take a very fast sample of the analog signal and then hold the value while the A/D operation is performed.

As shown in Fig. 6-8, the maximum amplitude uncertainty occurs when the slope of the signal is a maximum. If the maximum amplitude error is "a", it will be related to the maximum rate of change with time of the input signal by the relation:

$$a = \tau \left.\frac{dv(t)}{dt}\right|_{max} \tag{6-3}$$

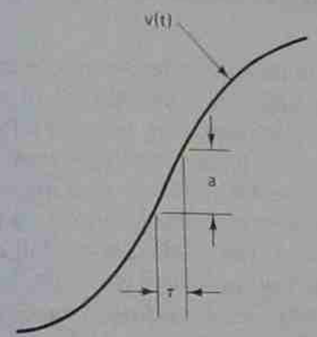


Fig. 6-8 Illustration of Aperture Time (τ) and Amplitude Uncertainty (a)

Consider the case of a sinusoidal signal, $A \sin \omega t$. The maximum amplitude uncertainty will occur at the zero crossing of the waveform, that is,

$$\begin{aligned} a &= \tau \frac{d}{dt}(A \sin \omega t)|_{t=0} = \tau A\omega \cos \omega t|_{t=0} \\ &= \tau A\omega \end{aligned} \tag{6-4}$$

Let the step size of the encoded signal be equal to "a", since this is the size of the worst amplitude uncertainty. If the encoder has n bits of resolution, it can distinguish between 2^n levels. The maximum peak to peak voltage that can thus be obtained is a $(2^n - 1)$ volts, where the -1 takes care of the beginning level which has no amplitude, therefore for a peak amplitude A,

$$A = \frac{a(2^n - 1)}{2} \tag{6-5}$$

Substituting (6-5) into (6-4) we obtain

$$\tau = \frac{1}{(2^n - 1)\,\pi f} \tag{6-6}$$

Equation (6-6) is graphically plotted in Fig. 6-9. Note that for a 4 kHz sinusoid with 8 bits of resolution, the maximum aperture time is 0.312 μs. Within this time, the A/D converter must encode the analog signal into 8 binary digits. For several channels sharing the same A/D converter, this time must be even further reduced.

6–6 AMPLITUDE SAMPLING

Because the encoder in the transmitter section of the PCM system converts continuous amplitude into a discrete set of amplitude levels, an error waveform known as quantizing noise results. Quantizing noise exists only when a signal is present, and appears much as white noise. Even when no signal is present, some idle channel noise will still exist at the encoder output because of residual noise such as thermal and shot noise at the encoder input. This residual noise will occasionally exceed the thresholds of the first few encoding stages at the low voltage levels, resulting in random output pulses. Excessively large applied signals should also be avoided since they will be clipped when they exceed the maximum threshold level and will introduce large amounts of distortion.

By maintaining uniform encoder step sizes, the quantizing noise would essentially be kept constant for all signal levels. The maximum possible error is the same for all sample amplitudes for equally spaced or linear quanta. This is fine for strong signals, but it results in very poor signal-to-noise ratios for the weaker signals.

In order to improve the S/N ratios for the weak signals, which are much more probable to occur in the case of speech signals, non-uniform quantization can be used. Rather than maintaining the same maximum possible error for all sample magnitudes, the spacing of decision levels, or step sizes, is reduced for the smaller sample magnitudes but increased for the larger sample magnitudes. That is, the step sizes are made approximately proportional to the signal amplitude, resulting in a nearly constant S/N ratio over the speech volume range. Figure 6-10 illustrates a non-uniform encoder transfer characteristic which shows how quantizing noise can be increased with signal size to keep a constant signal-to-mean quantizing noise ratio.

Another technique commonly employed to maintain a more or less constant S/N ratio over the dynamic signal range is companding. The companding process is called compressing at the transmitting end and

expanding at the receiving end. In compressing, the higher amplitude signals effectively obtain a lower gain than the weaker signals. In expanding, the reverse takes place.

The compandor devices are called compressors and expandors. Using this method, the signal is instantaneously compressed, and the higher amplitude signals are compressed more than the lower amplitude signals. The companded signal then enters an encoder which performs uniform quantization on the signal prior to coding.

At the receiver, expansion is performed after decoding is completed. Companding can improve the S/N ratio in voice circuit by up to 25 dB. It is of little value in data communications where a signal is fully OFF or ON, at 0, or in saturation. Each condition is equally probable and thus there is no advantage to using companders. The compander, because of its nonlinear characteristics, only adds intermodulation distortion.

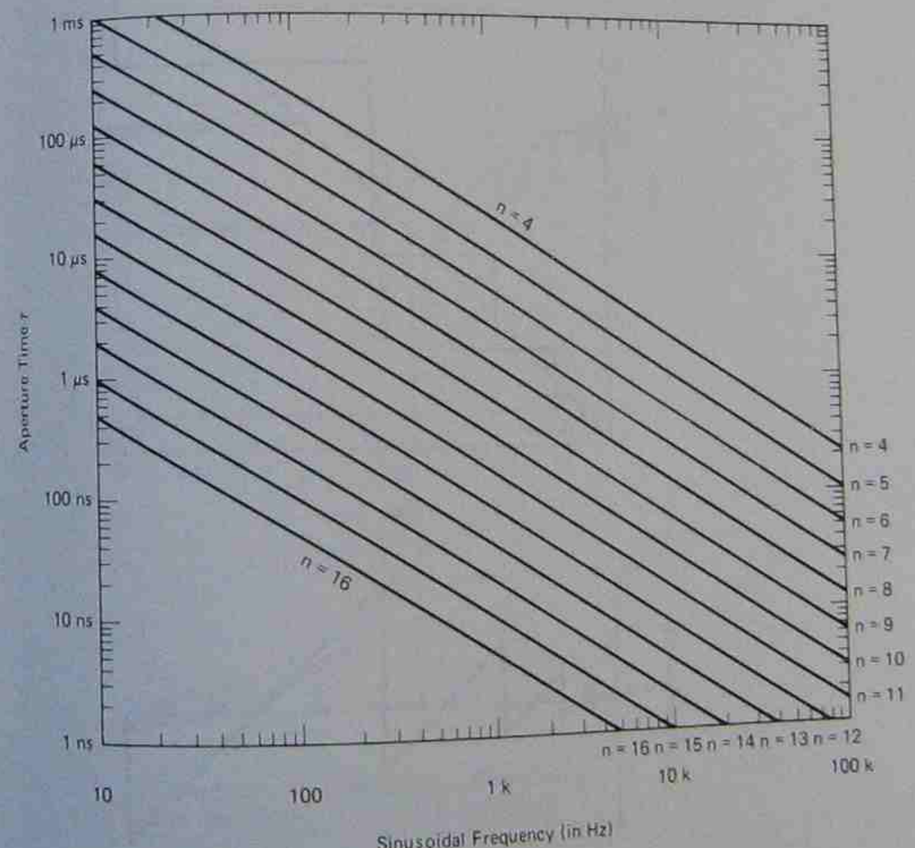


Fig. 6-9 Aperture Error as a Function of Sinusoidal Frequency

The D2 or D3 PCM-channel banks commonly in use throughout North America employ a compressor with a logarithmic gain characteristic. It follows an encoding law known as the mu law, Eq. (6-7), where the value $\mu = 255$

$$Y = \frac{V \ln(1 + \mu v/V)}{\ln(1 + \mu)}, \quad 0 \le |v| \le V \tag{6-7}$$

v = input voltage
V = maximum input voltage = maximum output compressed voltage

The dotted curve in Fig. 6-11 illustrates the companding effect of the mu law.

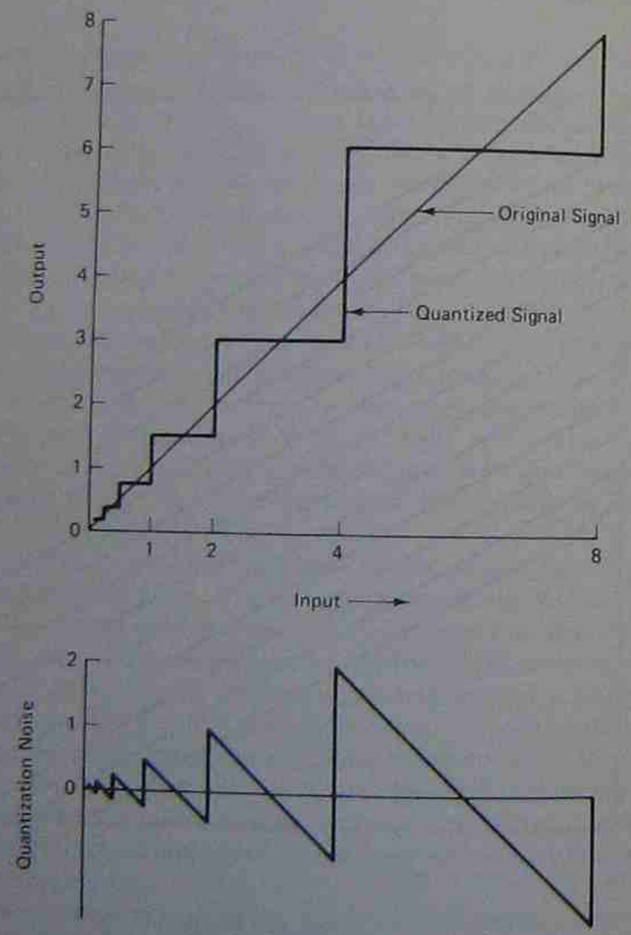


Fig. 6-10 Non-Linear Quantizing

In Europe, the CCITT's "A law" on the companding curve is followed. It is given by the expression

$$Y = \begin{cases} \dfrac{Av/V}{1 + \ln A}, & |v| \le \dfrac{V}{A} \\ \dfrac{1 + \ln Av/V}{1 + \ln A}, & \dfrac{V}{A} \le |v| \le V \end{cases} \tag{6-8}$$

where v = input voltage
V = maximum input voltage = maximum output compressed voltage

where $A = 87.6$ for a 13 segment approximation

Rather than follow a smooth logarithmic curve, the D2 or D3 channel banks assume a segmented approximation to the continuous function. A total of 15 linear segments, as illustrated in Fig. 6-11, make up the approximation: seven positive segments, seven negative segments, and one segment near zero, which consists of two co-linear segments on either side.

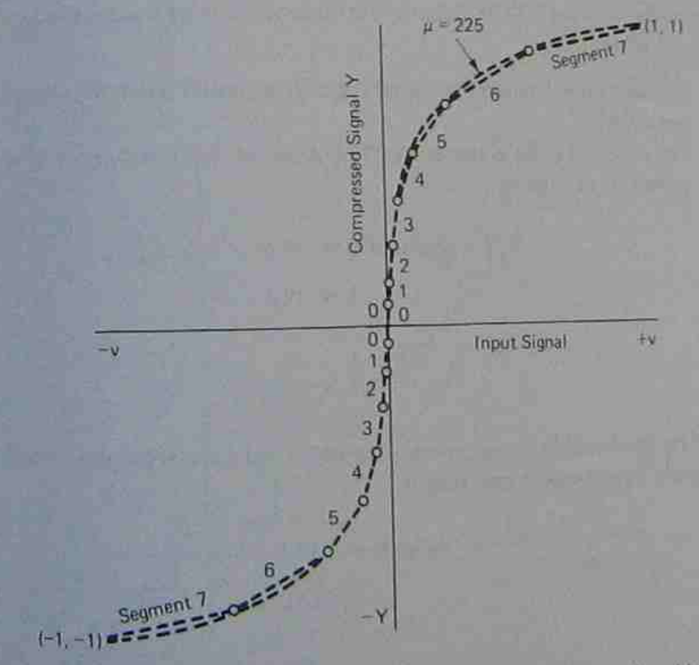


Fig. 6-11 Fifteen-Segment Compression Characteristic for the Mu Law ($\mu = 255$) Where the Maximum Voltage is Equal to 1

$$Y = \frac{\ln(1 + \mu v)}{\ln(1 + \mu)}$$

Within each segment there are 16 steps of equal size. The step sizes increase by a factor of 2 when moving from one segment to the higher adjacent one. In total, there are 255 discrete voltage levels for the D_2/D_3 type system, which require eight bits for coding.

The 8-bit code representing the signal is broken up into three groupings. As illustrated in Fig. 6-12, the first bit represents the sign (1 is positive; 0 is negative); the next three bits indicate which segment contains the sample; and the last four bits specify the level, up to 16 possible, within the segment. The various segments are counted sequentially from the lower levels towards the higher ones.

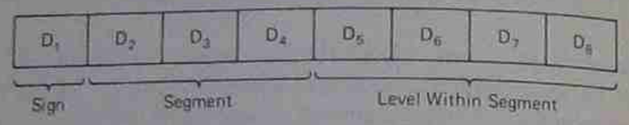


Fig. 6-12 Digit Pattern

As an example, consider the coding scheme used by the encoder found in the Lenkurt 9002B system. The code word is split up as shown in Fig. 6-12 where:

a. The sign bit is 1 when the signal is positive, and 0 when the signal is negative.
b. The segment code is represented by a sequential binary code in reversed form; that is

segment 0 = 111
1 = 110
.
.
7 = 000.

c. The level within the segment is represented by a sequented binary code in reversed form; that is,

level 0 = 1111
1 = 1110
. .
. .
15 = 0000

A pattern given by 0 111 1100 represents a negative voltage on Segment 0 at Step 3. Note that the segment, or step number, can be obtained by subtracting the actual code digits from 7(111) or 15(1111).

Thus, for a segment binary representation of 4(100), the actual segment is $7 - 4 = 3(011)$. Although this type of coding, frequently referred to as folded binary coding, may appear strange since the maximum number of 1's appear near the no-signal level, it does give a higher pulse pattern for the more probable lower speech levels where there is a danger of losing timing recovery in the receiver. If few pulses are transmitted, synchronization could well be lost.

Number the levels from 0 to +127 for the positive signals and from 0 to −127 for the negative levels, with level 0 representing the zero voltage level. An expansion of an encoded signal at low voltage levels is illustrated in Fig. 6-13. Note that 0 volts is represented by the 0 levels of both the + and −0 segment; that is, by the codes 11111111 and 01111111.

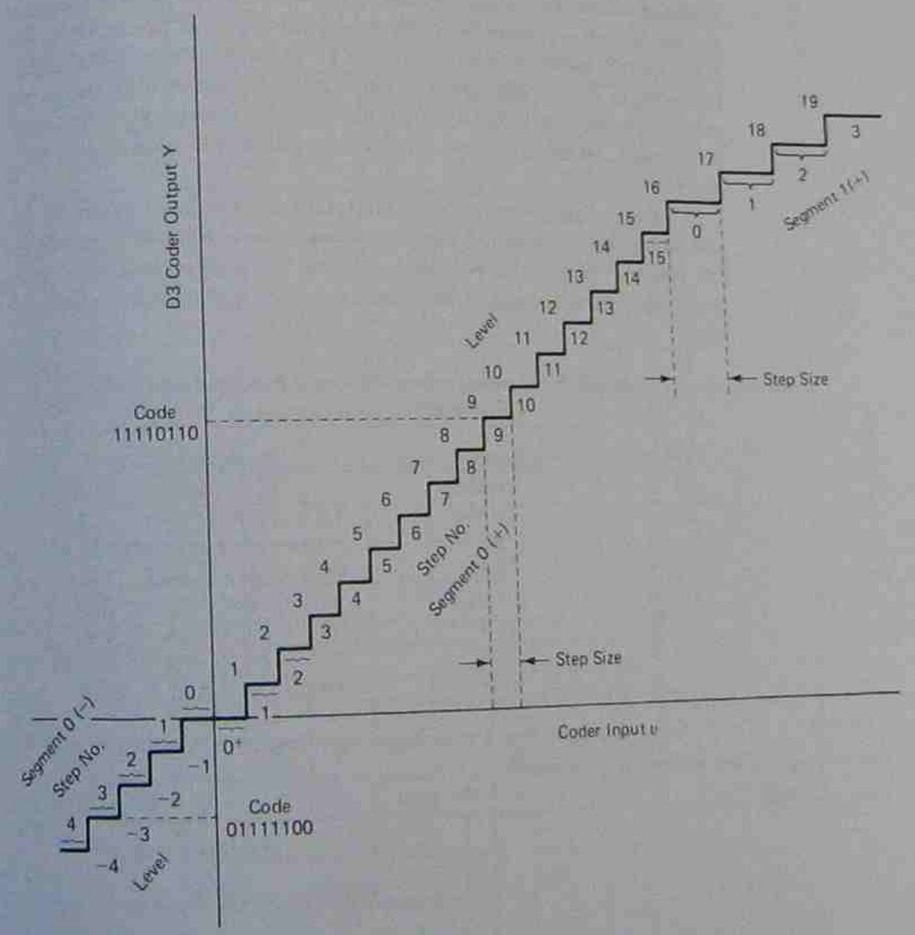


Fig. 6-13 D3 Coding Characteristic

This results in a total number of 255 permissible levels for the 8-bit code. Also note that Segment 1 has twice the step size as Segment 0.

For the positive samples, when the input signal resides between level n and $n + 1$, the code word output can be obtained by expressing $255 - n$ in binary form. For Level 9, for instance, the code number is $255 - 9 = 246$. This is represented by the code 1 111 0110.

This can also be interpreted as a positive signal (first bit is a 1) on Segment 0 (-7) at Step 9($15 - 6$). Similarly, for negative examples the output code word is given by $127 - n$, expressed in binary form. Thus, Step 3 on the 0 negative segment is represented by $127 - 3 = 124$(01111100).

If all the coding steps were to be of the same height as the smallest step for all the segments, a linear code requiring 13 bits would result. Codes in which step sizes increase by multiples of the smallest step are readily convertible from the compressed form to the linear form. The linear form is necessary if processing is to be done on the digital signal. For example, if two digital speech signals are to be added together, as in a conference call, each digital signal must first be converted to the linear counterpart, added, and then converted back to the compressed non-uniform code.

Table 6-1 gives the increment step sizes and the break points for the various segments. Table 6-2 tabulates some typical 8-bit PCM words for the Lenkurt 9002B channel bank. Figure 6-14 illustrates the corresponding positive section of the fifteen segment compression characteristic.

Table 6-1 Increment Step Size and Breakpoints for the Lenkurt 9002B Channel Bank

Segment	Step Increment (mV)	Breakpoint (mV)
0	0.475	$16 \times 0.475 =$ 7.6
1	0.963	$7.6 + 16 \times 0.963 =$ 23.0
2	1.93	$23 + 16 \times 1.93 =$ 53.9
3	3.88	$53.9 + 16 \times 3.88 =$ 116
4	7.69	$116 + 16 \times 7.69 =$ 239
5	15.4	$239 + 16 \times 15.4 =$ 486
6	30.9	$486 + 16 \times 30.9 =$ 981.0
7	63.7	$981 + 16 \times 63.7 =$ 2000

As an example, find the voltage, sign, segment, step, and level for the following code word

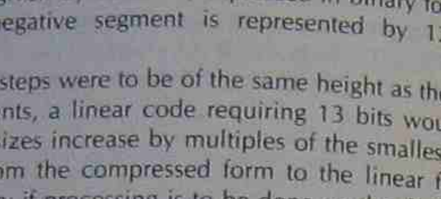
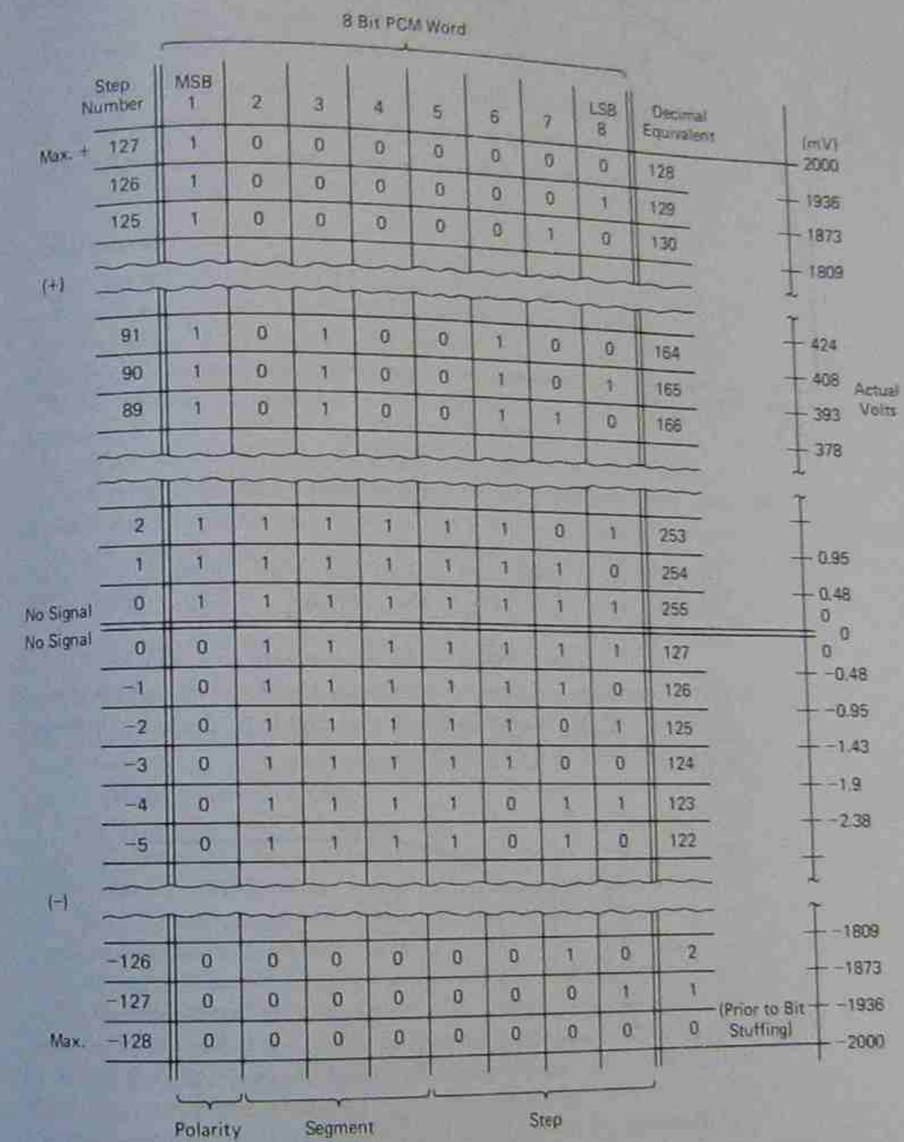


Table 6-2 Folded Binary Code



Step Number	MSB 1	2	3	4	5	6	7	LSB 8	Decimal Equivalent	(mV)
127	1	0	0	0	0	0	0	0	128	2000
126	1	0	0	0	0	0	0	1	129	1936
125	1	0	0	0	0	0	1	0	130	1873
										1809
91	1	0	1	0	0	1	0	0	164	424
90	1	0	1	0	0	1	0	1	165	408
89	1	0	1	0	0	1	1	0	166	393
										378
2	1	1	1	1	1	1	0	1	253	0.95
1	1	1	1	1	1	1	1	0	254	0.48
0	1	1	1	1	1	1	1	1	255	0
0	0	1	1	1	1	1	1	1	127	0
−1	0	1	1	1	1	1	1	0	126	−0.48
−2	0	1	1	1	1	1	0	1	125	−0.95
−3	0	1	1	1	1	1	0	0	124	−1.43
−4	0	1	1	1	1	0	1	1	123	−1.9
−5	0	1	1	1	1	0	1	0	122	−2.38
										−1809
−126	0	0	0	0	0	0	1	0	2	−1873
−127	0	0	0	0	0	0	0	1	1	−1936
−128	0	0	0	0	0	0	0	0	0	−2000

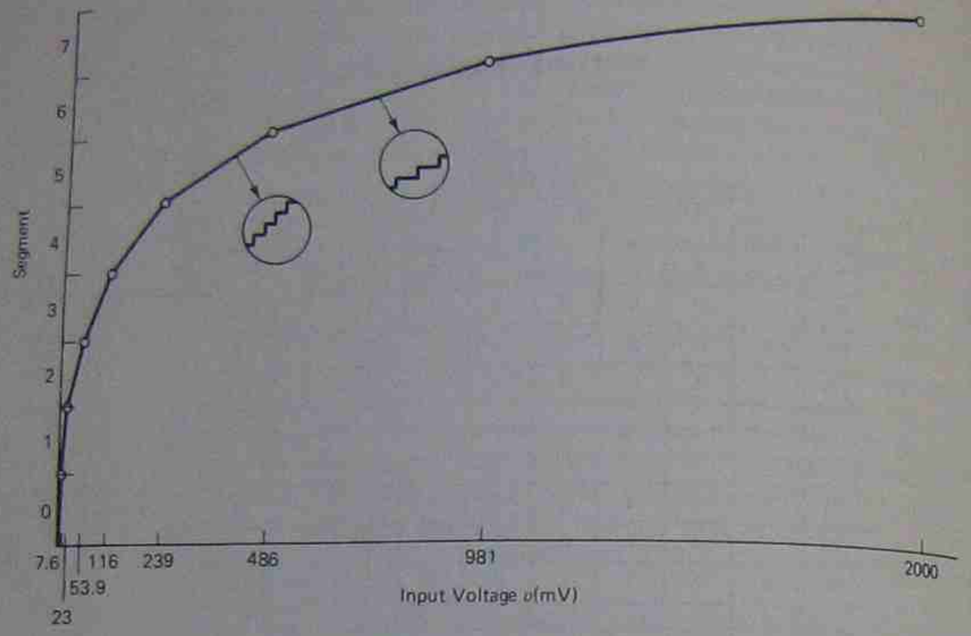


Fig. 6-14 Non-Linear Companding of Lentcurt 9002 B-Channel Bank

The voltage at Step 2 of Segment 5 from Table 6-1 is (239 + 2 × 15.4) mV = 269.8 mV. The voltage is on level 255 − 221(11011101) = 34

For 8-bit systems following the mu law with $\mu = 255$, the S/N ratio that is C-message weighted never exceeds 38 dB and remains at this level for about a 40 dB variation in signal level. Figure 6-15 shows the theoretical performance of such a terminal.

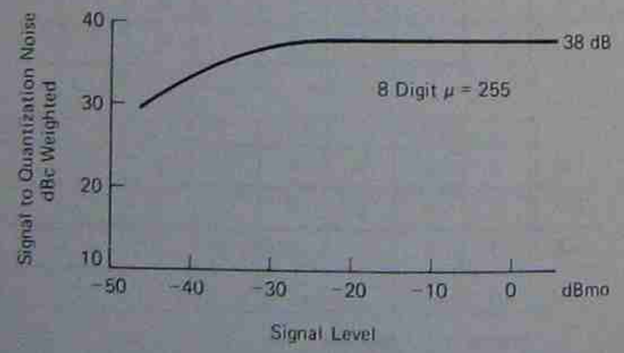


Fig. 6-15 Signal-to-Distortion Performance of an 8-Bit with $\mu = 255$ Compression Characteristic

6–7 ENCODERS AND DECODERS

After filtering, time quantizing (sampling), and amplitude quantizing the analog signal, the result is encoded into an binary signal. (See the block diagram of a PCM terminal in Fig. 6-5.) Over the years, a large array of encoders have been invented, which for the purposes of this text, can be broken down into three main categories:

1. Parallel encoders, which generate a whole character at a time
2. Serial encoders, which generate a digit at a time
3. Hybrid encoders, which generate several digits at a time but not a whole character

Parallel Encoder

The parallel encoder has the advantage of being very fast and able to readily utilize any code. Its chief drawback is that for a large number of amplitude levels, it requires a large number of comparators and cross-points. For n-bits resolution, it requires $2^n - 1$ comparators. A 4-bit converter, for instance, requires only 15 comparators, but an 8-bit converter requires 255.

A basic format of a parallel quantizer is shown in Fig. 6-16. Latch, D-type flip-flops store the binary outputs from the comparator. The comparator threshold points are spaced 1 LSB apart for the series resistance chain and voltage references. For an applied analog voltage, all comparators biased below the threshold turn ON and all those above it remain OFF. In addition, a serial register is added which is clocked at $n\ f_s$ to transfer the encoded signal to a serial train of two-level pulses.

Serial Encoder

One type of serial, or sequential encoder, uses a successive approximation technique which generates one bit at a time, with the most significant bit (MSB) determined first.

For example, take a 4-bit encoder with a range of 16. The MSB would represent one-half of the full range scale, or 8. The next MSB would represent one-fourth of the full scale range, or 4, and so on.

The successive approximation process is controlled by a programmed register as shown in Fig. 6-17.

The sample-and-hold circuit is inserted to hold the input signal constant while the encoding is in process. The analog input is initially compared with one-half of the full range scale by setting the MSB in the register to 1. If the analog input, X, is greater than this approximation, the first bit of the transmitted character is 1. This bit is also held in the programmed register and this weighting of 8 is held on the summing network.

At the beginning of the next digit period, the next weight of 4 is ap-

plied to the summing network. This causes the input to be compared to 8 + 4 or 12.

On the other hand, if the input is less than 8, the first character bit is 0. During the second digit period, the input would be compared to 4. This process is continued until the least significant bit is determined. A typical cycle, with the analog voltage setting at $10\frac{1}{2}$, is illustrated in Fig. 6-18.

Another type of serial encoder is one that scans the signal range one level at a time. The encoder shown in Fig. 6-19 employs an up/down counter which is clocked at a rate of $(2^n - 1)f_s$, where 2^n represents the number of levels for n bits. This assumes the counter resets at an extreme

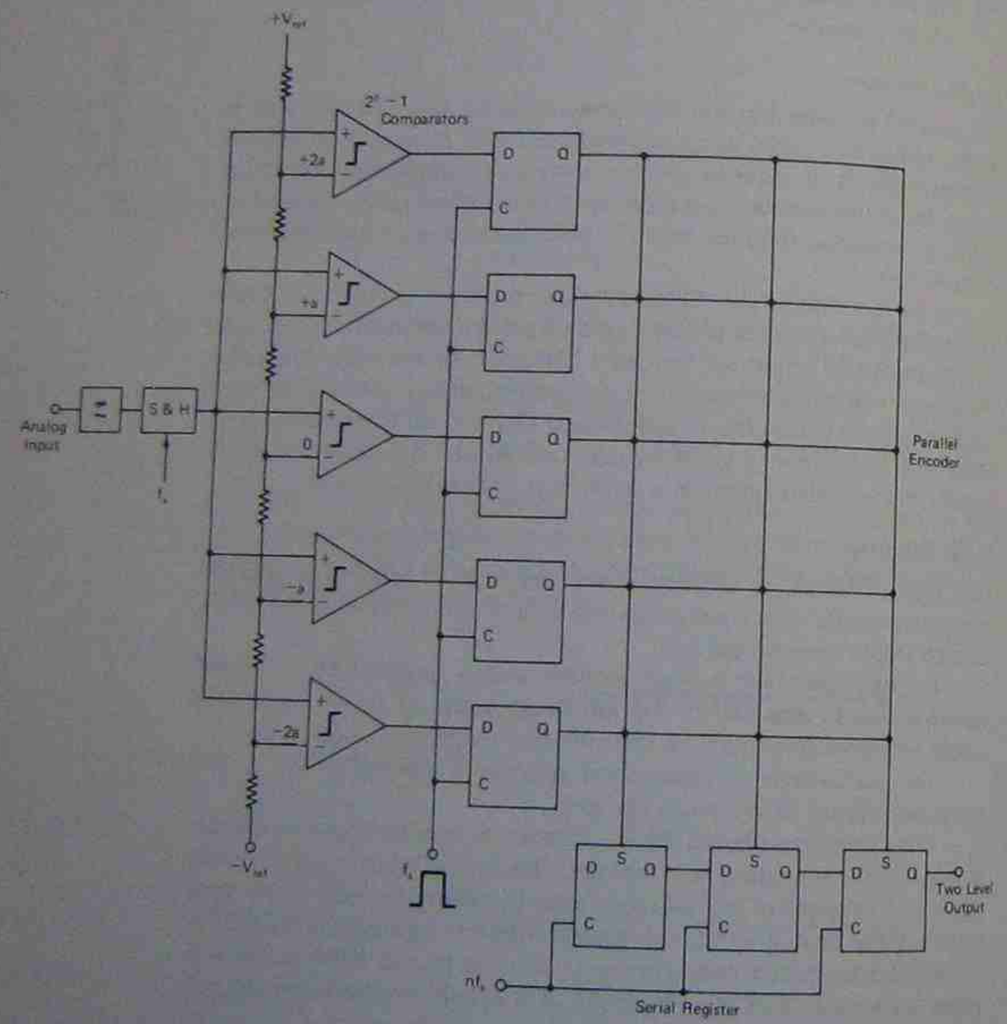


Fig. 6-16 Parallel Encoder for Arbitrary Code with Serial Register

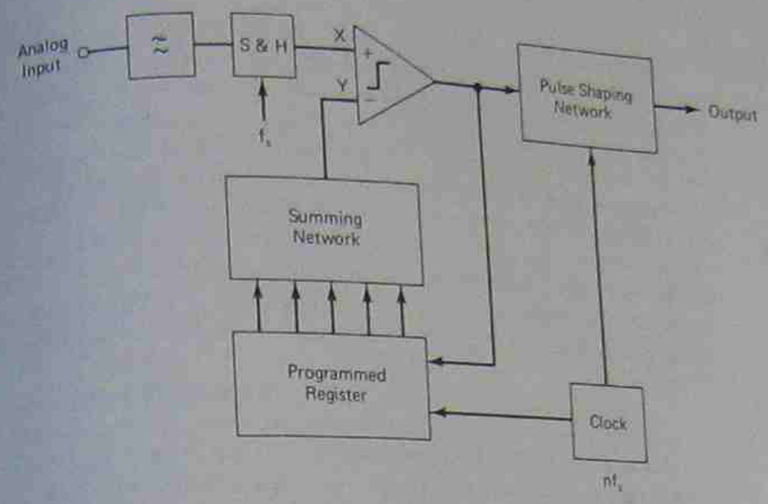


Fig. 6-17 Block Diagram of a Serial Encoder

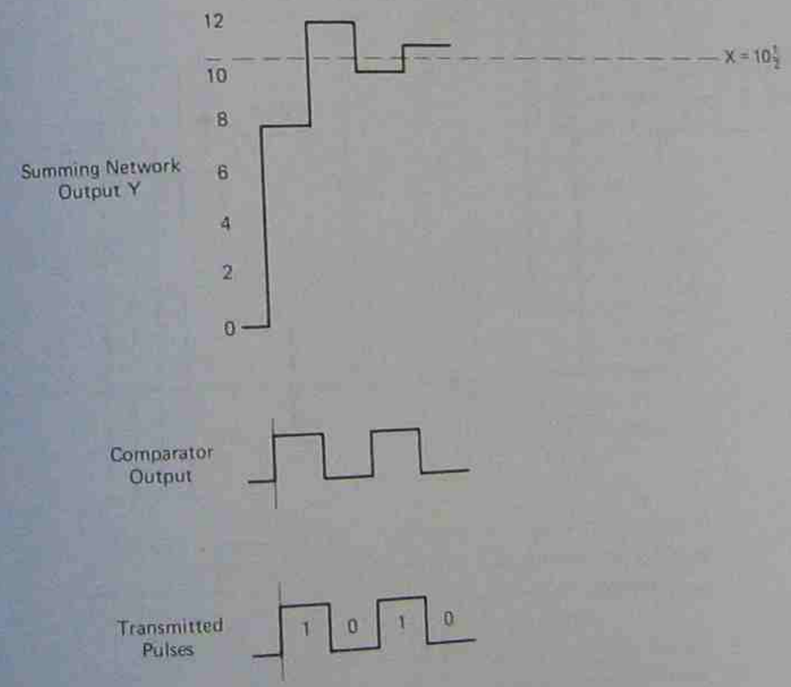


Fig. 6-18 One Operational Cycle of a Serial Encoder

positive or negative level. The clock rate could be halved if the counter was reset at the midway level at the beginning of a cycle. A 1 must be subtracted from the total number of levels, 2^n, because the extreme positive (or negative) level is one of the 2^n levels. Therefore, there are only $2^n - 1$ steps above (or below) the extreme for the A/D converter. If, for instance, an 8-bit code represents an analog signal sampled at a 8 kHz rate, the counter must operate at a $(2^8 - 1)8$k or 2.04 MHz rate. The output of the comparators, X and Y, controls the direction in which the counter must run or whether the counter output should remain stationary. If X = 1, the counter output should increase; if Y = 1, the counter output should decrease; and if both X and Y are 0, the counter output should remain fixed.

The digital output from the counter is applied to a D/A converter which in turn is applied to a difference amplifier at the encoder input. This negative feedback causes the output of the D/A converter to track the input signal and thus the counter output represents the encoded signal when a steady-state condition is achieved.

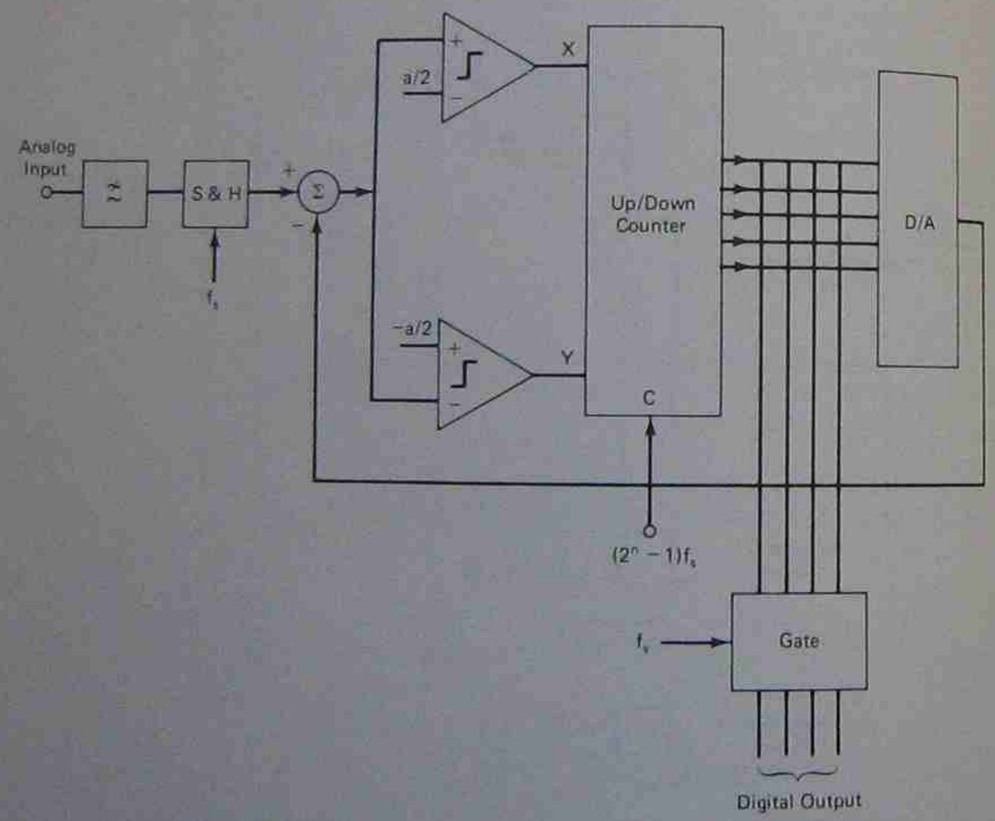


Fig. 6-19 Serial Encoder Employing a Counting Technique

As the number of distinguishable levels increases, the clock rate of the up/down counter rapidly rises. This is one of the chief disadvantages of a serial encoder. On the other hand, very few digital chips are needed, which results in a very reasonably priced encoder-decoder. Non-uniform quantizing, or companded PCM, can be generated by changing the weight for each digit of the circuit shown in Fig. 6-17. The weight of each digit is a function of the current binary decision together with the information stored about the previously determined digits. Figure 6-19 can be transformed into a non-uniform encoder by inserting a non-linear transfer device after the D/A converter and forcing the distance between adjacent levels to increase as the signal level increases. This is illustrated in Fig. 6-20, along with a corresponding decoder circuit.

Hybrid Encoder

A hybrid encoder is a combination of the parallel and serial encoders with expected accompanying compromises. It requires less equipment than a parallel encoder and can operate at lower clock frequencies than the serial encoder. One technique it uses is the subranging converter shown in Fig. 6-21 whereby k bits are determined simultaneously and the entire n bits are obtained in n/k steps.

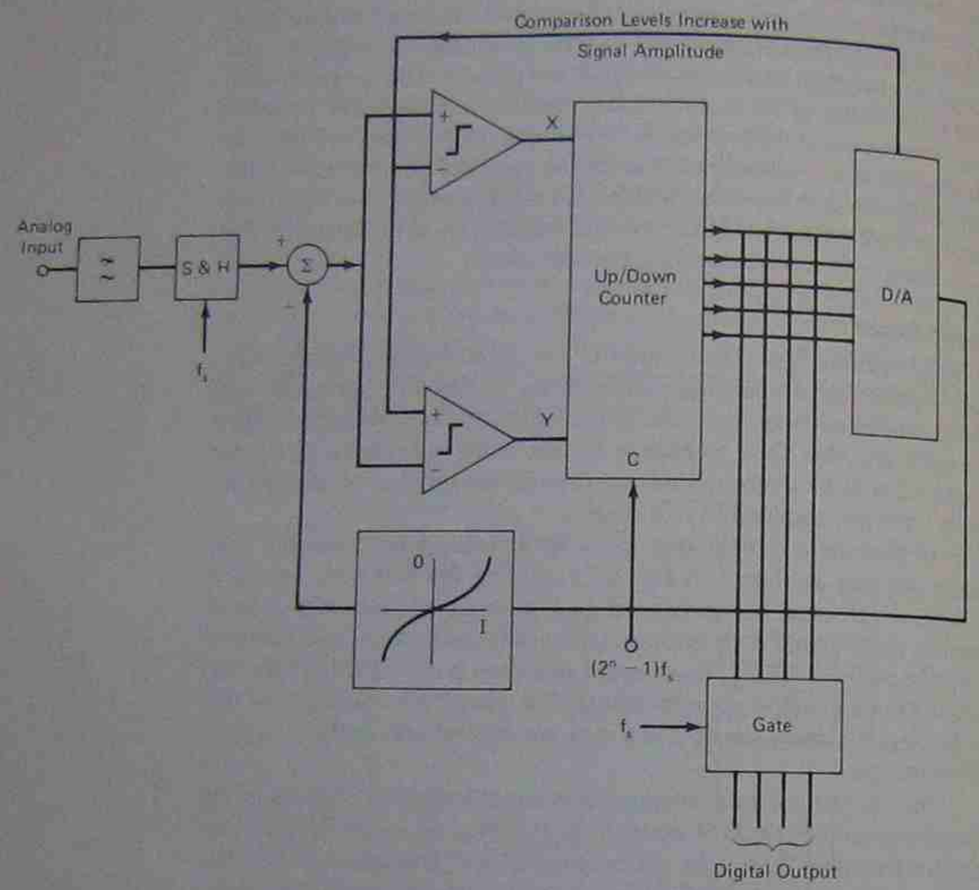
To illustrate this technique, consider a system which has four subranges per step as shown in Fig. 6-22. During the first step, the input voltage, which here is represented as a number equivalent to a level number, is compared with the various boundaries to determine between which boundaries it lies. The selected subrange is then divided into four more subranges and the process repeated. It takes 2 bits to determine one of four levels. After 4 steps, the 8 bits are determined in the case of a 256-level system.

The comparators are referenced at equally spaced intervals in the range between the two D/A converters. Initially, the lower D/A is set at zero and the upper D/A at the maximum of 256 V. The output of the comparators indicates in which subrange the signal lies; in this example, between 64 and 128 V. The reference voltage of 64 V is then applied to the lower D/A and 128 V are applied to the upper D/A. This process is repeated until the smallest range, in this case a range of 4 V, is attained.

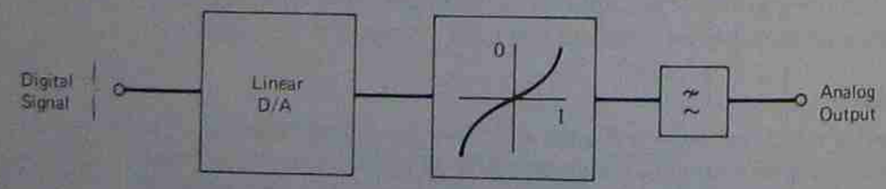
Decoders

The device which converts a digital word into its analog counterpart is known as a digital-to-analog converter. D/A converters are also referred to as DAC's and decoders. For each input code word applied to a decoder, a discrete analog voltage is produced at the output.

The most popular decoder in use is the weighted current source D/A illustrated in Fig. 6-23. Each transistor current source has a different emitter resistor with binary related values of R, $2R$, $3R$, . . . 2^nR, where n refers to the length of the code word. These current sources are



(a) Non-Uniform Serial Encoder



(b) Decoder for Non-Uniform Quantization

Fig. 6-20 Block Diagram for the Encoding and Decoding of Companded PCM

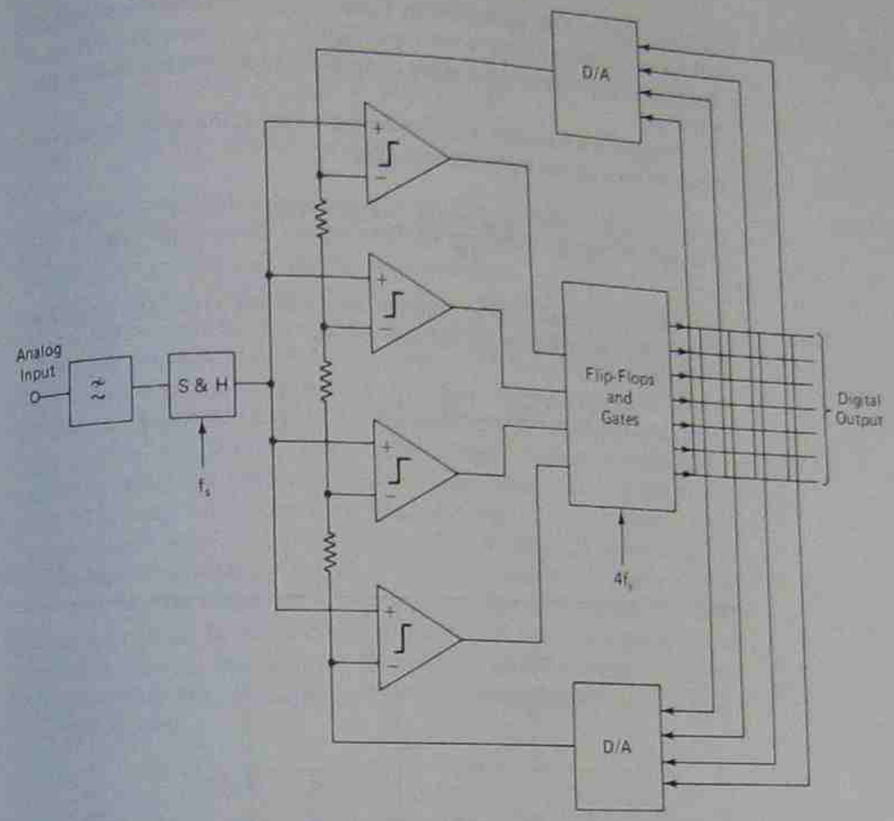


Fig. 6-21 Subranging Converter

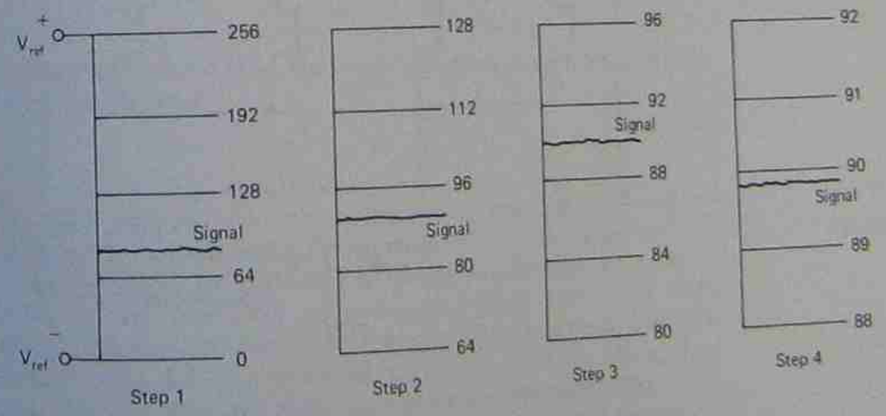


Fig. 6-22 Subranges for a Converter

switched ON or OFF by the binary signals applied to the control diodes connected to the emitters. If the input signal is low, no current flows. If the input is high, there is a current contribution towards the output summing line.

Assuming a 0.6 volt-drop from emitter to base, the total current contribution from all the transistor collectors is:

$$I_o = \frac{V_{ref} - 1.8}{R} + \frac{V_{ref} - 1.8}{2R} + \frac{V_{ref} - 1.8}{4R} + \cdots \frac{V_{ref} - 1.8}{2^n R}$$

Since the output voltage from the ideal current to voltage amplifiers is $-I_o R_f$, the resultant output voltage will be

$$V_{out} = \frac{-(V_{ref} - 1.8)}{R} R_f \left[1 + \frac{1}{2} + \frac{1}{4} + \cdots \frac{1}{2^n}\right] \qquad (6\text{-}9)$$

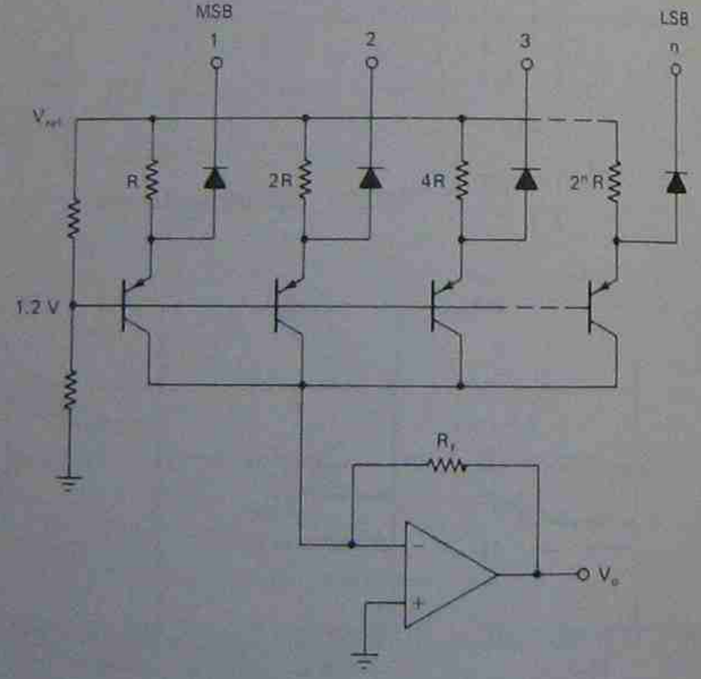


Fig. 6-23 Weighted Current Source Decoder

This can be expressed in the form of a digital word as

$$V_{out} = A\left[\frac{b_0}{2^0} + \frac{b_1}{2^1} + \frac{b_2}{2^2} + \frac{b_n}{2^n}\right] \qquad (6\text{-}10)$$

where A is the constant $\frac{-(V_{ref} - 1.8)}{R} R_f$ and $b_1, b_2, \cdots b_n$ are the bit coefficients which are quantized to be either 1 or 0. From Eq. (6-9) we can see that the most significant bit (MSB) should be applied to Terminal 1, and the LSB applied to Terminal n.

One concern of this network is the accuracy required of the resistors, particularly for the higher resolutions. If, for example, a typical R value of 10k is used in an 8-bit converter, and R is out by 1% (9.9 kΩ), the MSB error contributions would be more than that contributed by the LSB. To overcome this difficulty, a R-2R ladder network such as that shown in Fig. 6-24 is employed. The inputs operate single-pole-single throw switches which tie the shunt resistors either to ground or to the current summing line. Such a network employs only a narrow range of resistances which need to be matched and have identical thermal tracking properties. The absolute values of the ladder resistances are not critical.

As can be observed, the input resistance at every junction is the same and equal to R. Also, the current splits into equal parts at every junction, giving the desired binary division of currents, or binary-weighted currents, to either the summing line or ground. The net output voltage is thus:

$$\begin{aligned} V_o &= -I_o R_f \\ &= -\left(\frac{V_{ref}}{2R} + \frac{V_{ref}}{4R} + \frac{V_{ref}}{8R} + \frac{V_{ref}}{2^n R}\right) R_f \\ &= \frac{-V_{ref} R_f}{R}\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{2^n}\right) \end{aligned} \qquad (6\text{-}11)$$

As before, this circuit gives the proper weighting to each binary digit of the input code word.

6–7.1 Codecs

Monolithic integrated circuit chips that convert voice are now being produced on a massive scale. These chips, which are called CODECS (Coder-Decoders), are becoming so inexpensive that they are being installed in the individual channel units rather than as a single encoder/decoder unit that is time shared by all the channels. Thus, as illustrated in Fig. 6-25, each channel contains a coder for sampling, quantizing, and encoding. The digital outputs are then passed to a digital multiplexer where they are combined to form a single serial bit stream for

transmission. The inverse process occurs in the receiving section of the codec.

Single-channel codecs have the advantage of

1. Being less catastrophic than a time-shared system because the loss of a codec results in the loss of a single channel rather than all the channels in an entire channel bank, and
2. Employing digital multiplexing which tends to be simpler and easier to implement than analog multiplexing. By the proper combining of digital pulse streams, analog baseband signals of various bandwidths can be digitized and carried over the digital system.

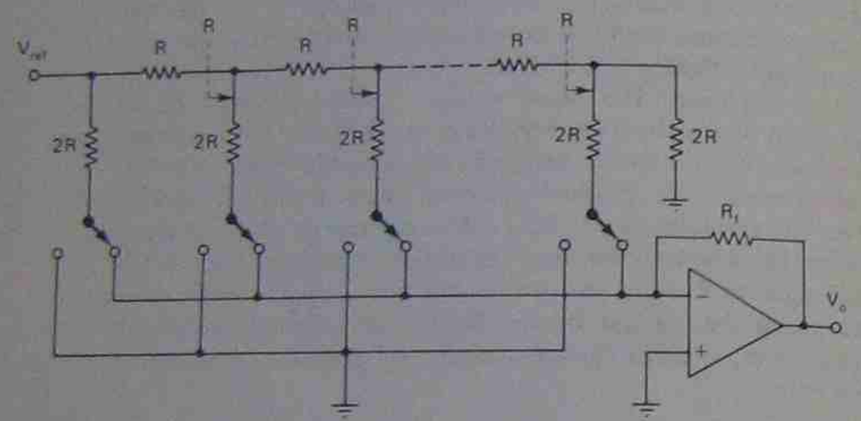


Fig. 6-24 R-2R Ladder Decoder

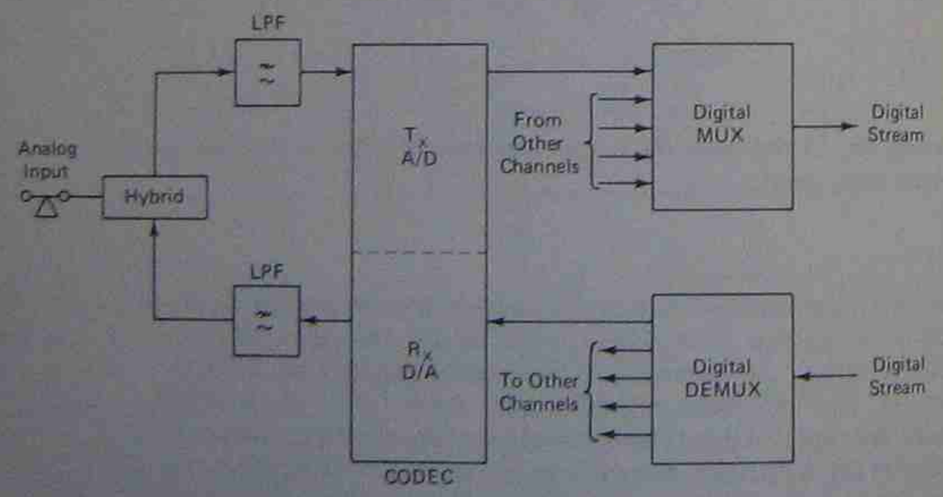
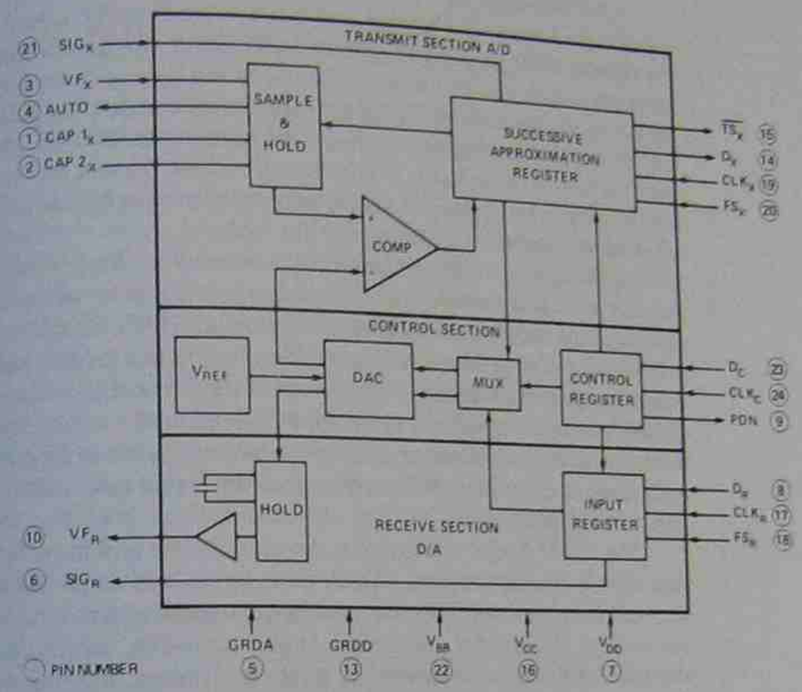


Fig. 6-25 PCM System Using One Codec per Channel

An example of a codec is the INTEL2910A PCM CODEC. This chip contains sample-and-hold circuits, an A/D converter (successive approximation type), a D/A converter, as well as some other logic and interface circuitry which make it suitable for full-duplex PCM operation. The block diagram of this Codec is shown in Fig. 6-26.



PIN NAMES

CAP 1_X, CAP 2_X	Holding Capacitor
VF_X	Analog Input
VF_R	Analog Output
D_R, D_C, SIG_X	Digital Input
SIG_R, D_X, $\overline{TS}_X$	Digital Output
CLK_C, CLK_X, CLK_R	Clock Input
FS_X, FS_R	Frame Sync Input
AUTO	Auto Zero Output
V_{BB}	Power (−5 V)
V_{CC}	Power (+5 V)
V_{DD}	Power (+12 V)
PDN	Power Down
GRDA	Analog Ground
GRDD	Digital Ground
NC	No Connect

Fig. 6-26 Block Diagram and Pin Names for 2910A Codec (Reprinted by permission of INTEL Corporation, Copyright 1979. All mnemonics, Intel Corporation, copyright 1979.)

The codec provides two major functions:

1. Encoding and decoding of analog signal (VOICE, dial tone, busy tone, ring-back tone)
2. Encoding and decoding of the signaling and supervision information (OFF or ON HOOK, rotary dial pulses, etc.)

On a non-signaling frame, the Codec encodes the incoming analog signal at the frame rate, FS_X, into an 8-bit PCM word which is sent out on the D_X lead at the proper time. A frame consists of a single sample of all channels, which will be further discussed in the next section. Similarly, the Codec fetches an 8-bit PCM word from the receive highway, D_R lead, and decodes an analog value which will remain constant on lead VF_R until the next receive frame. Transmit and receive frames are independent. They can be asynchronous (transmission) or synchronous (switching) with each other.

On a signaling frame, which occurs every sixth frame, the Codec transmit side will encode the incoming analog signal as previously described and substitute the signal present on lead SIG_X for the least significant bit of the encoded PCM word. Similarly, on a receive signaling frame, the Codec will decode the seven most significant bits and will output the least significant bit value on the SIG_R lead until the next signaling frame. Signaling frames on the send and receive sides are independent of each other, and are selected by a double-width frame sync pulse on the appropriate channel.

The 2910A Codec is intended to be used on line and trunk terminations. The call progress tones, which include the dial tone, busy tone, ring-back tone, and re-order tone, and the pre-recorded announcements, can be sent through the voice-path. Digital signaling, which includes off-hook and disconnect supervision, rotary dial pulses, and ring control, is sent through the signaling path.

Circuitry is provided within the Codec to internally define the transmit and receive timeslots. In small systems, this may eliminate the need for any external timeslot exchange. This feature can be bypassed and discrete timeslots sent to each Codec within a system.

In the power-down mode, most functions of the Codec are directly disabled to reduce power dissipation to a minimum. The Codec can be portioned into three major sections, as noted in Fig. 6-26: the middle or control section, the top transmit or A/D section, and the bottom receive or D/A section.

Control Section

The operation of the 2910A is determined by the control word, which is serially loaded as an 8-bit word through the D_C lead (data) and the CLK_C lead (clock). The loading is asynchronous with the other operations of the Codec, and takes place whenever transitions occur on the CLK_C lead. The D_C input is loaded during the trailing edge of the CLK_C input.

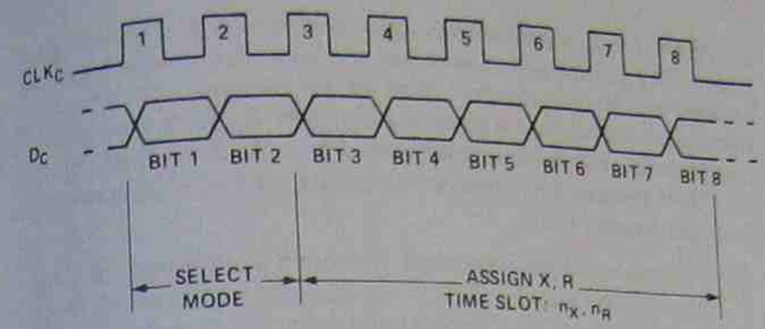


Fig. 6-27 The Control Word (Reprinted by permission of Intel Corporation, Copyright 1979. All mnemonics, Intel Corporation, Copyright 1979.)

The control word contains two fields. The mode select and the time Bit 1 and Bit 2 define whether the subsequent six bits apply to both the transmit and receive sides (00), the transmit side only (01), the receive side only (10), or whether the Codec should go into the standby, power-down mode (11). In the last case (11), the following six bits are irrelevant.

The last six bits of the control word define the timeslot assignment, from 000000 (timeslot 1) to 111111 (timeslot 64). Bit 3 is the most significant bit. Bit 8 is the least significant and the last into the Codec. The Codec will retain the control word (or words) until a new word is loaded in or until power is lost. This feature permits dynamic allocation of timeslots for switching applications. The bits are all tabulated in Table 6-3.

Table 6-3 Control Word Bit Allocations (Reprinted by permission of Intel Corporation, Copyright 1979.)

Bit 1	Bit 2	Mode
0	0	X & R
0	1	X
1	0	R
1	1	Standby

Bit 3	4	5	6	7	8	Timeslot
0	0	0	0	0	0	1
0	0	0	0	0	1	2
			•			•
			•			•
			•			•
			•			•
1	1	1	1	1	1	64

Each Codec is initially programmed via CLK_C and D_C for the desired transmit and receive timeslots. All Codecs tied to the same D_R bus will, in general, have different receive timeslots, although that is not a device re-

quirement. There may be separate buses for transmit and receive, or all Codecs may transmit and receive over the same bus, in which case the transmit and receive channels must be synchronous ($CLK_x = CLK_R$). There are no restrictions on timeslot assignment. A device may have the same transmit and receive timeslot even if a single bus is used.

There are several requirements for using the $CLK_C - D_C$ interface in the microcomputer mode.

1. A complete timeslot assignment, consisting of eight negative transitions of CLK_C, must be made in less than one frame period. The assignment can overlap a framing pulse as long as all eight control bits are clocked in within a total span of 125 μs (for an 8 kHz frame rate). CLK_C must be left at a TTL low level when not assigning a timeslot.
2. A dead period of two frames must always be observed between successive timeslot assignments. The two-frame delay is measured from the rising edge of the first CLK_C transition of the previous timeslot assigned.
3. Transmit only/receive only operation is not permitted. That is, the A/D and D/A converters will begin to operate only after timeslot assignments have been made to both channels. Special circuits which use only one channel should be physically disconnected from the unused bus. This allows a timeslot to be made to an unused channel without consequence.

Example of microcomputer control mode:

First load the two words 01000001 and 10000010 into the Codec as shown in Fig. 6-28. The transmit side is now programmed for Timeslot 2 and the receive side for Timeslot 3. The Codec will output a PCM word on the transmit PCM highway (bus) during the Timeslot 2 of the transmit frame, and will fetch a PCM word from the receive PCM highway during Timeslot 3.

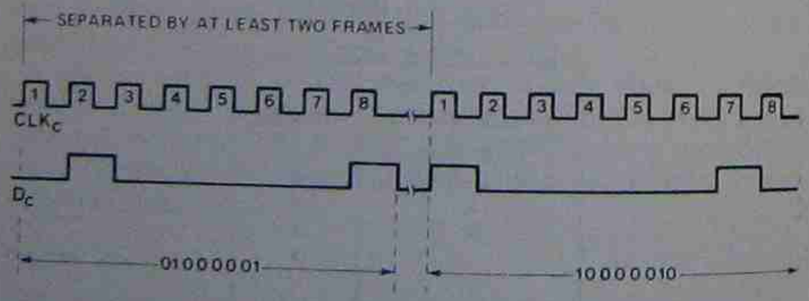


Fig. 6-28 Loading of Control Words into 2910A Codec (Reprinted by permission of Intel Corporation, Copyright 1979. All mnemonics, Intel Corporation, Copyright 1979.)

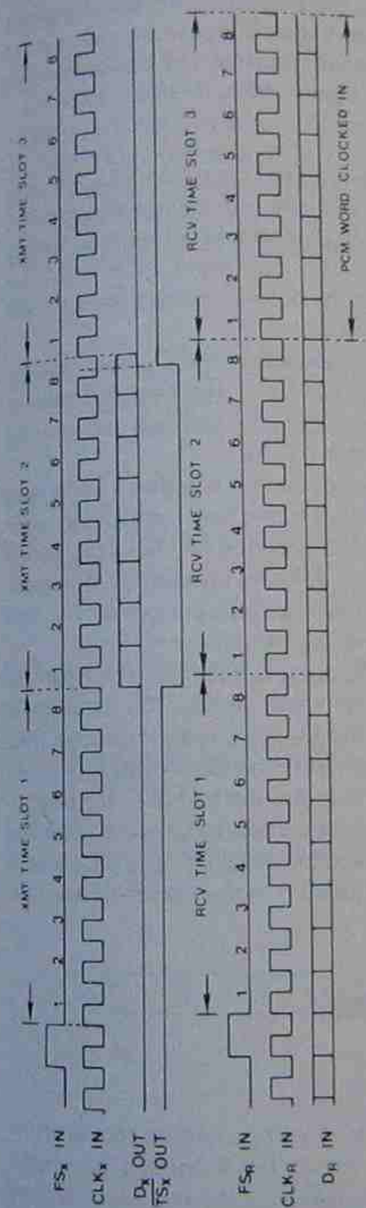


Fig. 6-29 Timeslot Allocations for the Control Words Shown in Fig. 6-28 (Reprinted by permission of Intel Corporation, Copyright 1979. All mnemonics, Intel Corporation, Copyright 1979.)

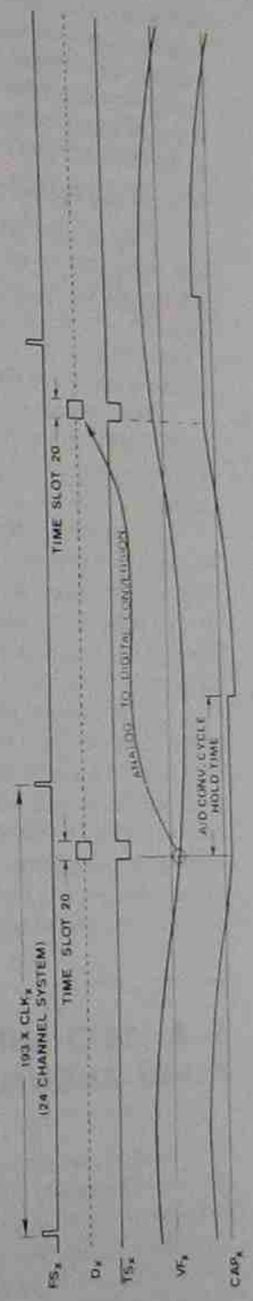


Fig. 6-30 Analog-to-Digital Conversion (Reprinted by permission of Intel Corporation, Copyright 1979. All mnemonics, Intel Corporation, Copyright 1979.)

In the microcomputer mode, each Codec performs its own timeslot computation independently for the transmit and receive channels by counting clock pulses, CLK_X and CLK_R. All Codecs tied to the same data bus receive identical framing pulses, FS_X and FS_R. The framing pulses reset the on-chip timeslot counters every frame; hence, the timeslot counters of all devices are synchronized. In this example, the Codec interface to the PCM highway will function as shown in Fig. 6-29.

Transmit and Receive Sections

The VF signal is applied to the sample-and-hold circuit on VF_X. It is sampled by an internal switch and held by an external capacitor connected between the $CAP1_X$ and $CAP2_X$ terminals. For an 8 kHz sampling system, 2000 pF would be quite adequate. The converted PCM (mu law) is then output to the D_X terminal when the proper timeslot occurs in the following frames, as illustrated in Fig. 6-30.

When receiving, the PCM word is fetched by the D_R lead from the PCM highway at the proper timeslot occurrence. The decoded value is held on an internal sample and hold capacitor. The buffered output signal on the VF_R lead has a range of about ± 2.2 volts.

The duration of the FS_X and FS_R pulses defines whether a frame is an information frame or a signaling frame. A frame synchronization pulse which is a full clock period in duration (CLK_X period for FS_X, CLK_R period for FS_R) designates a non-signaling frame. A frame synchronization pulse which is two full clock periods in duration (two CLK_X periods for FS_X, two CLK_R periods for FS_R) designates a signaling frame.

On the transmit side, when the FS_X pulse is widened, the eighth bit of the PCM word will be replaced by the value on the SIG_X input at the time when the eighth bit is output on the D_X lead. On the receive side, when the FS_R pulse is widened, the eighth bit of the PCM word is detected and transmitted on the SIG_R lead. That output is latched until the next receiving signaling frame. The remaining seven bits are decoded according to the value given in the CCITT G733 recommendation. The SIG_R lead is reset to a TTL low level whenever the Codec is in the power-down state.

6–8 24-CHANNEL SYSTEM FRAME ALIGNMENT

In a practical multi-channel PCM system, it is important for the receiving terminal to properly interpret which bits in the incoming bit stream belong to which channel, which bits are signaling bits, and which bits are synchronizing bits. The signaling bits relay signaling information such as ON HOOK or OFF HOOK condition, dialing, and BELL ringing. Since signaling is rather slow, it needs to be sampled less often than the speech information.

In the Lenkurt 9002B system, each of the voice frequency (VF) channels is sampled at 8000 times per second (8 kHz rate), and every sample is encoded into an 8-bit digital word. A frame consists of a single sample of all 24 channels, hence 24 digital words or 192 bits. Because frame synchronization is required to determine which bits make up Channel 1, 2, and so forth, a single synchronization bit or frame bit (s-bit) is added to give a total frame of 193 bits. This yields a total bit rate of 1.544Mb/s (total bits/frame × sampling rate). In terms of timing, each channel is sampled at an 8 kHz rate, yielding 125 μs ($^1/_{8000}$) frame time or time needed to sample all 24 channels. Each frame consists of 193 bits (192 channel bits + 1 framing bit), which gives a time slot of 648 ns for each bit or 5.18 μs for each 8-bit channel word. Figure 6-31 indicates the

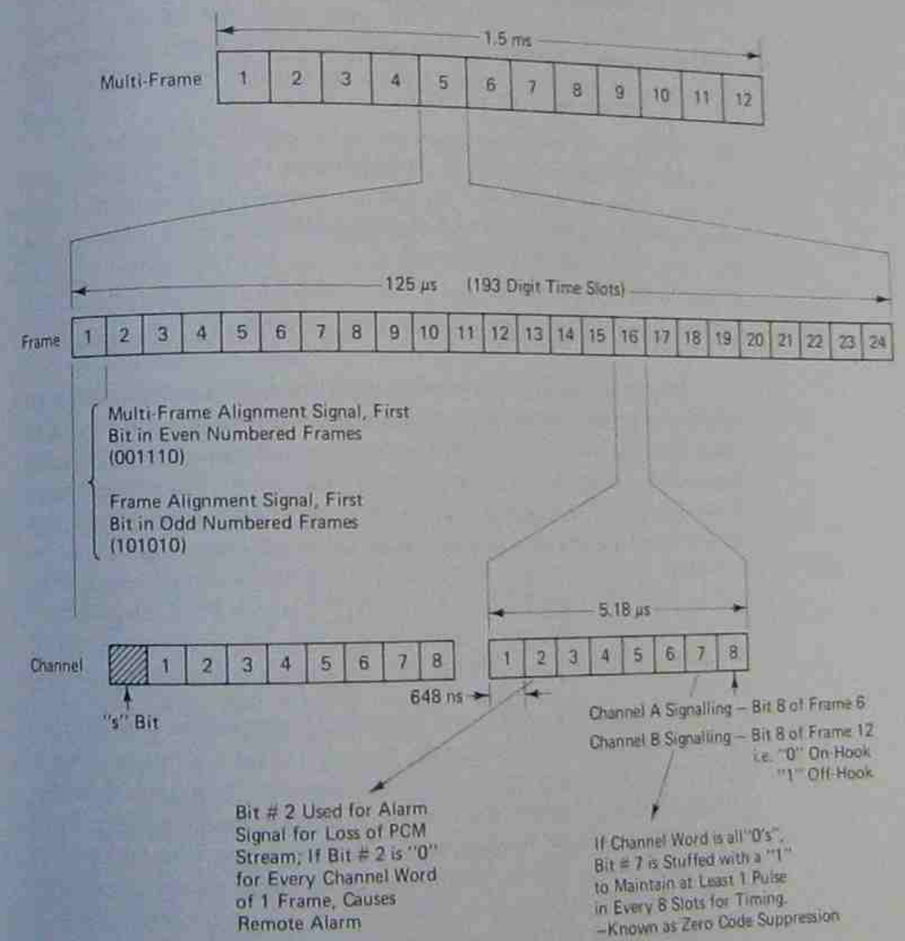


Fig. 6-31 24-Channel Time Slot Assignment

relationship of the multiframe, the frame, channel words, and bit sequences along with the synchronized information bits.

Two signaling channels are made available by stealing the eighth bit from Frames 6 and 12. This causes about a 2 dB reduction in the theoretical S/N ratio. Each signaling channel thus transmits at a rate of 666 b/s (1/1.5 ms), or one twelfth of the channel signal rate. This is quite adequate for monitoring the slower signaling information.

The following list gives the various time allottments in abbreviated form:

1 channel time slot = 8 digit timeslots

sampling rate of each channel timeslot = 8 kHz

$$1 \text{ frame time} = \frac{1}{8000} = 125\ \mu s$$

1 frame = 24 × 8 + 1 = 193 timeslots

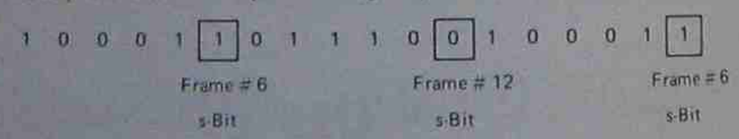
1 multiframe = 12 × 125 μs = 1.5 ms

$$1 \text{ digit timeslot} = \frac{125\ \mu s}{193} = 648 \text{ ns}$$

channel A and channel B signaling rate
= 1 bit/multi-frame/channel
= 666 b/s/channel

(combined signaling rate = 2 × 666 = 1333 b/s/channel)

A study of Fig. 6-31 shows that the s-bit must provide a recognizable frame synchronizing pattern to indicate which is Channel 1 in the sequence of channels. Its second function is to indicate when a signaling frame (A signaling channel or B signaling channel) occurs and which one is occurring. This is done by the interleaving of two separate logic patterns to produce the unique framing sequence of s-bits; that is,



In this sequence, every sixth frame bit is, alternately, 1 or 0. The A signal channel is enabled during the time slot of Bit 8 for every channel word occuring during a signaling frame (every sixth frame) which follows a frame s-bit of 1. Similarly, the B signal channel is enabled during the Bit 8 slot of every channel word during a signaling frame following a frame s-bit of 0. Hence, signaling information occurs only during the channel words of every sixth frame, or Frames 6, 12, 6, 12, and so forth. From the pattern of s-bits in a 12-frame sequence (called a superframe), a terminal

framing pattern identifies the locations of the time slot of Bit 1 of Channel 1 in each frame, and a signaling frame pattern identifies Frames 6 and 12 in which signaling information occurs.

Terminal framing has an alternating 1 and 0 s-bit pattern (101010) in odd-numbered frames between Frame 1 and Frame 11. Signal framing has an s-bit pattern of 001110 in even-numbered frames between Frame 2 and Frame 12. The signaling s-bit changes 0 to 1 at Frame 6 and from 1 to 0 at Frame 12. Since one signaling bit occurs every sixth frame or 2 bits occur in 12 frames (1.5 ms), the total signaling bit rate is 2 b/1.5 ms or 1333 b/s. Thus, each of the two signaling channels, A and B, has a bit rate of 666 b/s.

System supervisory information regarding operational status is also determined by examining Bit 2 of every channel word in every frame. If this bit is 0 for every word of one frame, an alarm is given (*REMOTE* alarm) to indicate that, most likely, no PCM bit stream is present. Bit 2 is used because, under normal conditions, this bit will contain a 1 less than 10% of the time. Its absence becomes a good indicator of the absence of PCM bits.

To ensure that timing information necessary for the receiver is not lost by a long string 0's representing a particular value, Bit 7 in every word can be used as a timing bit. The receiver timing depends upon detecting a 1 on the average in every 8-bit word, thus allowing a maximum of 15 consecutive 0's in any particular event. All bits in each word are monitored at the transmit side. If all the bits in a PCM word are 0, Bit 7 is forced to 1, which sidesteps the previously stated condition. Changing Bit 7 causes the least error in code value.

One of the practial problems associated with frame synchronization is how to regain framing synchronization within the system once it has been lost; that is after the system is turned on or if there is a break in the transmission path. Using the framing scheme shown in Fig. 6-31, once the receiver loses knowledge of the location of Bit 1 of Channel 1, it will remain lost because there is no inherent method to induce the bit stream of the frame to resynchronize. To solve this, an out-of-frame detector is used. This detector is made up of circuitry which examines every second frame (Frames 1, 3, 5, 7, and so on), hence every 386 bits (2 × 193 bits) are examined for the 010101 framing pattern. If a problem has occurred, either two 1's or two 0's will occur in the pattern: (0101001010) or (0101101010). The existence of either of these two situations causes an extra bit to be inserted in the frame after the location of the incorrect pattern. The detector examines the pattern until this condition is found again, and then adds another extra bit. This process continues until the correct pattern is regained. The receive patterm is internally generated in the receiver and is adjusted by the above process until it matches the incoming frame pattern. Thus, the system becomes synchronized. Synchronization status is indicated by the LOCAL alarm on the system.

6–9 PCM TERMINAL EQUIPMENT

The D3 PCM-channel bank can be applied to intertoll service as well as to direct and to toll connecting service, as illustrated in Fig. 6-32.

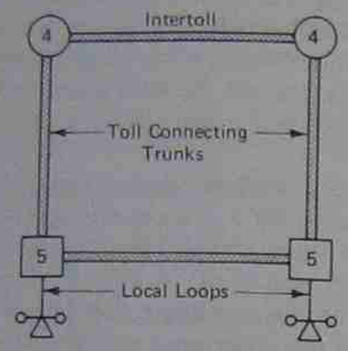


Fig. 6-32 Application of D2 Channel Bank

The Lenkurt 9002B PCM-channel bank is made up of 24 plug-in channel units, a common transmit unit, a common receive unit, and a common alarm unit. The basic block diagram outlining the relationship of these parts is shown in Fig. 6-33. Note that 24 analog signals are time division multiplexed, although only three channels are shown in the diagram.

Recall that any PCM system essentially must have the functions shown in Fig. 6-34.

The Lenkurt channel bank has all these blocks or functions buried in it. Referring again to Fig. 6-33, it can be seen that the handsets and signal source become the analog input signals into the system, usually with voice frequencies of 200 to 3400 Hz. Examination of the channel unit (CU) shows that it performs the basic function of band limiting the input signal through use of bandpass filter, and that it also performs sampling at a rate of 8 kHz, hence multiplexing.

The CU block is followed by the Common Transmit Unit (CTU), which performs quantizing and encoding functions. The resulting pulse stream is fed to the Common Alarm Unit (CAU) and then to the optional Office Repeater, if it is present. Basically, the CAU performs the task of setting up standard pulse levels and of detecting basic system failures, which will be seen later. Signals are then passed to the transmission, or span line.

On the PCM pulse stream going into the channel bank, the Office Repeater, if used, regenerates the supposedly low amplitude pulses to make them acceptable to the decoder. Again, the alarm unit basically functions as a signal monitor for obvious system failures, such as loss of

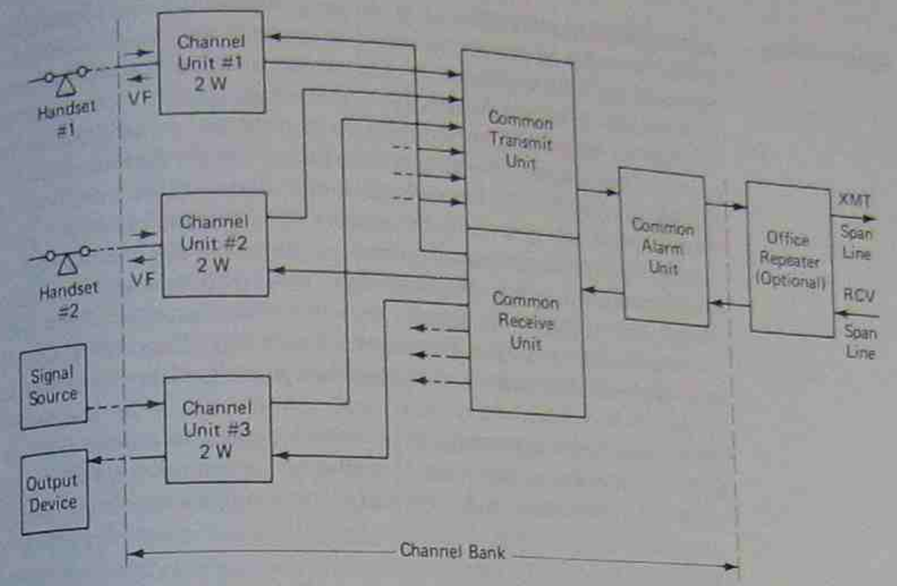


Fig. 6-33 Lenkurt 9002 B-Channel Bank

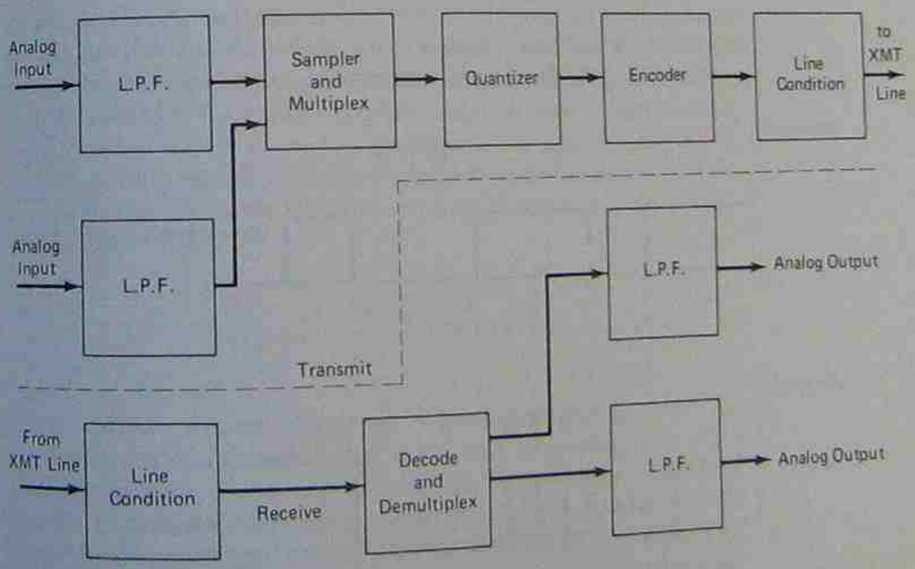


Fig. 6-34 Basic Functions of a PCM-Channel Bank

pulses. The PCM pulse stream is fed to the Common Receive Unit. At this point, the pulse stream is examined to determine timing information regarding the location of each bit (bit timing) and the start of the frame (first channel in the possible set of 24). This information is then applied to the decoder with the PCM pulse stream to produce a quantized analog pulse stream. It should be a replica of that originally produced by the Common Transmit Unit. Hence, the Common Receive Unit functions as a decoder while also generating the appropriate timing information.

The quantized analog pulse stream is passed on the receive-side of the channel unit which demultiplexes, or separates out, the pulses needed to reconstruct its signal. Reconstruction is accomplished by low-pass filtering of the pulses to form a facsimile of the original analog signal applied to that channel. Although the system has been illustrated as a basic PCM system, many additional refinements, techniques, and methods are used to accomplish the needed functions of timing, synchronization of frames, transmission of additional information (such as signaling and supervisory), and detection of system faults.

6–10 LINE PULSE FORMAT

Two basic digital signal formats are used in most PCM systems. These formats are illustrated in Fig. 6-35.

The 9002B system uses the RTZ format with a 50% duty cycle since it requires timing information on the receive end to ensure that each information bit is examined in its correct time interval. This allows the receiver to track any small changes in transmitter bit rate without causing errors. In the NRTZ format, a long string if 1's would result in virtually no

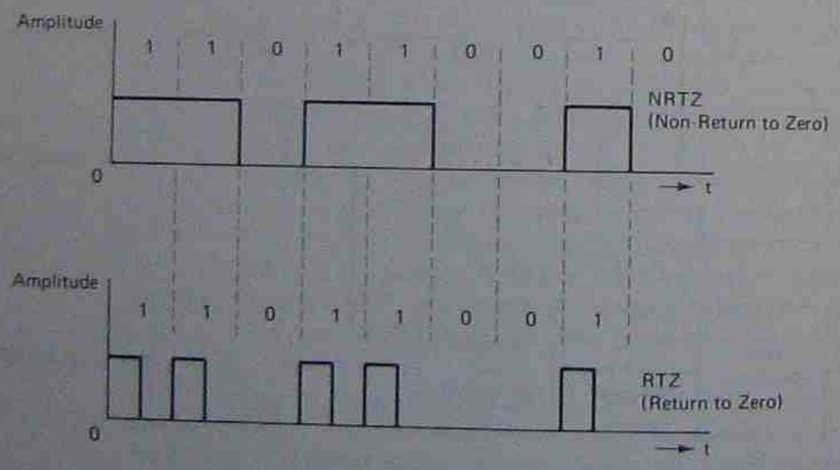


Fig. 6-35 Signal Formats

signal transition and hence no timing information would be obtained. The RTZ format ensures that a signal transition will occur regardless of the content of the information bits. This timing information is also necessary for operation of the regenerators, or repeaters, in the transmission line.

The signal formats shown in Fig. 6-35 are also known as unipolar pulses because all 1's are represented by only one voltage, or current polarity. Obviously, the ability the resolve a 0 or a 1 is influenced by the amount of noise present in the system.

Because unipolar pulses are biased towards a positive voltage, the net dc voltage level increases with the number of 1's transmitted. If a fixed threshold between a 1 and a 0 is used for decisions, a small amount of noise could cause a 0 to be interpreted as a 1 when a long string of 1's are present and causing the dc voltage level to reside near the threshold level.

In addition to timing information, the bandwidth or frequency content of the pulse stream is important in transmitting pulses. Since the only requirement for PCM signal information recovery is detection of either a 1 or 0, the minimum required bandwidth of a PCM signal is the fundamental frequency component of the pulse stream. This is illustrated in Fig. 6-36. Usually the worst, or largest bandwidth, occurs when the string alternates between 1's and 0's, i.e., 101010. This will produce the highest fundamental frequency. For example, a pulse width, t_p, of 324 ns in a RTZ-unipolar format has a fundamental frequency (bandwidth) of:

$$BW = \frac{1}{2t_p} = \frac{1}{2(324\text{ ns})} = 1.544\text{ MHz}$$

In an attempt to reduce both noise sensitivity and bandwidth, the scheme outlined in Fig. 6-36(b) is implemented. A bipolar Alternate Mark Inversion (AMI) format is used as the transmission format in the 9002B system. It quickly becomes apparent that for the same bit information, the fundamental frequency, or needed bandwidth, is only one-half of the unipolar situation; hence,

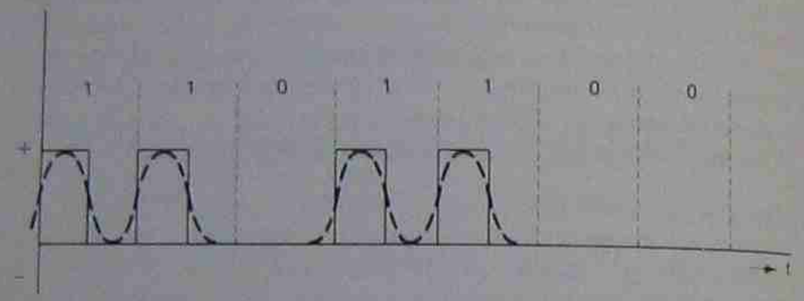
$$BW_{AMI} = \frac{1.544\text{ MHz}}{2} = 722\text{ kHz}$$

In addition, the net dc content, regardless of the bit content, becomes 0. Thus, much larger noise values can be present in the signal before the decision of a 1 or 0 becomes doubtful. Also, this signal now becomes transparent, or passes unmodified through transformers and capacitors. Of course, the penalty paid for these advantages is increased system complexity to process this format. One additional advantage, however, is that detection of errors due to noise becomes possible. This is

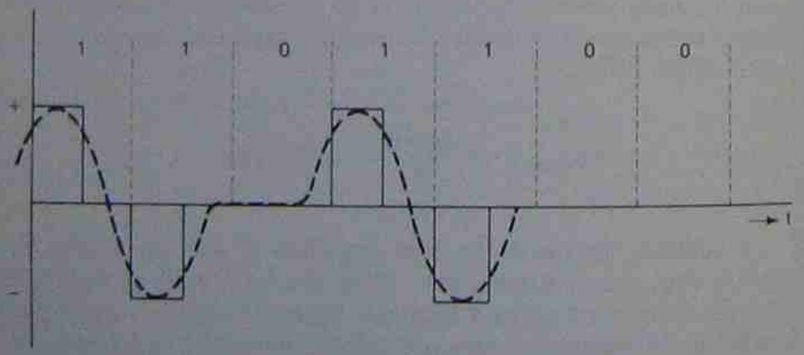
done by detecting bipolar violations, or the occurrence of two or more marks, 1's of the same polarity in a sequence.

The removal of the dc component from the unipolar pulses used in the terminal logic circuits is accomplished in the 9002B system by a bipolar converter in the common transmit unit. The conversion back to the unipolar pulse format from the bipolar line transmission format is performed by a converter in the common receive unit.

In North America, the original Western Electric T1 carrier system contained the D1-PCM-channel bank. It was suitable for transmission on direct trunks between Class 5 end offices, and between Class 5 end offices and Class 4 offices (toll-connecting trunks). It was, however, of unsuitable transmission quality for intertoll networks. The D2-channel bank was designed for this, and it has recently been upgraded to form the D3-channel bank. The major characteristics of the PCM-channel banks are shown in Table 6-4.



(a) Unipolar



(b) (Bipolar Alternate Mark Inversion; AMI)

Fig. 6-36 Line Code Formats

Table 6-4 Present PCM-Channel Bank Characteristics

Channel Bank Generation	Bell System	Lenkurt System	Characteristics
1st	D1	9001	• 24 channels VF • companding encoder, $\mu = 100$ • 7-bit voice endode (127 levels) • 1-bit signaling encode • pseudo random sampling sequence Time slot: 1 2 3 4 5 6 7 8 Channel unit: 13 1 17 5 21 4 16 7 • used with T1 repeatered line (1.544 Mb/s, RTZ, bipolar)
2nd	D2	9002A	• 24 channels VF • companding encoder, $\mu = 255$ • 8-bit voice encode, 83% of time • 1-bit signal encode, 17% of time • pseudo random sample sequence • uses T1 repeatered line
3rd	D3	9002B	• same as 2nd generation, except for straight sample sequence; e.g., Time Slot 1 is Channel 1.
4th	D4		• 48 channels VF • alternate bit interleaving (3.088 Mb/s plus control bits to give 3.152 Mb/s by multiplexing) • uses T1C repeatered line

6–11 NORTH AMERICAN MULTIPLEX HIERARCHY

Each type of signal source tends to have its own characteristic digit rate. In telephony, 64k b/s are required whereas in broadcast television about 60M b/s are required. The transmission media over which these signals must be transported also has its own characteristic digit rate. For a balanced two-wire line, 2M b/s is appropriate; a waveguide system may have a capacity in the region of 500M b/s.

Multiplexing is used to match such signals to a transmission path. A number of primary signal sources are combined to form a composite signal suitable for efficient use of the transmission media. By planning a suitable hierarchical multiplex structure, it is possible to combine various primary signals into identical formats that can be carried by the same

transmission equipment. Also, signals carried by low capacity transmission systems can be combined to form the basic unit of a higher capacity system.

PCM Multiplex Hierarchies

The Bell System, generally used as a standard in North America, established a multiplexing structure for PCM carrier systems and designated it a T-carrier system. The framework of this carrier system is a multiplexing hierarchy consisting of several levels based on the bit rate of the multiplex line signals. The first level, referred to as T1, is the basic building block of this hierarchy.

Figure 6-37 illustrates the North American multiplex hierarchy. Table 6-5 summarizes some of its major characteristics. Note that the T1C level does not have a designed line standard. In that case, assume the line standard of DS-1 with a slightly different bit rate. This was implemented to handle the newer duobinary PCM system. The T6 level is generally designed for fiber optic applications.

Table 6-5 Types of Digital Carrier Lines

Line Type	Line Signal Standard	Characteristics
T1	DS-1	• 1.544 M b/s • 24 TDM channels voice
T1C	—	• 2 time division multiplexed T1 lines • 48 TDM channels voice • 3.152 M b/s
T2	DS-2	• 4 time division multiplexed T1 lines • 96 TDM channels voice • 6.312 M b/s
T3	DS-3	• 28 time division multiplexed T1 lines • 672 TDM channels voice • 44.736 M b/s
T4	DS-4	• 168 time division multiplexed T1 lines • 4032 TDM channels voice • 274.176 M b/s
T5	DS-5	
T6	DS-6	

6–12 MEASUREMENT OF QUANTIZATION NOISE

To measure the quantization (S + N)/N ratio, a white noise signal is passed through a bandpass filter that acts as the baseband signal to the PCM channel. The white noise generator and filter are preferred over a

sine wave generator since all quantum levels are produced in a random fashion, which closely simulates the normal human voice. After passing through the PCM channel, the received signal and any associated noise are detected with a meter. A filter is then inserted (see Fig. 6-38) to receive the original baseband signal, which leaves only the noise. The ratio between the two measurements, usually expressed in decibels, is the signal-to-noise ratio (actually (S + N)/N ratio).

6–13 INTRODUCTION TO INFORMATION THEORY

The basic unit used when discussing information is the bit, an abbreviation of binary digit. It represents the speculation of one of two equally likely alternatives, much like the two sides of a coin in a coin-flip experi-

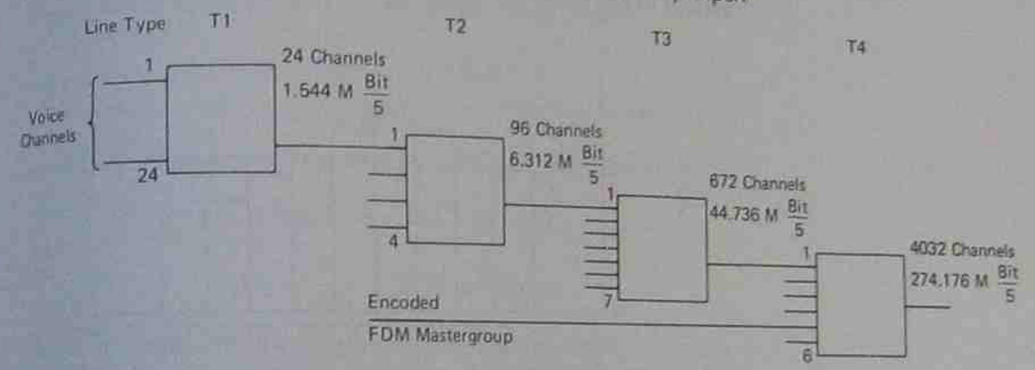
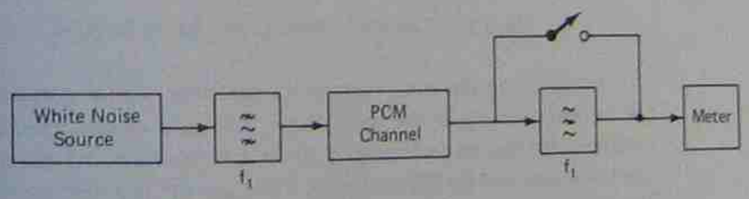


Fig. 6-37 North American Digital Multiplex Hierarchy



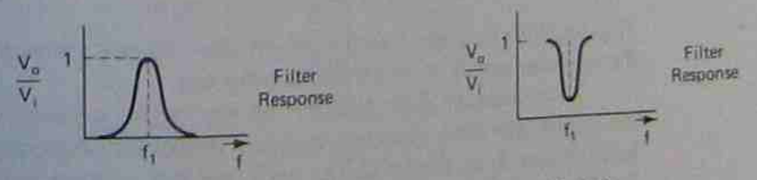


Fig. 6-38 Quantization Noise Measurement System

ment. If one has four equally likely alternatives, as for example four voltage levels, two bits would be required to determine the unique level. The four levels can be represented by the unique pairs 00, 01, 10, and 11.

If the number of levels are doubled to eight, three bits of information would be required. Similarly, sixteen levels would require four bits to determine the unique level. Extrapolating, the information content, H, is found to be:

$$H = \log_2 (\text{number of alternatives}) \qquad (6\text{-}12)$$

Applying Eq. (6-12) to a signal which can assume one of m equally likely voltage levels would result in $\log_2 m$ bits of information. The opposite is also true. If there are n bits of information 2^n different levels can be distinguished. Thus, with eight bits of information, $2^8 = 256$ different voltage levels can be distinguished.

Consider a message with a duration of T seconds over which there are n samples, each sample of which can assume one of m levels. This is illustrated in Fig. 6-39.

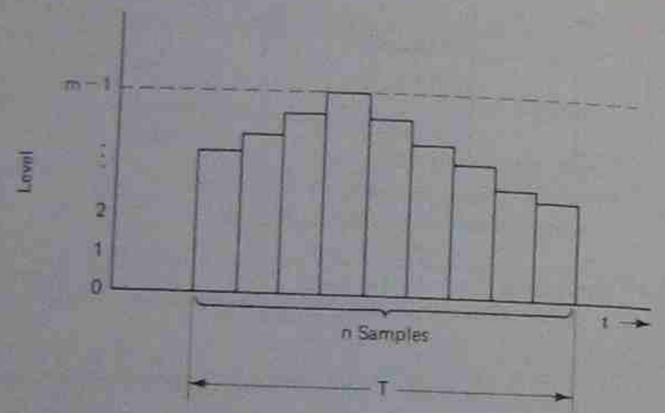


Fig. 6-39 Discrete Voltage Levels Being Sampled

In obtaining the total number of possible signal combinations, it can be observed that for the first sample, there are m possible signal combinations; for the second sample, there are $m(\cdot m) = m^2$ possible signal combinations; and for the n sample, there are m^n possible combinations. The information content of the message over a time T is

$$H = \log_2 m^n = n \log_2 m \qquad (6\text{-}13)$$

This equation indicates that the information content varies directly with the message length, n, as would be expected.

As an exercise, determine the number of bits required to represent the English alphabet, assuming an equal probability occurrence of each letter. If $n = 1$, as shown in Fig. 6-40, the bits needed would be $H = \log_2 26 = 4.7$ bits. Since fractional bits cannot be used, go to the next

highest integer, which is 5. This would leave $32 - 25 = 7$ levels for other characters.

Increasing the number of levels will increase the information transmitted. However, increasing the number of levels will also cause a decrease in the spacing between them which makes the signal more susceptible to noise and therefore to errors. For this reason, a two-level system is most commonly used. Chapter 8 will consider how the information capacity, or information rate, is related to the bandwith of the transmission medium.

Coding

A single, multilevel pulse can be broken up into several pulses while keeping the same number of information combinations. Consider a single pulse, 32-level system, which represents an information content of

$$H = \log_2 m^n = \log_2 32^1 = 6 \text{ bits}$$

If this is to be represented by binary pulses, it will require:

$$6 = \log_2 2^n = n \log_2 2 = n$$

$$n = 6 \text{ pulses}$$

If it is to be represented by a terniary system (3 levels) it will require:

$$6 = \log_2 3^n = n \log_2 3 = 1.583\, n$$

$$n = \frac{6}{1.583} = 3.79 \text{ or } 4 \text{ pulses}$$

The latter would actually permit 3^4, or 81 combinations.

Refer back to Eq. (6-13) to derive an expression for the information

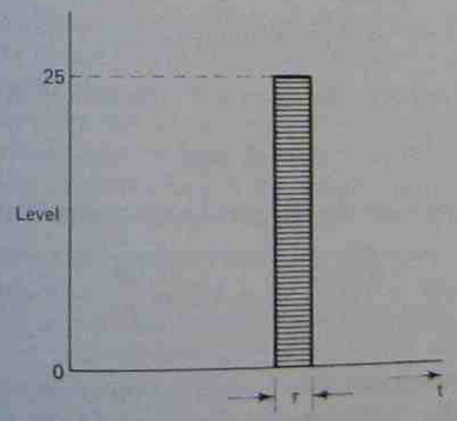


Fig. 6-40 Levels Reprinted to Represent the English Alphabet

rate, or information capacity, of the signal. Since H represents the information content over a time T, the information capacity will be

$$C = \frac{n}{T} \log_2 m \text{ bits} \tag{6-14}$$

From this result, it can be seen that the information capacity will be doubled if the number of samples per second are doubled. Doubling the number of distinguishable levels only causes a logarithmic increase in signal capacity. The number of distinguishable levels depends upon the noise power (N) and signal power (S). The greater the S/N ratio, the larger the number of levels that can be resolved.

The number of pulses that can be transmitted per second can be expected to relate linearly to the bandwidth. Chapter 8 will show that for an ideal low-pass transmission path of *bandwidth B, a maximum of 2B b/s can be transmitted* in a two-level system. If Eq. (6-14) is simplified to a two-level system where $m = 2$; $C = 2B = n/T \log_2 2 = n/T$, or $n/T = 2B$, then

$$C = 2B \log_2 (m) \tag{6-15}$$

From this expression we can see that for a given information capacity, number of signal levels (m) can be exchanged for bandwidth.

Example 6-1 In a channel bank the maximum signal-to-noise power ratio is 38 dB. Assuming a fixed noise level and a nominal 4 kHz voice channel, what is the information capacity of one voice channel?

Solution

$$\text{dB} = 10 \log \frac{S}{N}$$

where S and N are the signal and noise powers respectively.

$$\text{dB} = 20 \log \frac{V_S}{V_N}$$

and V_S and V_N are the signal and noise voltages respectively.

$$\text{For} \quad 38 \text{ dB} \quad \frac{V_S}{V_M} = \text{antilog} \frac{38}{20} = 79.4$$

Since the noise is approximately $^1/_{80}$th of the signal amplitude, 80 levels are distinguishable. The information capacity will be: $C = 2B \log_2 m = 2 \times 4\text{k} \log_2 80 = 50\text{k}$ b/s.

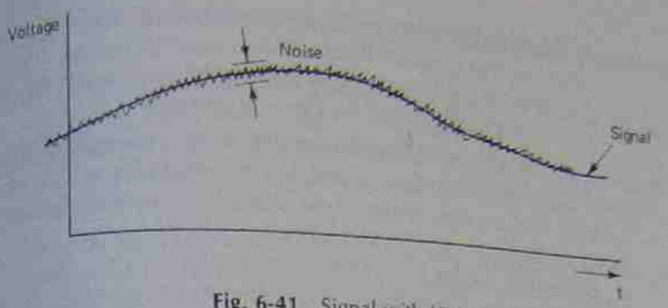


Fig. 6-41 Signal with Noise

PROBLEMS

1. a. Describe the operation of a simple delta modulator.
   b. Describe slope overload.
   c. Derive Eq. (6-1).
2. a. A 1 volt peak audio signal with a maximum frequency component of 3.3 kHz is to be transmitted by a simple delta modulator with a step size "a" of 1 mV. What is the minimum sampling frequency required to assure that slope overload does not occur?
   b. What can be done to decrease the sampling frequency of (a)? What effects would this have on the signal?
   c. What is the chief disadvantage of DM?
3. a. What is companded delta modulation, and why is it employed?
   b. Describe the cause of quantization noise.
   c. What precautions must be followed to prevent a telephone cable from acting as a 4 kHz low-pass filter?
4. Sony manufactures an audio digitizer which samples a signal at a rate of 44 kHz. On a video cassette recorder, it stores 32 bits for each sample taken.
   a. What is the storage rate required of this system?
   b. How many bits must be stored for a half-hour program?
   c. What is the advantage of recording digitally?
   d. Why is a *video* cassette required?
5. a. What is a commutator? Illustrate how it can be used to sample a signal.
   b. Sketch a channel-interleaving scheme for Time Division Multiplexing the following channels: Five 4 kHz telephone channels, and one 20 kHz music channel. Note the minimum times required between the samples for the two types of channels.

c. If 128 distinguishable levels are to be obtained for each signal, what channel capacity is required?
d. Estimate the minimum system bandwidth required if the minimum bandwidth is half the channel capacity.

8. A telephone T1 carrier consists of 24 voice channels multiplexed and PCM modulated. Each signal has a bandwidth of 300 Hz to 3.2 kHz. The encoder digitally encodes to 8 bits, and the rate of sampling meets the CCITT standard.
   a. Determine the bit rate of the T1 carrier.
   b. What baseband bandwidth is required to transmit this signal if the minimum bandwidth is half the channel capacity.
   c. Why does each voice channel require bandwidth limiting?
   d. Give three advantages of PCM over the present analog carrier system.
   e. If high fidelity music is to be encoded over each of the 24 channels, what changes would need to be made in the system? Draw a block diagram of a suitable transmitting section, clearly marking the signal frequencies and/or bit rate at each input/output.

7. a. Show that the rms quantization noise of the quantizing error voltage shown in Fig. 6-6(c) is $a/2\sqrt{3}$. For ease of calculation, assume some time periodicity T for the triangular waveform.
   b. For an n digit code, the number of available quantized signal levels will be 2^n. What is the peak signal amplitude available if the step size is "a"?
   c. Obtain the expression for the peak signal/rms, or quantization to noise ratio. Derive (Eq. 6-2) by converting the expression obtained to decibels.

8. What is the peak signal/rms, or quantization to noise ratio for a
   a. 7-bit binary code
   b. 8-bit binary code
   c. 16-bit binary code

9. a. Define aperture time. What effect does a sample hold circuit have on the aperture time when placed ahead of an A/D converter?
   b. A 15 kHz sinusoidal signal is to be digitized to a resolution of 12 bits. What aperture time must be used to give less than 1 bit of error?
   c. A 10 V peak-to-peak signal that changes at a maximum rate of 0.1 V/ms is to be digitized to a resolution of 8 bits. What aperture time must be used to give less than 1 bit of error?

10. a. What is companding, and why is it employed in voice communications?
    b. Why is companding not used when transmitting data?

   c. What type of mathematical characteristics does a compander generally have?

11. a. Why is a code that has a maximum number of 1's near the zero voltage level used in D_4/D_5 channel banks?
    b. Illustrate why 13 bits are required to represent a linear digital version of the 15 segment mu-law curve that is illustrated in Figs. 6-11 and 6-13.

12. Find the absolute voltage, signs, segment, step, and level for the code word 10100101 on the 9002B Lenkurt channel bank.

13. Plot the mu law for $\mu = 255$, and the A law for $A = 87.6$. Use linear graph paper and assume maximum input and output signal voltages of 1.

14. State the advantages and disadvantages of the (i) parallel, (ii) serial, and (iii) hybrid encoders.

15. In order to distinguish 256 levels in a voice channel, determine the minimal frequencies of the various circuit components of the encoder listed in the following table.

	Parallel Fig. 5-16	Serial Fig. 5-17	Non-Uniform Serial Fig. 5-20(a)	Hybrid Fig. 5-21
f_s				
Seq of Programmed Register Clock Frequency				
Up/Down Counter				
Flip-Flops and Gates				

16. a. Describe the basic theory behind D/A converters.
    b. What is the advantage of the R-2R ladder decoder over the weighted current source decoder of Fig. 6-23?

17. a. What is a codec?
    b. State two advantages of a single channel codec versus one codec shared by all channels.
    c. What control word should be fed into the INTEL 2910A Codec if both the receive and transmit ends are to be programmed for the time slot 23?
    d. What determines whether a frame is a non-signaling or a signaling frame? What type of information is applied to the SIGx input, and where does it appear in the output bit stream?

18. Describe what is meant by the following terms:
    a. frame
    b. multiframe
    c. signaling
    d. channel

19. a. Obtain the framing sequence (s) bits from Fig. 6-31.
    b. Dialing pulses from a telephone handset occur at a nominal rate of 10 pulses/s. If the bandwidth of these square wave pulses is 40 Hz, what bit rate do these pulses represent? Does this put a severe information load on the signaling channel?
    c. Describe how framing synchronization is obtained on the Lenkurt 9002B system.

20. a. Describe the formats of the NRTZ and RTZ signals. What is the advantage of the later format?
    b. What is the AMI code? Why is it employed on a line rather than the RTZ operation?
    c. Describe the North American digital multiplex hierarchy.

21. A message is sent using five pulses. The first pulse is restricted to two amplitude levels while the remaining pulses can assume any integer voltage level between -4 and $+4$ (inclusive).
    a. How many different messages can be sent using a single group of five pulses?
    b. If the message is transmitted as a binary sequence, what is the minimum information rate and transmission bandwidth (assuming no redundancy is added to the message) if the maximum transmission time is T seconds?

22. a. Determine the maximum information capacity of a 2-level, 3 kHz bandwidth system.
    b. If the number of levels is increased to three, what bandwidth would be required for the same information capacity?
    c. Which of the two systems, (a) or (b), is most immune to noise?
    d. How long would it take either of these systems to transmit a 10,000 bit message?

23. What is the information capacity of a 3 kHz bandwidth channel which has a 34 dB signal-to-noise ratio?

24. a. How many bits are needed to represent a single letter in the English language which consists of a total of 26 letters. Assume equal likelihood of occurrence for each letter.
    b. How many levels must each of two pulses have in order to obtain the minimal number of combinations that are suggested in (a).
    c. How many levels must each of three pulses have in order to obtain the minimal number of combinations that are suggested in (b).

25. A shaft position encoder is to have a resolution of at least 1°. If it can rotate at a rate of 1200 rpm, what bit rate must be transmitted if it is digitally encoded? Why is the gray code the most suitable for such an application?

26. A noise signal ranging up to 2 V peak-to-peak in amplitude and containing significant components up to a frequency of 15 kHz, is to be transmitted using PCM. If changes in amplitude of $^1/_{64}$ V are to be detected, what must the information capacity of the channel be?

27. The spatial frequency response of the eye, that is the ability of the eye to resolve detail, can be determined by placing a grating pattern on a screen. The screen should have luminance perpendicular to the bars and should be modulating sinusoidally about a fixed mean level. Fig. P6-27 shows the contrast sensitivity:

$$\left(\frac{\text{twice the mean luminance}}{\text{maximum luminance} - \text{minimum luminance}}\right)$$

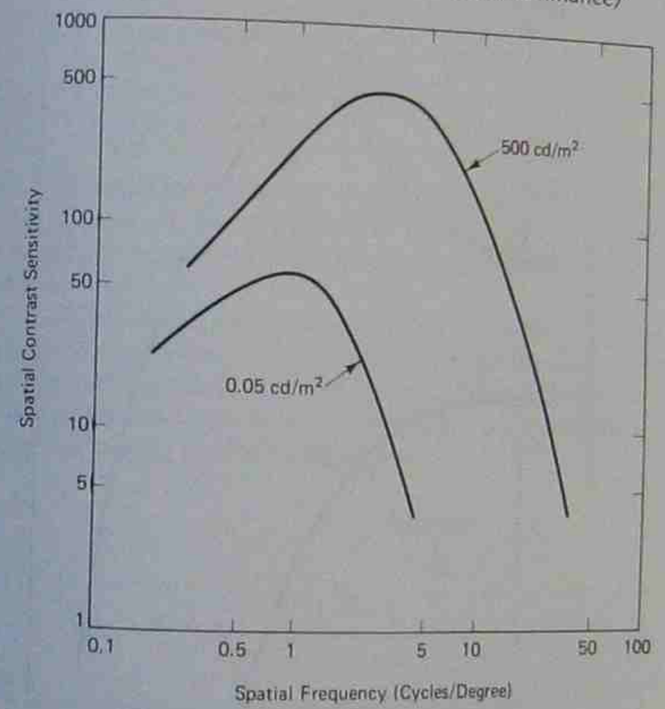


Fig. P6-27 Contrast Sensitivity of the Eye for Sine Wave Gratings. (After F. W. Campbell and J. G. Robson, "Application of Fourier Analysis to the Visibility of Gratings," *J. Physiology*, 197, 1968: pp. 551–566.)

of the eye for a sinewave grating. For a typical bright CRT display (500 cd/m²), the eye has a peak sensitivity at about 3–4 cycles/degrees with a fall-off in sensitivity at the higher and lower frequencies.

A facsimile of an A4 page (210 × 297 mm area) of written material is to be transmitted by low-speed scanning and digital transmission over a 3.4 kHz telephone line. The facsimile equipment is to produce the A4 page, and will be read from a distance of 40 cm.

a. Based on the eye response, what is the cut-off bandwidth? If the threshold or visibility is taken to be 30 cycles/degree, how many samples per degree would be required?
b. From a reading distance of 40 cm, what distance along the page does an arc of one degree subtend?
c. For a black and white document, only one bit per sample is required for reproduction without shades of grey. How many bits of information are required for every mm of the document?
d. For an A4 page, how many bits are required.
e. For a 1200 baud line (Bell system 202), how long does it take to transmit an A4 page, assuming 1 bit/baud.

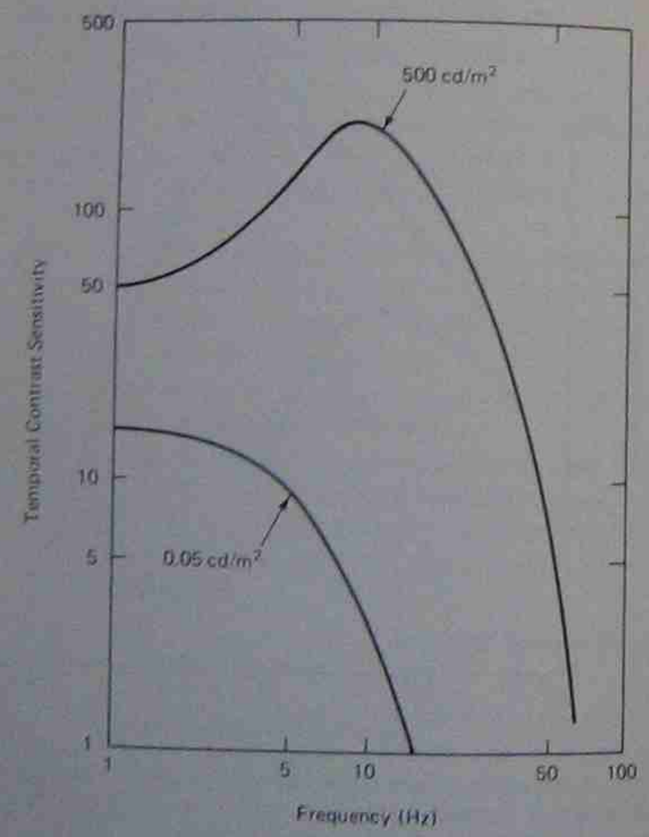


Fig. P6-28 Flicker sensitivity of the eye. (After H. De Lange, "Research into the Dynamic Nature of the Human Fovea-Cortex Systems with Intermittent and Modulated Light I, Attenuation Characteristics with White and Colored Light," *J. Opt. Soc. Am.*, vol. 48, No. 11, 1958: pp. 777–784.)

28. The temporal frequency response of the eye, or the ability of the eye to detect the flicker of a fluctuating light source, can be determined by observing a point source of light which is sinusoidally fluctuating about a mean value of luminence. Fig. P6-28 shows the temporal contrast sensitivity (twice the mean luminance/(max. luminance – min. luminance)) of the eye. The temporal contrast sensitivity indicates how frequently a picture must be projected onto a screen to produce little sensation of flicker. As can be seen from the response, the eye acts as a low-pass filter for temporal frequencies and peaks at about 8 Hz for 500 cd/m² of mean luminance.

The following table lists some CCIR-calculated values for the sampling frequency in each of the three dimensions of a 525-line television image viewed at six times the picture height.

Angle subtended by the eye (width × height)	12.7° × 9.5°
Horizontal spatial sampling frequency (samples/degree)	35
Vertical spatial sampling frequency (samples/degree)	51
Temporal sampling frequency (samples/second)	60

Considering the spatial and flicker sensitivity of the eye, (Figs. P6-27 and P6-28, respectively), are the spatial and temporal sampling frequencies adequate for a typical bright display of 500 cd/m²? Explain.

If 256 shades of grey in a black and white television set are to be distinguishable, what bit rate is required to transmit the 525 systems as outlined? What approximate vestigial bandwidth would be required to transmit this bit rate? How does this compare to the present analog of 5 MHz bandwidth?

7

Effects of Noise and Other Distortions on Digital Transmission

7–1 INTRODUCTION

Noise and distortion are always present in any communication system. All extraneous signals appearing at the channel output that are not due to the input signal are considered noise. This noise is due to such factors as thermal agitation, radio frequency interference, crosstalk, and erratic switch contacts. Noise can be broken down into the following categories:

- Quantization
- Thermal
- Transient
- Phase jitter

Keep in mind that the redundancy of human speech and the adaptability of the human ear make noise and distortion less of a problem in voice communication than in data transmission. If a noise burst makes a word indistinguishable, the listener can guess what the word should have been based on the context of the conversation. Machines are unable to do this; therefore, tight controls must be kept on noise levels to minimize errors. When transmitting data, few errors are generally permitted.

7–2 QUANTIZATION NOISE

Quantization noise is the result of the difference between the signal presented to a Codec and its equivalent quantized value. As suggested by Fig. 6-6, the root-mean-square (rms) quantization noise for uniform quantizing can be obtained by finding the rms voltage of a triangular waveform with a peak-to-peak value of step size "a." Figure 7-1 illustrates the triangular waveform of the quantizing error, assuming a period T.

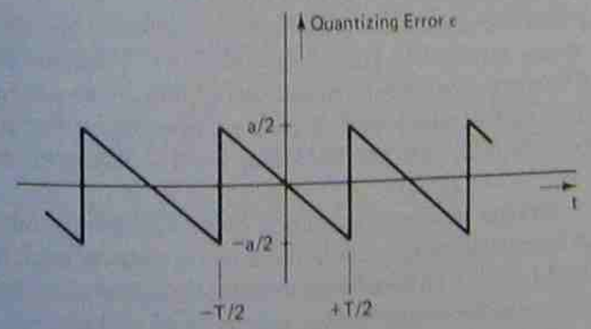


Fig. 7-1 Quantization Noise Waveform

The rms quantization noise is thus equal to:

$$\epsilon_{rms} = \sqrt{\frac{1}{T}\int_{-T/2}^{T/2}\left(-\frac{a}{T}t\right)^2 dt}$$

$$= \frac{a}{2\sqrt{3}} \tag{7-1}$$

Assume that there are M available quantized levels for a single pulse. Thus, the peak signal-to-rms quantization noise will be

$$\frac{V_S}{V_N} = \frac{Ma}{a/2\sqrt{3}} = 2\sqrt{3}\,M \tag{7-2}$$

Expressed as a power ratio, this is:

$$\frac{S}{N} = (2\sqrt{3}\,M)^2 = 12\,M^2 \tag{7-3}$$

or, in dB

$$\left(\frac{S}{N}\right)_{dB} = 10 \log 12\,M^2$$
$$= 10 \log 12 + 20 \log M$$
$$= 10.8 + 20 \log M$$

If the single pulse is represented by a group of n pulses, each having m levels, and if M is replaced by m^n, the following is obtained:

$$\left(\frac{S}{N}\right)_{dB} = 10.8 + 20 \log m^n$$
$$= 10.8 + 20\,n \log m \tag{7-4}$$

For the binary code $m = 2$, Eq. (6-2) is obtained.

For practical situations, the actual average power is always less than the peak power assumed in Eqs. (7-3) and (7-4). Thus, an expression developed for the average power would be helpful. As an example, consider a simple case in which each level is equiprobable. For an m-level system, the probability that any particular level is being considered is $1/m$.

For illustration purposes, picture a cubic die which has 6 faces. Since each face has an equal probability of being thrown, each face has a probability of 1 out of 6 or $1/6$ of being thrown. If, for instance, a voltage is always at a level a, the average power would be a^2. If, however, the voltage was at that level only $1/4$ of the time on the average, the average power would be $a^2/4$.

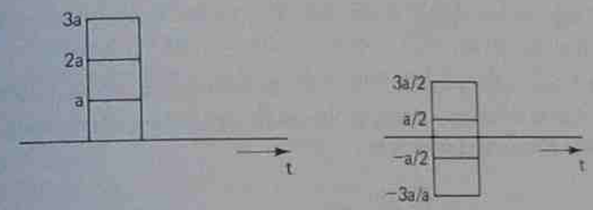
Consider the average power in a single-polarity pulse stream, positive only or negative only, as shown in Fig. 7-2(a). Consider also the average power in a bipolar pulse stream as shown in Fig. 7-2(b), each having four equally likely levels, and level spacings of "a." For the case of the single, four-level pulse stream, the average power is given by

$$S = \frac{1}{4}0^2 + \frac{1}{4}a^2 + \frac{1}{4}(2a)^2 + \frac{1}{4}(3a)^2$$
$$= 3.5a^2 \text{ watts} \tag{7-5}$$

For the four-level bipolar pulse stream, the average power is given by

$$S = \frac{1}{4}\left(\frac{-3a}{2}\right)^2 + \frac{1}{4}\left(\frac{-a}{2}\right)^2 + \frac{1}{4}\left(\frac{a}{2}\right)^2 + \frac{1}{4}\left(\frac{3a}{2}\right)^2$$
$$= 1.25\,a^2 \text{ watts} \tag{7-6}$$

As can be seen, the bipolar pulse stream transmits a message with substantially less power than the single polarity pulse stream. Because of this advantage and the fact that bipolar pulse streams can be designed without dc components, this examination will concentrate on the bipolar case only.



(a) Single Polarity Pulse (b) Bipolar Pulse

Fig. 7-2 Four-Level Pulse Streams

In general, for the bipolar pulse stream of M levels shown in Fig. 7-3, the average power will be

$$S = \frac{2}{M}\left[\left(\frac{a}{2}\right)^2 + \left(\frac{3}{2}a\right)^2 + \cdots \left(\frac{M-1}{2}a\right)^2\right]$$
$$= \frac{a^2}{2M}\left[1^2 + 3^2 + \cdots (M-1)^2\right]$$

The series in the brackets can be expressed by

$$\frac{M/2\,(M-1)\,(M+1)}{3}$$

$$\text{Thus} \qquad S = \frac{a^2}{12}(M^2 - 1) \tag{7-7}$$

The average signal-to-quantization noise power is thus reduced to

$$\frac{S}{N} = \frac{\frac{a^2}{12}(M^2 - 1)}{a^2/12} = M^2 - 1$$

or in dB

$$\left(\frac{S}{N}\right)_{dB} = 10 \log (M^2 - 1) \tag{7-8}$$

If the single pulse is represented by a group of n pulses, each having m levels, and M is replaced by m^n, the following is obtained:

$$\left(\frac{S}{N}\right)_{dB} = 10 \log (m^{2n} - 1) \tag{7-9}$$

For a binary code where $m = 2$, the following is obtained:

$$\left(\frac{S}{N}\right)_{dB} = 10 \log (2^{2n} - 1) \tag{7-10}$$

which for 8 bits reduces to 48.2 dB. This represents about a 10 dB reduction in S/N ratio when compared to the peak signal-to-quantization noise power under the same conditions.

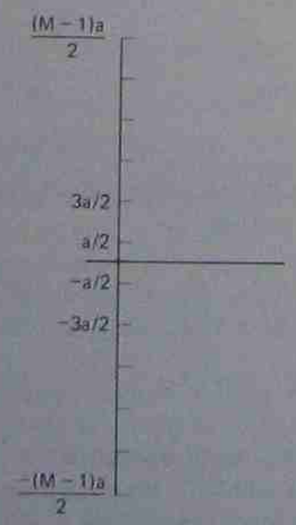


Fig. 7-3 Bipolar Pulse Having M Levels

7-3 EFFECTS OF GAUSSIAN-TYPE NOISE ON DIGITAL TRANSMISSION

Noise entering or contributed by a system results in errors in the detected signal. The error rate is dependent upon both the signal and noise amplitudes, as well as upon the characteristics of the noise. Noise containing large amplitude spikes causes more errors in PAM systems than, for instance, thermal noise containing the same average power.

For many applications, noise is considered Gaussian, even though in practice this is only an approximate assumption. Noise is Gaussian if it can be considered random and the result of many independent overlapping current or voltage pulses, such as the random motion of free electrons making up thermal noise in a resistor.

Gaussian noise has a probability density function which follows the familiar bell curve illustrated in Fig. 7-4. The larger amplitudes have less likelihood of occurrence than the lower ones. For example, voltage near zero volts has a greater probability of occurrence than the larger positive or negative voltages. When analyzing probable signals, the terms *variance* and *standard deviation* are frequently employed. The standard deviation, σ, is a measure of the spread of observed values. In electronics, it is called the rms voltage of the ac portion of the signal, or the rms signal voltage with the dc removed. The square of the deviation, that is, σ^2, is called the variance. The larger the variance, the larger the spread, and the more squat the density function, as shown in Fig. 7-4.

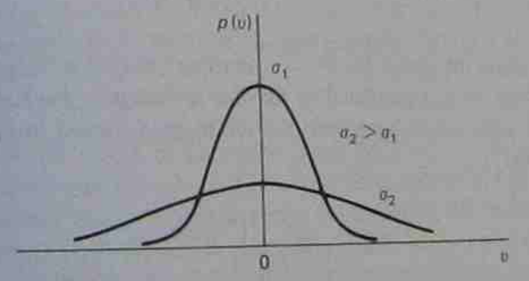


Fig. 7-4 Probability Density Function for Gaussian Noise Containing No dc

These curves can be expressed analytically as

$$p(v) = \frac{1}{\sqrt{2\pi}\sigma} e^{-v^2/2\sigma^2} \tag{7-11}$$

where $p(v)$ is the probability density of the voltage v which, when integrated over an interval, gives the probability or the percentage chance that the voltage exists in the interval.

σ is the standard deviation
σ^2 is the variance

To find the probability, for instance, that the voltage is greater than A volts, the shaded area of Fig. 7-5 must be obtained. Obviously, the area under the entire curve must be unity (1) since the probability that the voltage lies somewhere between $-\infty$ and $+\infty$ is 1, or 100%.

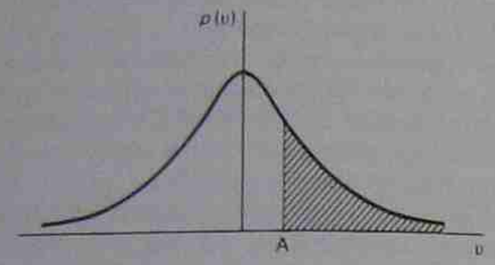


Fig. 7-5 Shaded Area Indicating the Probability that $v > A$.

Since it is difficult to integrate Eq. (7-11), tables and graphs have been developed to aid in determining the areas under the Gaussian density function. To analytically determine the probability that the voltage lies between 0 and $k\sigma$ volts, where k is some constant, the following integral must be solved.

$$prob\ (0 < v \leq k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^{k\sigma} e^{-v^2/2\sigma^2}\, dv \qquad (7\text{-}12)$$

This equation integrates Eq. (7-11) over the interval of interest.

When using a graphical or tabular technique, the function $\Phi(k)$ is generally used which represents the white area shown in Fig. 7-6 and is defined as:

$$\Phi(k) = \frac{1}{\sqrt{2\pi}} \int_0^{k} e^{-u^2/2}\, du \qquad (7\text{-}13)$$

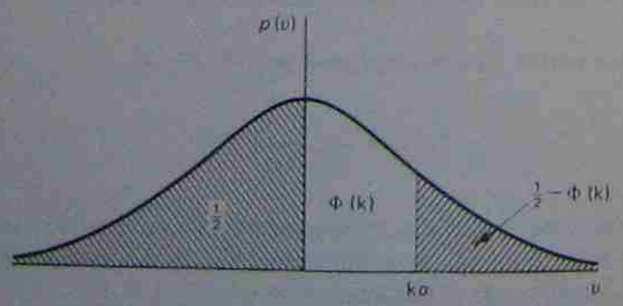


Fig. 7-6 Gaussian Density Function

Of special interest in this case is the tail probability, which represents the probability that the noise voltage exceeds $k\sigma$ volts. Since the total area under the density curve is unity, and symmetry indicates that the area under either side is one-half, the tail probability must be given by:

$$prob\ (v > k\sigma) = 1 - {}^1/_2 - \Phi(k)$$
$$= {}^1/_2 - \Phi(k) \qquad (7\text{-}14)$$

Figure 7-7 gives a plot of $^1/_2 - \Phi(k)$. For $k \gg 1$, this can be approximated by

$$\frac{1}{2} - \Phi(k) = \frac{1}{\sqrt{2\pi}k} e^{-k^2/2} \qquad (7\text{-}15)$$

Example 7-1

What is the probability that a 0.5 volt rms Gaussian noise signal exceeds a magnitude of 1 volt?

Solution

The standard deviation is given as 0.5 volts. In this case, the area in the two tails shown in Fig. 7-8 is to be found.

$$\text{Since} \quad k\sigma = 1$$
$$k = \frac{1}{\sigma} = \frac{1}{0.5} = 2$$

For $k = 2$, the area under each tail from Fig. 7-7 is $2.3 \times 10^{-2} = 0.023$. Therefore, the probability that the voltage exceeds 1 volt is $2 \times 0.023 = 0.046$ or 4.6%.

Consider the effect of Gaussian noise on a bipolar, or two-level, signal with permissible levels of $A/2$ or $-A/2$ as shown in Fig. 7-9. The decoder assumes that a positive pulse is present if the instantaneous signal exceeds 0 volts and that a negative pulse is present if the signal drops below 0 volts. Assume that either level is equiprobable; that is, each level has a 0.5 probability.

The decoder in Fig. 7-9 responds to the amplitude of the signal-plus-noise at the mid-pulse periods. For the pulse train shown, two errors occur.

To determine the error rate of such a pulse system, assume that Gaussian noise will ride on the two levels, $-A/2$ and $A/2$, as illustrated in Fig. 7-10.

What is the probability that the decision will be in error? When the transmitted voltage is $-A/2$ volts, an error will occur whenever the voltage exceeds the decision threshold level of 0 volts, represented by the shaded area of Fig. 7-10(b). When the transmitted voltage is $A/2$ volts, an

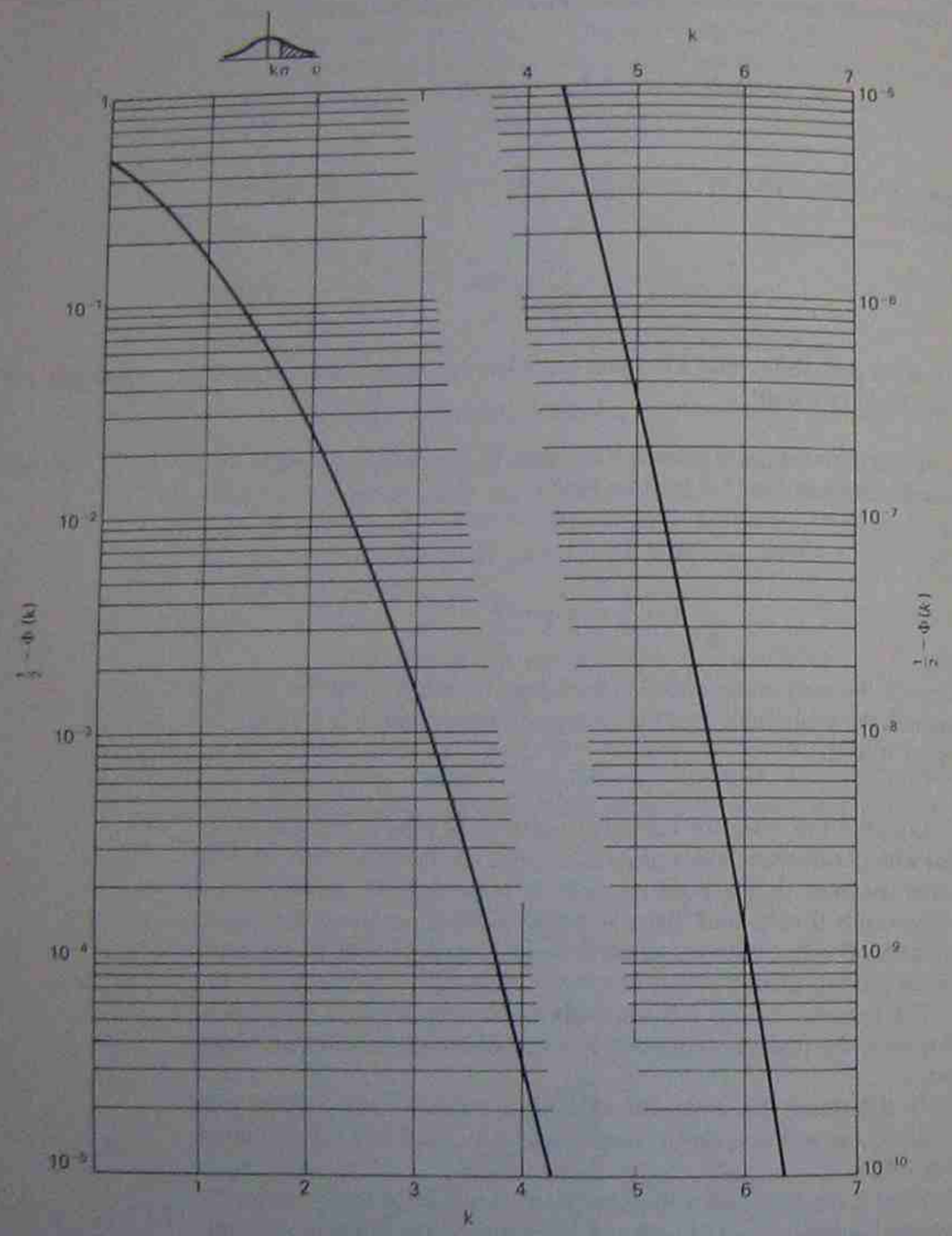


Fig. 7-7 Tail Area from $k\sigma$ to ∞ of the Gaussian Probability Density Function

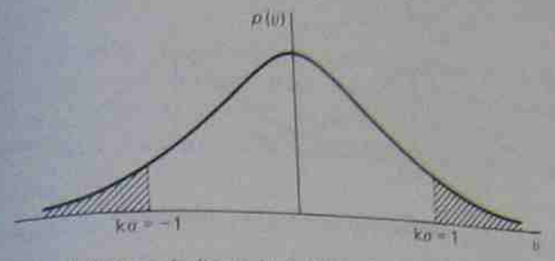


Fig. 7-8 Shaded Areas Indicate Probability that the Voltage Exceeds 1 Volt

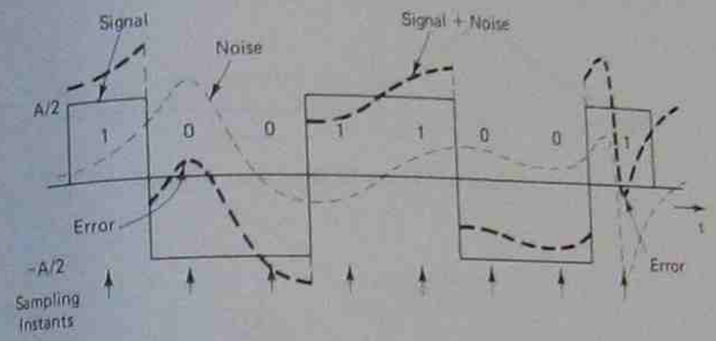


Fig. 7-9 Effect of Noise on Two-Level Pulses

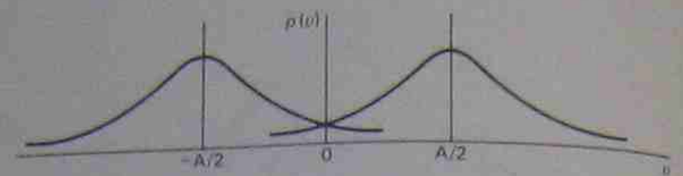
error will occur whenever the voltage drops *below* the decision threshold of 0 volts. This is represented by the shaded area of Fig. 7-10(c). Since either signal is equally probable, and the Gaussian distribution is symmetrical, the probability of error is given by:

probability of signal at $-A/2$ × tail probability
+ probability of signal at $A/2$ × tail probability
= ½ × tail probability + ½ × tail probability
= 2 × ½ × tail probability
= tail probability

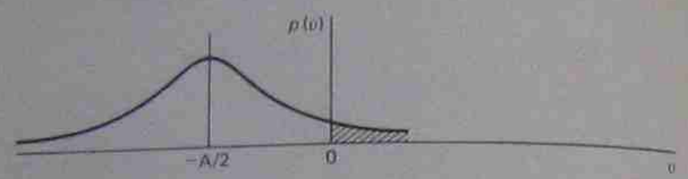
The equivalent tail probability has been redrawn in Fig. 7-11.

Since $k\sigma = A/2$, or $k = A/2\sigma$, the error probability can be determined as a function of peak signal $(A/2)$ to rms noise voltage (σ). Thus, Table 7-1 can be obtained with the aid of Fig. 7-7.

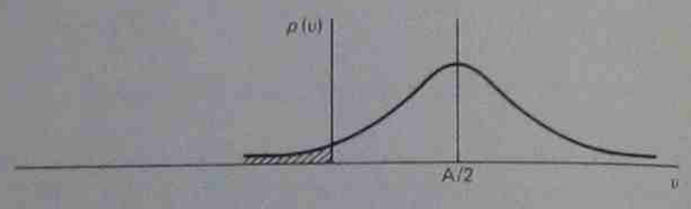
Figure 7-12 is a graphical representation of the error rate as a function of the peak signal-to-rms noise ratio. As can be seen from the graph, the error rate is very sensitive to the S/N ratio. For a 1 decibel change in S/N in the 16 dB range, the error rate changes by a factor of about 100.



(a) Gaussian Noise on a Two-Level System



(b) Probability of Error when Signal is at −A/2 Volts



(c) Probability of Error when Signal is at A/2 Volts

Fig. 7-10 Determination of Error Probability on a Bipolar Pulse Stream

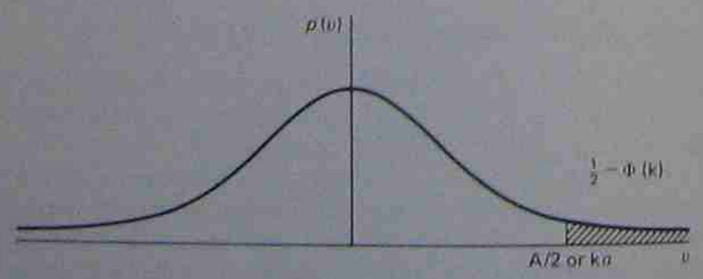


Fig. 7-11 Error Probability Shown by Shaded Area

7-4 TRANSIENT NOISE

The four major types of sporadic, or transient, disturbance in a communication system are impulse noise, gain hits, drop-outs, and phase hits. These noise types account for a substantial portion of errors which occur when digital signals pass through a switched network.

Impulse Noise

Impulse noise consists of irregular noise spikes of relatively high amplitude that tends to occur in sporadic groupings, or bursts. This type of noise is not overly obnoxious to listeners on voice lines, but it can seriously in-

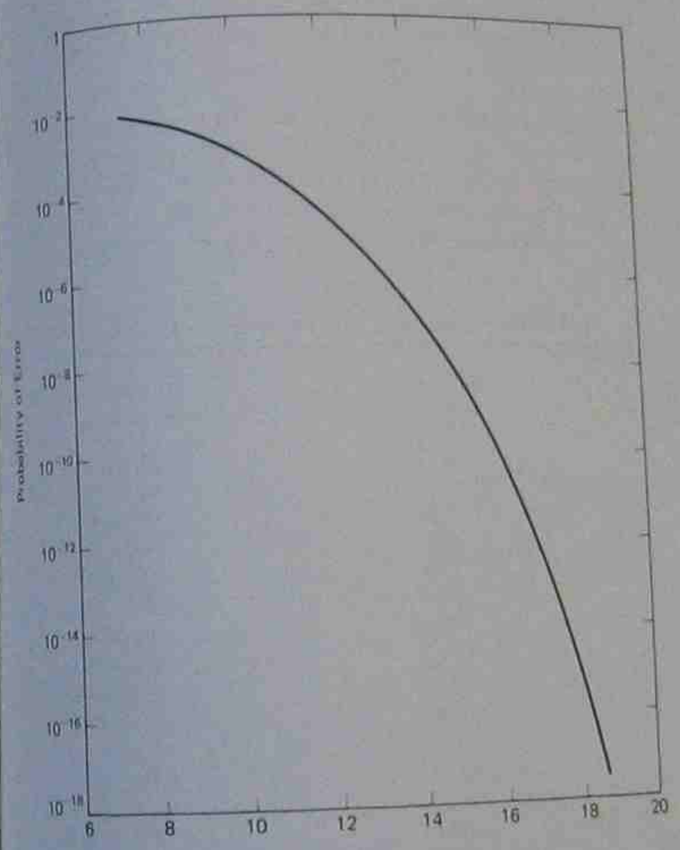


Fig. 7-12 Error Rate as a Function of S/N in dB

crease error rates on digital systems since it can cause loss of synchronization.

Impulse noise comes from natural causes such as lightning, and from man-made causes such as ignition systems and power lines. It can also come from telephone switching networks as a result of dialing, and of the action of electromechanical switches and relays. (Refer back to Sec. 2-7.1 and 2-7.2 for more details.)

Gain Hits

Gain hits are sudden decreases or increases in recent signal level that cause changes of less than 12 dB in the nominal receive signal level. Because such changes in signal level may appear as data, especially in systems that use amplitude modulation, the error rate of the systems can be significantly increased. Gain hits last at least 4 ms and can go on for hours. In the telephone industry, no more than 8 gain hits exceeding 3 dB from the nominal receive level should be detected over a 15 min interval.

Dropouts

Dropouts can be compared to severe gain hits. They are defined as signal decreases that are greater than 12 dB and last longer than 4 ms. Because synchronization can be lost under such circumstances, more data can be lost while trying to resynchronize and automatically equalize. Over a period of 30 min, no more than 1 dropout should occur over the telephone system.

Table 7-1 Error Rate Due to Gaussian Noise

k or $\frac{A}{2\sigma}$	20 log $\frac{A}{2\sigma}$	Error rate ($\frac{1}{2} - \Phi(k)$)
0	$-\infty$	0.5000
0.5	−6.02	0.3085
1	0	0.1587
1.5	3.52	0.0668
2	6.02	0.0227
2.5	7.96	0.0062
3	9.54	0.00135
3.5	10.88	0.00023
4	12.04	0.00003
4.5	13.06	0.00000355
5	13.98	2.97×10^{-7}
5.5	14.81	1.96×10^{-8}
6	15.56	1×10^{-9}
6.5	16.26	4.107×10^{-11}
7	16.90	1.305×10^{-12}
7.5	17.50	3.246×10^{-17}
8	18.06	6.315×10^{-16}
8.5	18.59	9.607×10^{-18}
9	19.08	1.142×10^{-19}

Phase Hits

A phase hit is a sudden change in the received signal phase, and thus frequency, which lasts for durations exceeding 4 ms. Phase lists can appear as false data in PSK and FSK systems. Phase hits occur: (1) when power supplies are changed; (2) when alternate transmission facilities with different propagation times are used; and (3) when the optimum receiver in a space diversity microwave system is used.

Phase Jitter

Variations in the phase of the carrier signal cause a smearing in the zero-crossing of received data pulses. For severe jitter, this may result in pulses moving into time slots allocated for neighbouring data pulses.

Phase jitter is caused by the 20 Hz ringing currents in adjacent channels, by the 60 Hz ripple and associated harmonics in the power supplies, and by insufficient filtering of image sidebands. Over the telephone system, phase jitter should not exceed 10° between 20 Hz and 300 Hz and 15° between 4 Hz and 20 Hz.

Special equipment, such as the Hewlett Packard 4943A Transmission Impairment Measuring Set, is available to measure transient noise and phase jitter. A block diagram of the transient noise measurement is illustrated in Fig. 7-13. A block diagram of the phase jitter measurement is shown in Fig. 7-14(a).

Because phase jitter is a low frequency phenomenon, the front-end, low-pass filter will pass the jitter and, at the same time, reduce the effects of noise and other interference. The voltage controlled oscillator (VCO) is phase-locked to the long term average of the detected signal. It provides

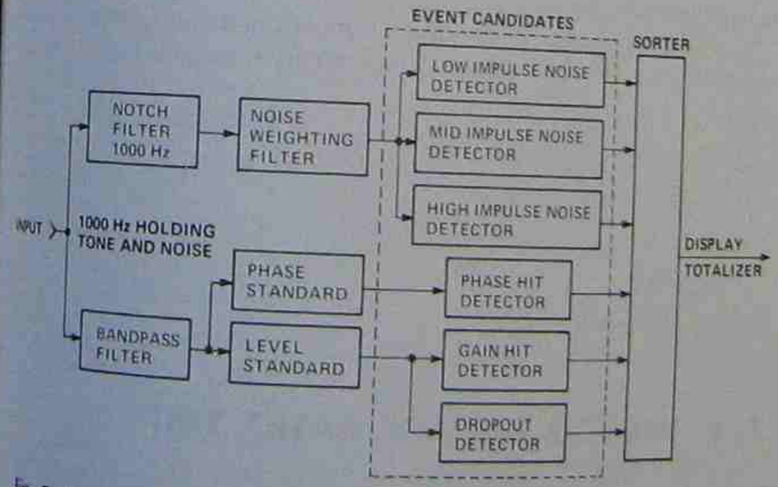
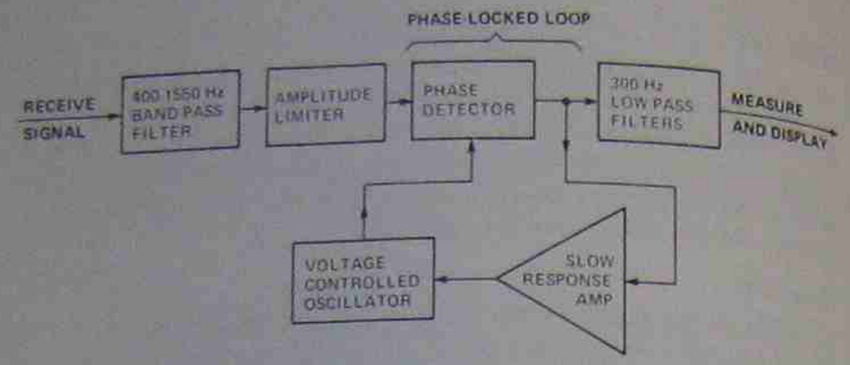
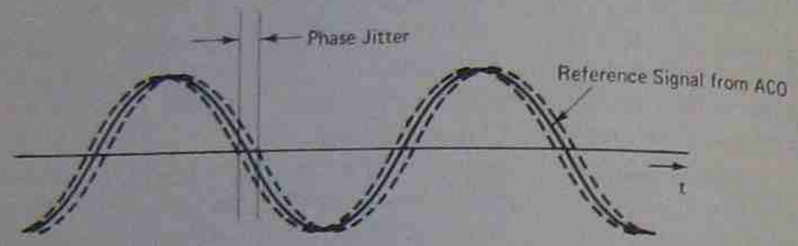


Fig. 7-13 Impulse Noise, Gain and Phase Hits, and Dropout Measurement (Courtesy of Hewlett Packard)



(a) Phase Jitter Measurement (Courtesy of Hewlett Packard)



(b) Phase Jitter on Holding Tone

Fig. 7-14

the reference signal phase of Fig. 7-14(b). Any phase deviation from this average will be detected by the phase detection and displayed. The phase signal at the 300 Hz low-pass filter output is frequency limited to a band between 20 Hz and 300 Hz. Below 20 Hz, the phase-locked loop is able to track the phase deviation.

In addition to the types of noise just mentioned, many other impairments may also be present in the communication system. In brief, these are:

a. envelope and delay distortion,
b. intermodulation distortion,
c. crosstalk,
d. echo, and
e. frequency shift.

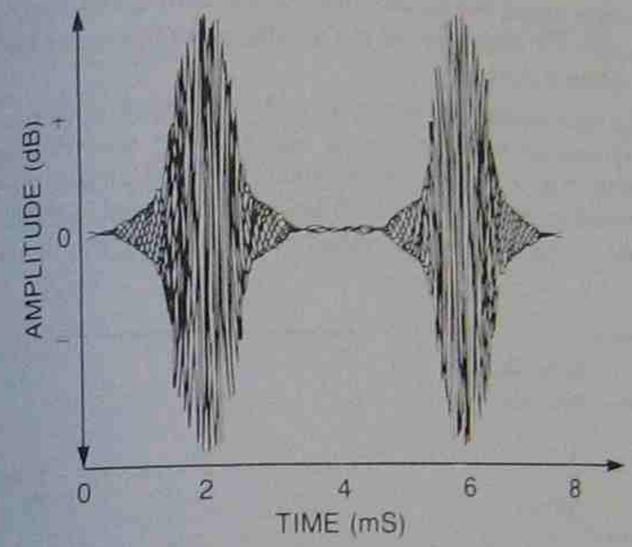
7–5 PEAK-TO-AVERAGE RATIO TEST

One measurement which indicates the fidelity of a channel is peak-to-average or (P/AR) measurement. It does not pinpoint the specific impairment which is causing problems, but it gives a good indication of the

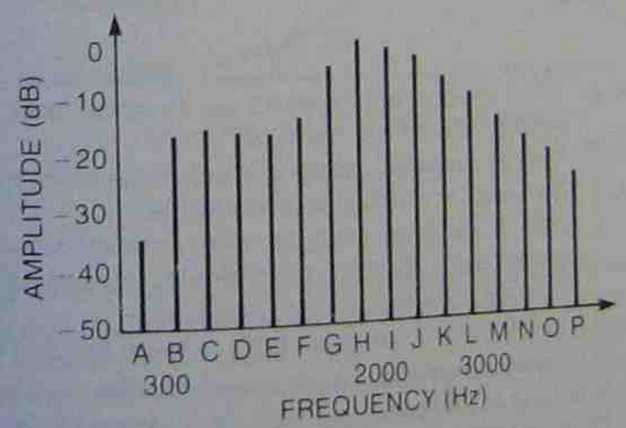
"health" of the system. If there is a decrease in the P/AR from one reading to another, there is a degradation in the system.

The P/AR is determined by applying a complex signal to the channel and then measuring the peak-to-average ratio at the receiver. This measurement is then compared to the P/AR value of 100 which indicates no signal degradation. Figure 7-15(a) shows a complex signal used to determine the P/AR. Figure 7-15(b) shows the frequency spectrum of the signal.

The P/AR values run in the range of 48 for a C_1-conditioned channel to 78 and 87 for a C_2- and C_4-conditioned channel.



(a) P/AR Transmit Signal Envelope



(b) P/AR Transmit Signal Frequency Spectrum

Fig. 7-15 P/AR Measurement Test Signal (Courtesy of Hewlett Packard)

PROBLEMS

1. a. Is quantization noise, in actual practice, a periodic triangular wave as illustrated in Fig. 7-1? Sketch the quantization noise for a sinusoidal signal.
   b. What is the peak signal-to-rms quantization ratio (in dB) for a 256-level system?
2. a. Find the average normalized power of a bipolar pulse train having five equally likely levels with spacings of "a."
   b. Find the average normalized power of a single polarity pulse train having five equally likely levels with spacings of "a."
   c. State the advantages of the bipolar pulse train over the single polarity pulse train.
3. If the peak voltage of a two-level and a multilevel system is fixed, what approximate noise penalty occurs in an m-level system as compared to a binary system? Figure P7-3 illustrates the levels permitted by each system, both having equal peak amplitudes of A volts. The penalty is determined by the ratio d_b/d_m.

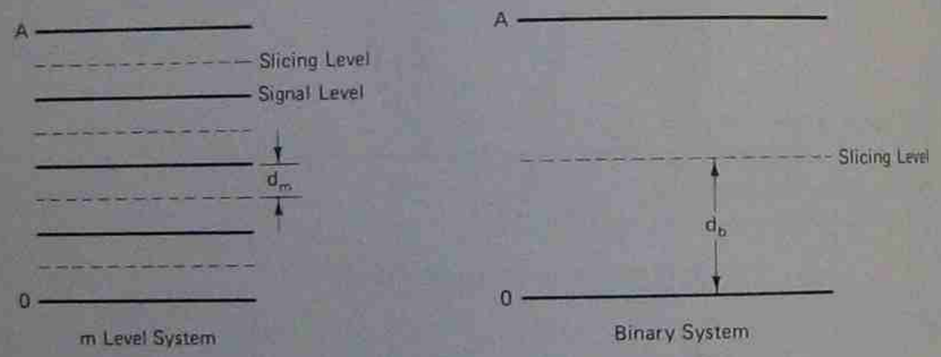


Fig. P7-3

4. a. What characteristics does Gaussian noise have, and under what conditions is it produced?
   b. Give an electronic concept of standard deviation and relate it to the Gaussian distribution characteristic.
   c. Why is Gaussian noise so often assumed when evaluating system noise characteristics? Would you consider the noise from a car distributor to be Gaussian?
   d. List some other significant types of noise in cable systems.
5. A two-level system (± 5 volts) experiences a Gaussian noise of 1 volt rms. For a 1-G-bit pulse train, how many bits on the average would be in error? When calculating the tail probability, use both Eqs. (7-15) and Fig. 7-7.

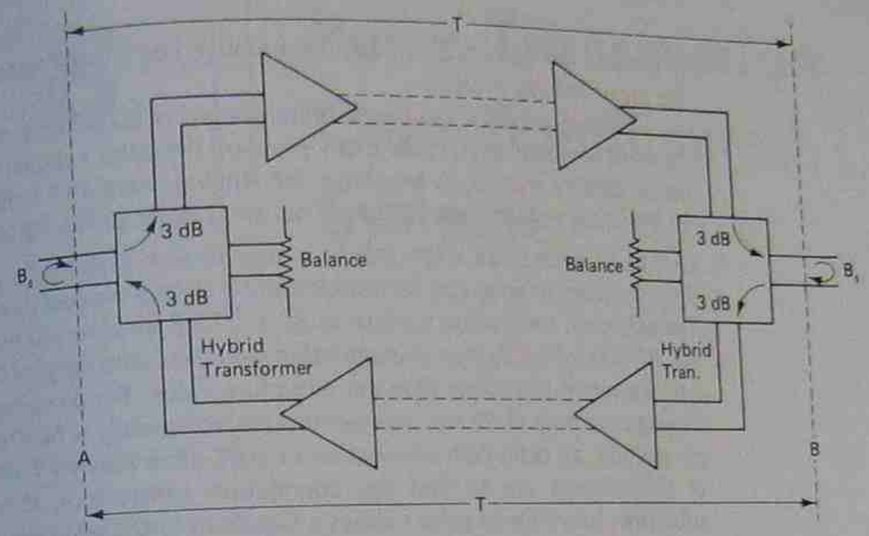


Fig. P7-6 (a)

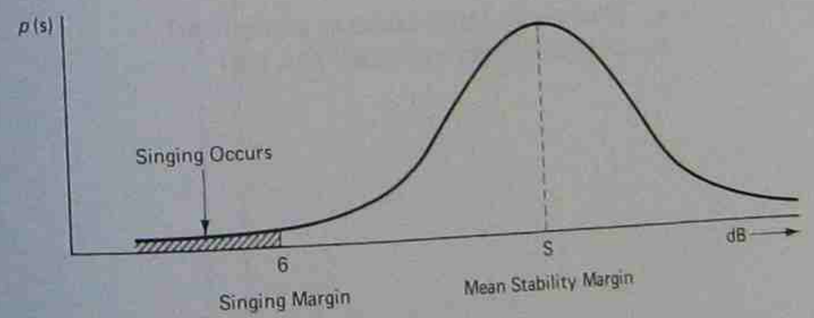


Fig. P7-6(b)

6. In order to pass signals in the same frequency band in both directions at once, hybrid transformers, as illustrated above, are used. B_s, which is known as the balance return loss, is the loss experienced due to mismatching. If the loss between the two wire points, A and B, is assumed to be T, the total loss around the loop is

$$\text{Loss} = 2\,(T + B_s)\ \text{dB}$$

In order to prevent the circuit from oscillating or singing, the loop loss should be greater than 6 dB.

$$\text{Thus} \quad 2\,(T + B_s) \geq 6 \text{ dB}$$

In general, the quantity $2T + 2B_s$ is symbolized by S and known as the *mean stability margin*.

If the distribution on the stability margin is Gaussian with a standard deviation of σ_s, how many standard deviations should the mean stability margin, S, be above the singing margin of 6 dB if 99.9% of the time the circuit should not sing? Refer to Fig. P7-6(b).

7. Due to an impedance mismatch in a transmission system, the reflected energy or echo can be troublesome if delays exceed 10 ms. The echopath attenuation is equal to $B_E + 2T$ dB as observed from Fig. P7-6(a), where B_E is an average value over the audio range of the balance return loss rather than the minimum value. For a one-way propagation time of 50 ms, representing approximately a 6000 mi connection, an echo path attenuation of 30.9 dB is needed if 50% of the subjects are to find the connection satisfactory. If the subscriber tolerance-to-echo follows a Gaussian distribution with a standard deviation of 2.5 dB, how many people will be satisfied by a loss of 33.4 dB? For longer distances, the attenuation necessary becomes excessive and echo suppressors are needed.
8. List the four major types of transient noise on a cable radio system, and illustrate with waveforms.
9. a. What are the major causes of phase jitter?
   b. Of what significance is the P/AR test?

8 Pulse Transmission Over Bandlimited Systems

8–1 INTRODUCTION

Since the signal spectrum of a pulse is large, bandlimiting can have a drastic effect on the shape of the received pulse. Since any channel is bandlimited in practice, severe degradation of the detected signal can occur unless approximate precautions are observed. This chapter will consider the effects of a filter on the time response of a pulse.

8–2 THE UNIT IMPULSE

The unit impulse is often used to obtain the frequency response of a network. The unit impulse, $\delta(t)$, is also called the Dirac impulse or delta function. It is a function with the following two properties:

(1) It has value when its argument is zero, otherwise it has a zero value; i.e.,

$$\delta(t) = \begin{cases} 0 & t \neq 0 \\ \infty & t = 0 \end{cases} \qquad (8\text{-}1(a))$$

(2) The area or integral of the function over all time is unity; i.e.,

$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1 \qquad (8\text{-}1(b))$$

More precisely, it can be stated that $\delta(t)$ has the property

$$\int_{-\infty}^{\infty} f(t)\, \delta(t - t_0)\, dt = f(t_0) \qquad (8\text{-}1(c))$$

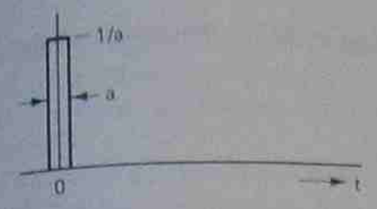
where $f(t)$ is a function continuous at t_0.

In other words, an impulse $\delta(t - t_0)$ is located at $t = t_0$ and is zero everywhere else. It has a unit area concentrated at the point t_0. One such function is illustrated in Fig. 8-1(a) and (c).

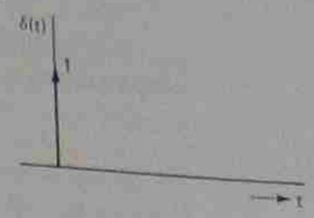
The graphical representation of $A\delta(t - t_0)$, or an impulse of weight A, is shown in Fig. 8-2(a).

The frequency spectrum of such an impulse can be found by taking the Fourier Transform. The result is a continuous spectrum of amplitude A, as shown in Fig. 8-2(b). Applying such an impulse to a network readily provides the frequency response of the system, since the input effectively contains all frequencies of equal amplitude.

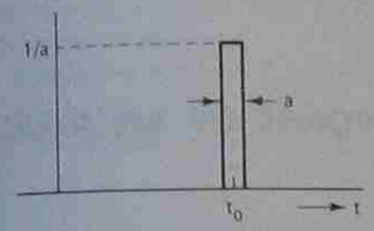
As an initial study of pulse propagation through a bandlimited transmission path, a very pecular impulse wave train is usually used. It consists of a periodic string of impulses spaced by the sampling period T. The in-



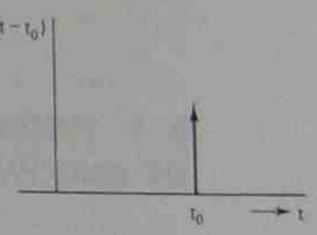
(a) The Unit Impulse at the Origin as $a \to 0$



(b) Graphical Representation of the Unit Impulse at the Origin

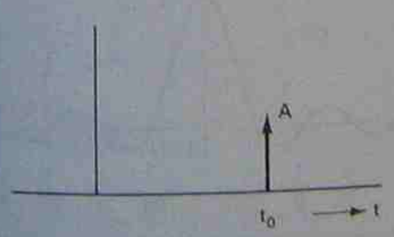


(c) The Unit Impulse at $t = t_0$ as $a \to 0$

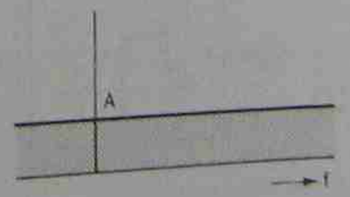


(d) Graphical Representation of the Unit Impulse at $t = t_0$

Fig. 8-1 The Unit Impulse



(a) $A\,\delta t(t - t_0)$



(b) Frequency Spectrum

Fig. 8-2 Graphical Representation of $A\delta(t - t_0)$ and Its Frequency Spectrum

verse of T, or $1/T$, is called the sampling frequency. Such an impulse wave train is illustrated in Fig. 8-3 and can be mathematically expressed as

$$\delta(t - nT) \qquad \text{where} \quad n \text{ is an integer} \tag{8-2}$$

This wavetrain will be used to examine the effects of bandpass filters on pulses.

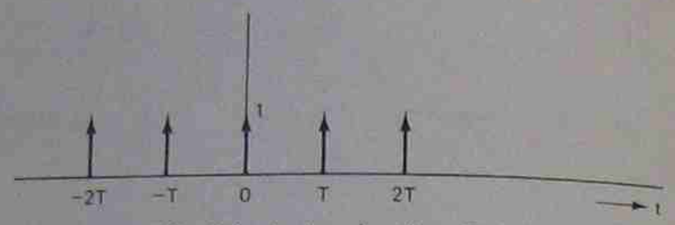
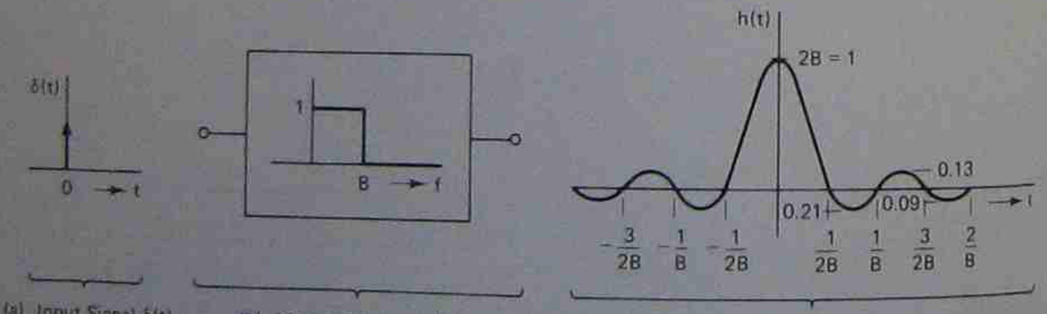


Fig. 8-3 An Impulse Wave Train

8–3 FREQUENCY RESPONSE OF AN IDEAL OR BRICKWALL FILTER

Visualize the system shown in Fig. 8-4. In this system, an impulse input of $\delta(t)$ is applied to a transmission channel which has a frequency response corresponding to that of an ideal, or brickwall, low-pass filter of bandwidth B. Although such a filter, and therefore such a response, is unrealizable due to the infinite roll-off of the filter skirts, the ideal gives the approximate response that would be realistically observed when a short pulse is applied to a bandlimited system.



(a) Input Signal $\delta(t)$ (b) Ideal or Brickwall Filter (c) Output Signal or Time Response

Fig. 8-4 Time Response of an Ideal or Brickwall Filter Due to Bandlimiting

As can be seen from Fig. 8-4(c), the result of bandlimiting an impulse signal is ringing, or time spreading, of the pulse.

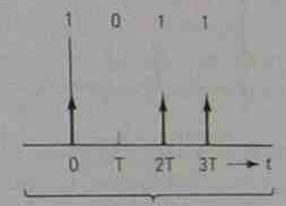
The output waveform is given by the expression

$$h(t) = 2B \text{ sinc } 2Bt = 2B\frac{\sin 2\pi Bt}{2\pi Bt} \tag{8-3}$$

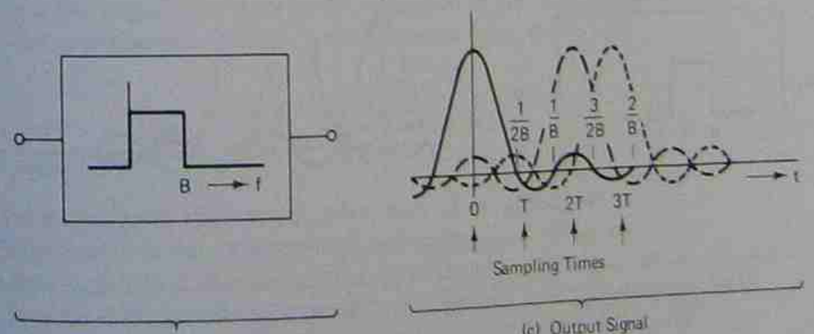
From this expression, it can be observed that the ringing tails decay at a rate of $1/t$ and that the amplitude passes through zero at points where $t = n/2B$ when n is an integer, since the sin function is zero whenever $2\pi Bt = n\pi$.

If a wavetrain of pulses is supplied to a bandlimited system, the tails of several pulses can overlap, thus interfering with the major pulse of interest at the detector. For the case illustrated in Fig. 8-5, ignoring any time delays through the system, spurious responses from the third and fourth impulses also appear at the sampling moment $t = 0$. This interference is commonly termed inter-symbol interference, (ISI). If by some method the rate of decay of the tail could be increased from the $1/t$ rate, the ISI would be reduced.

At sampling time $t = T$, the tails of the several pulses appear within the sampling slot, which results in an observed voltage instead of none at all. This ISI can seriously degrade the noise immunity of the system



(a) Input Impulse Wavetrain



(b) Ideal or Brickwall Filter (c) Output Signal

Fig. 8-5 Time Response of a 1011 Impulse Train after Bandlimiting

since the background ISI makes the system tolerant to a smaller amount of noise.

The effects of ISI can be minimized by:

1. Properly timing the pulse train and sampling intervals as they relate to the system bandwidth.
2. Pulse and channel shaping.

By timing the pulse train so that pulses occur or do not occur at intervals of $T = 1/2B$, and are detected at the same rate, the tails of the neighbouring pulses are all at zero volts and interference does not occur. At the appropriate sampling instants, only the desired pulse or no pulse will be present. This is illustrated in Fig. 8-6 where the sampling times occur at multiples of $T = 1/f_s = 1/2B$, or at a rate of f_s, which is twice the transmission bandwidth. This is known as the Nyquist rate.

$$\text{pulse rate} = 2B \tag{8-4}$$

where B is the low-pass channel bandwidth.

Thus, if the sampling at the receiver is performed at the correct instants, or at every T where $T = 1/f_s = 1/2B$, then no ISI occurs.

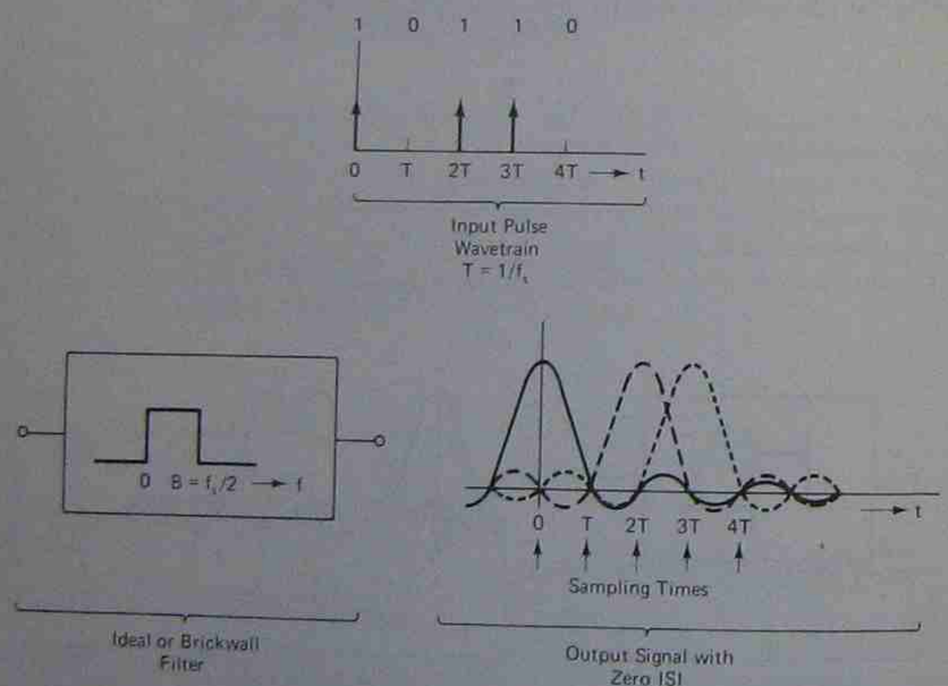


Fig. 8-6 Time Response of a 1011 Impulse Train after Bandlimiting, with Pulse Rate Equal to Twice the System Bandwidth ($f_s = 2B$)

Stating this somewhat differently, the bandwidth B must be equal to one-half the pulse rate, or

$$B = f_s/2 \tag{8-5}$$

This puts very severe restrictions on the sampling times at the receiver. A small time displacement would result in substantial spurious amplitudes at the sampling times. Extremely accurate and stable oscillators, which deviate no more than about 50° in phase from the incoming signal, must be used. Thus, rather expensive equipment is required. Drift due to temperature variations also constitutes a real problem.

Pilot tones can be used to generate the sampling frequency, but noise in the system can cause timing jitter in the actual pilot. In addition, pilot tones require the use of a separate, but parallel, transmission path in order to avoid crosstalk, and phase jitter can be introduced as a result of variations in gains of the parallel path.

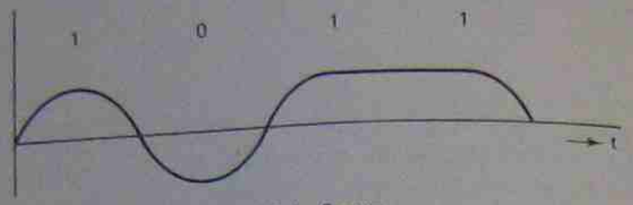
The most common solution is to obtain the timing information from the received pulse stream itself. As long as recurring pulses are received, the signal can control an oscillator. Figure 8-6 shows that the bandlimiting of the system prevents the sampling frequency from being transmitted. That is, since the pulse rate is $1/T = f_s$, the bandwidth limiting the ideal low-pass filter with a cutoff frequency of $f_s/2$ will not pass the sampling frequency f_s. To obtain the timing from the pulse stream, the signal is full-wave rectified, which results in doubling the first harmonic of the signal. Figure 8-7 shows the waveforms of a typical received signal before and after rectification. The rectified signal contains a spectral component at the pulse rate of $1/T$ or f_s which may be extracted by means of a sharply-tuned bandpass filter.

In reality, the tails of sample pulses that have a finite width do not vanish at multiples of T, even when $B = f_s/2$. This only occurs as the pulse width is forced to zero. If pulse widths are not too large, the ISI may still be acceptable. The response of an idealized low-pass filter with no delay to a rectangular pulse of width t_p is given by Eq. (8-6). Figure 8-8 shows the response for different pulse durations when $B = 4$ kHz.

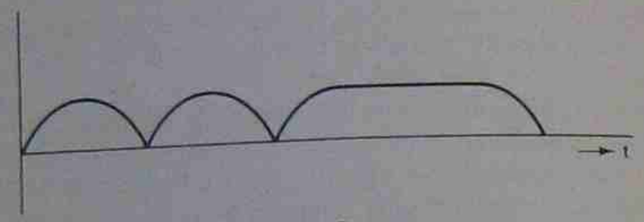
$$g(t) = \frac{t_p}{\pi} \int_0^{2\pi B} \frac{\sin \omega t_p/2}{\omega t_p/2} \cos \omega t \, d\omega \tag{8-6}$$

Table 8-1 gives the relative percentage of the sample g(0)'s amplitude that appears at the adjacent sample instants of $\pm T$. When the pulse duration is $t_p = 0.3\ T$, the percentage error is 0.76.

Because substantial undershoots and overshoots can appear at the adjacent sampling instants as a result of time shift in the signal and non-zero pulse durations, other system frequency characteristics have been developed to reduce these problems. Furthermore, the ideal filter is



(a) Received Pulse Stream

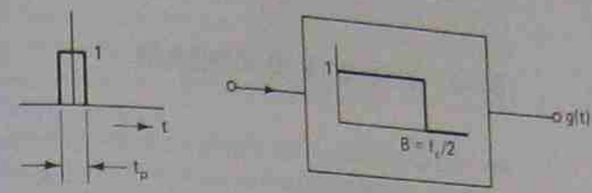


(b) Rectified Pulse Stream

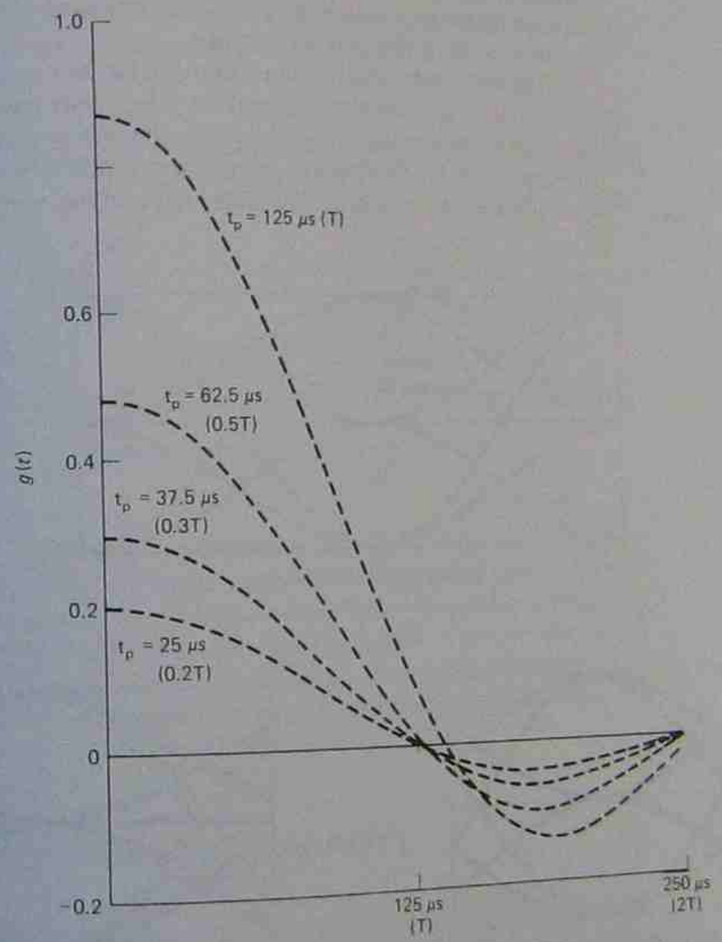
Fig. 8-7 Timing Extraction Employing Rectification

Table 8-1 Relative Percentages of Pulse Present in Terms of the Maximum Amplitude, at Adjacent Sampling Points

Pulse Duration	% error at $t = \pm T$
$0.2T$	0.33
$0.3T$	0.76
$0.5T$	2.1
T	8.7



(a) Circuit Configuration



(b) Response of a 4 kHz Low Pass Filter

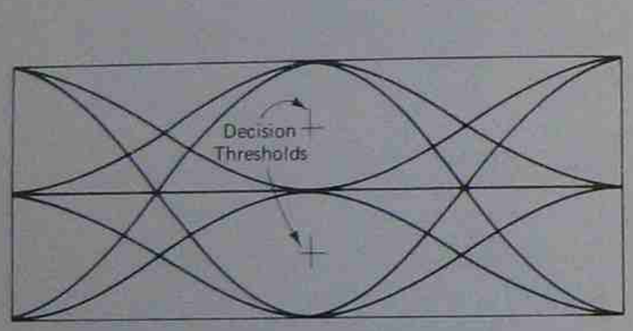
Fig. 8-8 Effect of Filtering on a Pulse of Finite Duration

unattainable in practice. Two solutions in present use that have been found acceptable are the raised cosine response and partial response encoding. These will be briefly considered in Secs. 8–6 and 8–7.

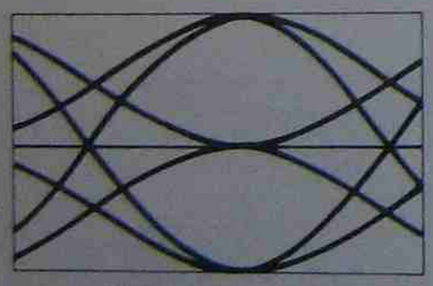
8–4 THE EYE DIAGRAM

In order to determine the effects of random noise, jitter, and so forth on the inter-symbol interference of an incoming pulse stream, the eye diagram is used. It is obtained by applying all the possible pulse combinations to the vertical amplifier of an oscilloscope. About two or three pulse durations are displayed along the horizontal axis. Figure 8-9(a) shows an ideal eye pattern of a three-level system with no noise. With noise, a pattern such as Fig. 8-9(b) or (c) is obtained.

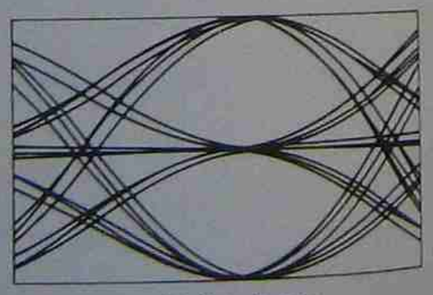
The term "eye diagram" comes from the fact that some areas should appear clear of any signal tracks, making it look very much like an eye. Eye diagram gives an indication of how much noise or error can be tolerated and where the threshold or comparison levels should be centered. The vertical distance between the threshold and the noisy signal tracks



(a) Ideal Three-Level System



(b) Slightly Noisy System



(c) Noisy System

Fig. 8-9 Eye Diagrams

indicate approximately how much additional noise can be tolerated, whereas the horizontal spacing gives an indication of the phase error or jitter that can be tolerated.

If multi-amplitude pulses are used, eye areas would be observed for each sampling instant. The eye centers represent the optimum threshold points for the detector. The ratio of the actual opening to the maximum possible opening represents the degradation caused by the inter-symbol interference in the absence of noise.

For an eye pattern of a two-level system, such as that shown in Fig. 8-10, the peak ISI is about 30%. The equivalent degradation in the S/N rate is 20 log (1 − 0.3) = 3.1 dB. This number must be added to the ideal S/N ratio obtained from Fig. 7-12 to determine the actual required S/N. Thus, if the probability of error is 10^{-8}, the peak signal/rms noise ratio (from Fig. 7-12) is ideally about 15 dB. The required peak signal/rms noise ratio must be increased to 15 + 3.1 = 18.1 dB.

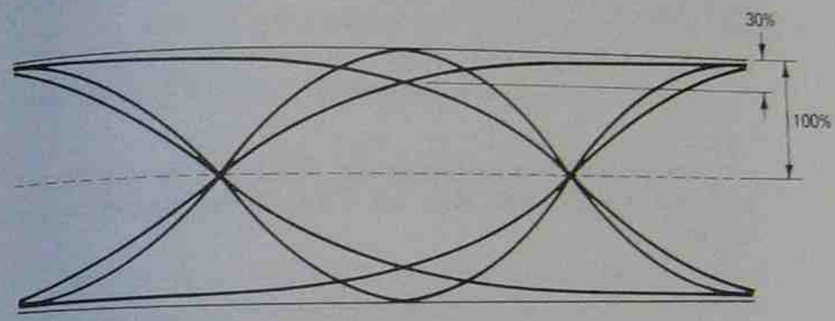


Fig. 8-10 Eye Pattern of a Two-Level System

The horizontal opening in the eye can also undergo serious degradation due to variations in the extracted timing, signals, threshold misalignment at the detector, and input pulse shape. Noise and crosstalk can, for instance, cause jitter in the received clock pulse. A percent jitter measurement can be made on the received data by monitoring the short-term movement of the data edges in relationship to a stable clock reference. It would be given by the ratio $\Delta t/T$ of Fig. 8-11.

8–5 INFORMATION CAPACITY

In Chap. 6, the information capacity for a series of uniformly separated pulses was obtained. Looking back at Eq. (6-14), n/T represents the time duration between pulse centers for a series of impulse type pulses that are uniformly separated by time intervals of T. Using this equation, the

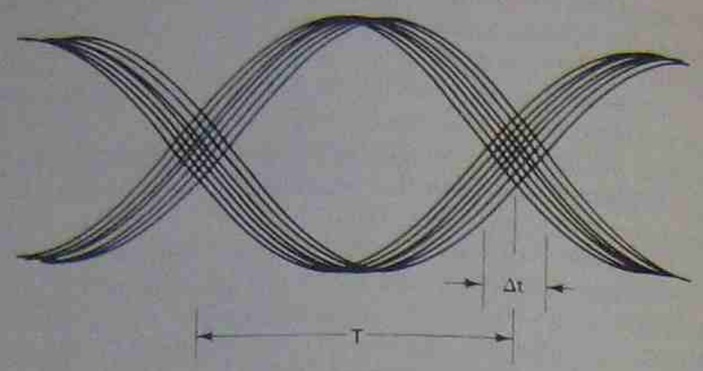


Fig. 8-11 Percent Jitter Measurement

channel capacity for an m-level system similar to that shown in Fig. 8-12 can be given by

$$C = \frac{1}{T} \log_2 m \text{ b/s} \tag{8-7}$$

If the pulse rate is at the Nyquist rate, or is twice the system bandwidth, then, according to Eq. (8-4), $1/T = 2B$ and the channel capacity becomes

$$C = 2B \log_2 m \text{ b/s} \tag{8-8}$$

From this expression, observe that in order to increase the information capacity, either the bandwidth or the number of signal levels must be increased. If the number of signal levels is 2^n, where $n = 1$ for binary systems, multi-level systems have, in principle, n-times the channel capacity of binary systems at the expense of an increased number of levels. This implies greater sensitivity to noise and more errors.

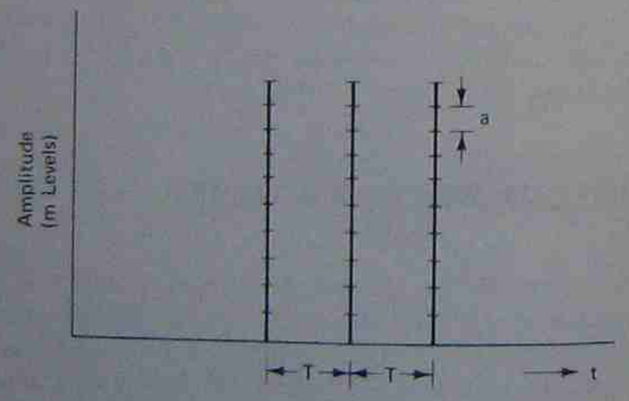


Fig. 8-12 m-Level Impulse Train

If the message consists of bipolar pulses of m levels, the average signal power is related to the "a" level spacing and to the number of levels by Eq. (7-7). From this expression, the relationship of m to the average signal power and the step size can be obtained. Thus,

$$m = \sqrt{1 + \frac{12S}{a^2}} \tag{8-9}$$

Replacing m in Eq. (8-8) by Eq. (8-9) produces

$$C = B \log_2 \left(1 + \frac{12S}{a^2}\right) \text{ b/s} \tag{8-10}$$

The relative importance of bandwidth and average signal power to information capacity is indicated in Eq. (8-10). If the system capacity is to double, for example, either the bandwidth can be doubled or the argument of logarithm $(1 + 12S/a^2)$ can be squared. For the latter case, the argument must be increased by

$$\left(1 + \frac{12S}{a^2}\right)^2 - \left(1 + \frac{12S}{a^2}\right) = \frac{12S}{a^2} + \frac{144S^2}{a^4} \tag{8-11}$$

Equation (8-10) can be developed a bit further to relate the capacity to the signal-to-noise ratio. First, relate the level spacing to the mean-noise power in the system by the expression

$$a^2 = k^2 N \tag{8-12}$$

where k^2 is a constant of suitable choice to give a low error rate
N is the mean noise power

Quantization noise is not included since it has been accounted for in Eq. (8-9).

By keeping the distance between the levels suitably large, the chances of error can be kept small, as shown in Sec. 7–3. For a given level of spacing and a mean-noise power of known noise distribution, it is possible to evaluate the probability of error.

Substituting Eq. (8-12) into Eq. (8-10) produces the following result:

$$C = B \log_2 \left[1 + \frac{12S}{k^2 N}\right] \text{ b/s} \tag{8-13}$$

Given the constant k for a system and the S/N ratio, it is thus possible to obtain the channel capacity.

Work done by Shannon* on information theory illustrates that over a white, bandlimited Gaussian channel, it should be possible to attain a channel capacity of

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \text{ b/s} \qquad (8\text{-}14)$$

where B is the channel bandwidth
S is the signal power
and N is the total noise within the channel.

In practice, this goal is far from obtainable because of the extremely complex encoding required and the long time delays. Instead, a $12/k^2$ value of 1/7 is more typical, which represents an increase in power of 10 log7, or 8.4 dB over that required for Shannon's case.

Assume time division multiplexing is used to insert pulses during the single sampling interval of T shown in Fig. 8-12. If n pulses or channels are inserted, the total channel capacity will go up to

$$C = 2nB \log_2 m \text{ b/s} \qquad (8\text{-}15)$$

This increase in the number of pulses transmitted increases both the required channel capacity and the effective transmission bandwidth (nB).

8–6 THE RAISED COSINE CHANNEL RESPONSE

Impulses are unattainable in practice and the ideal, or brickwall, filter cannot be built. Thus, when designing a digital communication channel, appropriate adjustments must be made to the impulse type signals and filters discussed in Sec. 8–3.

The desired end result is a filter response which is realizable and which produces minimum overshoots and undershoots in its ringing and which passes through zero at adjacent sampling times. When the combination of pulse shape, channel, and equalizer have an overall odd symmetric transfer function about $B/2$, as illustrated in Fig. 8-13, the response to an impulse results in small overshoots and undershoots and passes through zero at the sampling instances of $nT = n/B$. By selecting a suitable response, such as the solid curve in Fig. 8-13, the filter becomes realizable since there are no abrupt changes and the slopes are finite. In

* Shannon, C. E. *A Mathematical Theory of Communication*. BSTJ, vol. 27, pp 379–623, 1948.
Shannon, C. E. *Communication in the Presence of Noise*. Proc. IRE, vol. 37, p 10, 1949.

such a model, the tails of neighbouring pulses do not interfere with the main pulse. The transmission channel can compensate for any deviation of the pulse from the impulse function.

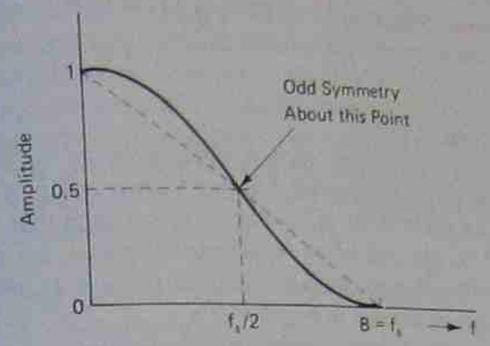
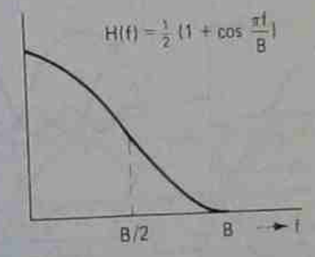
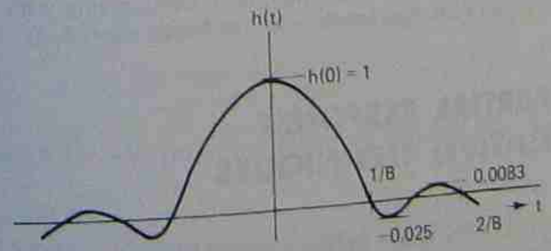


Fig. 8-13 Transfer Function for Realizing Minimum ISI

One acceptable solution to the problem of designing a communication channel is the raised-cosine response, which is shown in Fig. 8-14(a). Although such a response is not strictly realizable, it can be more closely approximated than the brickwall response. It has a gradual roll-off



(a) Frequency Response



(b) Impulse Response

Fig. 8-14 Raised-Cosine Channel Response

of up to about 1.5 times the Nyquist rate, but cuts off rapidly at twice the Nyquist rate. The raised-cosine frequency response is given by

$$H(f) = \frac{1}{2}\left[1 + \cos\left(\frac{\pi f}{B}\right)\right] \quad f < B$$
$$= 0 \qquad\qquad\qquad\qquad f > B$$

and shown in Fig. 8-14(a).

An impulse applied to such a filter gives an output, or corresponding impulse, response similar to that shown in Fig. 8-14(b).

ISI will not occur if the pulse train is timed such that at every time interval $T = 1/B$, the tails of all neighbouring pulses are at zero volts. Only the desired pulse or no pulse will be present at the appropriate sampling instants, as shown in Fig. 8-15.

The bandwidth of such a transmission channel is doubled. (Comparing it to the brickwall case, it goes to f_s, or twice the analog signal bandwidth called for in the Sampling Theorem.) The undershoot and overshoot errors, however, are drastically reduced. The slope of the ringing curve is also much gentler near the sampling times, which permits more jitter, or room for time displacement, than the sin x/x type of ringing present in the case of the brickwall filter.

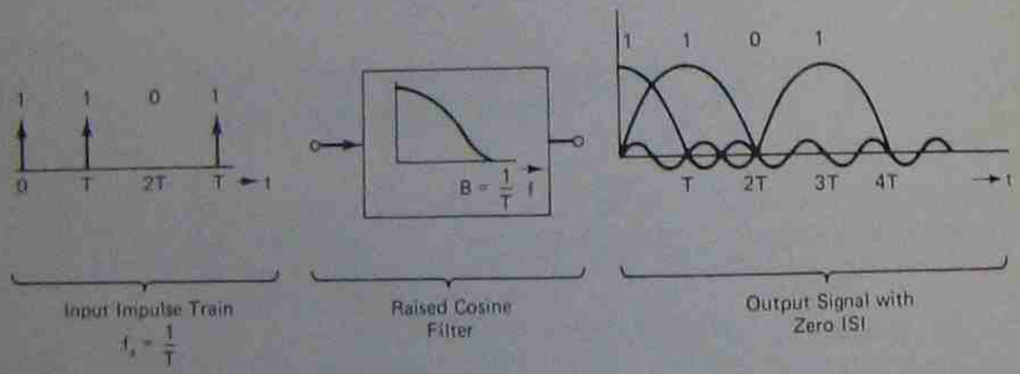


Fig. 8-15 Time Response of a 1101-Impulse Train after Bandlimiting with a Raised Cosine Filter in which Pulse Rate Equals Filter Bandwidth ($f_s = B$)

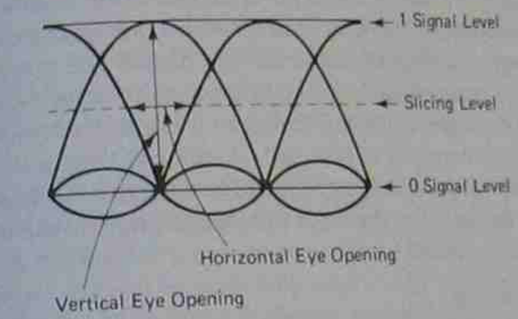
8–7 PARTIAL RESPONSE (CORRELATIVE) TECHNIQUES

In order to meet present-day high speed data requirements, a multilevel system which uses more than two signal states is necessary. If each level is represented by groups of m binary digits, where the total number of signal levels is 2^m, then the speed capability of the m system is m times

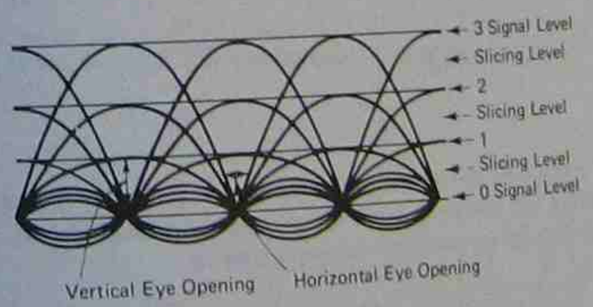
the speed of the binary system. Increasing the speed capability requires adding more levels, which makes the system more prone to error because of the increased sensitivity to noise.

The horizontal eye opening also deteriorates as more levels are added because the number of possible transitions between levels increases. The vertical eye opening remains unaffected, however.

Figure 8-16 illustrates this for a four-level system where three eyes exist—one between each pair of adjacent levels. Although the vertical eye is fully open, the horizontal eye is reduced over the binary case at the slicing level. The chief contributor to this deterioration is the transition between the extreme levels of 0 and 3. Transition between adjacent levels cause no deterioration.



(a) Theoretical Binary Impulse Eye Pattern



(b) Theoretical 4 Level Impulse Eye Pattern

Fig. 8-16 Eye Patterns of Binary and Multilevel Impulses

Some of the classes of partial response systems discussed in this section permit transitions to occur only between adjacent levels, and thus keep the inter-symbol interference lower than in multilevel systems. The level-coded partial response systems still incur noise penalties when com-

pared to the binary systems, much like the multilevel system described in Problem 7-3. Just as in the multilevel case, partial response systems have bit rates which exceed the Nyquist rate.

Because there is no correlation between the levels of a multilevel system, or, rather, because the past history of an encoded signal has no effect on its present state, each combination of m binary digits is associated with a unique or particular level. At the receiver end, each group of m binary digits represents the same level.

In partial response encoding, each signal level depends not only on the present input state but also on the previous input states. That is, there is a correlation between the levels. For this reason, partial response is frequently called corrrelative. At the receiver, however, each level is still treated independently from the rest. For example, at the transmitter in the encoder, each MARK (or SPACE) is assigned one of several permissible levels. The assignment depends upon the past history and present state of the signal. At the decoder, each level is uniquely associated with a MARK or SPACE which is independent of the past history of the waveform. Because distinctive patterns are transmitted, violation of these patterns at the receiver results in errors which can be readily detected. Redundant digits do not need to be added to the input since binary data at the transmitter provide for error detection.

Duobinary Technique

One technique for transmitting data is the duobinary. In context, the term *duo* indicates the doubling of the bit rate over the straight binary system. It is based on an end-to-end system transfer function of

$$|H(f)| = \begin{cases} 2 \cos \pi T f & f \leq \dfrac{1}{2T} \\ 0 & f > \dfrac{1}{2T} \end{cases} \qquad (8\text{-}17)$$

where T represents the time duration between applied impulses. The B bandwidth of the system is thus $1/2T$ or half the signaling rate. Alternatively, this transfer function can be expressed as

$$2 \cos \pi f T = |1 + e^{-j2\pi fT}| \qquad (8\text{-}18)$$

since $$\cos \frac{\theta}{2} = \sqrt{\frac{1 + \cos\theta}{2}}$$

The impulse response of such a system consists of the sum of two sinc functions separated by T, as illustrated in Fig. 8-17. As can be seen, the tails of the resulting output pulse are greatly reduced due to the cancellation of the individual sinc function tails. However, the resulting pulse is 1.5 times as wide as the impulse response of a brickwall filter.

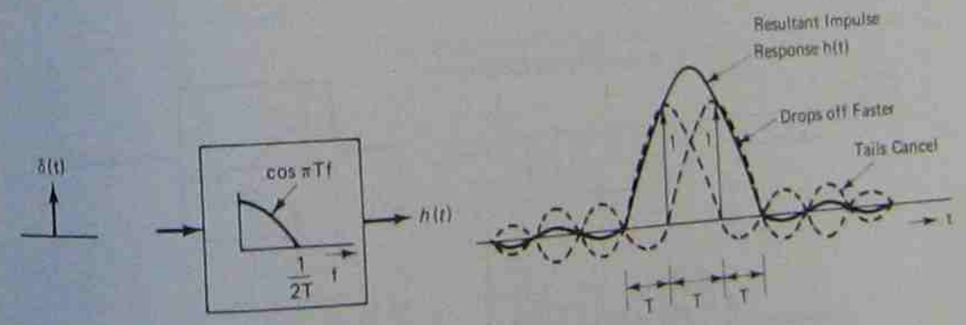
Consider the response for a binary input of 11, which is represented by $\delta(t)$ and $\delta(t - T)$. The input consists of two impulses separated by a time of T. The output, ignoring any time delays in the system, consists of $h(t)$ and $h(t - T)$ where $h(t - T)$ is a delayed version of the $h(t)$ function illustrated in Fig. 8-17. Each of these functions can, in turn, be represented by two sinc type pulses, numbered 1 for $h(t)$ and 2 for $h(t - T)$ in Fig. 8-18. The difficulty in drawing the response can be reduced by drawing the maximum amplitudes of the overlapping sinc functions and assuming that the tails cancel, as shown in the approximate waveform.

Figure 8-19 shows the response to a series of MARK and SPACE impulses. Note that for a full response on the positive side, two or more consecutive MARKS or 1's are needed. For a full response on the negative side, two or more consecutive SPACES or 0's are needed. To obtain 0, the mark and space must alternate. Note that a three-level system now exists—that is, +2, 0 and −2—and that transitions can occur only between adjacent levels for successive signaling elements. In other words, there must be at least one interval at 0 when going from a −2 to a +2 level.

In addition, if an impulse is at either the +2 or −2 level, it must be at the 0 level for an even number of baud intervals before returning to the same +2 or −2 level. If the impulse is at the 0 level for an odd number of baud intervals, it must be switching from +2 to −2 or from −2 to +2.

The signaling rate of the duobinary system is still at the Nyquist rate, that is the bandwidth is equal to $1/2T$, or one-half the signaling rate. Signal rates slightly exceeding the Nyquist rate can be used and reasonable results still obtained.

Figure 8-17 or Equation 8-18 shows that the duobinary response can be produced by using a digital delay circuit and brickwall filter. This is illustrated in Fig. 8-20(a). The effective addition of the delayed pulse intro-



Note: The Impulse Response Consists of Two Sinc Pulses Marked by a 1.

Fig. 8-17 Impulse Response of a Duobinary System

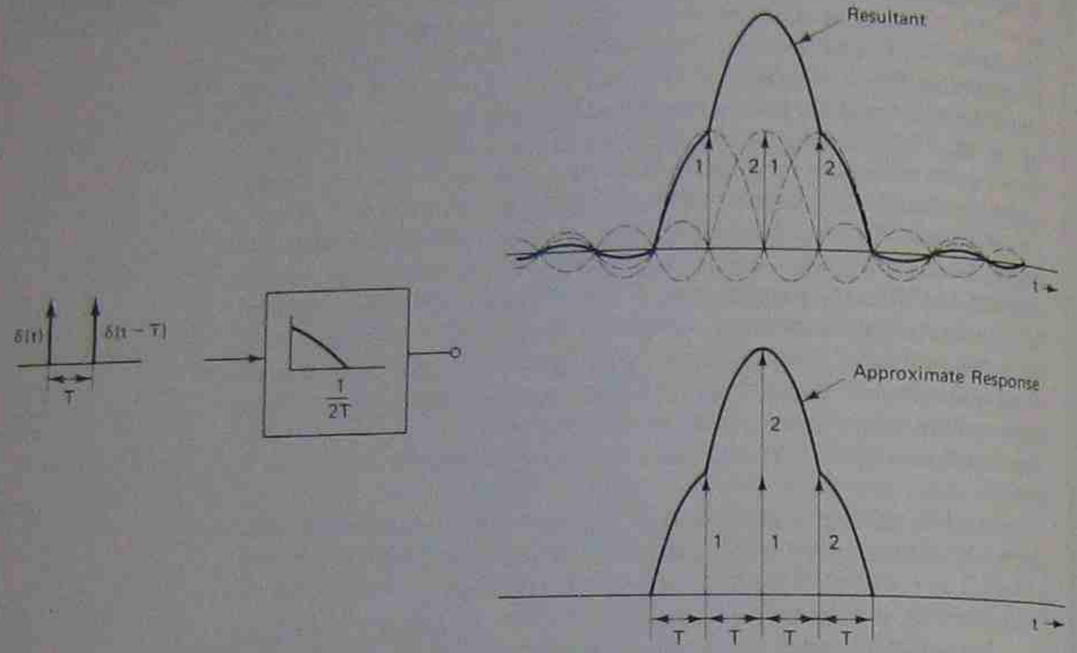


Fig. 8-18 Response of a Duobinary System to Two Consecutive Impulses

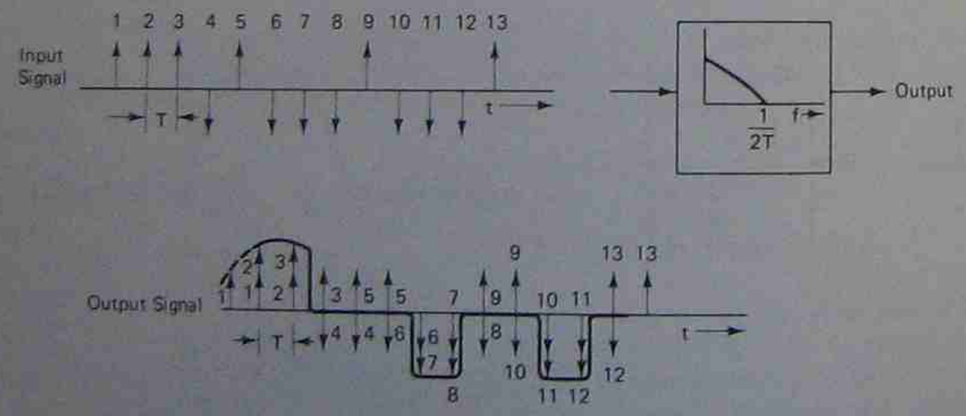


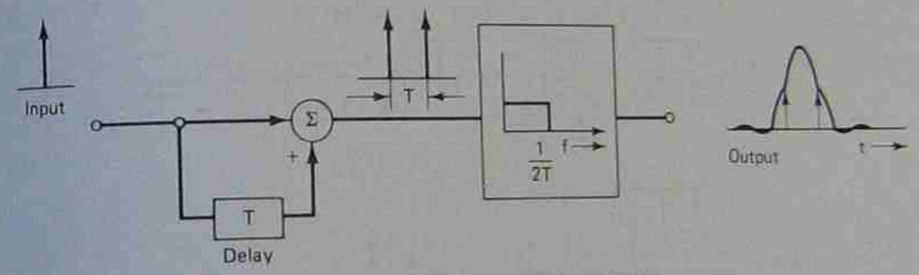
Fig. 8-19 Response of a Duobinary System to an Input Data Stream

duces a predictable amount of inter-symbol interference at the transmitter. At the receiver, it must be subtracted out again as the decoder in Fig. 8-20(b) illustrates.

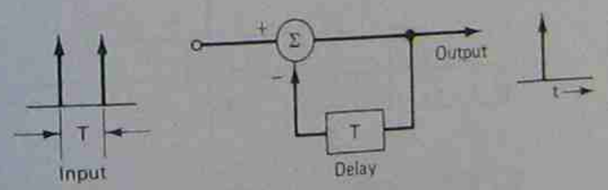
The various approximate waveforms for a typical data stream are illustrated in Fig. 8-21. Delays, other than those which occur in the digital delay circuits, are omitted.

One undesirable feature of the duobinary waveform is that the signal has three levels at each sampling instant and any level can be interpreted as a MARK or SPACE, depending upon the past history of the signal. It would be much more convenient to recognize the zero level as either a MARK or a SPACE and not have to keep track of the alterations. In addition, if an error were detected, immediate appropriate action could be taken instead of allowing the error to propagate into the next bit because of the delay circuit action. To solve these problems, a special precoder can be inserted before the encoder shown in Fig. 8-20(a), and a special post decoder can be inserted after the decoder shown in Fig. 8-20(b). These insertions are illustrated in Fig. 8-22.

The exclusive -OR gates perform Modulo 2 addition, which means the output goes low when the two inputs are the same and high if the inputs are different. Figure 8-23 illustrates the waveforms at the lettered



(a) Production of Duobinary Signal Employing a Digital Delay Line

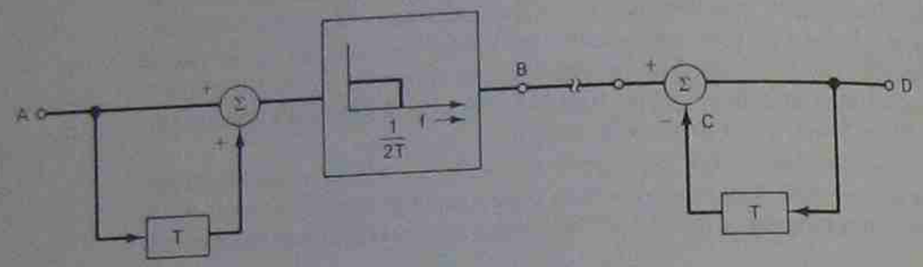


(b) Duobinary Decoder

Fig. 8-20 Encoding and Decoding of a Duobinary Signal

points in Fig. 8-22 for the data given. Notice that for the duobinary waveform (D) with pre-encoding, the amplitude levels can be directly interpreted without referring to the past history. The ±2 levels can be interpreted as 0 and the 0 level interpreted as 1.

An example of a system that uses the duobinary technique is the RD3 radio system. It produces two 45.5M b/s partial-data channels, one channel of which is in phase quadrature with the other. Block diagrams of the transmitter and receiver are illustrated in Fig. 8-24. The bandpass filter (BPF) selects the summed frequency components from the multiplier sections. The traveling wave tube (TWT) is a broadband power amplifier in common use in microwave and satellite systems. The partial-response



(a) Block Diagram

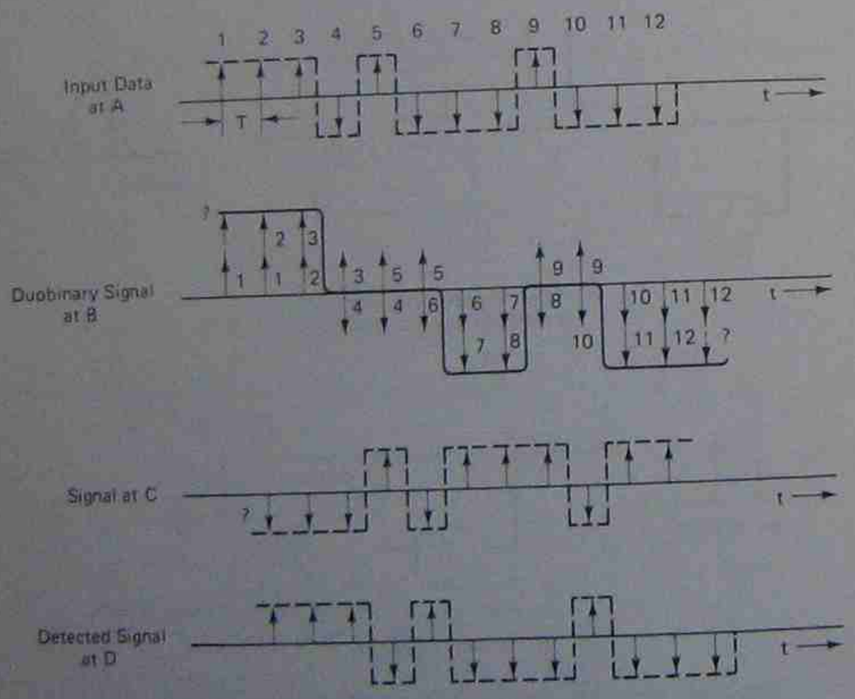


Fig. 8-21 Waveforms in a Duobinary System

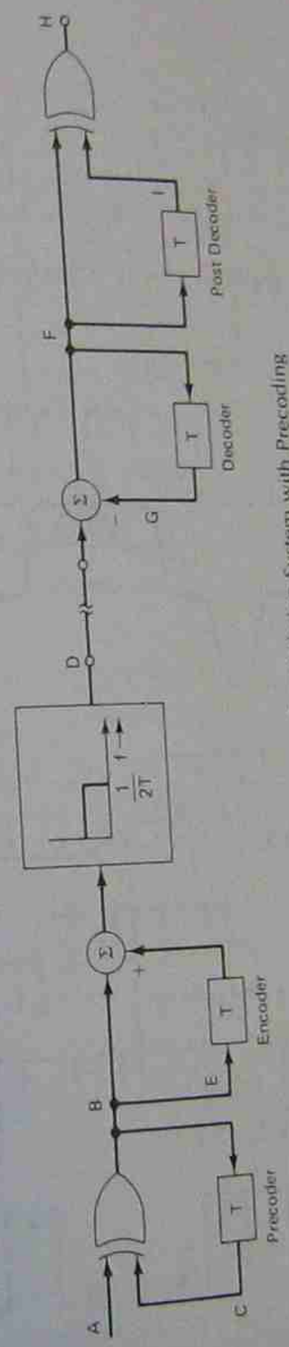


Fig. 8-22 Duobinary Transmission System with Precoding

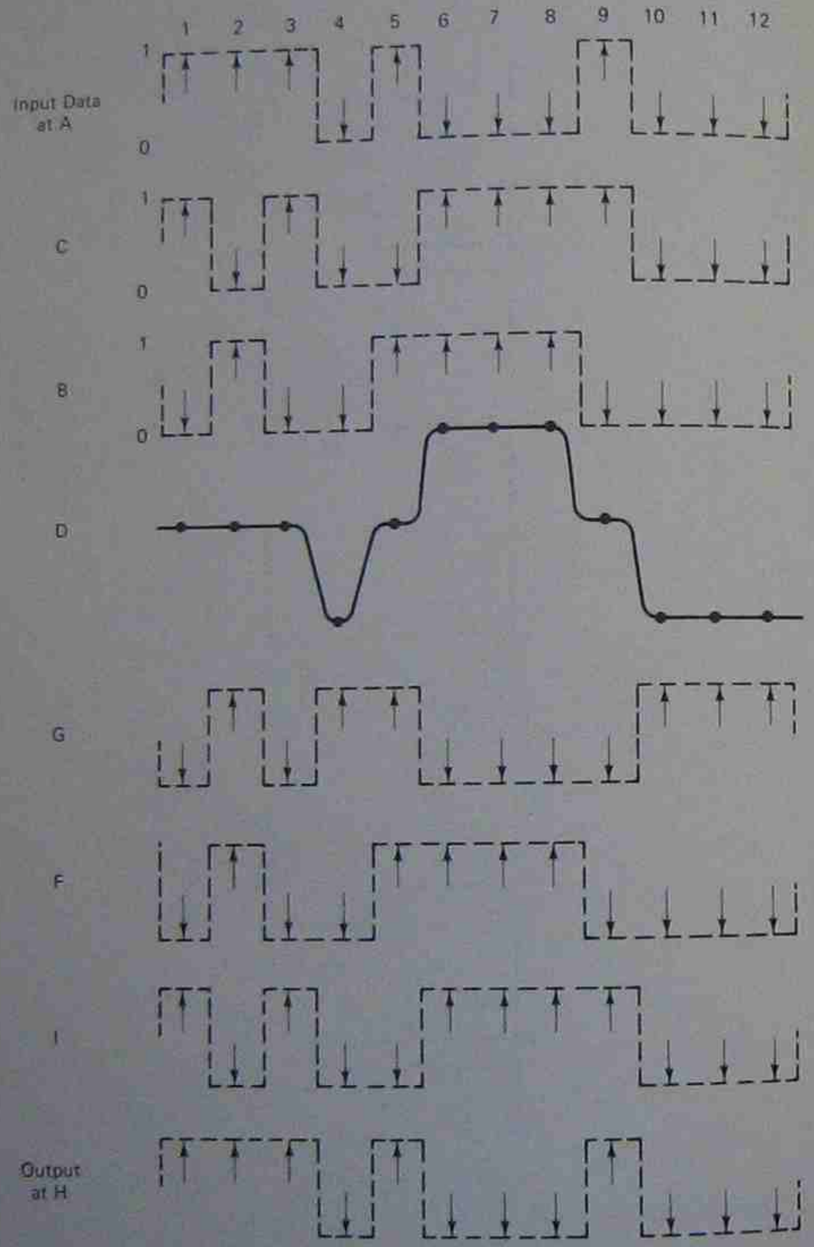
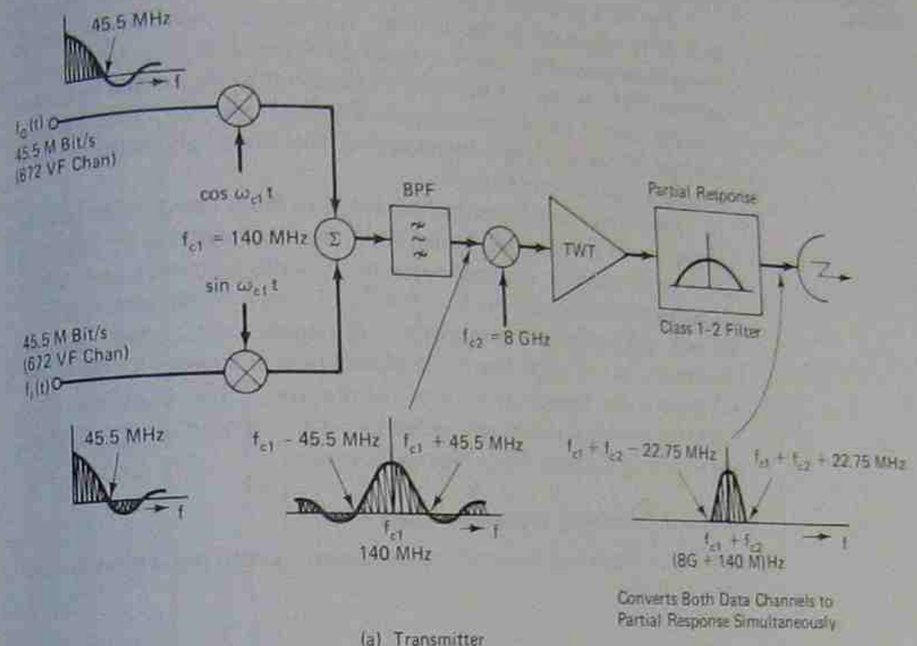
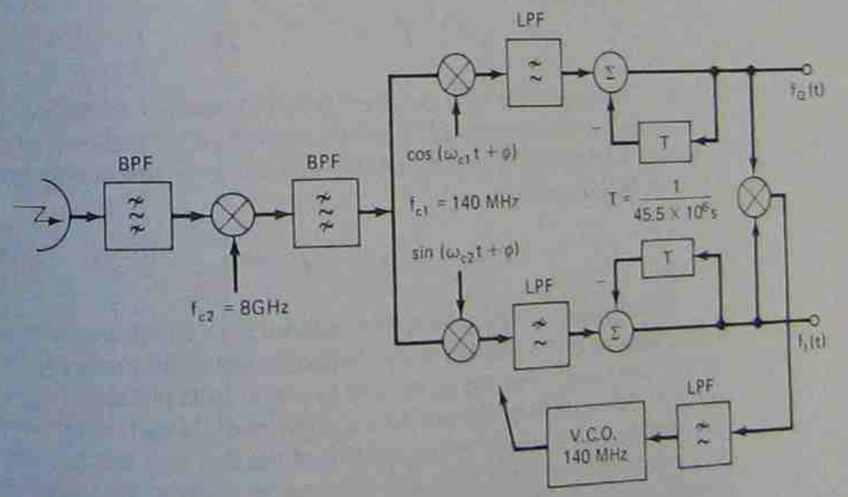


Fig. 8-23 Waveforms in the Duobinary System with Precoding (See Fig. 8-22)



(a) Transmitter



(b) Receiver

Fig. 8-24 Block Diagrams of an RD3 Radio System Employing Duobinary Encoding

filter circuit now has a bandwidth double that of the Nyquist, which is to be expected in a double sideband system.

The RD3 transmits a bandwidth of about 42 MHz rather than the 45.5 MHz suggested in Fig. 8-24(a). This slightly reduces the sample timing tolerance and causes a small S/N penalty. The RD3 system permits 1344 channels on the radiated bandwidth of a normal heavy microwave route. If the antenna is cross polarized, another 1344 channels, or a total of 2688 channels, are permitted. This is well above the capacity of normal analog microwaves.

The various correlative techniques have been categorized into the classes shown in Table 8-2. In this table, the bandwidth of the system is given by F, which is the equivalent of B in our nomenclature. When operating at the Nyquist rate, $F = B = 1/2T$. The k coefficients represent the peak amplitudes of the output sinc responses to an applied impulse. The binary, or 1-1, class has the brickwall type of transfer function. The 1-2 class is the duobinary system just discussed. The 4-3 class, often called the modified duobinary technique, is being used more frequently and will be elaborated on in the next section.

Modified Duobinary Technique (Class 4–3)

As stated in Table 8-2, the modified duobinary system has a transfer function equal to:

$$|H(f)| = \begin{cases} 2 \sin 2\pi fT & f \leq \dfrac{1}{2T} \\ 0 & f > \dfrac{1}{2T} \end{cases} \tag{8-19}$$

where T represents the time between adjacent applied impulses. The bandwidth B of the system is $1/2T$, which is equal to ½ the signaling rate. This also is called the Nyquist bandwidth. Alternatively, this transfer function can be expressed as

$$2 \sin 2\pi fT = |1 - e^{-j4\pi fT}| \tag{8-20}$$

The impulse response of a system with such a transfer function consists of the sum of two sinc functions of opposite polarities that are spaced by a distance of $2T$. (See the k_1, k_2, and k_3 coefficients of Table 8-2.) Figure 8-25 shows the two sinc functions and the result, which is the sum of the two sinc functions. As before, the tails of the two sinc functions tend to cancel. Ignoring these tails, the response to a series of MARK and SPACE impulses is shown in Fig. 8-26.

It can be concluded from Fig. 8-26 that the waveforms have three distinguishable levels—+2, 0, and −2—and no dc component at the

Table 8-2 C. R. Kretzmer, "Generalization of a Technique for Binary Data Communication," © 1966 IEEE. Reprinted, with permission, from *IEEE Transactions on Communication Technology*, vol. 14, p. 67, February 1966.

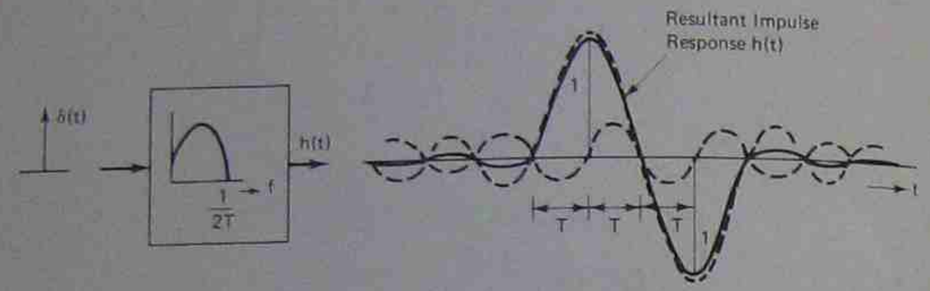
Class	k_1	k_2	k_3	k_4	k_5	h(t)	H(f) $[0 < f < F]$	No. of Rec. Levels
Binary (Ideal)	1						1	2
1 ($n = 2$)	1	1					$2 \cos \frac{\pi f}{2F}$	3
2 ($n = 3$)	1	2	1				$4 \cos^2 \frac{\pi f}{2F}$	5
3 ($n = 3$)	2	1	−1				$2 + \cos \frac{\pi}{F} f - \cos \frac{2\pi f}{F} + j\left[\sin \frac{\pi}{F} f - \sin \frac{2\pi f}{F}\right]$	5
4 ($n = 3$)	1	0	−1				$2 \sin \pi \frac{f}{F}$	3
5 ($n = 5$)	−1	0	2	0	−1		$4 \sin^2 \pi \frac{f}{F}$	5

*Entry in Figure is Area Under Curve.

sampling instants. This is a definite advantage for transmission systems which cannot carry dc.

Similar to the duobinary system, the modified duobinary signal can be produced with an ideal filter and delay circuits as illustrated in Fig. 8-27. Furthermore, the signal can be precoded and later post decoded at the receiver. This is illustrated in Fig. 8-28, where a 1 at the coder output is represented by both the +2 and −2 levels and a 0 by the 0 level.

Because the modified duobinary filter characteristic is difficult to approximate in practice, it can be more readily implemented by cas-



Note: The Impulse Response Consists of Two Sinc Pulses (of Oppostie Sign) and Separated by 2T. These are Marked by 1's.

Fig. 8-25 Impulse Response of a Modified Duobinary Encoder

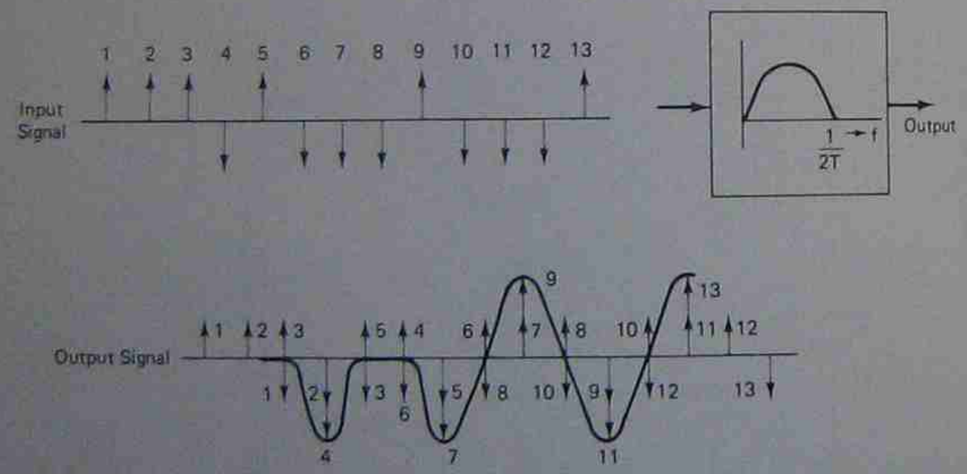
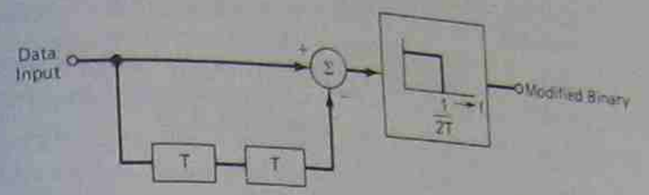
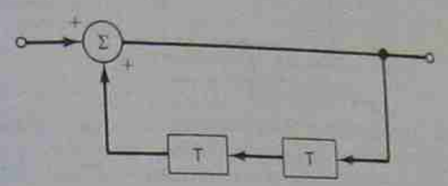


Fig. 8-26 Modified Duobinary Pulse Pattern

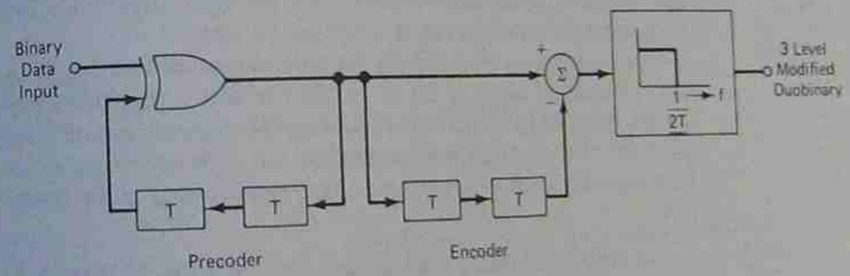


(a) Modified Binary Encoder

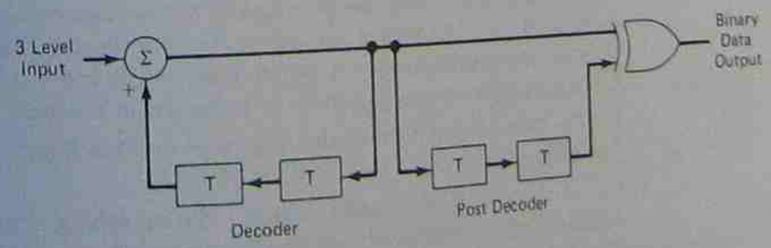


(b) Modified Binary Decoder

Fig. 8-27 Transmission and Detection of a Modified Duobinary Signal



(a) Tx. Section



(b) Rx. Section

Fig. 8-28 Modified Duobinary Transmission System with Precoding

cading a digital delay network with a duobinary filter. From Eq. (8-20), the transfer function can be rewritten as

$$1 - e^{-j4\pi fT} = (1 - e^{-j2\pi fT})(1 + e^{-j2\pi fT}) \qquad (8\text{-}21)$$

The term, $(1 + e^{-j2\pi fT})$, can be represented by the duobinary filter (see Eq. (8-18)). The term, $1 - e^{-j2\pi fT}$, can be represented by a network with a digital delay circuit, as illustrated in Fig. 8-29.

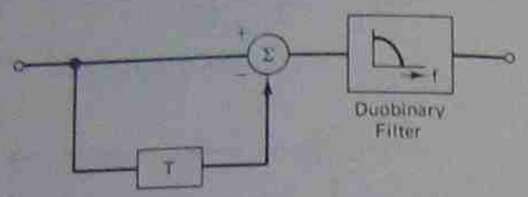


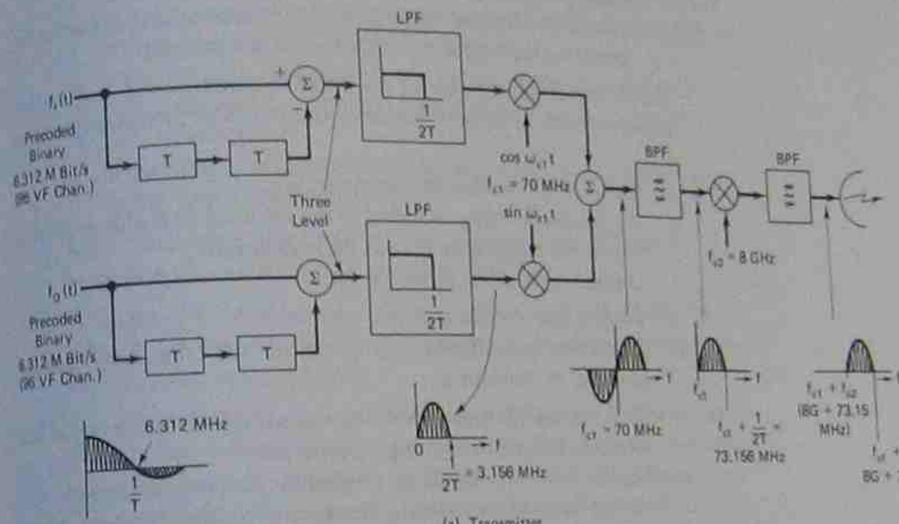
Fig. 8-29 Generation of Modified Duobinary Signal (Based on a drawing in Kamile and Feher, *Digital Communications: Microwave Applications*, © 1981, p. 151. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.)

GTE Lenkurt Inc. uses this modified duobinary technique in a system to produce a single sideband digital radio. A block diagram of an SSB system that uses modified duobinary encoding is shown in Fig. 8-30.

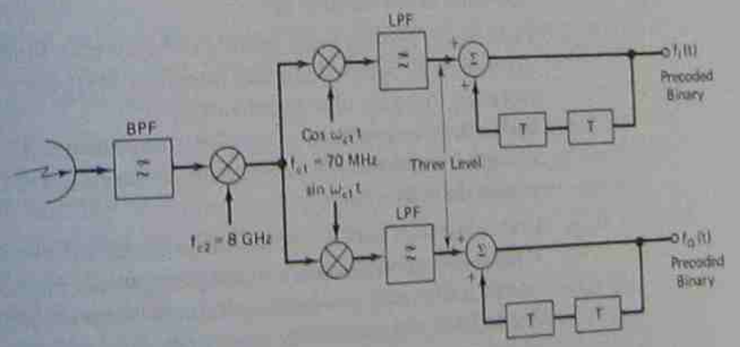
The modified duobinary system gives 4 bits per Hz of bandwidth (BW) compared to the RD3 system's 2 bits per Hz of BW. Both systems use quadrature modulation to obtain these bit rates.

As a chapter review, consider the transmission of a hi-fidelity signal with a frequency range of 20 to 15,000 Hz and a dynamic range of 100 dB. Assume that the human ear can detect changes of 0.5 dB in intensity level. Now compute the bandwidths for each of the transmission techniques discussed in this chapter. Assume encoding at the baseband level.

1. In the original analog form, a bandwidth of 20 to 15,000 Hz is required, or $B = 15$ kHz.
2. In PAM form, a sampling rate of twice the highest signal frequency is required, or 30 k samples per second. For the ideal or brickwall filter transmission response, the required bandwidth is $B = 15$ kHz. Of course, such a response cannot be achieved in practice.
3. For the raised cosine transmission response in PAM form, the required bandwidth would be $2 \times B$, or 30 kHz.
4. For a binary PCM-encoded signal, distinguishing between 100dB/0.5dB, or 200 amplitude levels, requires 8 bits, or binary pulses, since $2^8 = 256$. Therefore, the bit rate is $8 \times 30{,}000 = 240$k b/s. The brickwall response requires 120 kHz of bandwidth, or 2 b/Hz of bandwidth. The raised cosine response requires 240k Hz of bandwidth, or 1 b/Hz of bandwidth.
5. For a ternary PCM code transmitted over a channel with a raised cosine response, each sample requires 5 bits, or $3^5 = 243$. This represents a bandwidth of $5 \times 30{,}000$, or 150 kHz. One of the drawbacks of the ternary system is that it is more susceptible to noise than the binary encoded system.
6. For the PCM signal that is encoded in the duobinary or modified duobinary format, the required bandwidth is still the Nyquist bandwidth of 120 kHz. However, the Nyquist rate can be exceeded by up to 43% in marginal operation to give 2.86 b/Hz of bandwidth.



(a) Transmitter



(b) Receiver

Fig. 8-30 Block Diagram of an SSB Radio System Using Modified Duobinary Encoding

7. For the PCM signal encoded upwards in frequency to a DSB duobinary format, a bandwidth of 240 kHz is required, since the bandwidth is doubled due to the presence of both sidebands. For the PCM signal encoded upwards into an SSB modified duobinary format, a bandwidth of 120 kHz is required since only one sideband is present.

PROBLEMS

1. a. What are the frequency and time characteristics of the impulse function $\delta(t - t_0)$?
   b. Why is the impulse function a convenient testing pulse for determining the frequency response of a network?
2. Sketch the time response of an impulse function $\delta(t - T)$ which passes through a low-pass brickwall filter with a cut-off frequency of f_c.
3. a. Define Inter-Symbol Interference (ISI).
   b. If a delayed impulse is added to the impulse of Problem 2, what would the spacing be if zero ISI is desired?
   c. Discuss methods of reducing ISI.
4. a. What is the Nyquist rate of a channel?
   b. Determine the Nyquist rate of a channel with the characteristics outlined in Problem 2.
6. a. Give two methods of generating the sampling frequency at the receiver. Why is this timing so critical?
   b. Explain why the sampling frequency cannot be directly extracted from the incoming wavetrain to the receiver when operating at the Nyquist rate.
7. a. Explain why real-life input sample pulses with a finite width still experience ISI, even when the sampling frequency is equal to twice the low-pass filter bandwidth.
   b. Sketch the time response of a pulse of duration 0.3T that passes through a bandpass filter of cut-off frequency $1/2T$. Give the amplitude at the time $t = T$.
8. a. Give a hook-up circuit that would produce an eye diagram.
   b. What information can the eye diagram give?
   c. An eye pattern of a two-level system indicates a closure of 40% due to ISI. If the probability of error is 10^{-7}, determine the required peak signal/rms noise ratio.
   d. If Δt and T of Fig. 8-11 are 1 μs and 10 μs respectively, what is the percent jitter?

9. a. Determine the channel capacity of a system which carries a signal with a peak voltage of 1 and pulse spacings of 1 ms. Assume that amplitude variations of 3.9 mV can be detected. If the channel is operated at the Nyquist rate, what is its bandwidth if an ideal brickwall response is assumed?
   b. Each sample or pulse in (a) is converted into an 8-bit signal. Now what is the channel capacity? If the channel is operated at the Nyquist rate, what is its bandwidth if an ideal response is assumed?
   c. What is the chief disadvantage of the method given in (a) over that in (b) in regards to noise immunity?
10. a. A bipolar pulse stream of 1 W average power with step sizes of 1 mV is transmitted down a 4 kHz bandwidth channel. What bit rate can be sent down such a channel?
    b. If the bit rate is doubled, what is the increase in the average signal power? Assume no change in bandwidth. Is this reasonable in practice?
    c. If the bit rate of (a) is doubled by increasing the bandwidth, what is the new bandwidth?
11. State Shannon's information rate. Why is it so difficult for real systems to approach this theoretical rate?
12. A single, digitized voice channel requires an ideal bandwidth of 32 kHz. What bandwidth is required if 24 such channels are time-division multiplexed?
13. a. What difficulties are experienced in designing a brickwall filter?
    b. Discuss a feasible pulse-channel-equalizer frequency response which can be approximated in real life and which results in a reasonable ISI. Does the raised-cosine response meet this criterion?
    c. What bandwidth is required for the raised-cosine response in order to achieve minimal ISI? Compare this to the brickwall type of response.
14. For a specified probability of error, the partial response technique can transmit substantially more bits per second per hertz of bandwidth than the Nyquist-type systems that have no memory—that is, that ignore past history. Explain why.
15. a. Draw the impulse response of a duobinary system.
    b. If the bandwidth of a duobinary system is B, what should the spacing be between applied impulses to keep the ISI to a minimum?
    c. Sketch the duobinary response for the input sequence 1011100100011.

16. For the same input sequence given in Problem 15(c), sketch the waveforms at points A, B, C, and D for Fig. 8-21(a).
17. a. For the same input sequence given in Problem 15(c), sketch the waveforms at points A to H, inclusive, for Fig. 8-22.

b.

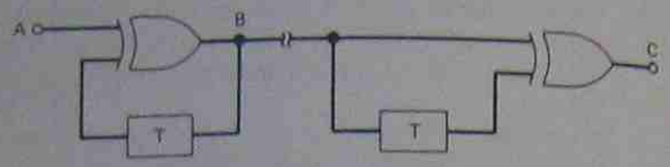
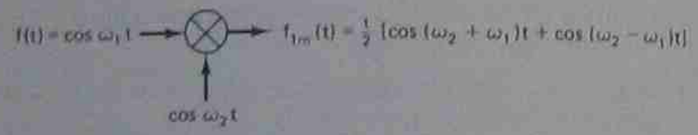


Fig. P8-17

Prove that C = A for the precoder/post decoder shown in P8-17.

18.



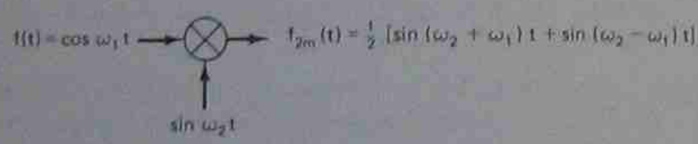


Fig. P8-18(a)

When multiplying a signal—$f(t) = \cos \omega_1 t$, for example—by the orthogonal signals $\cos \omega_2 t$ and $\sin \omega_2 t$, the modulated signals expressed in Fig. P8-18(a) are obtained. These modulated signals are then summed and transmitted. Using the following identities:

$$\cos x \cos y = \frac{1}{2}[\cos (x + y) + \cos (x - y)]$$

$$\sin x \cos y = \frac{1}{2}[\sin (x + y) + \sin (x - y)]$$

$$\sin x \sin y = \frac{1}{2}[\cos (x - y) - \cos (x + y)]$$

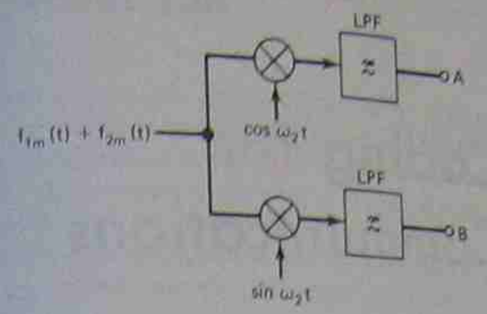


Fig. P8-18(b)

prove that the original signals can be separated out using the detector (complete with filter) of Fig. P8-18(b). Give the expressions for A and B.

19. Referring to Table 8-2, sketch the impulse response and frequency characteristics of a Class 2-3 System. Do the same for a Class 4-3 System. How many levels does each of these systems produce?
20. A hi-fidelity signal which has a 20 Hz to 20 kHz frequency range is to be transmitted, and 512 levels are to be distinguished. Obtain the transmission bandwidth for the following conditions:
    a. Transmission in its original analog form.
    b. Transmission in a PAM form via a channel having a brickwall and a raised-cosine response.
    c. Transmission in a binary PCM form via a channel having a brickwall and a raised-cosine response.
    d. Transmission in a ternary PCM form via a channel having a brickwall and a raised-cosine response.
    e. Transmission in a duobinary form.
    f. Transmission in a modified duobinary form over an AM and SSB system.

9 Coding for Communications

9-1 INTRODUCTION

Since computer systems use binary digits, coding systems have been developed to convert letter and number characters into binary notation. It is usually desirable to use a code which is both recognizable to humans and acceptable to computers. Only 6 bits are required to express numeric and alpha information, which allows for a total of 2^6, or 64, combinations. The extra combinations can be used for punctuation marks, control characters, and other special characters required for data communications.

Special bits, called parity bits, are occasionally added for error detection. In paper tape codes, for example, a check or parity bit is added to the end of a character to create either an even or an odd parity. If even parity is used, a parity bit is added to make the total number of 1 bits even; for odd parity, the sum of 1 bits with parity is odd.

For example, the following table shows the application of even parity:

Code Word	Parity Bit
0000	0
1000	1
1010	0
1100	0
0100	1
1011	1

In *odd* parity schemes, the parity bit is used to make up an odd number of ones. For the characters in the example just given, the parity bits would be the opposite of that shown.

Parity can detect, but not correct, single errors. It will miss double errors. For example, if the following code is transmitted in an even parity scheme:

Data	Parity
0110	0

but the following code is received:

Data	Parity
1110	0

then a parity error exists since an odd number of ones were received with

a 0 parity bit. However, if the same code is transmitted but the following code is received:

Data	Parity
1010	0

the error will *not* be detected, because the correct parity bit was received with an even number of ones.

This chapter will consider some of the more popular binary coding systems and their applications.

9–2 CODES

Baudot Code

One of the first codes developed which is still popular is the Baudot teleprinter code. (Refer back to Fig. 1-14.) In this code, each character is represented by five bits of equal length. This would normally allow for $2^5 = 32$ distinct characters. By using a shift arrangement, much like that of the upper and lower cases on a typewriter, the number of distinct characters can be doubled to 64. This is performed by the letters (LTRS) and the figures (FIGS) control characters. Depending upon which one of these two control characters was most recently used determines all of the subsequent series of characters. The LTRS shift indicates that all characters after it are alpha characters. All the succeeding characters after a FIGS shift are numerics and other special characters.

Since six of the possible combinations are used for function codes—LTRS, FIGS, SPACE, CARRIAGE RETURN, LINE FEED and BLANK—58 are left for characters. After taking into account the nine characters of the number system and the 26 characters of the alphabet, this leaves only 23 for special characters.

Because of the code being prefixed by FIGS or LTRS, when a baudot terminal is interfaced to a computer, the software must monitor the FIGS-LTRS status in order to interpret the data correctly. A receiver has no way to tell if an error has occurred, as a parity bit is not used.

The Baudot code is usually transmitted asynchronously, whereby the data bits of each character are preceded by a start bit and followed by a stop interval. The stop interval is 1.42 times the length of the other intervals. Because these start and stop bits add no information to the character, the time taken up by these bits reduces the net data throughput. Refer back to Fig. 1-18 for an illustration of the configuration of the letter J.

The Baudot code is used in telegraph communication on the keyboard printer and paper tape reader.

Binary Coded Decimal (BCD)

One of the earlier codes developed for data processing machines was the Binary Coded Decimal, shown in Table 9-5. Since there are no alpha characters, this code is not suitable for a communications code, only for the storing of numeric information. The BCD code is often hidden in other codes, such as EBCDIC and ASCII.

To represent a numeric greater than 9, other four-bit groupings are used; for instance 49 is represented by the code 0100 1001.

Extended Binary Coded Decimal Interchange Code (EBCDIC)

Another code developed later is the EBCDIC code. It is an extension of the BCD code to 8 binary bits, having the capability of accomodating 256 combinations. Table 9-1 illustrates the corresponding binary and hexadecimal equivalents for this code. In the hexadecimal code of base 16, numbers beyond 9 are represented by the letters A, B, C, D, E, and F. This is shown in the first two columns of the table.

The EBCDIC code can handle graphic and control characters. It can be used in data processing applications since it is very compatible with BCD-oriented data processing equipment. This code suffers somewhat from a communications viewpoint as no bits are set aside for parity detection.

American National Standard Code for Information Interchange (ASCII)

A more recent code, which has become the one most extensively used as a communication code, is the ASCII code. (See Table 9-2.) This code is a 7-bit code, with an eighth bit added for parity. The eighth parity bit, can be either odd or even parity, depending on what the user desires. Some terminal vendors are using this eighth bit for creating another 128 special characters from the keyboard rather than for parity.

As an example, using even parity, the letter A would be represented by 01000001. It would be 11000001 for odd parity.

The ASCII code and variations of the code are used by most computer terminal manufacturers. This code is used commonly in computer-to-computer and computer-to-terminal communications.

Note that the control functions, such as SOH (Start of Heading), can be detected by 0's in the first two bits. If both are not 0's, a character is being transmitted.

In order to transfer information properly, the receiver needs to know what is to be printed and how it is to be printed, and what actions are to be taken under special circumstances. In order to ensure message control, message heading and error detection and correction, control characters or control codes are used. Some of these control characters are explained in Table 9-3.

Table 9-1 EBCDIC Code Chart Courtesy Tektronix

HIGH / LOW	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
B4 B3 B2 B1	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	NUL	SOH	STX	ETX	PF	HT	LC	DEL		RLF	SMM	VT	FF	CR	SO	SI
0001	DLE	DC1	DC2	DC3	RES	NL	BS	IL	CAN	EM	CC		ITS	IGS	IRS	IUS
0010	DS	SOS	FS		BYP	LF	EOB/ ETB	ESC/ PRE			3M			ENR	ACK	BEL
0011			SYN		PN	RS	UC	EOT					DC4	NAK		SUB
0100	SP									¢	.	<	(	+		
0101	&									!	$	*	)	;	¬	
0110	-	/									,	%	_	>	?	
0111										:	#	@	'	=	"	
1000		a	b	c	d	e	f	g	h	i						
1001		j	k	l	m	n	o	p	q	r						
1010		~	s	t	u	v	w	x	y	z						
1011																
1100	{	A	B	C	D	E	F	G	H	I						
1101	}	J	K	L	M	N	O	P	Q	R						
1110	\		S	T	U	V	W	X	Y	Z						
1111	0	1	2	3	4	5	6	7	8	9						

BINARY — HEX — EBCDIC

Table 9-2 American Standard Code for Information Interchange

Bit Number																
								0	0	0	0	1	1	1	1	
								0	0	1	1	0	0	1	1	
								0	1	0	1	0	1	0	1	
b_7	b_6	b_5	b_4	b_3	b_2	b_1	Row \ Column	0	1	2	3	4	5	6	7	
0	0	0	0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p	
0	0	0	0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q	
0	0	0	0	0	1	0	2	STX	DC2	"	2	B	R	b	r	
0	0	0	0	0	1	1	3	ETX	DC3	#	3	C	S	c	s	
0	0	0	0	1	0	0	4	EOT	DC4	$	4	D	T	d	t	
0	0	0	0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u	
0	0	0	0	1	1	0	6	ACK	SYN	&	6	F	V	f	v	
0	0	0	0	1	1	1	7	BEL	ETB	'	7	G	W	g	w	
0	0	0	1	0	0	0	8	BS	CAN	(	8	H	X	h	x	
0	0	0	1	0	0	1	9	HT	EM	)	9	I	Y	i	y	
0	0	0	1	0	1	0	10	LF	SS	*	:	J	Z	j	z	
0	0	0	1	0	1	1	11	VT	ESC	+	;	K	[	k	{	
0	0	0	1	1	0	0	12	FF	FS	,	<	L	\	l	|	
0	0	0	1	1	0	1	13	CR	GS	-	=	M	]	m	}	
0	0	0	1	1	1	0	14	SO	RS	.	>	N	^	n	~	
0	0	0	1	1	1	1	15	SI	US	/	?	O	_	o	DEL	

Table 9-3 ASCII Control Codes

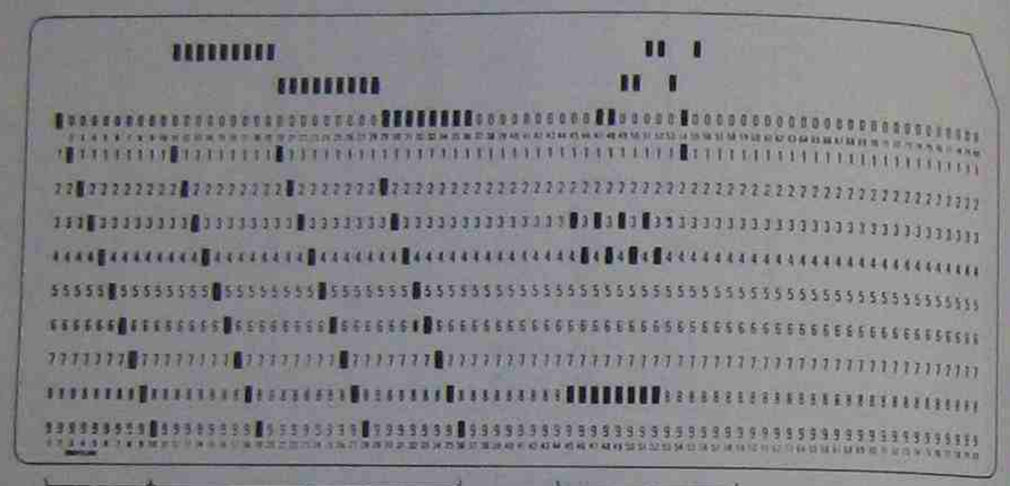
NUL (null)	All zeros character, used for fill.
SYN (synchronous idle)	Used in synchronous transmission for character synchronization.
SOH (start of header)	Used at the beginning to indicate routing information.
STX (start of text)	Used at the beginning of a sequence of characters which are to be referred to as text.
ETX (end of text)	Used at the end of text.
ETB (end of block)	Indicates end of a block of data.
EOT (end of transmission)	Used at end of transmission or end of call.
ACK,NAK	Sent by receiving station to the transmitting station to indicate successful (ACK) or unsuccessful (NAK) reception of a message.
DLE (data link escape)	Changes the meaning of a limited number of contiguously following characters.
ENQ (enquiry)	Used as a request for a response from a remote station; typical response may be address or status content of station's buffer.
CAN (cancel)	Disregard the accompanied data.

Hollerith Code

The Hollerith code shown in Table 9-4 is the most widely used code for punched card applications. It is a 12-level code that consists of two parts: zone and data bits. Only one of the three zone bits and one of the nine data bits are on any one code.

Punched cards are frequently used for customer billing and for entering information into a computer for batch processing.

Table 9-4 Hollerith Punched-Card Code



Numeric Codes

Numeric codes represent number values by using binary digits. The most obvious numeric code is the simple binary number system. The following list outlines some special characteristics of numeric codes. Several of the more common numeric codes are tabulated following the list.

1. Weighted Codes. Some numeric codes give a *weight,* or value, to each bit position so that the value of the encoded number is found by adding the weights of the 1's in the code word. For example, look at the 8421 BCD code in Table 9-5. The code for 6_{10} is 0100. The subscript represents the base or radix of the number system.

$$0110 = 0 \times 8 + 1 \times 4 + 1 \times 2 + 0 \times 1$$
$$= 6_{10}$$

   Weighted codes are convenient for analog-to-digital and digital-to-analog conversion.

2. Self-complementing codes. An example of an unweighted code is

the BCD Excess 3. This code is obtained by adding 3 to the equivalent of 8421 BCD. No direct weighting exists for this code, but it has the advantage of being self-complementing. If the bits in a code word are complemented (0's become 1's and vice versa) the resulting code word represents the so-called 1's complement. This is parallel to the 9's complement in the decimal system, which is obtained by subtracting the number from 9.

In Table 9-5, for example, 0101 Excess 3 represents 2_{10}. The 1's complement is 1010, which represents 7_{10}. Note that the 9's complement of 2_{10} is 7_{10}, i.e., 9_{10}-2_{10}.

Self-complementing codes are useful in performing arithmetic operations on the digitized data. When wishing to subtract a number for instance, computers usually find its complement and then add. The reader may wish to refer to some basic digital text to review this technique.

3. Exact count codes. One error checking method involves using a code that has groups of bits in which every code word has the same number of 1's. This is called an exact *count* code, and the Biquinary code shown in Table 9-5 is an example of it. In the Biquinary code, each of the two bit groups contains a single 1 for every code word. Note that Biquinary is weighted. Other error checking and correcting codes, such as the Hamming Code, are more complex.

Table 9-5 Several Numeric Codes

Decimal	8421 BCD	Excess 3	4221 BCD	Biquinary		Gray
	Weights 8421	—	4221	50	43210	—
0	0000	0011	0000	01	00001	0000
1	0001	0100	0001	01	00010	0001
2	0010	0101	0010	01	00100	0011
3	0011	0110	0011	01	01000	0010
4	0100	0111	1000	01	10000	0110
5	0101	1000	0111	10	00001	0111
6	0110	1001	1100	10	00010	0101
7	0111	1010	1101	10	00100	0100
8	1000	1011	1110	10	01000	1100
9	1001	1100	1111	10	10000	1101
10	—	—	—	—		1111
11	—	—	—	—		1110
12	—	—	—	—		1010
13	—	—	—	—		1011
14	—	—	—	—		1001
15	—	—	—	—		1000

4. Unit distance codes. When an adjacent code word differs by only one bit position, the code is called a unit-distance code. The most important example is the Gray code shown in Table 9-5. For example, in going from the decimal 3 (0011) to the adjacent code word of decimal 4 (0110), only one of the bits has changed. This is different from the BCD case where three bits are changed, i.e. 0011 to 0100. The Gray code can be extended beyond four bits, and algorithms have been developed to convert a binary number of any length to the Gray code and vice-versa.

   The Gray code is convenient when positional indications are required—for example, a shaft rotary position. When the shaft encoder is at the binary position 3 (0010) in BCD, if the second most significant bit is changed slightly before the two least significant bits, the output would indicate 7 (0111) in BCD—a highly erroneous number. In the Gray code, however, it would indicate the adjacent position.

In most codes, redundant or superfluous information is added to the original data in order to increase the probability of detecting an error. If enough redundancy is added, it is even possible to correct errors. However, as more redundant bits are added, the efficiency of the coding scheme is decreased. Efficiency is defined as the ratio of information bits to total bits. Thus, for the Baudot teletypewriter code described earlier in this section, the system efficiency is

$$\eta = \frac{\text{Information pulses/character}}{\text{Transmission pulses/character}} \qquad (9\text{-}1)$$

$$= \frac{5}{7.42} = 76.4\%$$

For the seven-bit ASCII code, which has an eighth bit for parity

$$\eta = \frac{7}{8} = 87.5\%$$

9–3 ERROR DETECTION AND CORRECTION

Error control methods fall into two basic categories: (1) error detection with retransmission, and (2) forward-acting error correction.

The first method involves use of a backward channel in the system. When an error is detected, a retransmission request (ARQ) is sent back to the transmitter.

The second method not only detects error but, by a proper coding

technique, also corrects the error via the receiving circuitry. Forward error correction is the only viable technique in situations where one transmitting source is connected to many receivers. (If each receiver, upon detecting an error, requested the transmitter to retransmit, the information throughput in the system would be exceptionally small because of all the time tied up in retransmitting.) Forward error correction is used on some space probes, where retransmission is often impossible. The long time delays would require excessive storage facilities in the probe's electronic circuitry.

To detect errors, a parity bit can be added to the bits describing the character. Even parity is obtained when one bit is added to the characters that have an odd number of 1 bits. If a bit is inverted upon transmission, the combination will result in an odd number of 1 bits, which indicates an error. Double errors will not be detected, as even parity will still be obtained at the detector.

By further increasing the redundancy, more errors can be detected. The negative effect is that system efficiency is reduced. Efficiency, in this sense, is defined as the transmission of the largest number of characters regardless of errors. One code, the 3-out-of-7 binary code, for instance, allows only those sequences which have four 0's and three 1's to be transmitted.

The IA3, which is widely used in radio telegraphy, is an example of such a 3:4 constant ratio code. The IA3 code detects odd errors but not all even errors, (for example, when a 1 replaces a 0 and a 0 a 1). The code allows 35 combinations out of 2^7, or 128, possibilities for an efficiency rate of $35/128 = 27\%$.

By using large blocks divided into many characters, it is possible with both row and column parity to detect not only all single but also most double and triple errors. In addition, all single errors can be corrected.

A complete character, known as a block check character (BCC), may be added to the end of every block of information to give columnar or longitudinal, parity checking. In Table 9-6(a), for instance, a block of information that begins with a STX (start of text) character is encoded in ASCII with even parity. The ETB (end of transmitted block) character indicates that the character following is the block check character.

When a single error occurs, as in the case of the (1) in the received message of Table 9-6(b), an error occurs in the row and column parity checks. By checking for even parity in each row and column, it is possible to locate which bit is in error.

In low-speed applications (teletypewriters, for example), the message is often sent in duplicate and compared at the receiver to see if a match occurs. If the messages match, it is assumed that an error has not occurred. If there is no match, a request is sent to the sending station for a retransmission of the signal.

In high-speed synchronous applications, such as computer-to-computer data transfer, data is frequently transmitted in blocks with bits added for parity checking. The receiving station performs parity checks. If sent and received information matches, it is assumed that no error has occurred, and a positive acknowledgement (ACK) signal is sent to the transmitter terminal. If parity does not check, a negative acknowledgement (NAK) is sent to the transmitter, requesting a retransmission of the erroneous block.

For any particular application, there is an optimum block length for net data throughput (NDT), or the rate at which data bits arrive at the receiver end. If block lengths are extremely long there is a high probability of error and therefore retransmission requests are very likely. On the other hand, if block lengths are extremely short, the NDT suffers because a great deal of time is spent on signal control, or ACK and NAK. A typical NDT curve is shown Fig. 9-1.

Hamming Distance

In detecting errors, the list of code words is restricted. If a message is received which is not on the restricted list, it must be in error.

An important property of a code is its Hamming distance, which is equal to the number of bits by which two code words differ. Thus, the Hamming distance between the words 1 0 0 1 0 0 and 1 0 1 0 1 0 is

Table 9-6 Block Check Character

CHARACTERS		STX	N	A	I	T	I	S	O	K	ETB.	BCC. ← Column Parity Bits
Message Bits	b_1	0	0	1	1	0	1	1	1	1	1	1
	b_2	1	1	0	0	0	0	1	1	1	1	0
	b_3	0	1	0	0	1	0	0	1	0	1	0
	b_4	0	1	0	1	0	1	0	1	1	0	1
	b_5	0	0	0	0	1	0	1	0	0	1	1
	b_6	0	0	0	0	0	0	0	0	0	0	0
	b_7	0	1	1	1	1	1	1	1	1	0	0
Character Parity Row Parity Bits	b_8	1	0	0	1	1	1	0	1	0	0	1

(a) Transmitted Message Block

		N	A	I	~	I	S	O	K		
	b_1	0	0	1	1	0	1	1	1	1	1 ← Column Parity Check
	b_2	1	1	0	0	0	0	1	1	1	0
	b_3	0	1	0	0	1	0	0	1	0	0
	b_4	0	1	0	1	(1)	1	0	1	1	0 1 X
	b_5	0	0	0	0	1	0	1	0	0	1 1
	b_6	0	0	0	0	0	0	0	0	0	0 0
	b_7	0	1	1	1	1	1	1	1	1	0 0
Row Parity Check	b_8	1	0	0	1	1	1	0	1	0	0 1
						X					

(b) Detected Message Block

X Indicates an Error in the Parity Check

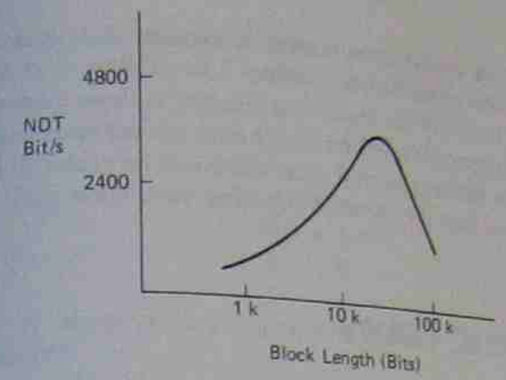


Fig. 9-1 Net Data Throughput

three, since the second, third, and fourth bits, counting from the right, are different.

A Hamming distance of one between code words, such as in the Baudot code, does not permit error detection since all combinations are valid. In the ASCII code with even parity and a Hamming distance of two, single errors can be detected since a change in any one bit produces an error code. If a code has a Hamming distance of three, any single error will be closer to its original code word than that of another code word. For instance, if the valid codes of a three bit word are 000 and 111—for a Hamming distance of 3—a change in one bit will still more closely resemble the valid code word and the bit in error can be corrected.

This concept is readily illustrated by the use of the geometrical models of Fig. 9-2. The vertices of the three dimensional cubes represent the valid or invalid code words. In Cube a, the Hamming distance is 1 and all the code words are valid. An error in any code word resembles another valid combination, thus not allowing for any detection of errors.

In Cube b, the Hamming distance is 2, and error detection is possible since any single error will result in an invalid code word. Since the

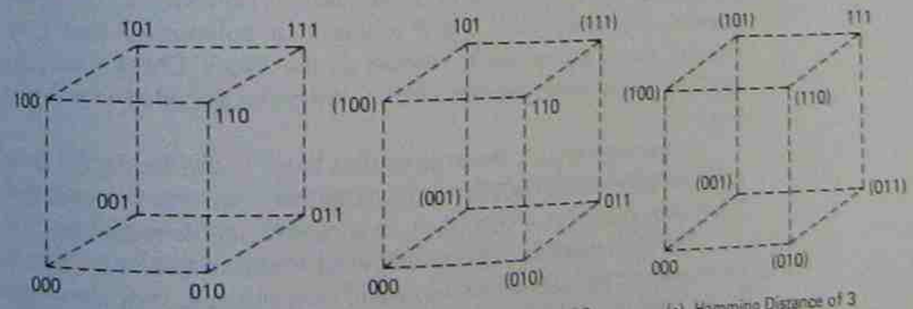


Fig. 9-2 Geometric Models of Three-Digit Code Words. (Invalid Code Words In Parentheses.)

distance from the invalid combination to adjacent valid code words are identical, no error correction is possible.

In Cube c, where the Hamming distance is three, both single and double error detection is possible. Single error correction is now also possible, since the resulting invalid code word will be closer to its original valid code word than the value of any other valid code word.

9–4 CYCLIC CODES

In forward-error correction schemes, bits are added to the data in a way that allows for error detection and correction by the receiver. One such scheme is the combined row-column parity check discussed previously. This technique, however, is not foolproof since it is possible to have double bit errors in rows occur simultaneously with double bit errors in columns, which results in undetected errors. This is of particular concern in communication transmission systems where noise tends to be impulsive and bursts of errors result.

The cyclic codes are one set of codes which are very effective in detecting error bursts. The use of a feedback mechanism, by which any transmitted code word is very much dependent upon the previous history of the message, makes it highly unlikely for an error burst to produce an acceptable or undetected received, encoded message. The hardware required to perform cyclic coding is minimal, but the code efficiency is somewhat poor, as several check bits must be added to the message bits. The cyclic redundancy check code known as the CRC-12 has a 99.955% chance of detecting error bursts of 12 bits in length.

Cyclic redundancy codes add several bits (up to about 25) to the shifted message encoded bits. The total data, or message, block is usually considered one lengthy binary polynomial, $P(x)$. At the transmission end, this message block is divided by a fixed binary polynomial, called the generator polynomial $G(x)$, which results in a polynomial plus a remainder, $R(x)$. This remainder, known as the Block Check Character (BCC), is added to the message block to form the encoded transmitted block.

At the receiver end, the received data block is divided by the same generator polynomial, $G(x)$. If there is a remainder, an error has occurred. Either error correction is performed, or retransmission is requested.

The serial data bits of the various code words in CRC's are usually represented as polynomials for the sake of convenience. They are written in the form $x^n + x^{n-1} + \cdots + x^2 + x^1 + x^0$, where x^n represents the highest order bit present and x^0 represents the least significant bit. For example, the binary word 10101 is represented by $x^4 + x^2 + 1$. The degree of the polynomial is one less than the number of bits in the code word.

The addition or subtraction used in calculating the polynomial is called Modulo 2 addition and is performed by exclusive–OR gates. In this arithmetic

$$\begin{aligned} 0 \oplus 0 &= 0 \\ 0 \oplus 1 &= 1 \\ 1 \oplus 0 &= 1 \\ 1 \oplus 1 &= 0 \end{aligned} \qquad (9\text{-}2)$$

Subtraction is identical to addition. Adding two polynomials, for example, results in

$$\begin{array}{r} x^4 \quad + x^2 + x \\ x^3 + x^2 + x \\ \hline x^4 + x^3 \end{array}$$

Since multiplication is nothing but continued self-addition, when the two polynomials above are multiplied, the result is

$$\begin{array}{r} x^4 + x^2 + x \\ x^3 + x^2 + x \\ \hline x^7 \quad + x^5 + x^4 \\ x^6 \quad + x^4 + x^3 \\ x^5 \quad + x^3 + x^2 \\ \hline x^7 + x^6 \quad + x^2 \end{array}$$

Dividing $x^5 + x^3 + x$ by $x^2 + x + 1$ results in

$$\begin{array}{r} x^3 + x^2 + x \\ x^2 + x + 1 \overline{)\, x^5 \quad + x^3 \quad + x} \\ \underline{x^5 + x^4 + x^3} \\ x^4 \\ \underline{x^4 + x^3 + x^2} \\ x^3 + x^2 + x \\ \underline{x^3 + x^2 + x} \end{array}$$

If a remainder is present, it can always be expressed to a degree one less than the degree of the divisor.

The code words mentioned above can be defined as follows:

$M(x)$ Originating message polynomial consisting of m bits
$G(x)$ Generator polynomial
$T(x)$ Transmitted polynomial consisting of t bits
$E(x)$ Error polynomial

The encoded message, or transmitted polynomial, is obtained in several steps. First, the originating message, $M(x)$, is shifted c bits to the left, where c is the number of check bits to be added. The value of c is equivalent to the degree of the generator polynomial. The shifted message can be analytically expressed as $M(x)x^c$.

This shifted message is next divided by the generator polynomial, $G(x)$. The remainder of this division operation is added to the shifted message to generate the transmitted polynomial, $T(x)$. This polynomial can always be exactly divided by $G(x)$. Thus,

$$\frac{M(x)x^c}{G(x)} = Q(x) + \frac{R(x)}{G(x)} \tag{9-3}$$

where $Q(x)$ is some quotient and $R(x)$ is the remainder, which has a degree one less than the generator polynomial. Multiplying through by $G(x)$

$$M(x)x^c = Q(x)G(x) + R(x) \tag{9-4}$$

Since subtraction is identical to addition, then

$$T(x) = M(x)x^c + R(x) = Q(x)G(x) \tag{9-5}$$

Thus, it is proven that T(x) can always be exactly divided by $G(x)$.

Consider the following example:

1. Let Message polynomial $M(x) = 101101$ $(x^5 + x^3 + x^2 + 1)$, and Generator polynomial $G(x) = 11001$ $(x^4 + x^3 + 1)$. $G(x)$ contains 5 bits and is of degree 4. Therefore, the $M(x)$ must be shifted by $c = 4$ bits. The following shifted polynomial is obtained by multiplying $M(x)$ by x^4:

$$M(x) \cdot x^4 = (x^5 + x^3 + x^2 + 1)\,(x^4) = x^9 + x^7 + x^6 + x^4$$

2. Dividing this product by the generator polynomial $G(x)$ gives

$$\begin{array}{r|l} & x^5 + x^4 + x^2 \\ x^4 + x^3 + 1 & x^9 + x^7 + x^6 + x^4 \\ & x^9 + x^8 + x^5 \\ \hline & x^8 + x^7 + x^6 + x^5 + x^4 \\ & x^8 + x^7 + x^4 \\ \hline & x^6 + x^5 \\ & x^6 + x^5 + x^2 \\ \hline & + x^2 \end{array}$$

remainder →

The remainder, $R(x) = x^2$, is added to $M(x)x^4$ to give the transmitted polynomial

$$x^9 + x^7 + x^6 + x^4 + x^2$$

Thus, the transmitted code word is

1011010100

If there is no error at the receiver, this message will produce no error when divided by the same generator polynomial.

As a check

$$\begin{array}{r|l} & x^5 + x^4 + x^2 \\ x^4 + x^3 + 1 & x^9 + x^7 + x^6 + x^4 + x^2 \\ & x^9 + x^8 + x^5 \\ \hline & x^8 + x^7 + x^6 + x^5 + x^4 \\ & x^8 + x^7 + x^4 \\ \hline & x^6 + x^5 + x^2 \\ & x^6 + x^5 + x^2 \\ \hline & 0 \end{array}$$

no remainder

To implement this particular code, the feedback shift register shown in Fig. 9-3 is used. The number of shift register gates required is equivalent to the order of $G(x)$ and c.

After initially clearing the shift register, the message, or dividend, is entered. The highest power of x corresponds to the first digit which arrives. Since no feedback occurs until a 1 dividend digit arrives at the far or

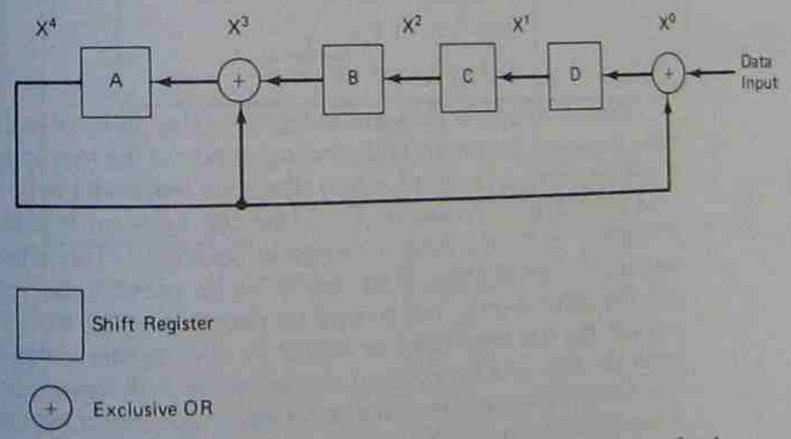


Fig. 9-3 Circuit to Obtain Remainder Bit With $G(x) = x^4 + x^3 + 1$

left end, the first four operations are merely simple shifts. Whenever a 1 digit arrives at the left end, a 1 is subtracted from the second and fifth digits on the next shift. This is identical to the process just outlined. The message bits are followed by c 0's. What is left in the register forms the remainder, which is appended to form the transmitted message. For the message and generator polynomial previously given, the sequences in the shift register are given in Table 9-7.

Table 9-7 Shift Register Sequence With $G(x) = x^4 + x^3 + 1$

Shift No.	A	B	C	D	
Initial Clear	0	0	0	0	
1	0	0	0	1	1
2	0	0	1	0	0
3	0	1	0	1	1
4	1	0	1	1	1
5	1	1	1	1	0
6	0	1	1	0	1
7	1	1	0	0	0
8	0	0	0	1	0
9	0	0	1	0	0
10	0	1	0	0	0

Entry of Data

Entry of C = 4 Zeros

Remainder Added to the Shifted Message

$M(x) = 1\ 0\ 1\ 1\ 0\ 1$

$T(x) = 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 0$

The full encoder is illustrated in Fig. 9-4. The delay of 4 shifts (c shifts, in general) is necessary to allow computation of the remainder. Initially the switch is in position 1, which allows the remainder to be calculated after $M(x)x^c$ is divided by $G(x)$. After the message is shifted by adding 4, or c, 0's, the switch is moved to position 2. This allows the remainder, which is located in the register, to be added to the message bits. The identical circuit can be used for decoding. After the received message $R(x)$ has been added or shifted in, the register contains the remainder. If an error has occurred, this remainder will be non-zero.

Often, higher-order generator polynomials are used. One example is the CRC-16 code that has a generator polynomial of $G(x) = x^{16} + x^{15} + x^2 + 1$. It can be implemented by the shift register shown in Fig. 9-5.

Hamming Codes

As discussed in the previous section, the ability to detect errors is related to the Hamming distance between code words. This, in turn, is related to the ability to correct errors. Table 9-8 outlines these relationships.

As an example, consider a case in which the Hamming distance is 3. If a single error occurs, correction can be attempted by trying to change each bit. If the wrong bit is changed, a double error will occur, which is detectable. Only the erroneous bit, when changed, will result in a correct combination.

One of the first cyclic codes devised for error correction was developed by R. W. Hamming of Bell Laboratory. The Hamming code can correct for single errors and detect double errors in a block.

To develop a Hamming code for single-error correction, assume that c check bits are added to the message, M, to form the transmitted word, T, of t bits. With c check bits, 2^c possible error patterns can be indicated. If one of these combinations indicates no error, then $2^c - 1$ combinations remain to indicate where the error is located. Therefore, the length of the transmitted code word, t, must be equal to or less than $2^c - 1$, or

$$t + 1 \leq 2^c \tag{9-6}$$

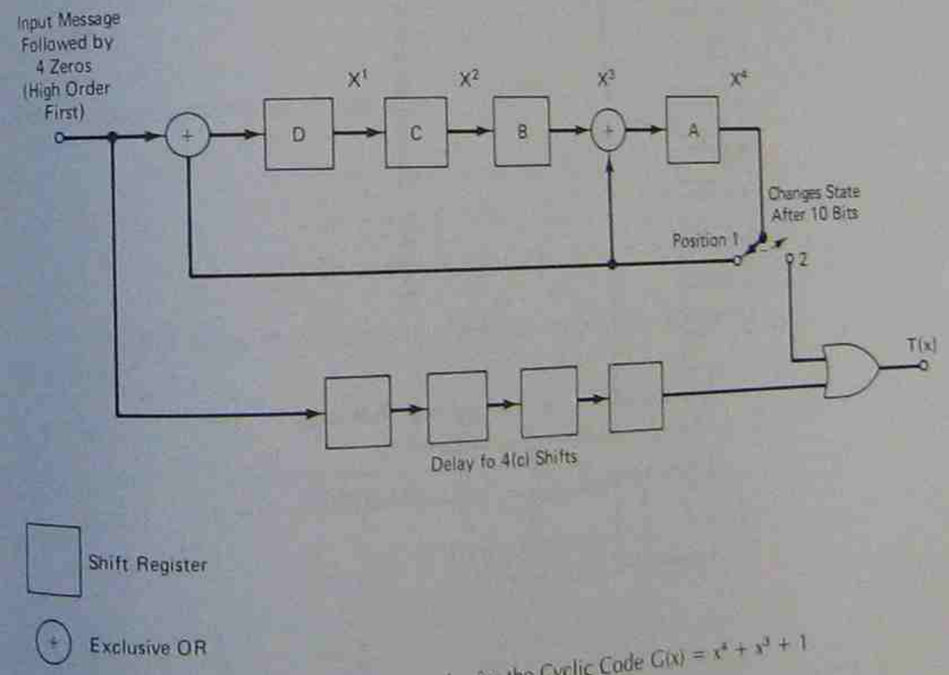


Fig. 9-4 Complete Encoder for the Cyclic Code $G(x) = x^4 + x^3 + 1$

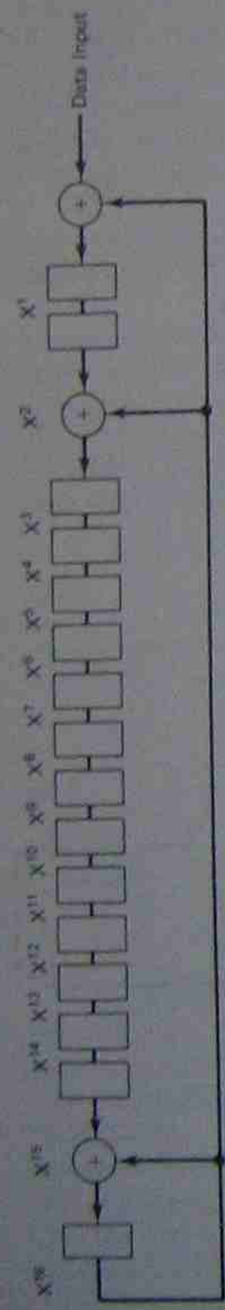


Fig. 9-5 Generation and Detection of the CRC-16 Cyclic Code

Table 9-8

Hamming Distance	Model	Errors Detected	Errors Corrected
1	□ □	0	0
2	□ * □	1	0
3	□ * * □	2	1
4	□ * * * □	3	1
5	□ * * * □	4	2
d Even		d-1	$\frac{d-2}{2}$
d Odd		d-1	$\frac{d-1}{2}$

□ Legal Code Word
* Illegal Code Word
↶ Error Correction

For convenience, transmit the maximum allowable code word, or

$$t + 1 = 2^c \tag{9-7}$$

If the message consists of m bits, then

$$t = m + c \tag{9-8}$$

$$m + c + 1 = 2^c$$

$$m = 2^c - c - 1 \tag{9-9}$$

Table 9-9 gives the number of check bits required for a message m bits long.

Table 9-9 Number of Check Bits Required for Single-Error Correction

c (check bit)	m (message length)
2	1
3	4
4	11
5	26
6	57
7	120

As an example of how such a scheme is implemented, number the transmitted code word of 7-bit length from right to left. From Table 9-9, $m = 4$ and $c = 3$. In the 7-bit code shown in Table 9-10, the check bits, or parity bits, are c_2, c_1, c_0, and the message bits are m_3, m_2, m_1 and m_0. (The subscripts give the weighting of a particular bit.) Now consider the rationale for the location of the check bits.

Table 9-10

m_3	m_2	m_1	c_2	m_0	c_1	c_0	bit notation
7	6	5	4	3	2	1	bit location

Upon detection, the check and message bits are checked for parity. If *binary weighting* is applied to the three parity checks of p_2, p_1, and p_0, and p_0 is the LSB, the following will be obtained for even parity:

$p_0 = 1$ when the error is in location 1, 3, 5, 7

$p_1 = 1$ when the error is in location 2, 3, 6, 7

$p_2 = 1$ when the error is in location 4, 5, 6, 7

If each parity is required to check one unique check bit in addition to the three message bits, then by observing the above, the c_0 check bit should be in location 1, c_1 check bit in location 2, and c_2 check bit in location 4. The other locations are excluded because the bits in these locations are repeated.

At the encoder, even parity for a binary weighting on the check bits is obtained when

$$c_0 \oplus m_0 \oplus m_1 \oplus m_3 = 0$$
$$c_1 \oplus m_0 \oplus m_2 \oplus m_3 = 0$$
$$c_2 \oplus m_1 \oplus m_2 \oplus m_3 = 0$$

Since subtraction is the same as addition

$$c_0 = m_0 \oplus m_1 \oplus m_3$$
$$c_1 = m_0 \oplus m_2 \oplus m_3 \qquad (9\text{-}10)$$
$$c_2 = m_1 \oplus m_2 \oplus m_3$$

Example 9-1 Determine the Hamming code word which corresponds to the message $m_3, m_2, m_1, m_0 = 1101$. Use even parity.

$$c_0 = 1 \oplus 0 \oplus 1 = 0$$
$$c_1 = 1 \oplus 1 \oplus 1 = 1$$
$$c_2 = 0 \oplus 1 \oplus 1 = 0$$

The coded word, T, is 1100110.

Assume that a single error has occurred during the transmission and that the received message is

$$\underbrace{1}_{m_3}\ \underbrace{1}_{m_2}\ \underbrace{0}_{m_1}\ \underbrace{1}_{c_2}\ \underbrace{1}_{m_0}\ \underbrace{1}_{c_1}\ \underbrace{0}_{c_0}$$

Checking for even parity:

$$p_0 = c_0 \oplus m_0 \oplus m_1 \oplus m_3 = 0 \oplus 1 \oplus 0 \oplus 1 = 0$$
$$p_1 = c_1 \oplus m_0 \oplus m_2 \oplus m_3 = 1 \oplus 1 \oplus 1 \oplus 1 = 0$$
$$p_2 = c_2 \oplus m_1 \oplus m_2 \oplus m_3 = 1 \oplus 0 \oplus 1 \oplus 1 = 1$$

Thus, bit 4 ($p_2p_1p_0 = 100$) is in error and the correct word is 1 1 0 0 1 1 0.

Odd parity, can be used as well as even parity. For longer messages, the check bits are located in positions which have a power of 2; that is,

$$2^0, 2^1, 2^2, 2^3, 2^4$$

Although the error detection and correction schemes outlined are very efficient, they are often incompatible with the types of codes required in a typical communication system. For instance, to assure synchronization of the receiver at all times, the encoded signal should undergo transitions continuously. In addition, due to hybrid transformers in a cable system, dc cannot be transmitted and, therefore, the number of 1's and 0's cannot be balanced out to give a 0 dc level. These considerations conflict with the efficient coding schemes just described.

9–5 LINE CODES

As just discussed, ideal line codes have their own desired properties. The following list outlines major considerations related to these properties.

1. All cable systems, and most other communication systems, will not

allow dc transmission due to either transformer action or, more likely, some low frequency, cut-off restriction. In the case of NRZ (non-return-to-zero) coding (see Fig. 9-6), this restriction forces all the positive and negative elements to have equal probabilities. For this reason, there should not be long strings of element sequences that maintain the same polarity.

2. Because any code adds redundancy, the code efficiency should be as high as possible. Increasing the element rate, or increasing the number of distinguishable levels, requires either an improvement in the S/N ratio or an increase in the bandwidth, both of which are costly.

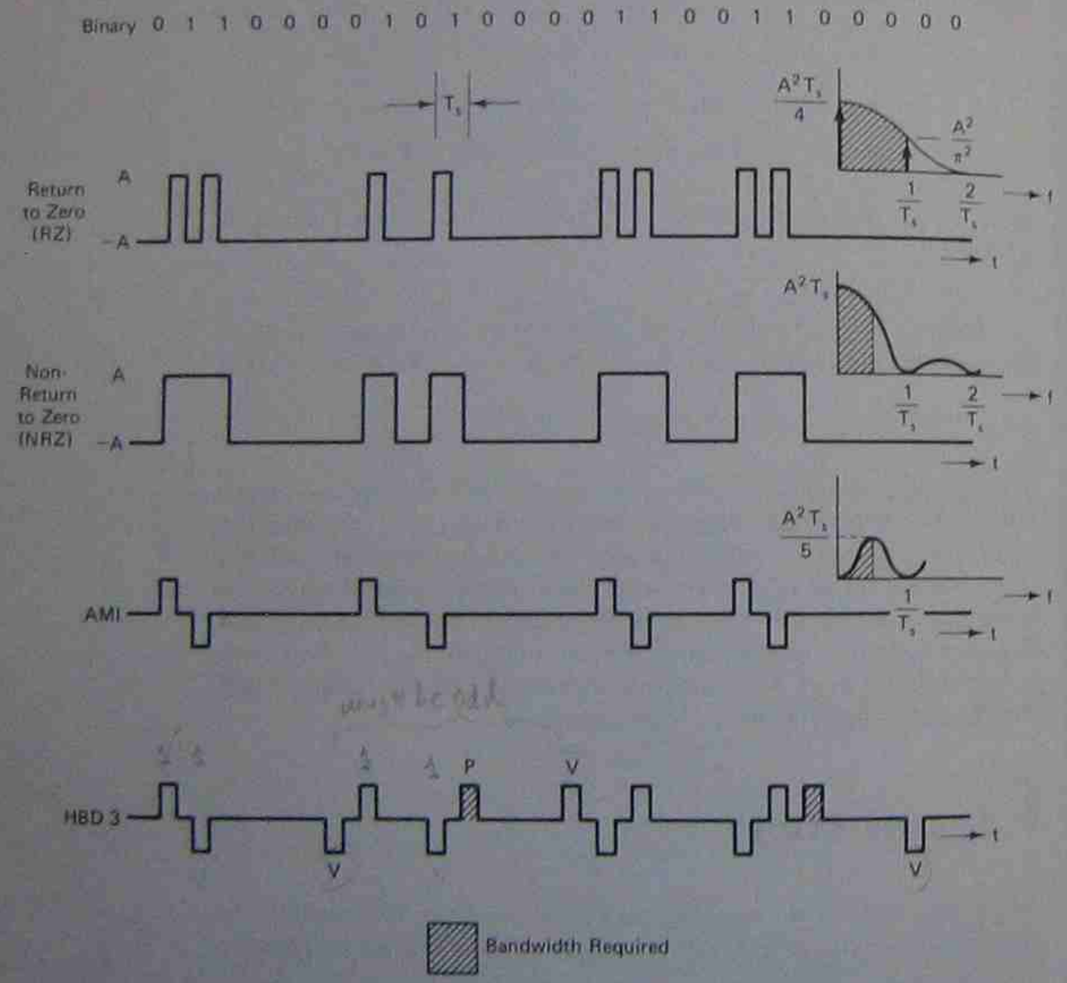


Fig. 9-6 Typical Line Codes with their Power Spectra

3. In order to obtain synchronization or timing at the receiver, the receiver usually locks in on the carrier being received. The derivation of the timing can be obtained only if the line signal undergoes a sufficient number of 0 crossings; that is, the transmitted signal should always undergo transitions.

Some of the line codes which take into account these considerations are the Alternate Mark Inversion (AMI) code, also known as the Bipolar, and the High Density Bipolar code (HDB3).

The AMI code is a three-level code (+, −, and 0). The 0 level represents the binary zero, and the alternative + or − represents the binary 1. Figure 9-6 shows the implementation of the AMI code, whereby the polarity of the alternate pulse representing marks are inverted. AMI suffers from excessive redundancy, but it readily detects single errors. A single error causes either an extra pulse to appear on the line or a pulse to be removed. In either case, the result is two successive pulses with the same polarity.

If a long sequence of 0's is transmitted, no transition can occur, and the AMI code can cause a loss of synchronization at the receiver. To overcome this difficulty, AMI is modified to the High Density Bipolar code (HDB). The HDB3 will allow a maximum of three 0's to occur. After that, a violation pulse is added. The rules for transforming from AMI to HDB3 are:

1. Every fourth 0 in a sequence of 0's is converted to a violation pulse *V*. Successive *V* pulses are of opposite polarity.
2. The number of pulses between successive *V* pulses must be odd to ensure that each *V* pulse is opposite in sign to the preceding pulse. If the insertion of a *V* pulse does not result in a bipolar violation, as in the case of an even number of pulses between successive *V* pulses, an additional pulse (*P*) is inserted at the location of the first 0.

Decoding is achieved by replacing every violation pulse *V* and the three digits preceding *V* by 0's. The HBD code is quite redundant and has a low noise immunity.

9–6 CHANNEL THROUGHPUT AND EFFICIENCY

Given a channel with a certain bandwidth and bit error rate, there is an optimum operating condition that will provide for a maximum net data throughput (NDT). The NDT is a measure of the usable characters detected at the receiver. It distinguishes between the total number of bits in the text, including all redundant bits, and the number of correct bits detected in a specified period of time.

Consider as an example the technique of requesting retransmission (ARQ) of a block of information after an error is detected. After a block of information is sent, the transmitting terminal stops and waits for a reply from the receiving terminal regarding whether or not the block was received with no error. If the block is received with no detected error, a positive acknowledgement (ACK) is sent to the transmitter and the transmitter is free to send a new block of information. If an error is detected, a negative acknowledgement (NAK) is returned, requesting a retransmission of the block. In the case of half-duplex transmission, where a channel is operating in both directions but not simultaneously, this technique results in a certain dead time, which reduces the NDT.

Figure 9-7 illustrates the procedure which uses a block-by-block data transfer scheme. In this method, a positive or negative acknowledgement is expected after a message block is transmitted. Special time-out, or shut-down, procedures follow if no response is received. If an ACK signal is received, the next block of information is transmitted; a NAK signal induces a retransmission of the error block.

The dead time, T, between the end of the transmission of one block and the beginning of the transmission of the next block may include:

1. *Modem turn-around times.* (Generally in the case of half-duplex operation.) This delay varies with the type of modem. The most significant time delay occurs between the Request to Send signal in the Data Terminal Equipment (DTE) and the Ready for Sending response in the Data Communication Equipment (DCE).

   Modems with adaptive equalizers also have a period of adjustment during which a short, predetermined pulse pattern is transmitted. This period, which can take up to 100 ms, must be included in the modem turn-around time.

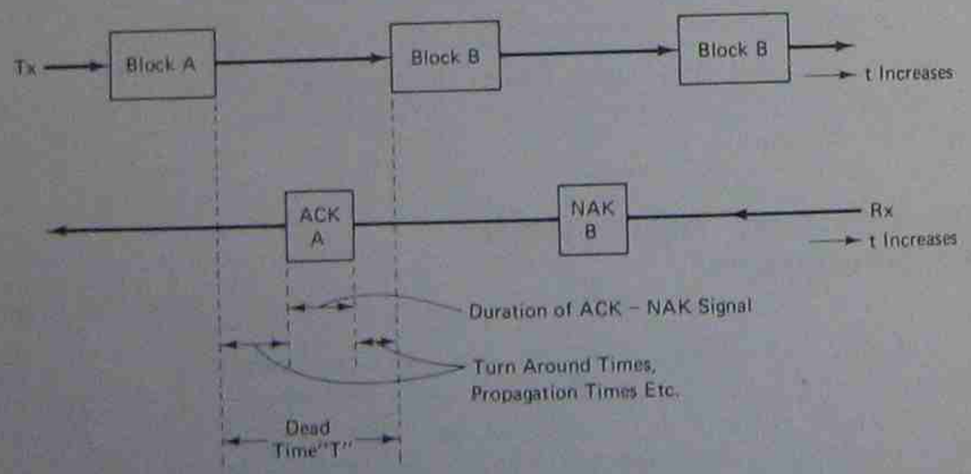


Fig. 9-7 Illustration of Dead Time and Turn Around Time on a Block-by-Block Data Transfer Scheme

2. *Loop propagation time of the system.* Typical propagation times are:

   Unloaded cable = .02 ms/km

   polyethylene coaxial line = 5μs/km

   radio/microwave propagation = 3.33 μs/km

   A link to a geostationary satellite placed 35,800 km above the equator at a slant height of 41,200 km (5° elevation) would require a propagation time of 275 ms from the earth station to the satellite and back again. Thus, the total delay for a satellite loop between two earth stations is around 550 ms, depending upon the exact location of the earth stations. Additional propagation delays occur in the modem filters, equalizer circuits, and other reactive circuits. These additional delays add a few milliseconds to the total delay.

3. *Transmitter and receiver delays.* The DTE at the receive end must check for errors and assemble the message after receiving the last character. It then returns an acknowledgement that must be interpreted by the DTE at the transmit end for appropriate action. These delays are minimal and usually ignored.

4. *Length of the return message.* This message may vary from 1 bit to 32 bits. Depending upon the bit rate of the return channel, this delay may or may not need to be taken into account. On a 75 baud backward channel, for instance, 32 bits would take 32 b/75 b/s, or 426 ms.

5. *Removal of echo suppressors.* Echo suppressors are often added on long transmission paths to attenuate echos from the far end. Since it takes about 150 ms to remove an echo suppressor, a turn-around time of 300 ms per block can be expected. On full-duplex operation, these echo suppressors are removed during the starting-up procedure and remain OFF for the duration of the hook-up.

A typical curve showing how the NDT varies with block length is illustrated in Fig. 9-8.

Small block lengths reduce the NDT because the relative percentage of dead time compared to available transmission time is large, even though the probability of error in any individual block is extremely small. On the other hand, extremely long block lengths have a high probability of error which results in a high likelihood of retransmission requests, which results in a reduction of NDT.

Data rates can be increased by using a circuit with a reverse, or secondary, channel in a duplex system. In such a system, a primary channel with a high signaling capacity and a secondary channel that flows in the opposite direction are present simultaneously. (In a full-duplex channel, the signaling rate capability is the same in both directions.) In data trans-

mission, the information block is put on the primary channel in the forward direction, and the ACK or NAK signals are placed in the secondary or reverse channel.

One half-duplex system with a backward channel for error control is the 1200-baud modem, which fits the CCITT standard for the general switched telephone network. It transmits data through the use of a 2100 Hz tone that represents a SPACE, or binary 0, and a 1300 Hz tone that represents a MARK, or binary 1. This method of modulation is called frequency shift keying (FSK). The backward channel, (MARK, 390 Hz; SPACE, 450 Hz), is used for error control and can handle up to 75 baud. Fig. 9-9 illustrates the frequency spectrum used by both the forward transmitted data and the reverse channel.

In an error-free circuit, the throughput efficiency of the data transmission system can be calculated by obtaining the ratio of the time taken

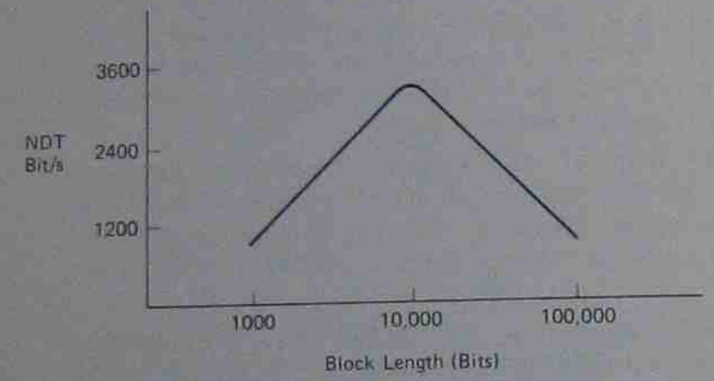


Fig. 9-8 Net Data Throughput as a Function of Block Length

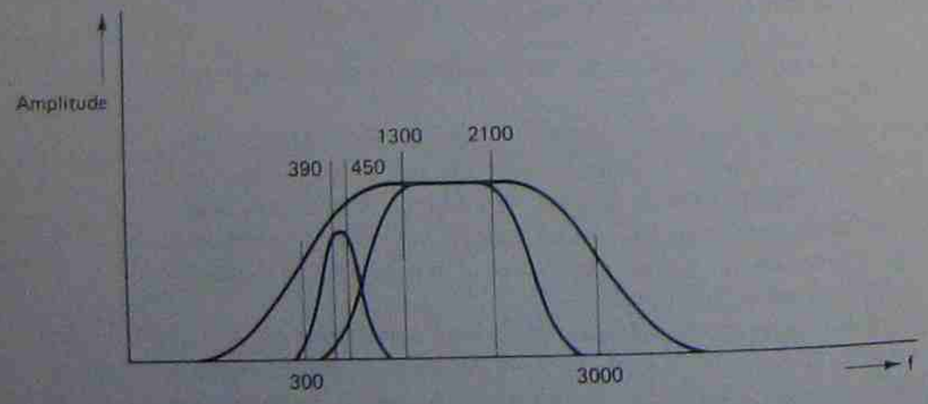


Fig. 9-9 Frequency Spectrum Utilization of a 1200-Baud Half-Duplex System with Backward Channel

in transmitting a data block (B/R) to the sum of the data block times and dead time T. This is given by:

$$\eta = \frac{\frac{B}{R}}{\frac{B}{R} + T} \times 100\% \qquad (9\text{-}11)$$

where η = the throughput efficiency
B = the block length in bits
T = the total dead time
R = the input data rate in b/s

Equation (9-11) ignores the losses due to parity and synchronization bits sent with each data block. If desired, these can be taken into account separately.

Example 9-2

Obtain the throughput efficiency of the 1200-baud half-duplex system shown in Fig. 9-9. Assume the following conditions:

B = 1,000 bits

Length of backward or return message = 16 bits

Length of path employing polyethelyene cable = 800 km

Modem turn-around time = 200 ms

Accumulation of modem delay times, Tx and Rx-delays = 12 ms

For this system R = 1200 b/s

Time taken for the 16 bits of the return message at 75 b/s = $^{16}/_{75}$ = 213 ms

Solution

Time taken for propagation delay (loop delay) = 5μs/km × 2 × 800 = 8 ms

Time taken for modem turn-arounds = 400 ms

Time taken for modem Rx + Tx delay = 24 ms

Thus T = (213 + 8 + 400 + 24) ms = 645 ms

$$\eta = \frac{\frac{1000}{1200} \times 1000}{\frac{1000}{1200} \times 1000 + 645} = 56.4\%$$

If errors are present, efficiency is further reduced by a factor of 1 − P(e),

when $P(e)$ is the block error probability. If one block in 100 is in error, then $P(e) = .01$ and the throughput efficiency is

$$56.4\ (1 - .01) = 55.8\%$$

Example 9-3 Consider a full-duplex, four-wire system in which a continuous transmit-and-receive channel is available. This system avoids any modem turn-around problems. Assume the same conditions presented in Example 9-2. Also assume a return channel of 1200 b/s. Since there are no turn-around problems, the following T and throughput efficiency is obtained:

$$T = \left(\frac{16}{1200} \times 1000 + 8 + 24\right) \text{ms} = 45.3 \text{ ms}$$

$$\eta = \frac{\frac{1000}{1200} \times 1000}{\frac{1000}{1200} \times 1000 + 45.3} = 94.8\%$$

This marks a significant increase in efficiency compared to the system with modem turn-around time involvement.

By increasing the data signaling rate, the time taken to transmit each block, (B/R), will be reduced and T will become proportionally larger. This will result in a reduced throughput efficiency. To compensate for this, it would be desirable to use longer blocks for the higher data rates. Longer blocks, however, have a greater probability of error. For this reason, some optimum block length such as that illustrated in Fig. 9-8 can be obtained for any particular system.

With this block-by-block transmission, the transmission efficiency is relatively poor for long dead times. For example, for a half-duplex satellite interconnection with a total loop delay of 550 ms, the efficiency of Example 9-2 is reduced to 41%.

An alternative method that gives greater throughput efficiency is a continuous mode of transmission whereby data blocks are transmitted without interruption, unless a negative acknowledgement is observed. When a NAK is transmitted back to the transmitter, the error block is retransmitted. This technique, which requires a full display system, is illustrated in Fig. 9-10.

The continuous block transfer system avoids the dead time of the block-by-block scheme but requires much more storage, or buffering. In block-by-block transmission, only one block at a time needs to be stored at either end. The next block is not sent until the first block has been correctly received. In continuous block tramsmission, when one of the

blocks is not received correctly, all of the subsequent blocks must be stored until the error block is acknowledged. Such a system can run out of buffer space very quickly, particularly if several blocks are in error or if one of the input lines is error prone. This can cause the entire system to slow down or cut off, which can also happen when equipment problems or severe signal fading occurs on one line. A method of resolving this problem by limiting the storage at each end to 7 blocks is discussed in the next chapter.

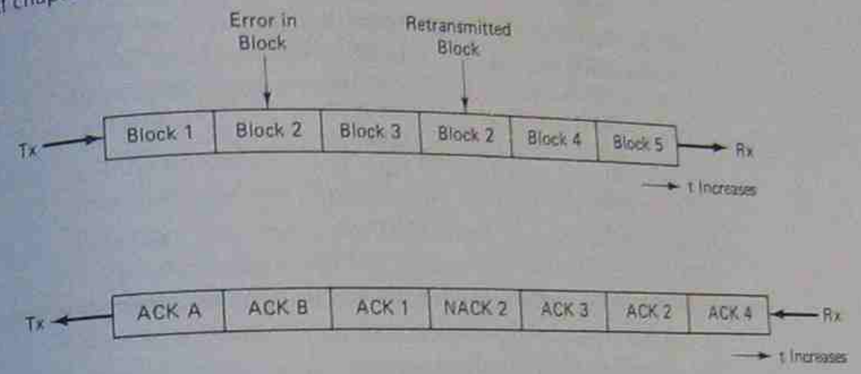


Fig. 9-10 Continuous Block Transfer Scheme

PROBLEMS

1. Indicate which of the following codes are weighted and which are not weighted.
   a. 8421 BCD
   b. BCD Excess 3
   c. 4221 BCD
   d. Biquinary
   e. Gray Code
2. What is the weighting of the following codes?

a. Decimal	Code
0	0000
1	0001
2	0011
3	0100
4	0101
5	0111
6	1000
7	1001
8	1011
9	1100

b. Decimal	Code
0	00000
1	00001
2	00011
3	00111
4	01111
5	10000
6	11000
7	11100
8	11110
9	11111

3. Which of the codes mentioned in Problems 1 and 2 are self-complementing?
4. Fill in the parity column in each of the following tables to add the indicated parity bit to each word.

a. Code	Even Parity
01001	
01100	
10000	
11111	
01100	
10100	
11101	
01110	
00000	
10101	

b. Code	Odd Parity
1000001	
1000010	
1000011	
1000100	
1000101	
1000110	
1000111	
1001000	
1001001	
1001010	

5. Briefly explain what each of the following terms means, and name a code which exemplifies each term.
   a. Exact Count Code
   b. Unit Distance Code
6. The following message is to be transmitted in ASCII with odd parity for the character check (row, or horizontal, parity bit) and with even parity for the block check character (column, or vertical, parity check). Give the transmitted message block.

S	S	S	S				S			S			E	B
Y	Y	Y		A	L	L	P	I	S	P	G	O	T	C
N	N	N	X				A			A			X	C
							E			E				

BCC Accumulation

7. Why is the net data throughput lower when the block length is exceptionally short or long?
8. Why are cyclic codes effective in detecting error bursts?
9. The message 1001001010 is to be transmitted in a cyclic code with a generator polynomial of $G(x) = x^2 + 1$.
   a. How many check bits does the encoded message contain?
   b. Obtain the transmitted code word.
   c. Draw the encoding arrangement to obtain the remainder bits.
   d. The decoder has the same arrangement as the encoder. After the received code word is clocked into the decoder input, what should be the content of the register stores?

10. Sketch the encoding/decoding arrangement for the 16-bit generator polynomial $x^{16} + x^{12} + x^5 + 1$. (CCITT recommendation is V4.)
11. a. What is the "Hamming distance"?
    b. For a Hamming distance of 5, how many errors can be detected? How many can be corrected?
12. a. Determine the Hamming coded word for the message $m_3, m_2, m_1, m_0 = 1010$. Employ even parity and apply binary weighting to the three check bits.
    b. If the received message is 1110010, what is the correct transmitted word, assuming that only a single error may occur?
    c. Regardless of the message length, where should the check bits be located?
13. a. List the properties a line code should have.
    b. Describe how well the AMI code meets the conditions of (a).
14. a. Describe why the continuous block transfer scheme has a better throughput efficiency than the block-by block data transfer scheme. List one disadvantage of the continuous block transfer scheme.
15. List several of the chief contributors to the dead time in the block-by-block transfer scheme.
16. a. Obtain the throughput efficiency of a half-duplex system under the following conditions:

    Block length (B) = 5000 bits

    Forward (and backward) data rates (R) = 2400 b/s

    Length of return message = 16 bits

    Length of single satellite loop = 82,400 km

    Modem turn around time = 100 ms

    Accumulation of modem delay times, Tx and Rx delays = 12 ms

    b. If the data signaling rates of (a) are increased to 4800 b/s, what happens to the throughput efficiency?
    c. Based on (b), what should happen to the block length as the signaling rate is increased? Since the block error probability rises as the block length is increased, is there some optimum block length? Explain.
17. Why is a greater throughput efficiency obtained in a full-duplex mode than in a half-duplex mode in block-by-block transmission?

10 Network Protocol*

** It is a pleasure to have Mr. Tom McGovern as the author of this chapter on network protocols. He is presently with the Computer Systems group at the Northern Alberta Institute of Technology in Edmonton.*

10–1 INTRODUCTION

Network is a broad term similar in meaning to system. Both are used to describe organized complexity. A network implies size, remoteness, and interrelationship. Data Communication Networks are concerned principally with the transfer of information between computers and computer terminals. As the number of terminals and computers in a network increases, problems concerned with the design, efficiency, cost, and control of such a network become paramount.

To begin with, a network may be described as an interrelated group of nodes. These nodes may be terminals or junctions. Terminals are any devices connected to the network, including computers and computer terminals, for the purpose of sending or receiving information. Junctions perform such functions as completing a link between two terminals (switching), deciding on which order the various information transfers should take place (scheduling), and allocating temporary storage of information along the route of its travel (buffering). Figure 10-1 illustrates the basic components of a network and the interrelationship of the nodes.

10–2 SWITCHING

The need for sophistication in protocols developed as networks grew in complexity. The first connections between terminals were made *point-*

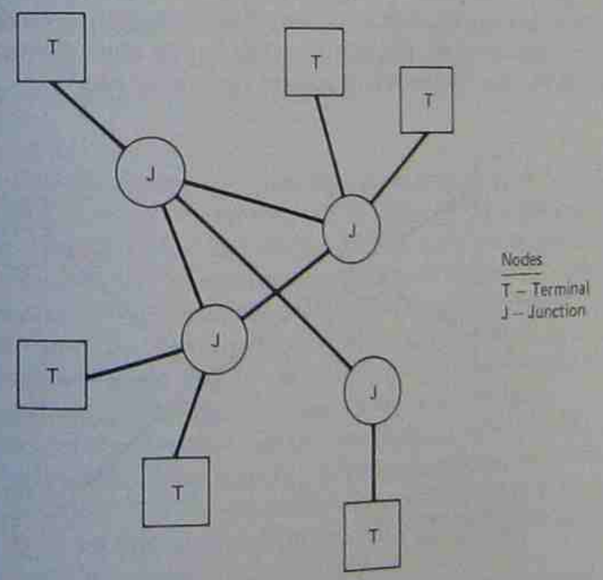
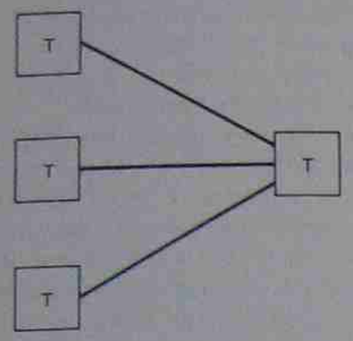


Fig. 10-1 Network

to-point, as shown in Fig. 10-2. In point-to-point connection, each terminal is connected to any other terminal by an individual and permanent line. The obvious disadvantages of this system include cost and underutilization of lines; that is, the user pays for the line whether it is being used or not.

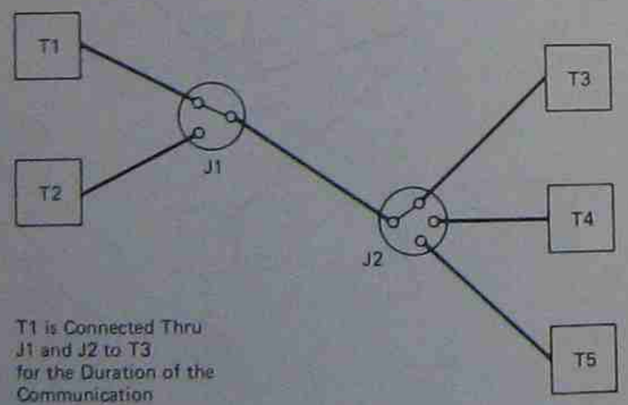


T – Terminal

Fig. 10-2 Point-to-Point Connection

Circuit and Multipoint Switching

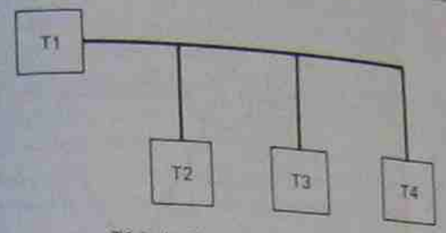
One solution found to the drawbacks of point-to-point switching was *circuit switching*, which is shown in Fig. 10-3. In this system, a switching center is created to provide a path between two terminals for the duration of the communication. The user pays only for the time connected and for sharing the lines and switching centers with other users.



T1 is Connected Thru
J1 and J2 to T3
for the Duration of the
Communication

Fig. 10-3 Circuit Switching

An alternative improvement over point-to-point connection is a *multipoint*, or multidrop, network that facilitates a one-to-and-from-many terminal connection. This is shown in Fig. 10-4. One terminal controls the data transfer between itself and all other terminals by means of polling.



T1 is the Controlling Terminal

Fig. 10-4 Multipoint Network

Both circuit switching and multipoint networks have their drawbacks. Circuit switching times can be high and therefore present a problem for interactive systems. Most users of a remote inquiry system would expect an almost immediate response—for example, in checking customer credit ratings. Even a delay of a few seconds, when experienced on a regular basis, becomes intolerable.

In multipoint systems, polling implies a programming overhead; that is, time must be spent checking each terminal to find out if there is a requirement for data transfer from a particular terminal. This is accomplished by the execution of a program in the controlling terminal. The cost of polling is normally distributed among the terminal users. Another problem of multipoint systems is underutilization of terminals due to the fact that there is only one connecting line.

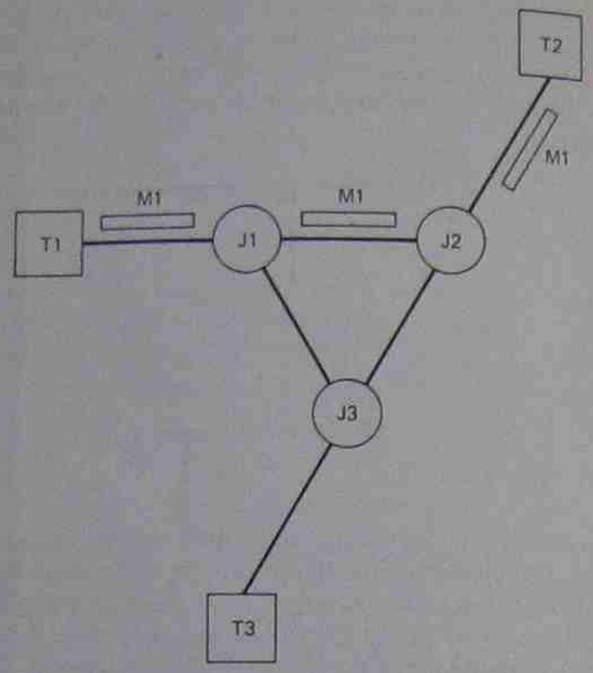
Packets

Two recent developments in computer communication systems that herald the development of new methods of information handling are the emergence of world-wide data networks and the use of packet switching.

The establishment of standards within countries and across continents is essential to the development of such data processing applications as electronic funds transfer. The network of the future must be accessible to a wide variety of users, and it must perform its own housekeeping functions; i.e., it must be intelligent and use its resources efficiently.

The second development, packet switching, is a refinement of message switching that involves transferring messages between terminals as units of information. (See Fig. 10-5.) The message travels through the network with a header that precedes the message and identifies the destination. Subsequent messages may take entirely different routes.

In packet switching, information is transferred between terminals in



Message M1 is Transmitted from Terminal T1
to Terminal T2 via Junctions J1 and J2

Fig. 10-5 Message Switching

discrete, variable-length units called packets. Figure 10-6 shows that each packet consists of:

a. a header that contains control data and the destination address, and

b. the message or data field, which is transparent to the network.

A Frame Check Sequence (FCS) is attached to the packet for error control at the local level.

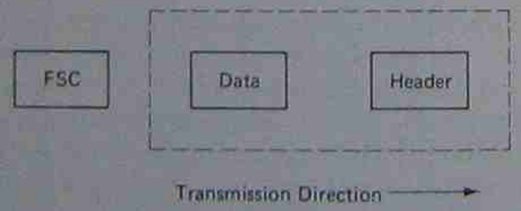
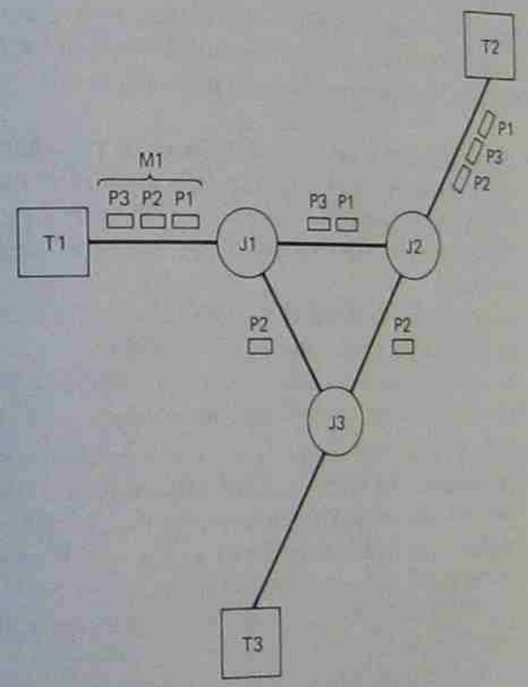


Fig. 10-6 The Packet

Transparency is a term frequently used in computer systems to describe a lack of awareness. For example, the user is often unaware of how

much the operating system does to satisfy a request made in a high-level language. It can be said that the operating system is 'transparent' to the user. In this regard, it can be claimed that the network and the software which drive the network have no interest in the actual information being transmitted. The network is concerned about the source, the destination, the quantity, and the accuracy of the information, but the information itself is transparent to the network.

Packets are transmitted via specialized computers which interpret header information, check for accuracy, and route the data. Packet switching, as shown in Fig. 10-7, does not physically link two terminals. (Compare this with circuit switching.) The network facilities are charged primarily on the basis of quantity of information rather than connection time. Packet switching systems are becoming less sensitive to distance with respect to charging than they were in the past.



Message M1 Becomes Packets P1, P2, P3
P1 and P3 are Transmitted from T1 to T2 via J1, J2
P2 is Transmitted from T1 to T2 via J1, J3, J2
The Packets are Returned to their Original Sequence
at Terminal T2

Fig. 10-7 Packet Switching

10–3 PROTOCOL LEVELS

As networks become more complex and wide-reaching and as the requirements for interlinking grow, the need for international protocol standards grows proportionately.

Protocol is defined as a "set of rules." Most familiarly used in a military or diplomatic sense, it is now applied to data communication systems. There are protocols for making connections, transferring messages, and performing enough other functions that they may be considered as a hierarchy of levels or layers. The establishment of a connection between two terminals would obey the lowest level of protocol. The transfer of a file of information to solve a specific problem would follow the protocol of the highest level in the hierarchy, and so on.

Protocol may be defined in terms of its level of "closeness" to the hardware in the same way that computer languages are categorized as high or low level. Four levels of protocol can be identified. (See Fig. 10-8.) Because of variations in terminology, it is necessary to provide more than one name for each level, but these different names refer to the same area of concern. The first mentioned names in the following list will be used in all future references to a particular level.

1. *Physical Level* (Line Interface). Specifies the characteristics of the medium used for linking communication terminals. The standard interface developed by the Electrical Industry Association and Bell System is called RS-232-C (C representing the latest revision) which contains:
   a. the electrical signal characteristics
   b. the interface mechanical characteristics
   c. a functional description of the interchange circuits
   d. specifications for particular applications
2. *Data Link Control* (Frame Level, Line Control). Specifies the use of procedures that ensure the accuracy of data transfer between the network and the sending or receiving station. Between network and station, data link control provides for synchronizing, monitoring and control, and error detection and correction.

   Three major forms of data link control protocol are presently in use. They are:
   a. IBM's Binary Synchronous Communications Protocol (BISYNC). Uses special characters for control of message transmission. See Fig. 10-9.
   b. DEC's Digital Data Communication Message Protocol (DDCMP). A generalized method of data link control that includes message numbering information and a field to keep track of the number of characters in the message. See Fig. 10-10.

   c. X.25 Protocol. Developed by the International Consultative Committee for Telephony and Telegraphy (CCITT) to define the rules governing the use of a public packet switching network. Developed jointly by France, Japan, the United States, the United Kingdom, and Canada. In addition, all nine countries of the European Common Market adopted X.25 as the basis for the proposed packet-switching Euronet Network. In view of its po-

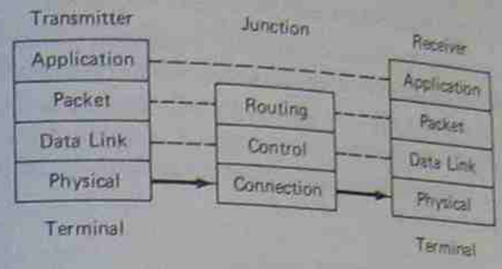
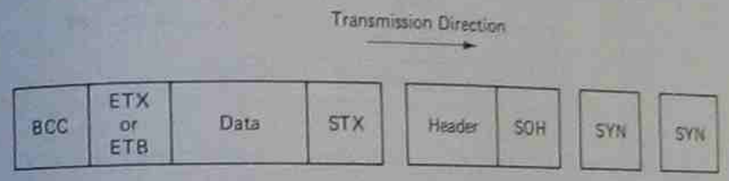
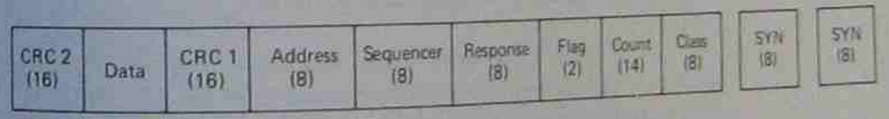


Fig. 10-8 Levels of Protocol



SYN – Synchronisation Character
SOH – Start of Header
STX – Start of Text (Data)
ETX – End of Text
ETB – End of Text Block
BCC – Block Check Character

Fig. 10-9 BISYNC (IBM's Binary Synchronous Communications Protocol)



The Bracketed Numbers Show the Allocation in Bits for Each Control Field for a Total of 12 Bytes (96 Bits) of Control Information.
The Maximum Number of Bytes in the Data Field is 16,383

Fig. 10-10 DDCMP (Digital Equipment Corporation's Digital Data Communications Message Protocol)

tential adoption as a world-wide standard, X.25 will be used to illustrate the concepts, functions, and interrelationships required of network protocols. All transmissions between stations are organized in blocks of information, or frames, with a standard basic structure. See Fig. 10-11.

3. *Packet Level* (Network Level, Information Level). Specifies the manner in which the user establishes, maintains, and clears calls through the network. Also specifies the manner in which control information and user data are structured into packets.
4. *Process Level* (Application Level). Is concerned with the meaning of the communication in terms of who sends the information, to whom it is sent, and for what purpose. Normally communication relates to data collection, inquiries, or changes in existing data. This level of protocol involves the data processing standards of the organizations using the network.

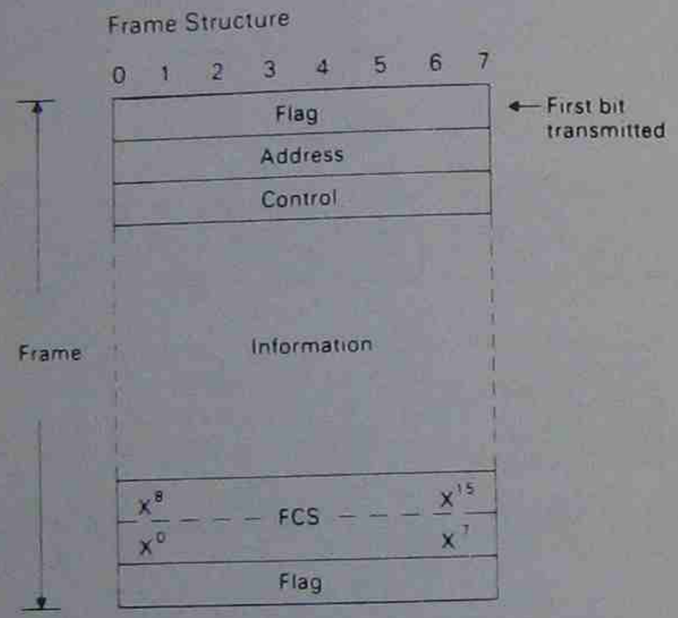


Fig. 10-11 Frame Structure

10–4 DATA LINK CONTROL

The physical level of protocol has a widely accepted standard specification, namely, RS-232-C. The second level of protocol, or data link control, has the several standards listed previously. Two of these standards, BISYNC and X.25, will be discussed in this section.

IBM's Binary Synchronous Communications, a more traditional approach to protocol, is still in common use. Refer back to Fig. 10-9 for the format of a BISYNC message.

SYN is a synchronization character that is used to keep the sending and receiving terminals in step. SYN is added by the sender and removed by the receiver.

SOH (Start of Header) tells the receiver that the information following is control information which relates to addressing and sequencing of user data.

STX (Start of Text) identifies the start of the user's information.

There are two kinds of information that may be transferred:

1. Characters from a standard character set, such as ASCII or EBCDIC. The user's information is then terminated by: EXT (End of Text) or ETB (End of Text Block). See Fig. 10-12.

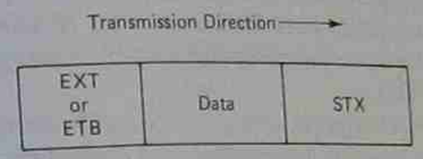


Fig. 10-12 Standard Character Set

2. Data that may contain control characters, from an analog to digital converter, for example. To prevent user data from being "recognized" by the receiver as control characters, BISYNC uses DLE (Data Link Escape). See Fig. 10-13.

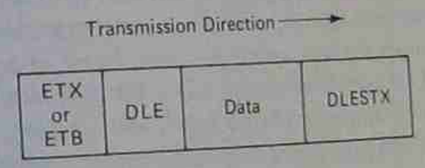


Fig. 10-13 Non-Standard Text

When DLE is first encountered by the receiver, the data that follows is treated as user data; that is, it is not tested for control characters. Any control characters present in the data are "transparent" to the receiver. The only problem, of course, is that DLE is also used to signal termination of user's data. If a pattern of bits occurs in the user's data that is equivalent to a DLE, the transmitter adds a second DLE to indicate that the first one is data. The receiver removes the second occurrence of all pairs of DLE's. The end of user's data is detected by means of DLE ETX or DLE ETB.

BCC is a block check character used for detection of errors which

have occurred during the transmission of the user's data. BISYNC is limited to serial-synchronous, half-duplex lines. Error checking is by VRC/LRC if the code used is ASCII.

Vertical Redundancy Checking (VRC) is used to check each character in a message by means of a parity bit and usually accompanies Longitudinal Redundancy Checking (LRC). Cyclic Redundancy Checking (CRC) is used for messages coded in EBCDIC.

The BISYNC sequence of operation is illustrated in the flowchart of Fig. 10-14.

The transmitter attempts to make contact with the receiver by sending an enquiry signal (ENQ). The receiver may respond with an ACK (acknowledgement) to say "you may send the data." The data to be sent is handled in blocks and the transmitter keeps track of these by a numbering system. The transmitter initializes the block number to 1 and transmits the first block. The receiver acknowledges receipt (ACK). The transmitter checks to see if all blocks have been sent. If not, it sends the next block until all ten blocks have been sent. After all blocks are sent, the transmitter will terminate the communication by sending an end of transmission signal (EOT).

The flowchart in Fig. 10-14 is an oversimplification of the situation, but it does cover initiation, termination, and control of a two-way communication. It contains one obvious flaw. If the receiver sends a NAK signal back to the transmitter indicating an error in the message it received, the flowchart enters an endless loop. This problem can easily be solved by means of a retry counter which would limit the number of times a transmitter would request to send a particular receiver, or the number of times a transmitter would send a particular block of data.

There are many other problems associated with accurate data transmission that have been solved in various ways. For example, to prevent a block of data from being lost completely, two forms of acknowledgement are used, ACK0 for even blocks and ACK1 for odd blocks.

There are contingencies associated with delays incurred in the system. Timeouts are used to create pauses before resending data and re-seeking acknowledgement of accurate transfer.

Figure 10-15 is an improved version of the Fig. 10-14 flowchart. It explains the sequential operation of BISYNC protocol and incorporates the retry counter, busy receiver, and timeouts.

10–4.2 X.25

X.25-compatible equipment is called Data Terminal Equipment (DTE). The processors in a Public Packet Switching Network (PPSN) are called Data Circuit Termination Equipment (DCE). X.25 protocol defines communication between DTE's and DCE's along the lines illustrated in Fig. 10-16. The implementation of this standard will be considered at both the Data Link Control level of protocol and at the Packet Level.

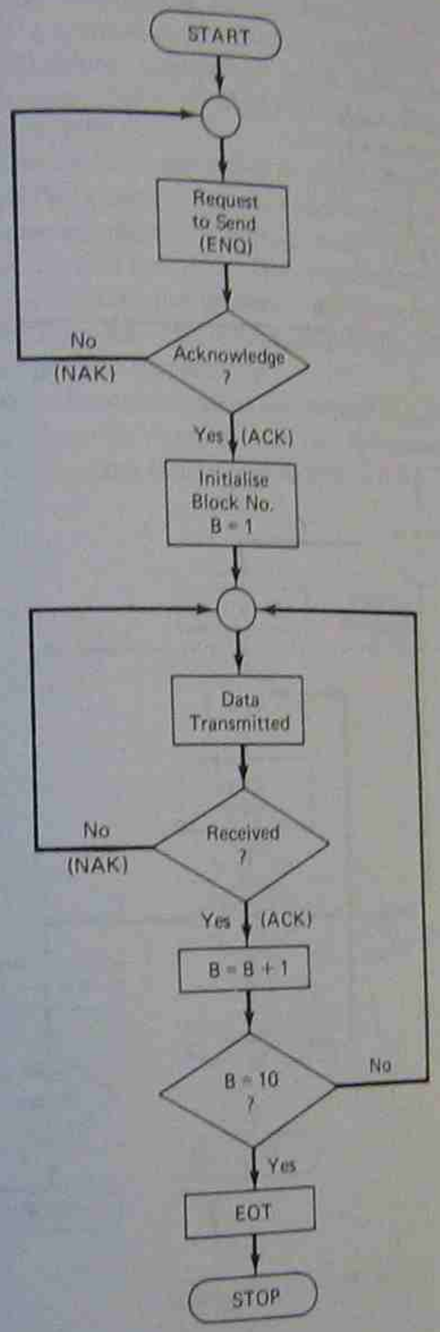


Fig. 10-14 BISYNC Flowchart (Simple)

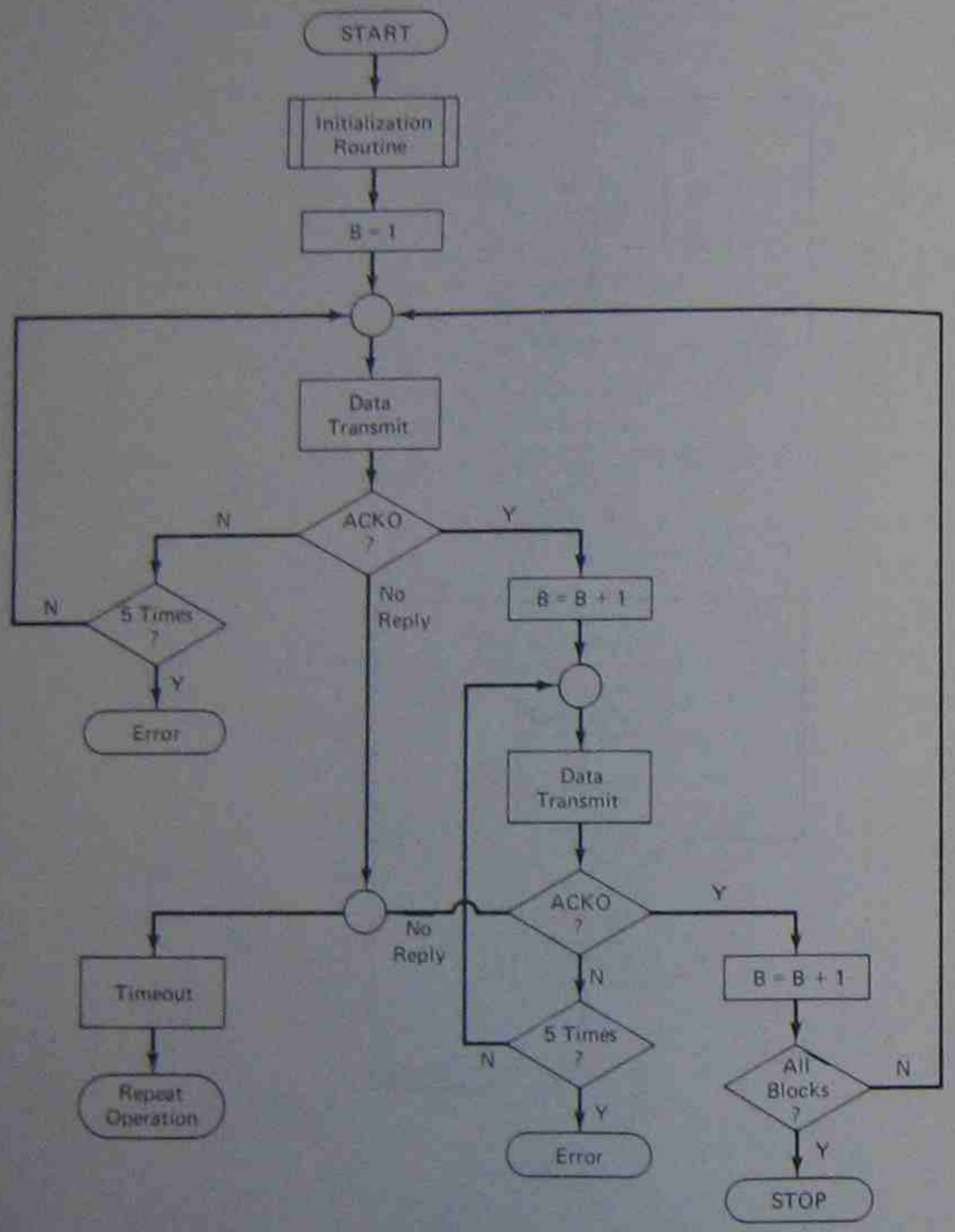


Fig. 10-15 BISYNC Flowchart (Improved)

The prime function of data link control is the transfer of data between the user (DTE) and the network (DCE). This transfer must include control transmissions for initiating, sequencing, checking, and terminating the exchange of user information. For this level of protocol, the CCITT X.25 recommends a procedure compatible with the High-Level Data Control (HDLC) procedure standardized by the International Organization for Standardization (ISO). IBM's Synchronous Data Link Control (SDLC) is a variant of HDLC.

The control procedures described in this section subscribe to the principles of a new ISO class of procedures for a point-to-point balanced system. The link configuration is a point-to-point channel with two stations. Each station has two functions. The primary function is to manage its own information transfer and recovery. The secondary function is to respond to requests from the other stations. In essence, these functions are commands and responses, respectively.

The data link control procedures are defined in terms of commands and responses and may be thought of as two independent but complementary transmission paths that are superimposed on a single physical

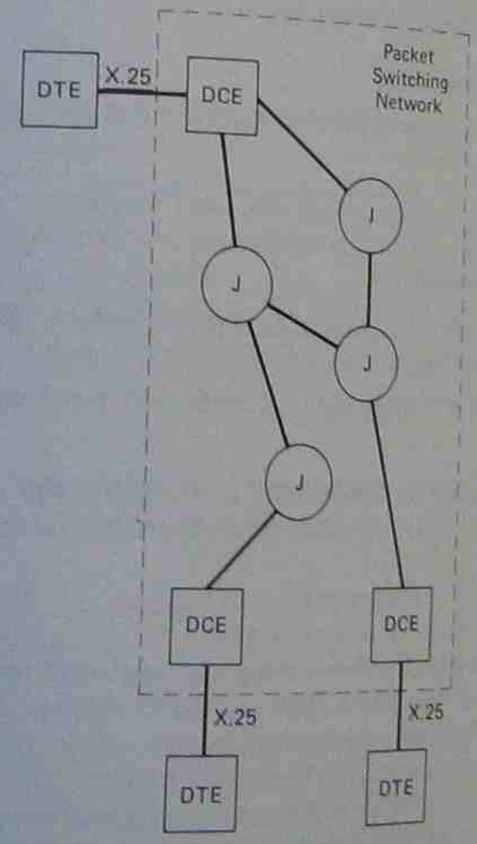


Fig. 10-16 PPSN (Public Packet Switching Network)

circuit. The DTE controls its transmissions to the DCE; the DCE controls its transmissions to the DTE.

Synchronization

To transmit information across a link, the link must be synchronized. This is accomplished by enclosing each frame with unique bit patterns, called flags. (Refer back to Fig. 10-11). Once the receiver recognizes the flag, it knows it is in step with the transmitter and in a position to accept the content of the frame. Similarly, when the receiver recognizes the second flag, it knows that all the information has been received. The receiver is now in a position to check the transmission and act upon the information it contains.

The potential occurrence of a flag bit pattern within the information is prevented by a mechanism similar to that described in the BISYNC protocol. The occurrence of a flag bit pattern in the information is made transparent to the receiving station. The transmitting station examines the frame content between the two flags, which includes the address, control, and FCS (Frame Control Sequence) data, and inserts a 0 after all sequences of five contiguous 1 bits to insure tht a flag sequence is not simulated. The flag bit pattern is 01111110. The receiving station removes any 0 bit which directly follows five 1's.

Commands and Responses

There are three variations on the frame structure of the X.25 data link control protocol:

1. Information Frames, which are used to perform an information transfer, and may also be used to respond to a correctly transmitted information frame.
2. Supervisory Frames, which are used to perform link control functions responding to received information frames.
3. Unnumbered Frames, which provide additional link control functions.

The commands and responses listed in Fig. 10-17 are identified in the frame control field which immediately follows the flag and address fields of Fig. 10-11.

Link Control

The link is controlled by means of the two bytes that immediately follow the opening flag. *The address byte* is used to distinguish between commands and responses made by the stations at either end of the link. Frames containing commands transferred from the DCE to the DTE will contain address "a." Frames containing responses transferred from the DTE to the DCE will contain address "a." Frames containing commands from the DTE to the DCE will contain address "b." Frames containing

Information Frames	
Commands	
I frames	
Supervisory Frames	
	Responses
	RR – Receive Ready RNR – Receive Not Ready REJ – Reject
Unnumbered Frames	
Commands	*Responses*
SARM – Set Asynchronous Response Mod	UA – Unnumbered Acknowledge
DISC – Disconnect	CMDR – Command Reject

Fig. 10-17 Commands and Responses Used in the Frame Control Field, Listed by Frame Type

responses from the DCE to the DTE will contain address "b." This is illustrated in Fig. 10-18.

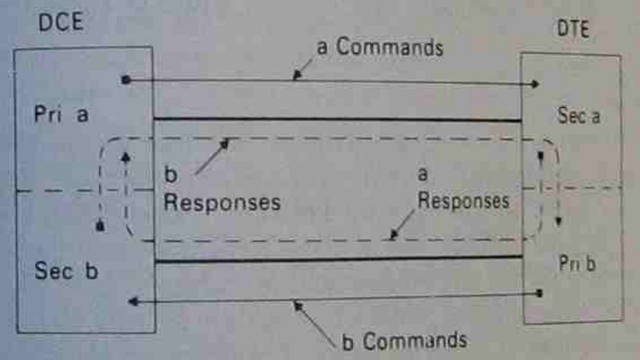
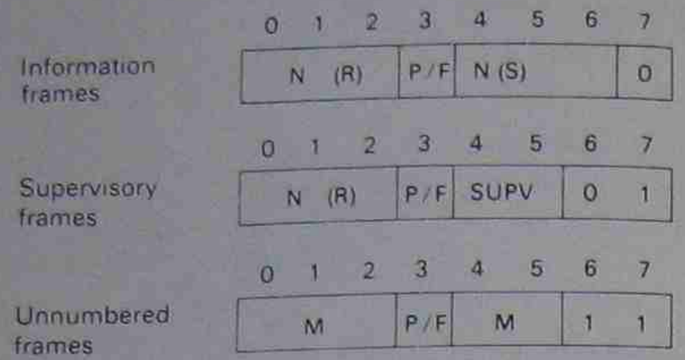


Fig. 10-18 Four-Wire, Full-Duplex Link

The *control byte* contains specific command and response identification as well as sequence numbers, if applicable. The contents of the control field depend on the frame type; that is, information, supervisor, or unnumbered. The control field also contains a Poll/Final bit, supervisory bits, and, in the absence of sequence numbers, five control bits, as shown in Fig. 10-19.

Operating a balanced link between a DTE and a DCE involves initiating the link, transferring the information, and terminating the link. The following sections will explain these steps in detail.



Where:

N(S) – Sequence number of the frame containing this control field

N(R) – Sequence number of the next frame the transmitter expects to receive

P/F – Poll/Final bit (Primary sets Poll, Secondary sets Final)

SUPV– Up to four supervisory functions

M – Additional link control functions

Fig. 10-19 Frame Control Field, in Detail, (Listed by Frame Type)

Initiation

Each primary function, or command station, is independently responsible for initiating the secondary function, or response, of the other station. This is carried out by Set Asynchronous Response Mode (SARM), which is used to place the addressed station in the asynchronous response mode, thus enabling the two-way simultaneous asynchronous exchange of information. At the same time SARM is sent out, the primary function sets its no-response timer, T1. When the secondary function receives SARM, it replies with UA (Unnumbered Acknowledge). Receipt of UA by the primary function confirms initiation in *one* direction, and T1 is reset. In order to confirm initiation in both directions, the primary function of the "other" station also sends out a SARM. This is shown in Fig. 10-20. The time base is horizontal; i.e., it runs from left to right.

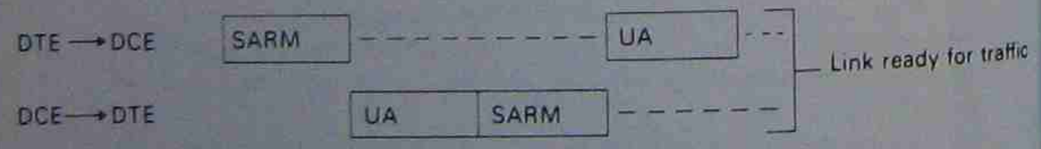


Fig. 10-20 Setting Up the Link (No errors)

If T1 should expire before UA is received, the primary function will resend SARM and restart timer T1. This action continues until UA is correctly received (see Fig. 10-21), or the link is placed in the system recovery mode. The number of retrys is specified when the link is implemented. The system recovers from transmission errors by means of cyclic redundancy checking. It recovers from procedural errors, such as a secondary function receiving a command not in its repertory by referring to the next highest level of protocol.

Information Transfer

Information transfer is carried out via the information frame control field format. Both stations maintain send and receive counters which cycle through 0-7 in order to keep track of the number of outstanding frames awaiting acknowledgement up to a maximum of seven.

When a station has an information frame to transmit, it places the current value of its send-sequence counter in the N(S) bits of the control field, with the N(S) count one higher than the last information frame transmitted. The transmitter also places the value of its receive-sequence counter into the N(R) bits of the control field. This receive sequence count, N(R), is only incremented when an information frame with a valid frame check sequence (FCS) and the expected send sequence count N(S) has been received from the other station. In this way, the N(R) count in the transmitter information frame represents the send-sequence number of the next frame that the transmitter expects to receive. By implication, the transmitter acknowledges frames numbered up to and including N(R) – 1. Figure 10-22 shows a typical two-way transfer without errors.

Note that the information frames are designated by three parameters: I (information frame), N(S) (send-sequence count) and N(R) (receive-sequence count). For example, the first information frame in the DTE→DCE directions contains 1, 0, 0, in the control field, which means this is an information frame of sequence number 0 whose transmitter is expecting an information frame of sequence number 0. The supervisory frames are designed by type: RR (receive ready), RNR (receive not ready), and REJ (reject).

A frame with an invalid FCS is discarded by the receiver. Subsequent information frames may have a valid FCS but the send-sequence count N(S) of the received information frame is not one higher than the last correctly received information frame. In such a case, one or more frames may have been lost. The receiver will discard the information content of this frame and all subsequent out-sequence frames and will transmit a supervisor/reject (REJ) frame with the N(R) count equal to the first missing frame. (The control field of out-of-sequence frames with valid FCS is retained.)

The REJ will request retransmission of the incorrect information frame and all subsequent frames sent by the transmitter (received correctly or not) in the interval between the frame in error and receipt of the

REJ frame. To avoid confusion, it is mandatory that no more than one REJ be outstanding at any one time. This implies that once a receiver has set a REJ condition, the receiver will not accept any incoming frames until it recognizes the retransmitted information frame indicated in the N(R) of the REJ response. If an REJ response is received while an I frame is being transmitted, the station may abort the I frame in progress. See Fig. 10-23.

A receiving station may receive two or more consecutive information frames with the same send-sequence number, N(S). This could occur if an acknowledgement is lost in the reverse direction of transmission. Upon detection that the N(S) of the information frame just received is identical to the N(S) of the previously received frame, the receiving station must discard the received frame and return a REJ with the N(R) sequence count one higher than the N(S) sequence count of the last valid I frame received. For an illustration of this, see Fig. 10-24.

The Receive Ready (RR) and Receive Not Ready (RNR) responses are available to the secondary stations for the following uses:

1. The station is temporarily short of space to buffer incoming messages and transmits RNR to inform the other station.
2. The station wishes to acknowledge the receipt of valid frames, but has no information to send at this time. Therefore, it sends RR with the appropriate N(R) count in the control field.

Termination

Once the link is no longer required for information transfer, a disconnect command can be used to terminate, or close, the link.

Error Detection and Recovery

CRC is the detection method recommended by the X.25. The following generator polynomial is used:

$$x^{16} + x^{12} + x^5 + 1 \qquad (10\text{-}1)$$

10–4.3 Implementation

Protocol for data link control may be implemented using software or hardware. Recent developments in microprocessors have produced hardware data link controllers such as the Motorola MC6854 chip and the Intel 8273 Programmable Protocol Controller. Both these chips provide an efficient data communications interface to support HDLC and SDLC protocols. They both contain such features as automatic FCS generation and checking, flag detection and synchronization, and full-duplex operation.

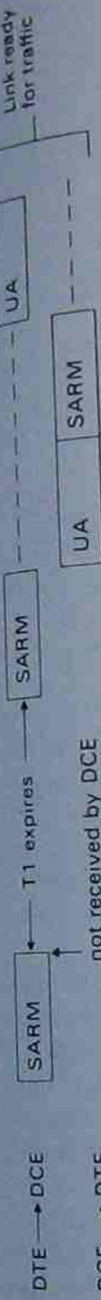


Fig. 10-21 Setting Up the Link (Errors)

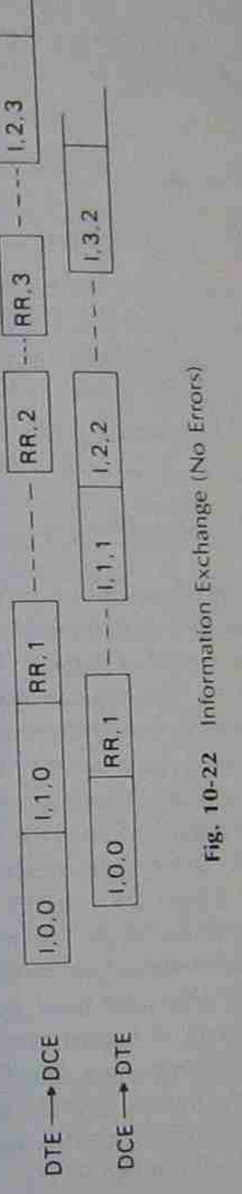


Fig. 10-22 Information Exchange (No Errors)

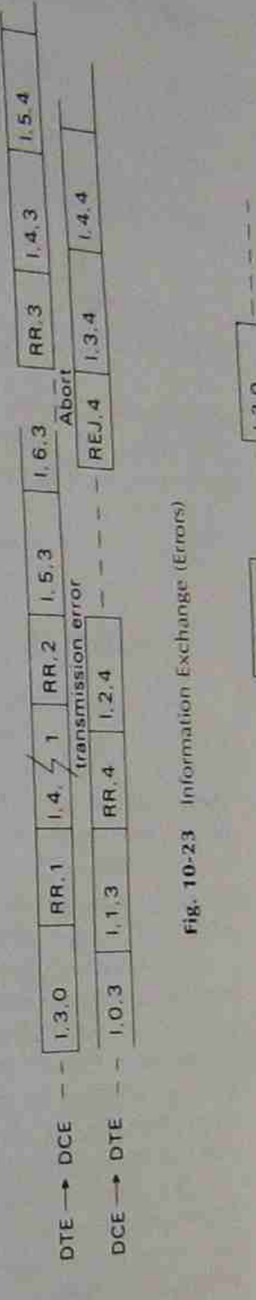


Fig. 10-23 Information Exchange (Errors)

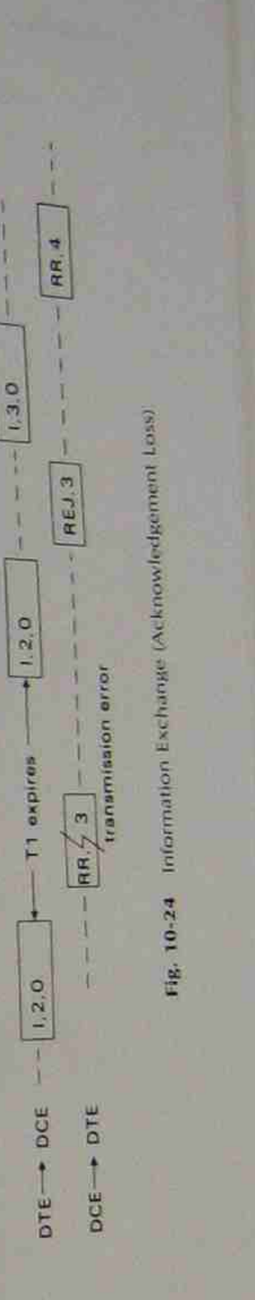


Fig. 10-24 Information Exchange (Acknowledgement Loss)

10–5 PACKET LEVEL

This level of protocol specifies the way in which users establish, maintain, and clear calls in the network. It also specifies the way in which user's data and control information are structured into packets. The relationship between the user data packet and the frame structure of the X.25 data link control protocol is shown in Fig. 10-25.

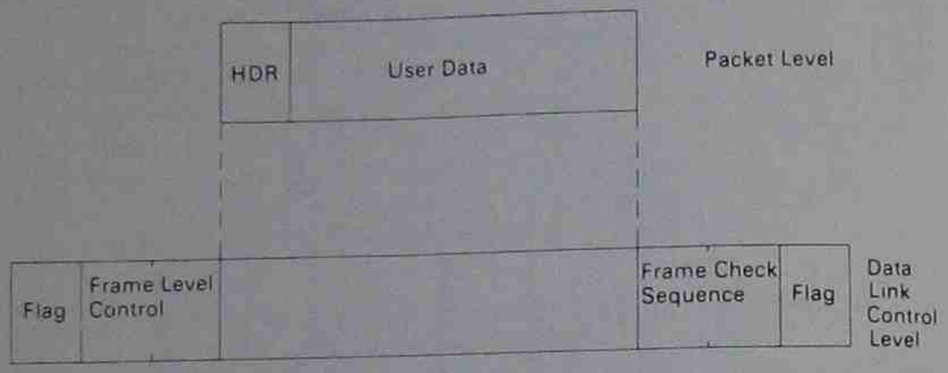


Fig. 10-25 Frame/Packet Relationship

A fundamental concept of packet switching is that of the *virtual circuit*. Just as circuit switching permits a physical link between two stations, so does a virtual circuit permit a logical bidirectional association between two stations (DTE's). This logical link is assigned only when packets are being transferred. A *permanent virtual circuit* is a permanent association between two DTE's, much like a private line. A *switched virtual circuit* is a temporary association between two DTE's and is initiated when a DTE sends a call request packet to the network.

At the packet level, DTE's may establish simultaneous communication with a number of other DTE's in the network. This is accomplished on a single physical circuit by means of packet-interleaved Asynchronous Time Division Multiplexing (ATDM). This type of multiplexing is similar to the straightforward time division multiplexing concept of improving the efficiency of information transfer by interleaving messages into the gaps of other messages. See Fig. 10-26(a).

Dynamically allocating the bandwidth to active, virtual-circuit ATDM further improves the transfer rate. This technique is referred to as statistical multiplexing and is illustrated in Fig. 10-26(b). The time previously allocated to non-active terminals is now used on a first come, first served basis.

The interleaved packets are transferred from the DTE's to the DCE's using the frame format of the data link control protocol. Each packet contains a logical channel identifier to relate the packet to a switched or

permanent virtual circuit. Figure 10-27 illustrates a typical packet transfer between the DTE and the DCE.

The logical channel identifier (LCI) is used strictly at the local level of the network, or between DTE's and DCE's. The LCI's are selected at each DTE and are used independently of all other DTE's. The LCI's associated with permanent virtual circuits cannot be used for any other purpose. The LCI's used with switched virtual circuits are all free initially and can be used by the DTE to originate new calls or to receive incoming calls from the network.

Call Establishment and Clearing

There are several packets associated with setting up a communication link between two DTE's. These packets include:

1. A Call Request Packet sent by the calling DTE to the network. This packet includes the logical channel identifier; the address of the DTE called; a facility field for reverse charging, priority, or user restrictions; and user data of 16 bytes.
2. A Call Connected Packet sent to the calling DTE to confirm that the call has been accepted.
3. A Clear Indication Packet sent to the calling DTE to indicate that the call was not established and to provide one of several reasons for this, such as Number Busy.

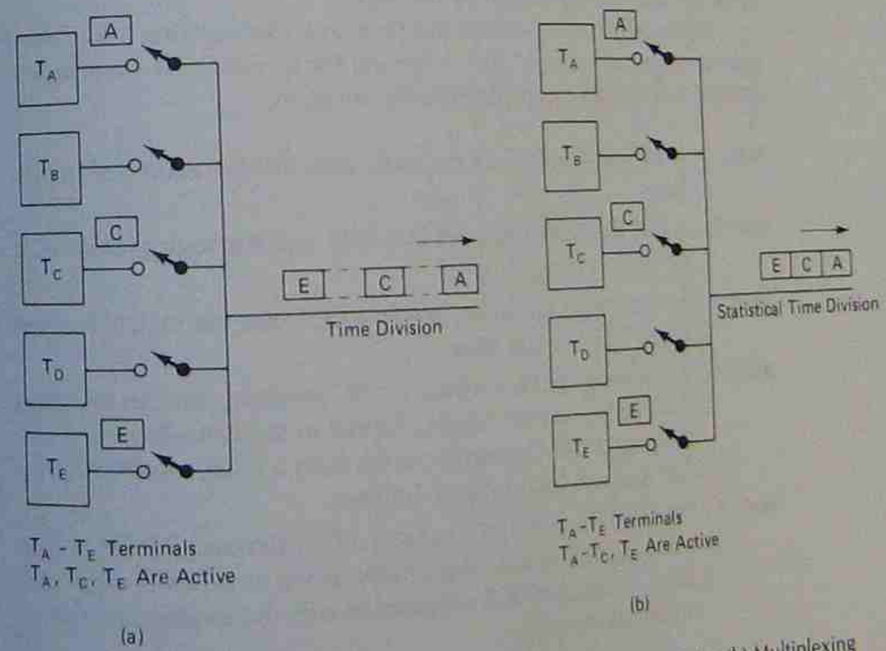


Fig. 10-26 Time Division (a) and Statistical Time Division (b) Multiplexing

4. An Incoming Call Packet received by the called DTE on an available logical channel as a result of a Call Request Packet. (The formats are very similar.)
5. A Call Accepted Packet sent by the called DTE that indicates a positive response to a call request. The network uses the Call Accepted Packet to generate a Call Connected Packet for the calling DTE.
6. A Clear Request Packet sent by either DTE to free the logical channel after data transfer is complete.

A sequence diagram showing the stimulus/response relationship required to establish a call between two DTE's is given in Fig. 10-28.

Data Transfer

Data Transfer can take place only after the virtual circuit has been established. The format of data packets is shown in Fig. 10-29.

The third byte in Fig. 10-29 is almost identical to that of the frame format except that the poll/final bit is replaced by a more data bit (M). P(S) is the packet send-sequence number. (Only Data Packets are numbered.) P(R) is the receive-sequence number used to locally confirm the receipt of packets. M is used in a full data packet to indicate continuation of the data in the next packet. Type-Data distinguishes this format from all the others. The maximum data field length is normally initialized to 256 bytes, but this may be revised by the user.

Figure 10-30 offers another approach to understanding information transfer at the packet level. The following list of events at each logical channel will help in comprehending the diagram.

State 1 A logical channel is in the ready state if there is no call in existence.

State 2 A Call Request Packet sent by a DTE puts the logical channel in the DTE Waiting State.

State 3 An Incoming Call Packet sent by a DCE puts the logical channel in the DCE Waiting State.

State 4 A Call Accepted Packet from a DTE, resulting from an Incoming Call Packet, puts the logical channel in the Data Transfer State. Similarly, a Call Connected Packet from a DCE, resulting from a Call Request, leads to Data Transfer.

State 5 When a DTE and a DCE transfer a Call Request Packet and an Incoming Call Packet, respectively, at the same time, a call collision occurs. The DCE will proceed with the request and cancel the incoming call.

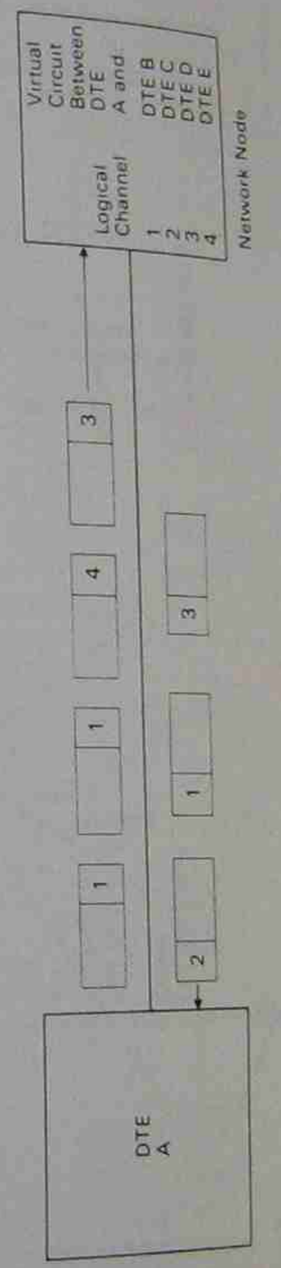


Fig. 10-27 Packet Flow

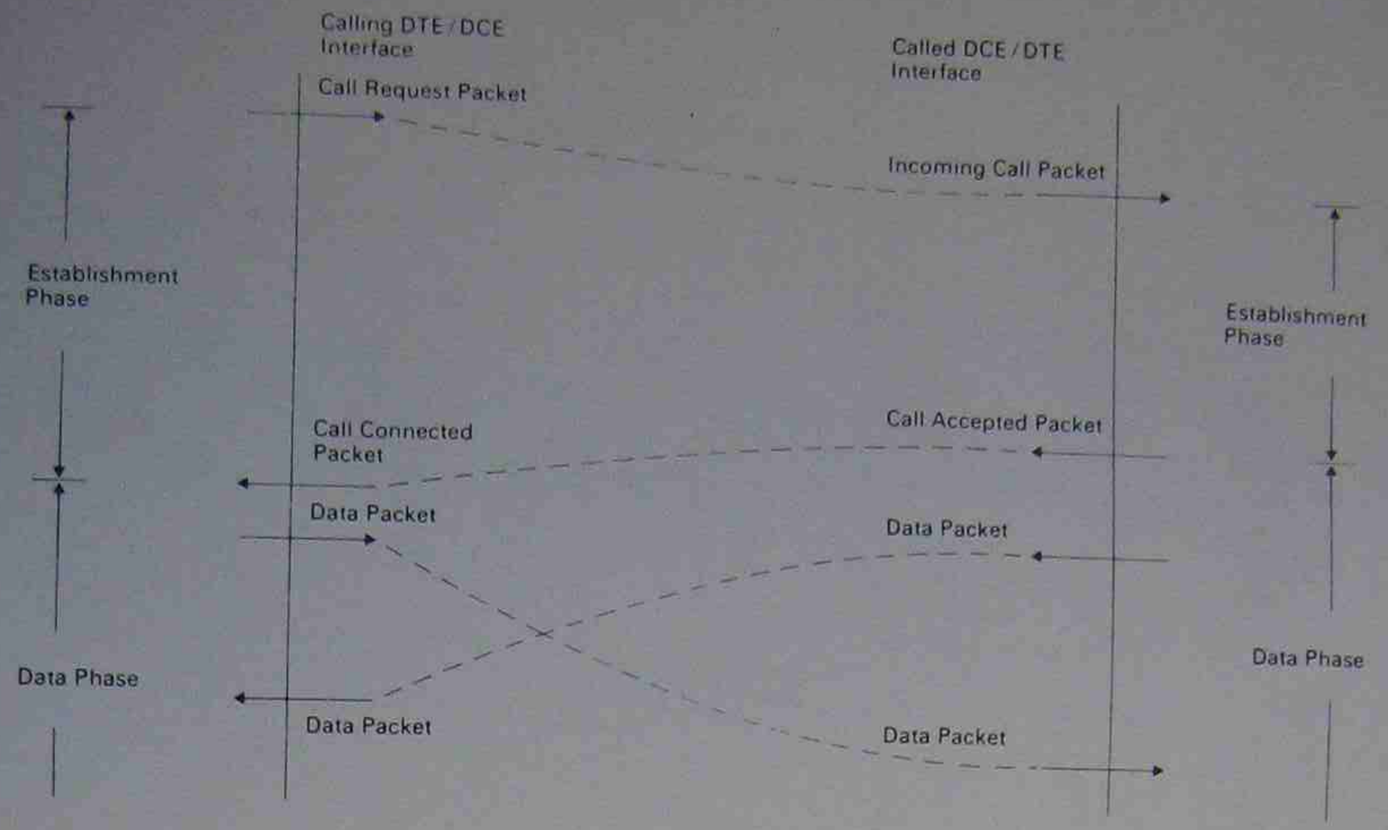


Fig. 10-28 Illustration of Call Establishment

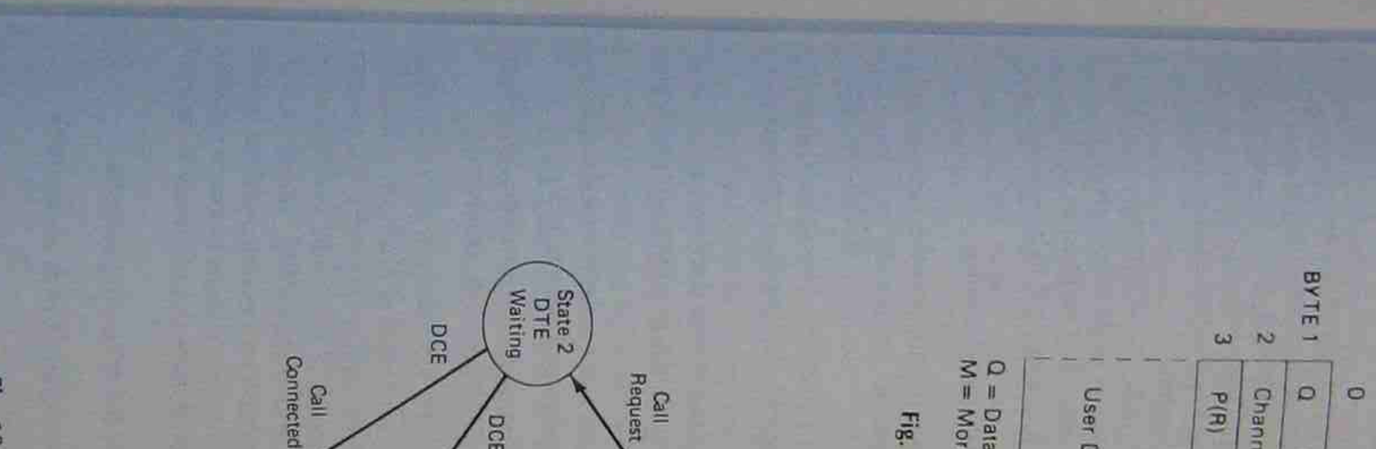


Fig. 10-29 Data Packet Format

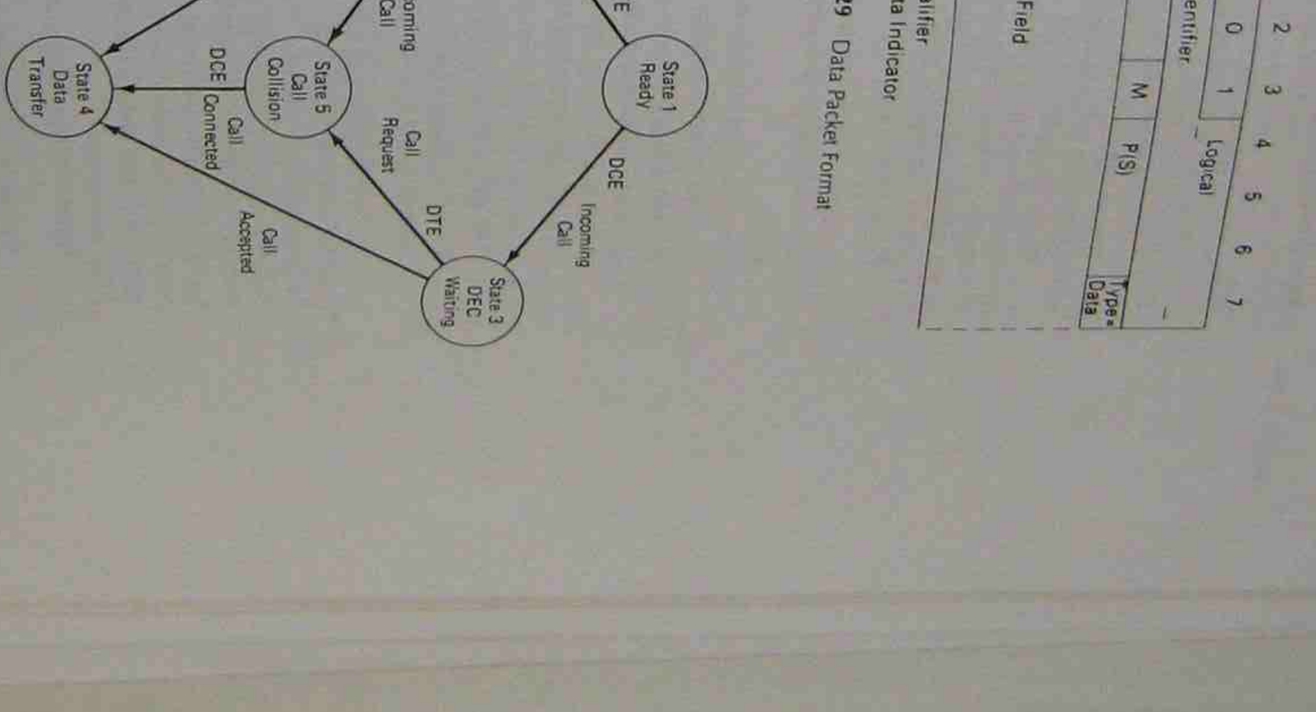


Fig. 10-30 Call Setup State Diagram

10–6 CONCLUSION

Public packet switching networks use point-to-point synchronous circuits. Non-packet switching networks use a physical channel as an increment of dedicated bandwidth. This bandwidth may be permanently leased as a private line, or periodically accessed via circuit switching. Both may result in inefficient use of facilities because the bandwidth is reserved whether or not data is being transmitted.

In packet switching, all user data is converted to discrete, variable-length units called packets. Packets of data from many users are dynamically interleaved over shared network transmission facilities and routed to their destinations. Bandwidth is allocated to a user only when he is actively transmitting data.

Packet switching provides the following benefits:

1. Universality. Resource sharing that is built into the network allows both small and large organizations to participate in a nationwide system.
2. Communication Interconnection Flexibility. The specification of a Standard Network Access protocol facilitates communications between user systems.
3. Low Cost. The ability to share transmission facilities and to share switching and control equipment results in lower communication costs for a large segment of data users.
4. Accuracy. Powerful error detection and correction facilities provide users with virtually error-free, end-to-end data transfer.
5. Reliability. Standby facilities, alternate network routes, and network monitoring and control result in high network availability.

Even with the universal adoption of X.25, there are still users critical of packet switching technology. The main thrust of criticism comes from the proponents of datagram, a network which transmits disconnected packets without any knowledge or concept of messages, connections, or flow control. The packet assembly and disassembly, sequence numbering, flow control, and virtual connection processing are done by computer DTE's. The problem with the datagram system is that it requires a more coordinated and cooperative set of users than is likely to exist in a public environment.

Packet switching is criticized for the following reasons:

1. The interface is complex and was drafted with undue haste.
2. There are flaws in the design.
3. There are standards which are open to different interpretation.
4. X.25 does not provide end-to-end integrity of data flow.

These criticisms are partially answered by CCITT recommendations X.28 and X.29. X.28 applies to the interface between the user and the terminal. X.29 applies to the interface between a remote handler and an application program, or an access method, located in a computer.

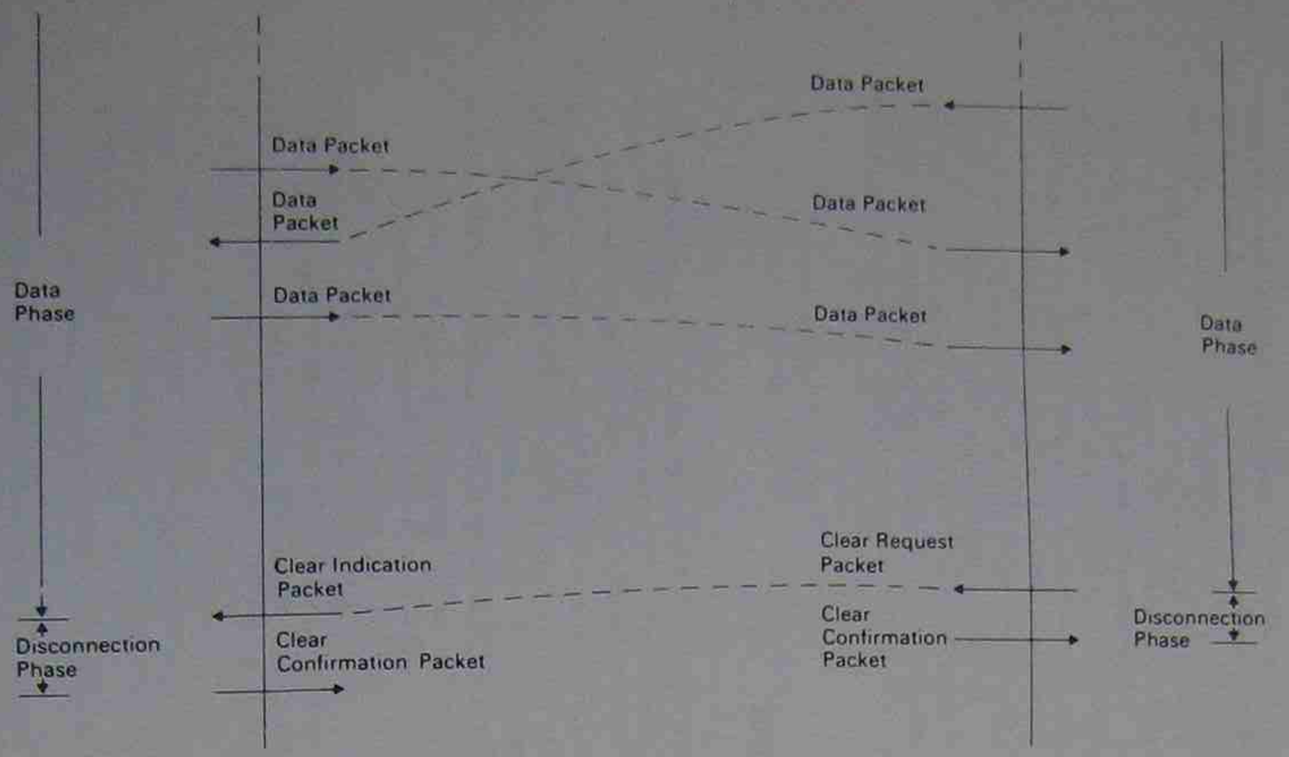
The establishment of public data networks on a national and international scale is fraught with technical, economic, and political difficulties. Considerable progress has been made by the CCITT in setting standards to meet customer demands on an international basis. It is not surprising that there are problems which still need to be solved.

New developments create their own set of new problems but the possiblities for improving communication systems are unlimited. Possibilities for the future include incorporating broadcast satellite facilities (packet satellite), sharing one wideband channel among many radio stations (packet radio), and compressing voice transmission by packet switching.

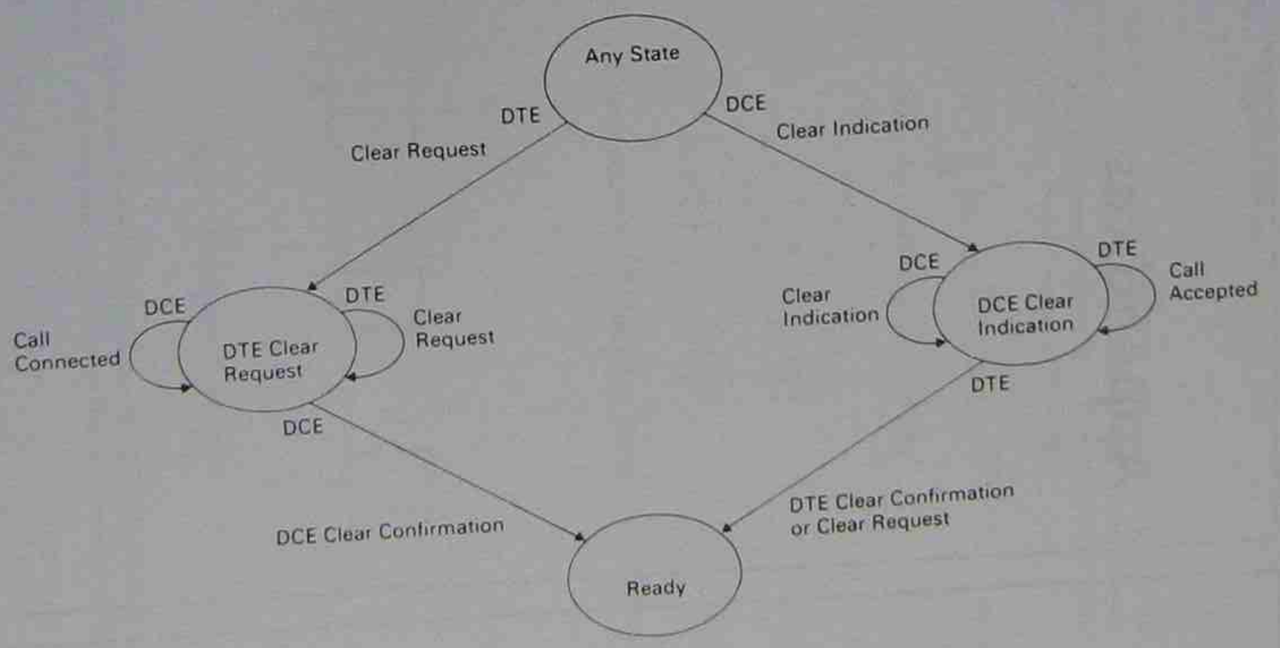
PROBLEMS

1. Each protocol level can be described by a unit of information, such as the frame or the packet (excluding the physical level). What unit of information would be used to describe the next level up?
2. Assuming they were being used for control purposes, how would you make the numeric characters of the EBCDIC code transparent to the receiver?
3. Draw a detailed flowchart for the initialization routine of Fig. 10-14.
4. Using Fig. 10-31, answer the following questions:
   a. Which signal is used by the DTE to indicate clearing?
   b. Which signal is used by the DCE to indicate clearing?
   c. When the logical channel is in the DTE clear request state which signal does the DTE use to free the logical channel?

   Note: After transferring a clear request packet a DTE may receive other types of packet before the channel is freed by the DCE.



(a) Data Transfer and Clearing



(b) Call Clearing

Fig. 10-31 Illustration of Data Transfer and Clearing (a) and Call Clearing (b)

Appendices

Appendix A. Fourier Series

Any periodic function with a finite number of axis crossings over some fixed time interval can be represented by an infinite trigonometric series. Such a function $f(t)$ having a period T where $f(t + T) = f(t)$, may be represented by a series called the Fourier series as indicated:

$$f(t) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2\pi_{vt}}{T} + b_n \sin \frac{2\pi nt}{T} \right) \qquad (A\text{-}1)$$

where the coefficients are given by

$$a_0 = \frac{1}{T} \int_{-T/2}^{T/2} f(t)\, dt \qquad (A\text{-}2a)$$

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos \frac{2\pi nt}{T} \qquad (A\text{-}2b)$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin \frac{2\pi nt}{T} \qquad (A\text{-}2c)$$

and where $1/T$ is the fundamental frequency of the periodic function.

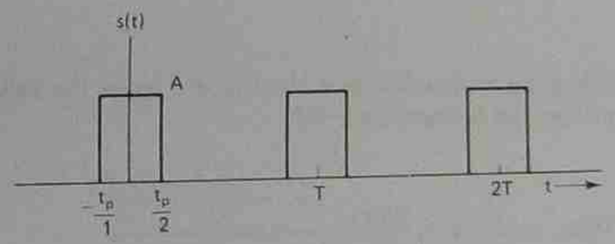


Fig. A-1 Rectangular Wave Train

As an illustration, find the Fourier series of a periodic train of pulses of amplitude A and duration t_p as shown in the accompanying figure.

Thus,

$$a_0 = \frac{1}{T} \int_{-t_p/2}^{t_p/2} A\, dt = \frac{A}{T} t \Big|_{-t_p/2}^{t_p/2} = \frac{At_p}{T}$$

This component represents the dc term in circuit theory.

$$a_n = \frac{2}{T}\int_{-t_p/2}^{t_p/2} A\cos\frac{2\pi nt}{T}\,dt$$
$$= \frac{2A}{T}\frac{T}{2\pi n}\sin\frac{2\pi nt}{T}\Bigg|_{-t_p/2}^{t_p/2}$$
$$= \frac{2At_p}{T}\frac{\sin\dfrac{\pi nt_p}{T}}{\dfrac{\pi nt_p}{T}}$$

The sinc function is defined as

$$\text{sinc } x = \frac{\sin\pi x}{x}$$

Thus $\quad a_n = \dfrac{2At_p}{T}\,\text{sinc}\,\dfrac{nt_p}{T}$

Because of even symmetry, $b_n = 0$

Thus the Fourier series for $S(t)$ becomes

$$S(t) = \frac{At_p}{2} + \frac{2At_p}{T}\sum_{n=1}^{\infty}\sin\frac{nt_p}{T}\cos\frac{2\pi nt}{T} \qquad (A\text{-}3)$$

By expanding the sin function as a MacLaurin's series, the value of the sinc x function can be found at $x = 0$.

$$\text{Thus,}\quad \text{sinc } x = \frac{\sin\pi_x}{\pi_x} = \frac{\pi_x - \dfrac{(\pi x)^3}{3!} + \dfrac{(\pi_x)^5}{5!}}{\pi x}$$
$$= 1 - \frac{(\pi x)^2}{3!} + \frac{(\pi x)^4}{5!}$$

$$\lim_{x\to 0}\text{sinc } x = 1$$

The Fourier series for $S(t)$ has a dc term of $At_p/2$ and harmonics of $1/T$ having an envelope given by $\dfrac{2At_p}{\pi}\,\text{sinc}\,nt_p/T$. This is illustrated in Fig. A-2. The sinc nt_p/T goes through zero whenever $\sin\pi nt_p/T = 0$ or when $\pi nt_p/T = m\pi$. This is equivalent to $n/T = m/t_p$.

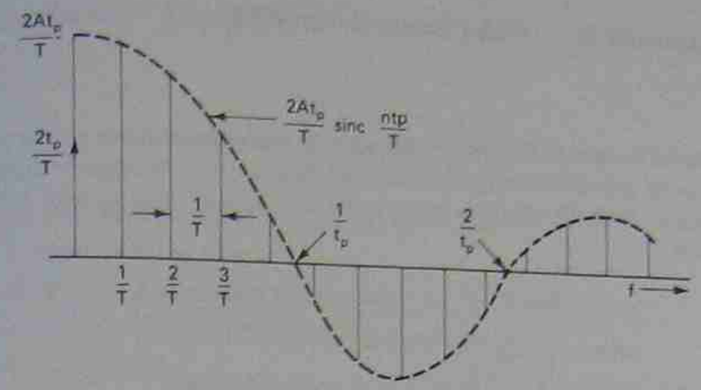


Fig. A-2 Fourier Series of a Rectangular Wave Train (Spectrum Graph)

Appendix B. Noise Power Ratio (NPR)

The NPR ratio is defined as the ratio (in dB) of the noise in the test channel with all channels loaded with white noise to the noise in the test channel with all channels except the test channel loaded with white noise.

$$\text{NPR} = 10 \log \left(\frac{N_G \times \dfrac{\text{channel slot bandwidth}}{\text{base bandwidth}}}{N} \right)$$

where N_G = power due to noise generators (mW)
N = noise power due to intermodulation and thermal noise

Dividing and multiplying the term in the brackets by the test tone level of the telephone channel, S

$$\text{NPR} = 10 \log \left(\frac{\dfrac{S}{N} \times \dfrac{1}{\text{base bandwidth}} \times \dfrac{N_G}{S}}{\text{channel slot bandwidth}} \right)$$

$$= 10 \log \frac{S}{N} - 10 \log \left(\frac{\text{base bandwidth}}{\text{channel slot bandwidth}} \right) + 10 \log \frac{N_G}{S}$$

therefore

$$10 \log \frac{S}{N} = \text{NPR} - \log \frac{N_G}{S} + 10 \log \left(\frac{\text{base bandwidth}}{\text{channel slot BW}} \right)$$

The term $10 \log N_G/S$ is given in the CCIR-loading equations of (3-6).

Appendix C. Group Delay

A traveling wave on a transmission line can be expressed by $\cos(\omega t - \beta x)$* where:

- ω represents the angular frequency ($\omega = 2\pi f$)
- t represents the time
- β represents the phase shift constant per unit length of line
- x represents the position in the transmission line

Consider two closely spaced angular frequencies of ω_1 and ω_2, each a distance of ω_m (modulating angular velocity) from a carrier angular velocity of ω_c, as illustrated in Fig. C-1, so that

$$\omega_1 = \omega_c - \omega_m$$

$$\omega_2 = \omega_c + \omega_m$$

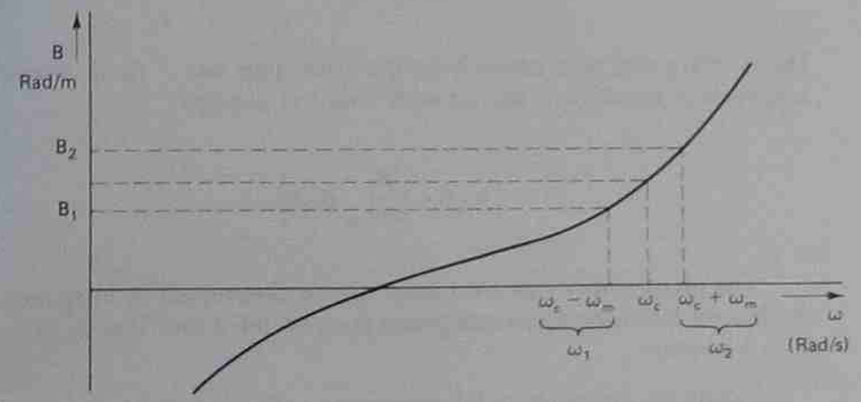


Fig. C-1

Let the sum of the two frequencies be represented by equal sinusoidal voltages:

$$v = \cos(\omega_1 t - \beta_1 x) + \cos(\omega_2 t - \beta_2 x)$$

$$= 2 \cos \left[\frac{(\omega_1 + \omega_2)}{2} t - \frac{(\beta_1 + \beta_2)}{2} x \right] \times \left[\cos \frac{(\omega_2 - \omega_1)}{2} t - \frac{(\beta_2 - \beta_1)}{2} x \right] \qquad (C\text{-}1)$$

* From W. Sinnema, *Electronic Transmission Technology: Lines, Waves and Antennas.* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1979), Chapter 3.

by use of the trigonometric identity.

$$\cos x + \cos y = 2 \cos \frac{(x + y)}{2} \cos \frac{(x - y)}{2}$$

The phase constants can be evaluated around ω_c by using the Taylors series. Thus:

$$\beta_1 = \beta_c - \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m + \frac{1}{2!} \frac{d^2\beta}{d\omega^2}\bigg|_{\omega_c} \omega_m{}^2 + \cdots \qquad \text{(C-2a)}$$

$$\beta_2 = \beta_c + \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m + \frac{1}{2!} \frac{d^2\beta}{d\omega^2}\bigg|_{\omega_c} \omega_m{}^2 + \cdots \qquad \text{(C-2b)}$$

Substituting Eq. (C-2) into (C-1) and keeping ω_m small,

$$v = 2 \cos \left(\omega_m t - \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m x \right) \cos(\omega_c t - \beta_c x) \qquad \text{(C-3)}$$

This is interpreted as a carrier traveling wave, $[\cos (\omega_c t - \beta_c x)]$, whose amplitude or envelope is varying with time and position:

$$\cos \left(\omega_m t - \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m x \right)$$

The phase velocity for the carrier can be determined by imagining an observer following a constant phase point on the wave. That is, $\beta x - \omega t$ = a constant.

Taking the derivative of this expression with respect to time:

$$\frac{d}{dt}(\beta x) - \frac{d}{dt}(\omega t) = \frac{d}{dt}(\text{constant})$$

$$\beta \frac{dx}{dt} - \omega = 0$$

$$v_p = \frac{dx}{dt} = \frac{\omega}{\beta} \qquad \text{(C-4)}$$

Similarly, the phase velocity, v_g, of the envelope (signal frequency ω_m) is given by $v_g = d\omega/d\beta|_{\omega_c}$ and is called group velocity. If a constant

phase point, $\beta x - \omega T = 0$, for example, continues to be followed, the time delay of the carrier is given by

$$T = \frac{\beta x}{\omega} = \frac{b}{\omega} \qquad \text{(C-5)}$$

where b represents the total phase shift for a line of length x. Similarly, the time delay in the envelope, commonly called the envelope or group delay is given by:

$$T_g = \frac{d\beta}{d\omega}\bigg|_{\omega_c} x = \frac{db}{d\omega}\bigg|_{\omega_c} \qquad \text{(C-6)}$$

where b represents the total phase shift for a line of length x.

by use of the trigonometric identity.

$$\cos x + \cos y = 2 \cos \frac{(x + y)}{2} \cos \frac{(x - y)}{2}$$

The phase constants can be evaluated around ω_c by using the Taylors series. Thus:

$$\beta_1 = \beta_c - \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m + \frac{1}{2!} \frac{d^2\beta}{d\omega^2}\bigg|_{\omega_c} \omega_m{}^2 + \cdots \tag{C-2a}$$

$$\beta_2 = \beta_c + \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m + \frac{1}{2!} \frac{d^2\beta}{d\omega^2}\bigg|_{\omega_c} \omega_m{}^2 + \cdots \tag{C-2b}$$

Substituting Eq. (C-2) into (C-1) and keeping ω_m small,

$$v = 2 \cos \left(\omega_m t - \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m x \right) \cos(\omega_c t - \beta_c x) \tag{C-3}$$

This is interpreted as a carrier traveling wave, $[\cos (\omega_c t - \beta_c x)]$, whose amplitude or envelope is varying with time and position:

$$\cos \left(\omega_m t - \frac{d\beta}{d\omega}\bigg|_{\omega_c} \omega_m x \right)$$

The phase velocity for the carrier can be determined by imagining an observer following a constant phase point on the wave. That is, $\beta x - \omega t$ = a constant.

Taking the derivative of this expression with respect to time:

$$\frac{d}{dt} (\beta x) - \frac{d}{dt} (\omega t) = \frac{d}{dt} (\text{constant})$$

$$\beta \frac{dx}{dt} - \omega = 0$$

$$v_p = \frac{dx}{dt} = \frac{\omega}{\beta} \tag{C-4}$$

Similarly, the phase velocity, v_g, of the envelope (signal frequency ω_m) is given by $v_g = d\omega/d\beta|_{\omega_c}$ and is called group velocity. If a constant

phase point, $\beta x - \omega T = 0$, for example, continues to be followed, the time delay of the carrier is given by

$$T = \frac{\beta x}{\omega}$$

$$= \frac{b}{\omega} \tag{C-5}$$

where b represents the total phase shift for a line of length x. Similarly, the time delay in the envelope, commonly called the envelope or group delay is given by:

$$T_g = \frac{d\beta}{d\omega}\bigg|_{\omega_c} x$$

$$= \frac{db}{d\omega}\bigg|_{\omega_c} \tag{C-6}$$

where b represents the total phase shift for a line of length x.

Appendix D. The XR-567 Touch-Tone Decoder

The XR-567 consists of a phase-locked loop, a quadrature phase detector, a voltage capacitor, and an output logic circuit. Figure D-1 shows the block diagram of the XR-561. Figure D-2 illustrates its component connections.

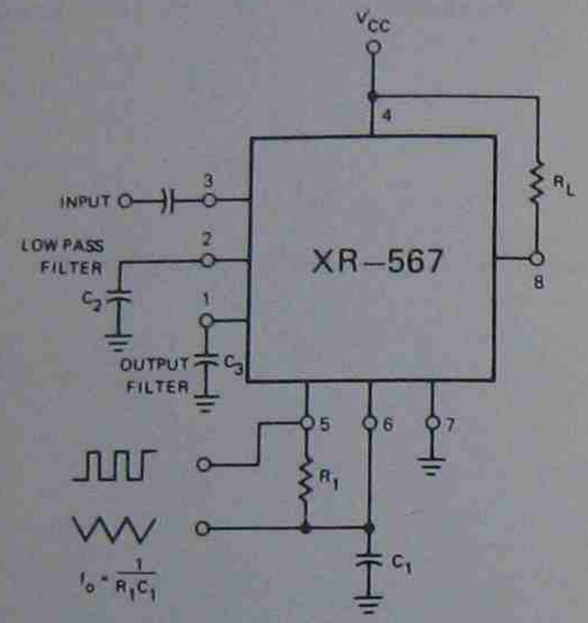


Fig. D-1 XR-567 Block Diagram (Courtesy of EXAR Integrated Systems, Inc.)

The nominal free-running frequency of the oscillator within the PLL can be approximated by:

$$f_o = \frac{1}{R_1C_1} \qquad \begin{matrix} R_1 \text{ in ohms } (\Omega) \\ C_1 \text{ in farads (F)} \end{matrix}$$

If the incoming signal is within the PLL's capture bandwidth centered around f_o, the phase-locked loop will be locked in on the incoming tone. This bandwidth, expressed as a percentage of f_o, can be approximated by:

$$BW = 1070\sqrt{\frac{V_i}{f_oC_2}} \qquad \begin{matrix} V_i \text{ in volts rms} \\ C_2 \text{ in microfarads} \end{matrix}$$

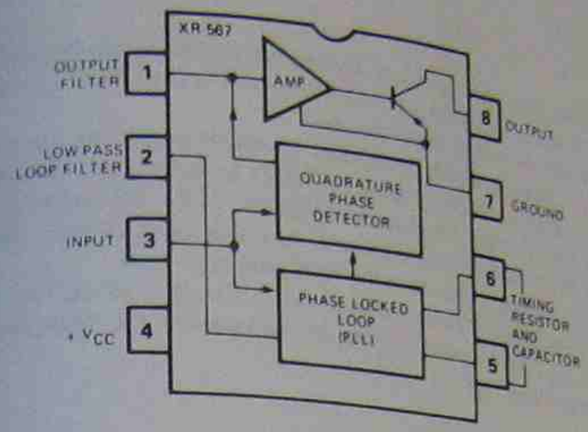


Fig. D-2 XR-567 International Block Diagram (Courtesy of EXAR Integrated Systems, Inc.)

where the input voltage V_i must be greater than the threshold voltage of about 20 mV rms.

With the PLL locked on an input tone, the quadrature phase detector undergoes an output voltage shift. This is transferred to the output logic circuit, which forces the output "open-collector" power transistor to its saturated value of about 0.5 volts. The load resistance R_L connected to the collector (see Fig. D-3) determines the current level. A typical input and output waveform is shown in Fig. D-4; normally the output is in a

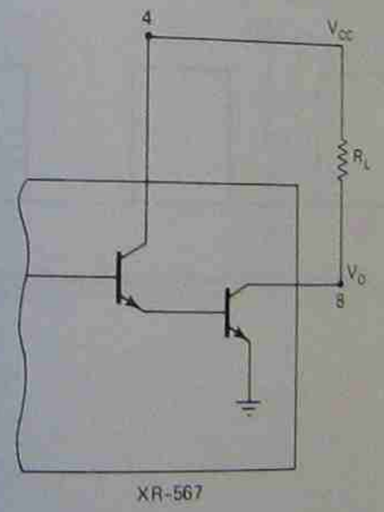


Fig. D-3 Output Driver

high state, but when a tone is applied which lies within the decoder's capture range, the output goes to a low state. The capacitor C_3 in Fig. D-1 determines the output response speed.

To obtain a dual-tone decoder in which two tones are simultaneously present, the circuit shown in Fig. D-5 can be used. The R_1C_1 values of each decoder are chosen so that one of the desired tones is selected. When both tones are present, the collector of the tone decoders goes low and the output from the NOR gate goes high.

A touch-tone decoder can be made up of several XR-567's as shown in Fig. D-6. When appropriate tone pairs are present, the respective NOR gate output goes high.

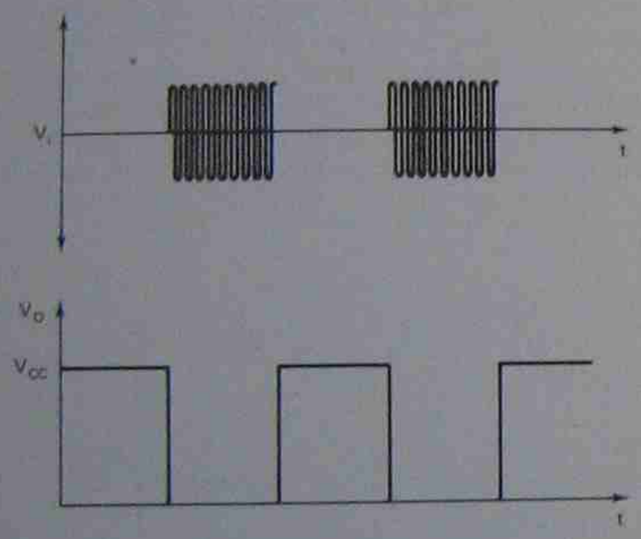


Fig. D-4 Response to Input Tone Bursts

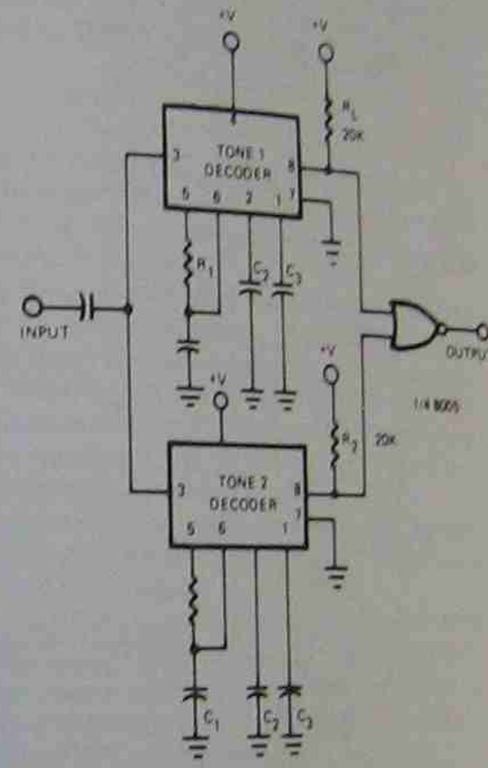


Fig. D-5 Two-Tone Detector. (Courtesy of EXAR Integrated Systems, Inc.)

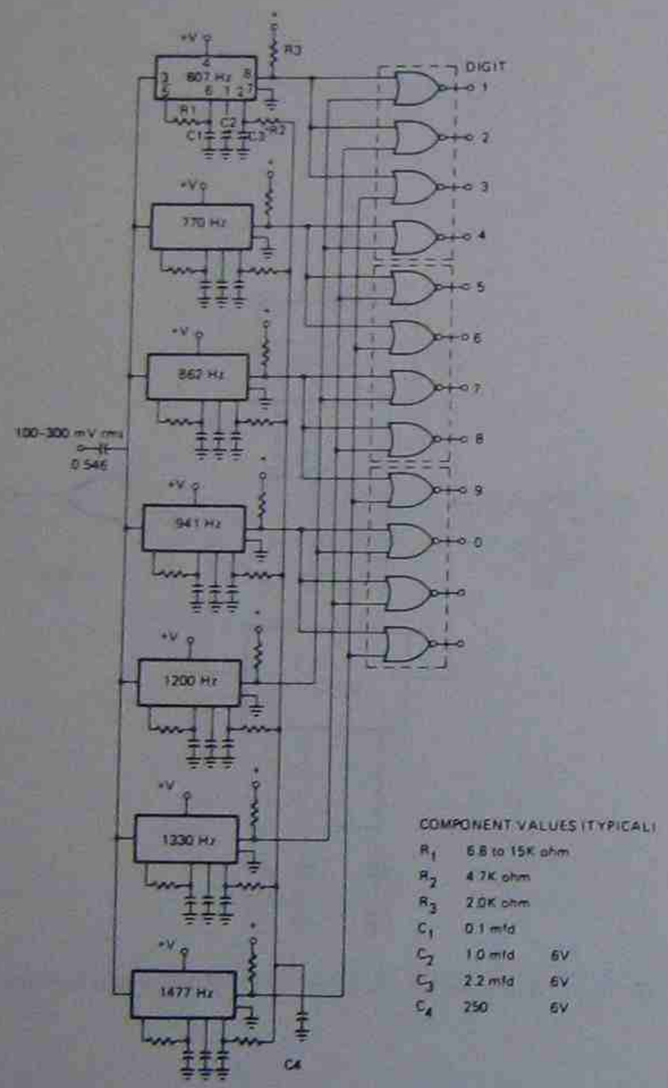


Fig. D-6 Touch-Tone Decoder (Courtesy of EXAR Integrated Systems, Inc.)

Appendix E. The Double-Balanced Ring Modulator

The high amplitude carrier, $\cos \omega_c t$, of the double-balanced ring modulator shown in Fig. E-1 turns the diodes OFF and ON at a carrier frequency rate. The transformer removes the diodes and the dc component.

When the carrier is positive, diodes A and B act as shorts and the output $v_o(t) = m(t)$, assuming no transformer losses and a 1:1 transformer turns ratio. When the carrier is negative, diodes C and D act as shorts and the modulator reverses the polarity of input signal, $v_o(t) = -m(t)$. If this switching action is represented by the switching function $S(t)$ where $m(t)$ is multiplied by $+1$ when $\cos \omega t > 0$ and by -1 when $\cos \omega t < 0$, Fig. E-2 shows that the output or DSBSC signal can be expressed as:

$$v_o = m(t) \times S(t) \tag{E-1}$$

Appendix A shows that the frequency content of the square wave, $S(t)$, consists of all the odd harmonics of $\omega_c/2\pi$ or f_c as $S(t)$ is a square wave with no dc component. It rolls off with the normal sinc response as illustrated in the $S(f)$ plot of Fig. E-2(b).

Since multiplying sinusoids in the time domain gives sum and difference frequencies in the frequency domain, the resultant frequency spectrum of the DSBSC signal consists of sidebands of $m(t)$ around the odd harmonics of the switching function $S(t)$. The harmonics of $S(t)$ are not present at the output because the $m(t)$ applied to the diodes has the dc content removed by the transformer. When multiplying a sinusoid by a zero, one obtains a null in that component.

To obtain high carrier rejection, small resistors of a few hundred ohms are usually placed in series with each diode in order to obtain equal resistance in all legs when the diodes are forward biased.

SSB is obtained by placing a bandpass filter after the balanced modulator. Fig. E-3 shows the resultant frequency spectrum. This figure also illustrates the difficulties experienced when low frequencies are present, in that the lower frequency components of the desired sideband experience greater attenuation than the mid and high frequency components.

Index

A law, 253
Activity factor, 114
Acoustical coupling, 145–146
Aliasing, 211–217
Alternate mark inversion (AMI), 281–282, 372–373
Amplitude modulation, 12–14
Amplitude shift keying, 148
Aperture distortion, 219–221
Aperture time, 248–251
Application level protocol, 389–390
ASC11 code, 353, 355
Asynchronous time division multiplex, 402–403
Asynchronous transmission, 27
  advantages and disadvantages, 30
  receiver, 29
Attenuation constant, 51–55
Attenuation distortion (frequency response), 115, 131
  compensating network, 123
Balanced line, 141
Balance return loss, 99–100, 103
Balanced modulator, 66–68, 425–427
Bandwidth of telephone channel, 42
  of AM,DSBSC,SSB,VSB, 14–15
  voice, 7
Baud, 143–145
Baudot code, 25–26, 28, 352
BCD code, 353
Bell system, 202, 103, 113
  modem frequency utilization, 147
BISYNC, 31, 389–394
Bit, 17
Block-by-Block data transfer, 374
Block check character (BCC), 359–360
Blocking, 38
Capacitive coupling, 106, 142
C weighting, 8, 9
CCIR, 11
CCITT, 7, 11, 32
Channel bank (Lenkurt), 278–280
  characteristics, 283
Characteristic impedance, 51–54, 99
CODEC, 267–268
  INTEL 2910A, 269–274
Codes, ASC11, 353, 355
  Baudot, 25–26, 28, 352
  BCD, 353
  Cyclic, 362–368
  EBCDIC, 353–354
  Excess 3, 357
  Gray, 149–150, 358
  Hamming, 367–371
  HBD 3, 372–373
  Hollerith, 356
  IA3, 359
  Numeric, 356–357
Common control switching, 73–75, 78–80
  relative merits, 80
Companding, 122–124, 250–257
Conditioning, 120–122
Continuous block data transfer, 378–379
Correlative techniques *See* Partial response
Crossbar switch, 73–75
Crosstalk, 60, 105–107, 321
Cyclic codes, 362–368, 400
Data Access Arrangement, 146, 183
Data communication equipment, 162, 395
Data communication networks, 383–387
Data line, 3, 93
Data link control, 388, 390–401
Datapak, 93
Dataroute, 94
Data set *See* Modem
Data terminal equipment, 162, 395
Data transfer rate, 145
Data transmission, 24–26
  channel, 94
  distances using various terminal equipment, 140
Decibel, 4–6, 51
Decoder, R-2R ladder, 267–268
  weighted current source, 266
Delay distortion, 115–123
Delta modulation, 239–242
Dialing, 41
  on step-by-step equipment, 71–72
  pulse train, 46
Dibit, 149
Differential operation, 142
Differential phase modulation, 160–162
Digital transmission, advantages and disadvantages, 23–24
Digital carrier lines, 284
Digital data communication message protocol (DDCMP), 388+389
Digital switching, 84–91
  advantages, 90
Direct distance dialing (DDD), 92
Dispersion, 131
Distortion in digital data transmission, 132–133
  aliasing, 211–217
  aperture, 219–221
Distributed carrier system, 59–60
Double-sideband, 14, 61, 64–66
Dropouts, 308
Duobinary technique, 332–340
Duplexer, 181
EBCDID code, 353–354
Echo, 17
  delay, 102, 375
  return loss, 100
  suppressor, 103, 105
Efficiency, of codes, 358
  of data transmission systems, 376–378
EIA, 12
  RS-232-C, 32, 141, 162–167
Electronic switching system (ESS), 80–84
Encoders, companded, 264

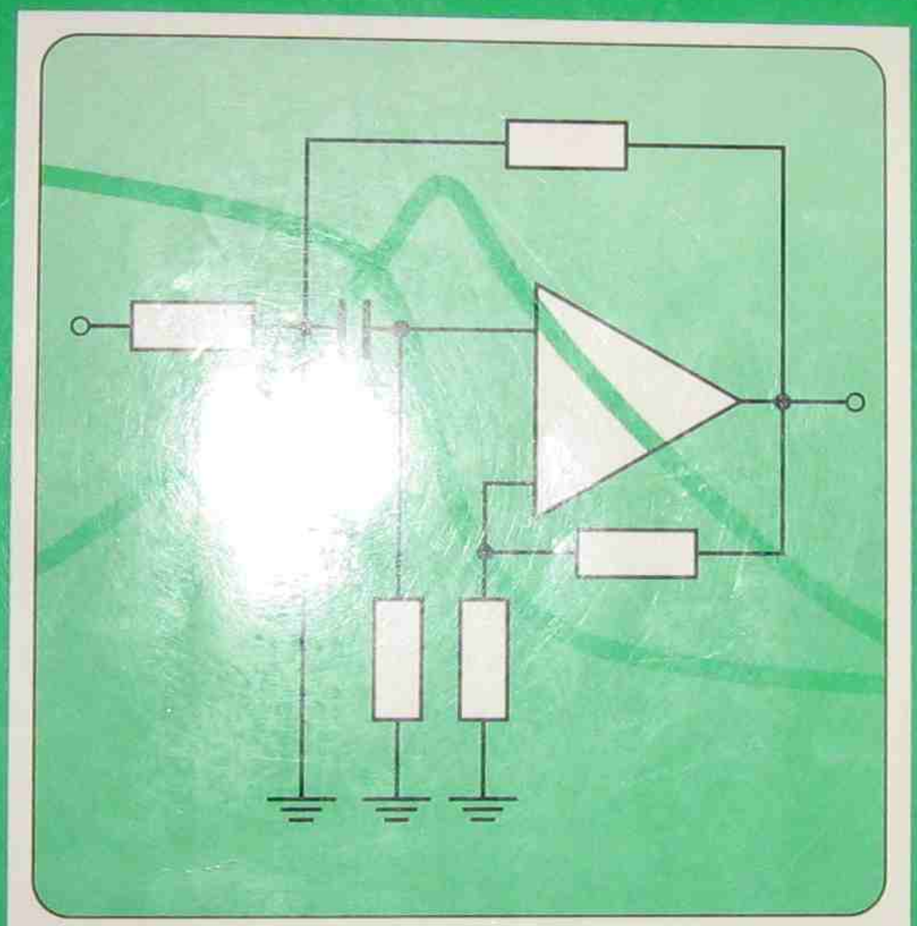
Encoders (*cont'd.*)
  hybrid, 263–265
  parallel, 259–260
  serial, 259–263
Envelope delay, 115–123, 417–419
  compensating network, 123
  measurement, 117–120
Equalizer attenuation, 115
  automatic, 152
  delay, 120
Error correction/detection, 358–371
Error rate due to gaussian noise, 303–308
EXAR 567 touch-tone decoder, 420–424
Excess 3 code, 357
Exclusive-Nor gate, 160
Eye diagram, 324–325, 331
Ferreed switch, 82
Ferrod sensor, 81
Filters, Bessel, Butterworth, Chebyschev, Elliptic characteristics, 222–225
  ideal or brickwall, 318–320
  low pass, 135, 212
Foldover distortion *See* Aliasing
Foreign exchange, 93
Four-wire circuits, 41–42, 60
Four-wire terminating set, 98–100
Fourier series of a pulse train, 134–135, 207–209, 413–415
  single pulse, 218
Frame, 275–277
  data block structure, 390
  structure of X.25, 396–401
Frequency division multiplex, 13–16, 61
  CCITT group, 65–67
  multiplex groups, 67–69
Frequency modulation, 12–15
Frequency shift keying, 148
  spectrum, 183–185, 376
Frequency spectrum diagrams, 135
  of FSK, 183–185
  of speech and voice band data, 152
Full duplex, 33
  modem, 147, 378
Gain hits, 308
Gaussian noise, 301–308
Gray code, 149–150, 358
Group, 61, 65, 69
  delay *See* Envelope delay
  velocity, 418
Half-duplex, 33
  modem, 147, 376–377
Hamming code, 367–371
  distance, 360–361
Harmonics, in transmission, 134–136
HBD3 code, 372–373
High level data control (HDLC), 395
Hollerith code, 356
Hybrid, 45
Hybrid (*cont'd.*)
  transformer *See* Four-wire terminating set
IFRB, 12
Impulse noise, 114–115, 307–308
Inductive coupling, 106, 142
Information (channel) capacity, 325–328
Information theory, 285–288
Intelligibility, 6–7
Intermodulation noise, 105, 108–111, 114
  measurement, 111–112
International Organization for Standardization (ISO), 395
Intersymbol Interference (ISI), 319–320
  ISI degradation, 325
ITU, 11
Junction, 383
Lenkurt 9002B channel bank, 278–280
  coding characteristics, 254–258
  time slot assignment, 275–280
Loading, 55–58
  coils, 55
Local loop *See* Subscriber loop
Line codes, 371–373
Line drivers, 140–142, 181
Lumped carrier system, 59–60
Measurement of impulse noise, gain and phase hits, dropouts, 309
  phase jitter, 310
  quantization noise, 284–285
Message switching, 385–386
Modems, 32, 140, 145–162
  commercial types, 151
  Motorola MC6860, 181–182
  Rixon T103A, 165–180
  turn around time, 374
Modified duobinary technique, 340–345
Modulation, 2
  AM, 12–14
  ASK, 148
  Delta, 239–242
  Digital AM, 22
  DPSK, 160–162
  FM, 12–14
  FSK, 148
  PAM, 224–231
  PCM, 20, 85, 242–248
  PDM (PWM), 230–235
  PM, 12
  PPM, 230–236
  PSK, 148–160
  QAM, 148, 150
Mu law, 252–253
Multicom, 93
Multiframe, 275–276
Multiplex hierarchy in North America, 283–284
Multipoint network, 385
Natural sampling, 204–215
Negative impedance repeater, 98

Neper, 51
Net data throughput (NDT), 360–361, 373, 375–376
Network protocol, 382
Nodes, 383
Noise, 105
  C weighting, 8–9
  factor, 113
  figure, 113
  gaussian, 301–308
  impulse, 114–115, 307–308
  intermodulation, 105, 108–111, 114
  psophometric weighting, 7–9
  quantization, 244–246, 250–252, 297–298
  shot, 112
  thermal, 105, 112–114
  white noise loading, 108, 111
Noise power ratio, 111–112, 416
Non-return-to-zero code, 152–154, 204, 280, 372
Number systems, 18–20
Numeric codes, 356–357
Nyquist rate, 320
Office, end or central, 39, 46–47, 105
Packet level protocol, 389–390, 402–409
Packet switching, 385–387, 395, 402–409
Parallel data transmission, 27
Parity, 351
Partial response, 330–345
Peak-to-average ratio test, 310–311
Phase constant, 51–54
Phase hits, 309
Phase jitter, 309–310, 321, 325–326
Phase modulation, 12
Phase shift keying, 148–160
Phase velocity, 52–55, 418
Psophometric weighting, 7–9
Physical level protocol, 388
Pilot tone, 68–69, 321
Point-to-point network, 384
Power in a multilevel pulse stream, 298–300
Private branch exchange (PBX), 41–42
Private lines, 121
Probability density function, 301–304
Progressive control switching, 71–73
Propagation constant, 51–54
Propagation times of cable and microwave systems, 375
Protocol levels, 388–390
Proximity effect, 131
Pulse amplitude modulation, 224–231
Pulse code modulation, 20, 85, 242–248
Pulse duration modulation, 230–235
Pulse position modulation, 230–236
Pulse transmission, 135
  effects of bandwidth limiting, 136
Quantization noise, 244–246, 250–252, 297–298
  measurement of, 284–285
Radiation losses, 131
Quadrature amplitude modulation, 148, 150
Raised-cosine characteristic, 328–330
Range extenders, 58, 60
RCA CD4051 multiplexor chip, 229
Reference equivalent, 101–103
Return-to-zero code, 280–281, 372
Ringing, 45
Rixon T103A data set, 165–180
Sample and hold, 205, 215–220, 249
Sampling, amplitude, 250
  rate for voice communication, 244
  theorem, 10, 207
  time, 204–219
Serial data transmission, 27
Shannon limit, 328
Shot noise, 112
Sidetone, 43, 45
Signal-to-noise ratio, 9, 64
  due to aliasing, 220
  due to quantization noise, 298–300
  for simple and adaptive delta modulation, 241
  improvement when companding, 124
  relationship to error rate for gaussian noise, 307–308
  using an 8 bit mu law encoder, 258
Signaling, 41, 129
  office-to-office, 50
  touchtone, 47–48
Signaling speed *See* Baud
Simplex operation, 32–33
Sinc function, 134, 207–208, 414
Singing, 101
Single-ended operation, 141
Single sideband, 13–14, 61, 64–68
Sink, 2
Skin effect, 130
Slope overload, 240–241
Source, 2
Special telephone services, 92–94
Speech, 6–7
Standard deviation, 301
Step-by-step switch, 70–72
Strowger switch, 70–71
Subscriber loop, 39, 41, 47, 56
Subscriber radio system, 58
Supervisory control, 73
Switched lines, 121
Switches, common control, 73–75
  crossbar, 73–75
  ferreed, 82
  space, 76–77
  step-by-step, 70–72
  time, 86
  strowger, 70–72
Switching, common control, 73–75, 78–80
  data network, 384
  digital, 84–91
  electronic (ESS), 80–84
  packet (message), 385–387, 395, 402–409

Switching (*cont'd.*)
  progressive control, 71–73
  space, 76–77
  time, 86–91
Switching office hierarchy, 39–40, 105
Synchronization, frame, 277
  with flag bit pattern, 396
Synchronous transmission, 27, 30
  advantages and disadvantages, 30
Tandem office, 41–42
Telegraph, 15
Telemetering, 93
Teletype, 32, 137
  TELEX, TWX, 92
Telephone network, 10–12
  500 type set 43–45
  touch tone set, 49
Terminals, 383
Thermal noise, 105, 112–114
Time division multiplex, 21, 224–226, 402–403
Timing extraction, 322
Tip-ring-sleeve, 45
Toll offices, 39
Touch-tone decoder, 48, 420–424
  signaling, 47–48
  telephone, 49
Transducer, 2
Transmission bridges, 47
Transmission line theory, 51–57
  balanced and unbalanced line, 141
Trunk circuits, 41–42, 47, 59–60
  loss, 105
Twenty milliampere loop, 32, 137–139
Two-wire circuits, 42, 60
Unbalanced line, 141
Unipolar pulses, 280–281
Unit impulse, 316–318
USART/UART, 31–32, 137–139
  COM 2502H, 185–192
  INTEL 8251A, 193–197
Variance, 301
Varistor, 45
Velocity of propagation, 115, 131–132
Vestigial sideband, 14
Via net loss, 103–105
Virtual circuit, 402–403
Voicecom 1, 92
Volume unit, 10
Wavelength, 3
White noise loading, 108, 111
Wide area telephone service (WATS), 92
X.25 protocol, 389, 392–401
Zero transmission level point, 8–9

ACTIVE FILTER DESIGN

TAFE NSW LIBRARY NETWORK
3 5555 07073483 7



ALLAN WATERS

621.
381532
WATE

MACMILLAN NEW ELECTRONICS
INTRODUCTIONS TO ADVANCED TOPICS

Macmillan New Electronics Series
Series Editor: Paul A. Lynn

R. F. W. Coates, *Underwater Acoustic Systems*
Paul A. Lynn, *Radar Systems*
A. F. Murray and H. M. Reekie, *Integrated Circuit Design*
Dennis N. Pim, *Television and Teletext*
M. J. N. Sibley, *Optical Communications*
Martin S. Smith, *Introduction to Antennas*
P. M. Taylor, *Robotic Control*
Allan Waters, *Active Filter Design*

Series Standing Order

If you would like to receive future titles in this series as they are published, you can make use of our standing order facility. To place a standing order please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address and the name of the series. Please state with which title you wish to begin your standing order. (If you live outside the United Kingdom we may not have the rights for your area, in which case we will forward your order to the publisher concerned.)

Customer Services Department, Macmillan Distribution Ltd
Houndmills, Basingstoke, Hampshire, RG21 2XS, England

Active Filter Design

Allan Waters

Nottingham Polytechnic

Macmillan New Electronics
Introductions to Advanced Topics

MACMILLAN

LIBRARY

© Allan Waters 1991

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No paragraph of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988 or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, 33-4 Alfred Place, London WC1E 7DP.

Any person who does any unauthorised act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

First edition 1991

Published by
MACMILLAN EDUCATION LTD
Houndmills, Basingstoke, Hampshire RG21 2XS
and London
Companies and representatives
throughout the world

Printed in Hong Kong

British Library Cataloguing in Publication Data
Waters, Alan
Active filter design.
1. Electric filters. Design
I. Title
621.3815324
ISBN 0-333-48861-X
ISBN 0-333-48862-8 pbk

Dedicated to Kay, Duncan and Catherine

Contents

Series Editor's Foreword x

Preface xi

1 Introduction **1**
1.1 Filter circuit concepts 1
1.2 Advantages of active over passive circuits 2
1.3 Component quality 3
1.4 Filter responses 4
1.5 Basic filter theory 4
1.6 Magnitude and frequency scaling 8
1.6.1 Magnitude scaling factor (K_m) 8
1.6.2 Frequency scaling factor ($K_f = \omega_o/\omega_n$) 9
1.7 Effect of complex signals 10

2 Filter Types **12**
2.1 Introduction 12
2.2 The Butterworth approximation 13
2.3 The Chebyshev approximation 18
2.4 Conclusions 25
Problems 26

3 The Operational Amplifier **27**
3.1 Introduction 27
3.2 The ideal op-amp 27
3.3 Two important op-amp configurations 29
3.3.1 The inverting circuit 29
3.3.2 The non-inverting circuit 30
3.4 The non-ideal op-amp 32
Problems 36

4 Basic Filter Circuits **39**
4.1 Introduction 39
4.2 The voltage-controlled-voltage-source first-order filter 39
4.2.1 Gain and phase response 43

ST. GEORGE LIBRARY Technical College

4.3 The voltage-controlled-voltage-source second-order filter
4.3.1 The low-pass filter 43
4.3.2 Gain and phase shift of the circuit 45
4.4 The VCVS high-pass filter 51
4.4.1 Gain and phase response 52
4.5 The VCVS band-pass filter 57
4.5.1 Gain and phase response 57
4.6 The infinite gain multiple feedback (MFB) filter 61
4.6.1 The multiple feedback low-pass filter 62
4.6.2 Gain and phase shift of the circuit 63
4.6.3 The multiple feedback high-pass filter 66
4.6.4 The multiple feedback band-pass filter 68
4.7 Higher-order filter circuits 72
4.8 Conclusions 74
Problems 80
80

5 Biquadratic Filter Circuits **83**
5.1 Introduction 83
5.2 The biquadratic circuit 83
5.3 The band-pass circuit 88
5.4 The band-reject, notch or band-elimination circuit 91
5.5 Switched capacitor filters 97
5.6 All-resistor active biquad circuits 105
5.7 Conclusions 108
Problems 109

6 Passive Filter Circuit Design **113**
6.1 Introduction 113
6.2 The coefficient matching technique 113
6.3 The design of doubly terminated *L*–*C* filters 114
6.4 Frequency transformations 123
6.4.1 Low-pass to high-pass 124
6.4.2 Low-pass to band-pass 126
Problems 128

7 Gyrator-based Filter Circuits **130**
7.1 Introduction 130
7.2 Properties of an ideal gyrator 130
7.3 The generalised impedance converter – GIC 132
7.4 Conclusions 140
Problems 140

8 Sensitivity **142**
8.1 Introduction 142

8.2 Definition of sensitivity 142
8.3 Sensitivity of the VCVS filter 143
8.4 Sensitivity of K, Q and ω_0 148
8.5 Conclusions 151
Problems 152

Appendix: The Use of Computer Aided Design Methods 153
A.1 Introduction 153
A.2 The CAD approach – the MICROCAP way 153
A.3 The laboratory procedure 155

Bibliography 158

Index 160

Series Editor's Foreword

The rapid development of electronics and its engineering applications ensures that new topics are always competing for a place in university and polytechnic courses. But it is often difficult for lecturers to find suitable books for recommendation to students, particularly when a topic is covered by a short lecture module, or as an 'option'.

Macmillan New Electronics offers introductions to advanced topics. The level is generally that of second and subsequent years of undergraduate courses in electronic and electrical engineering, computer science and physics. Some of the authors will paint with a broad brush; others will concentrate on a narrower topic, and cover it in greater detail. But in all cases the titles in the Series will provide a sound basis for further reading of the specialist literature, and an up-to-date appreciation of practical applications and likely trends.

The level, scope and approach of the Series should also appeal to practising engineers and scientists encountering an area of electronics for the first time, or needing a rapid and authoritative update.

Paul A. Lynn

Preface

My principal objective in writing this book is to present the basic concepts of active filtering in a way which, I hope, will be readily understood by undergraduate and BTEC HND students, and may also be found acceptable by postgraduate students who are unfamiliar with the subject.

Once the basic concepts have been assimilated, it is possible to comprehend the extensive published literature which covers all aspects of active filters. Indeed, I am indebted to many sources for much of the information in the book, not least the references given at the end of the book which have influenced me considerably.

The structure of the book is based on analysis, followed by synthesis; this approach has been favourably received by a succession of engineering students. While not claiming to treat all topics exhaustively, my aim has been to introduce, concisely, each underlying principle involved in the comprehensive design of a chosen filter. The books and articles in the Bibliography will provide more than adequate further information for those readers requiring a more descriptive and analytically detailed approach to the circuits.

It is my experience with undergraduates that achieving a successful design is paramount either for part of a complex project, or more mundanely for boosting confidence. The finer points and subtleties of the design procedure may then be studied later.

Chapter 1 introduces the concept of filtering and miscellaneous details, such as component quality, scaling and so on. The second chapter introduces the reader to the Butterworth and Chebyshev responses; this is followed, in chapter 3, by a brief overview of the operational amplifier.

Chapter 4 brings together the ideas in the preceding chapters and applies them to the popular VCVS and MFB circuits, along with a brief review of higher-order circuits.

Chapter 5 introduces the biquadratic circuits and includes a résumé of the switched-capacitor filter (SCF) and the all-resistor biquad.

Chapter 6 is devoted to an appreciation of passive circuits and concentrates on the synthesis technique. It also includes a brief discussion of the principle of frequency transformation.

Chapter 7 uses the passive circuit realisations of chapter 6 and introduces the reader to the gyrator, GIC and FDNR.

The final chapter shows how circuits behave for changes in component values – the so-called sensitivity analysis.

The book concludes with a short appendix outlining a computer aided design (CAD) technique which it is felt can considerably assist filter design.

Each chapter contains worked examples. In addition, there are problems, with answers, included at the end of most chapters.

Finally, I should like to record here my appreciation to those who have supported me in the writing of this book. I am indebted to my family for their continued support and encouragement, Barbara Sandford who produced the typescript, and Lynnda Aucott who drew the diagrams.

Allan Waters

1 Introduction

1.1 Filter circuit concepts

A filter is an electrical network which is designed specifically to modify, in a prescribed way, an electrical signal applied to its input terminals. An elementary example would be a signal having only low-frequency and high-frequency components. A *low-pass* filter would permit only the low frequencies to be transmitted, the high frequencies being suppressed or attenuated. The converse is true for passing the high frequencies while suppressing the low frequencies by means of a *high-pass* filter.

Other filters may be designed to pass a range of frequencies about a given selected frequency, rejecting frequencies outside the selected band. Conversely, a filter may be designed to reject a narrow band of frequencies while accepting all other frequencies outside the selected band. It may be said that electrical and electronic filter networks have become an indispensable part of electronic and modern communication systems.

The fundamental principles of electrical wave filters were outlined by Wagner in Germany and Campbell in the USA around the year 1915. Since those early beginnings, the *state-of-the-art* has encompassed many new ideas and techniques associated with modern network theory and feedback analysis. Examples include the work of Bode and Black on the stability of amplifiers and that of Darlington in the application of modern network theory to electrical filters.

The choice of a particular type of filter from the large range of filters must be influenced by several factors, some of which are: complexity, ease of tuning, economics, compatability with existing circuitry and whether or not power supplies may be required.

By referring to figure 1.1 it can be seen that active filter realisations are more advantageous at lower frequencies than is the case for passive types, although the reverse is true at higher frequencies. Furthermore, at frequencies close to or within the microwave range, only waveguide or coaxial cable type filters are possible.

Mention must be made of *digital filtering techniques* since such filters are becoming widely used and of increasing importance. Simply stated, these circuits perform the filtering task using analogue and digital elements including A/D converters, D/A converters, shift registers, multipliers and multiplexers, and were indeed first implemented using digital computers. Digital filters far exceed the analogue circuits where high-order filters and multiplexing are required, since

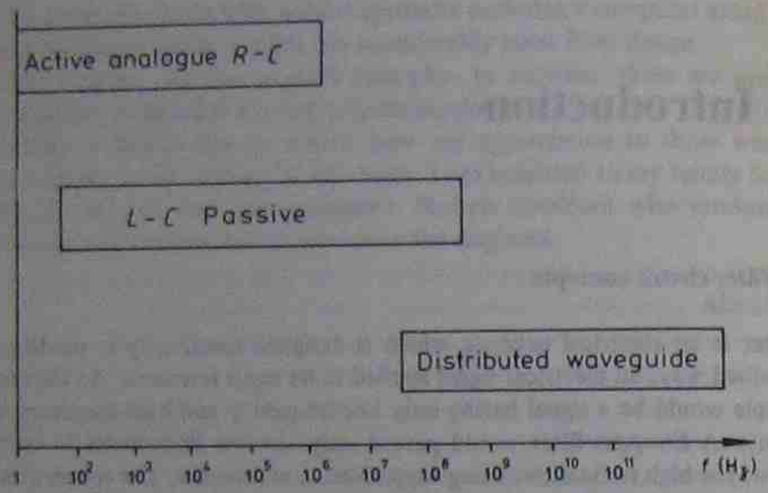


Figure 1.1 Comparison of frequency range of filters

tuning may be simply achieved by adjusting the coefficients of a mathematical algorithm. A disadvantage is that the filtering of high frequencies is limited by the time delays within the circuitry, but improvements are being made with the introduction of faster semiconductor materials such as gallium arsenide.

Other types of filters are the *piezo-crystal* type which presents very stable resonant frequencies, very high Q-factors and small power losses; *the surface acoustic wave* SAW type used in the majority of TV receivers and, finally, *mechanical* filters, a popular type being based on the principle of magneto-striction. This book will concentrate almost wholly on the R-C types with a brief mention of the passive types and their op-amp, gyrator equivalents.

1.2 Advantages of active over passive circuits

Passive filters are constructed from combinations of resistance and capacitance or inductance and capacitance and can be designed to cover a wide band of frequencies (typically 10 Hz to 500 MHz). They have the advantage over active circuits in that they do not require an external power supply and that they have low sensitivities to component variation. At low frequencies where cost and size are important considerations, there arises the problem of large coils having low Q factors. The coils usually require efficient magnetic cores and special winding techniques to enable a reasonable level of quality to be achieved, all of which results in increased costs.

Active filters are constructed from resistors, capacitors and an active source which is usually an operational amplifier. The main advantages over the passive

types are that expensive and bulky coils are eliminated, circuit gain may be realised, and high input and low output impedances give the amplifier good isolating properties which are useful when cascading networks to produce higher-order filters.

The main disadvantages are that a power supply is required, that above about 50 kHz the op-amp has a considerably reduced gain (although more expensive devices are available to the designer for frequencies up to 500 kHz) and that the circuit sensitivities to component changes are worse than the passive counterparts. At high frequencies, coils are of considerably improved quality, cheaper and physically much smaller, thereby making the passive L-C circuit a more viable proposition at such frequencies.

1.3 Component quality

Clearly it is of paramount importance to select good-quality components having a high stability over the dynamic working range. Care must therefore be taken in the choice of op-amps, resistors and capacitors when constructing active filter circuits. The op-amp is considered in chapter 3 where the important user design data is *gain-bandwidth product*, *slew rate* and whether internal or external *compensation* is to be employed.

The types of resistors used are: *carbon-composition*, *carbon-film* and *metal-film*.

Carbon-composition resistors are probably the most widely used where resistance values are maintained around ±10 per cent by the manufacturing process, although values within ±5 per cent are obtained by using an automatic sorting technique.

Carbon-film resistors are more stable than the composition type and have very good temperature coefficients of resistance with good ageing characteristics. Typical tolerances are ±5 per cent which makes this resistor well suited to general-purpose applications.

Metal-film resistors are the most widely used for precision requirements where tolerances within the range ±1 per cent to ±0.1 per cent are available. They exhibit better temperature stability and ageing characteristics than the carbon film type.

Choice of type of capacitor may be made from film (polyester types), mica, ceramic or electrolytic depending on factors such as frequency range, temperature stability and cost.

Film types should be considered for filters operating below a few hundred kiloherz and up to 125°C. They are normally available with standard tolerances of ±1 per cent, ±10 per cent and ±20 per cent.

Mica capacitors are more expensive than the film types but possess far superior properties such as temperature coefficient, where a few parts in a million may be achieved. These capacitors may be used well into the VHF region.

Ceramic capacitors have greater dielectric constants than the film and mica types and may be constructed to provide positive or negative temperature coefficients. This property makes these devices particularly useful for temperature compensation of active and passive circuits. They also exhibit minimum parasitic inductance and power dissipation over a range from low audio to high radio frequencies which make them attractive at radio frequencies. *Aluminium electrolytic capacitors* are usually employed as by-pass devices or for non-precision timing applications at low frequencies and are in general unsuitable for active or passive circuits. The *tantalum* form, however, may be used at low frequencies where low selectivity filters are required.

1.4 Filter responses

There are five basic forms of filter-response and these are shown in figure 1.2 where a second-order function for $H(j\omega)$ is assumed.

From a consideration of the gain and phase responses in figure 1.2, it can be seen that the denominator expressions for $H(j\omega)$ are similar in form but have differences in order to accommodate the different design requirements. For example, the low- and high-pass filters must involve not only the pass and stop-band information, but also the type of response – whether Butterworth or Chebyshev. The latter is controlled by the coefficients b_1 and b_0. The band-pass, band-reject and all-pass circuits must involve design data based on Q-factor, bandwidth, centre frequency and centre frequency gain.

Finally, it should be noted that the total phase shift for the circuits is 180°. This does not necessarily imply 0–180°, as will be seen in chapter 4.

1.5 Basic filter theory

A filter may be considered to be composed of lumped elements as shown in figure 1.3 where the input and output signals are continuous functions of time.

Using the Laplace operator s, where s is a complex variable, we may write the output/input relationship for the filter $H(s)$ in the form:

$$\frac{V_o}{V_i}(s) = H(s) \tag{1.1}$$

where $H(s)$ is called the *transfer function* of the filter.

We usually require to know how the filter responds to changing frequency so that equation (1.1) is modified by writing $s = j\omega$:

$$\frac{V_o}{V_i}(j\omega) = H(j\omega) \tag{1.2}$$

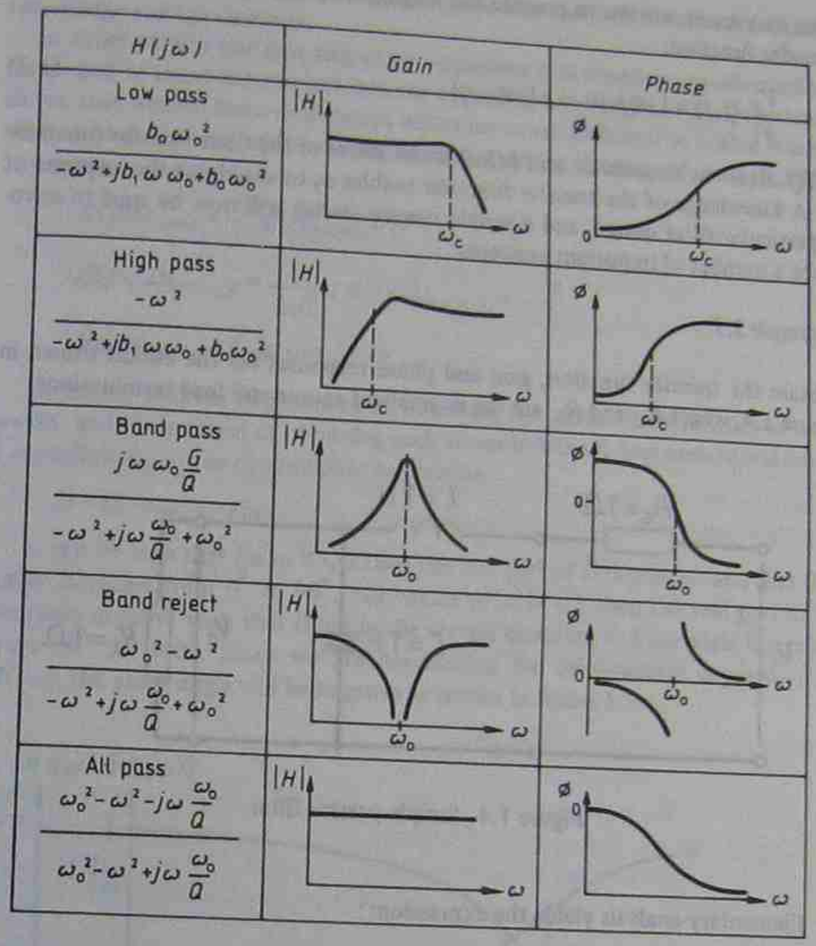


Figure 1.2 Basic filter responses

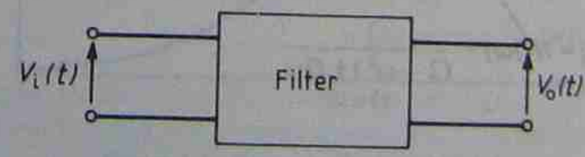


Figure 1.3 Two-port filter circuit

This may be re-written to provide the *magnitude* and *phase* components of the transfer function:

$$\frac{V_o}{V_i}(j\omega) = |H(j\omega)| \exp[j\phi(j\omega)] \tag{1.3}$$

$|H(j\omega)|$ is the *magnitude* and $\phi(j\omega)$ is the *phase* of the filter transfer function.

A knowledge of the transfer function enables us to anticipate the response of a particular filter design, and a simple *passive* circuit will now be used to introduce a number of important concepts.

Example 1.1

Obtain the transfer function, gain and phase responses for the circuit shown in figure 1.4, where R_S and R_L are the normalised source and load terminations.

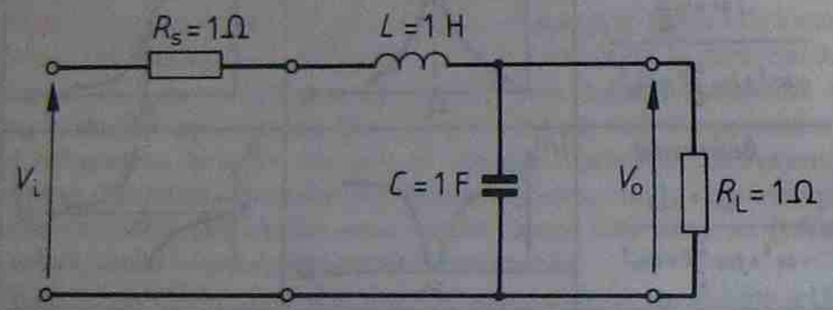


Figure 1.4 Simple passive filter

Elementary analysis yields the expression:

$$\frac{V_o}{V_i}(j\omega) = \frac{R_L}{R_L + R_S - \omega^2 LCR_L + j\omega(L + CR_LR_S)} \tag{1.4}$$

which becomes upon insertion of the component values:

$$\frac{V_o}{V_i}(j\omega) = H(j\omega) = \frac{1}{(2-\omega^2) + j2\omega}$$

giving

$$|H(j\omega)| = \frac{1}{\sqrt{\{(2-\omega^2)^2 + 4\omega^2\}}} \tag{1.5}$$

$$\phi(j\omega) = -\tan^{-1}\left(\frac{4\omega^2}{2-\omega^2}\right)$$

Notice the circuit is that of a second-order filter since it is constructed using two energy storage elements.

In order to plot the gain and phase responses, it is necessary to select values for ω and to insert these values into the appropriate equations (1.5). Experience shows that usually three well-chosen values for ω are sufficient to enable reasonable response curves to be drawn when considering reasonably simple filter structures.

$$|H(j0)|_{\omega=0} = \frac{1}{2};\ \phi(j0)_{\omega=0} = 0°$$

$$|H(j\sqrt{2})|_{\omega=\sqrt{2}} = \frac{1}{2\sqrt{2}};\ \phi(j\sqrt{2})_{\omega=\sqrt{2}} = -\frac{\pi}{2}$$

$$|H(j\infty)|_{\omega=\infty} = 0;\ \phi(j\infty)_{\omega=\infty} = -\pi$$

Difficulty is often experienced in obtaining the phase characteristics of networks, and the method of obtaining such characteristics is best understood from a consideration of the denominator expression:

$$D = (2-\omega^2) + j2\omega$$

It can be seen that for $\omega \leqslant \sqrt{2}$ then the real part of D is positive and that the angle increases from 0° to 90°. For values of $\omega \geqslant \sqrt{2}$ then the real part of D becomes negative such that D lies in the second quadrant and the angle increases from 90° to 180°. Since we are considering the denominator expression of $H(j\omega)$, the phase angle will be negative as shown in figure 1.5.

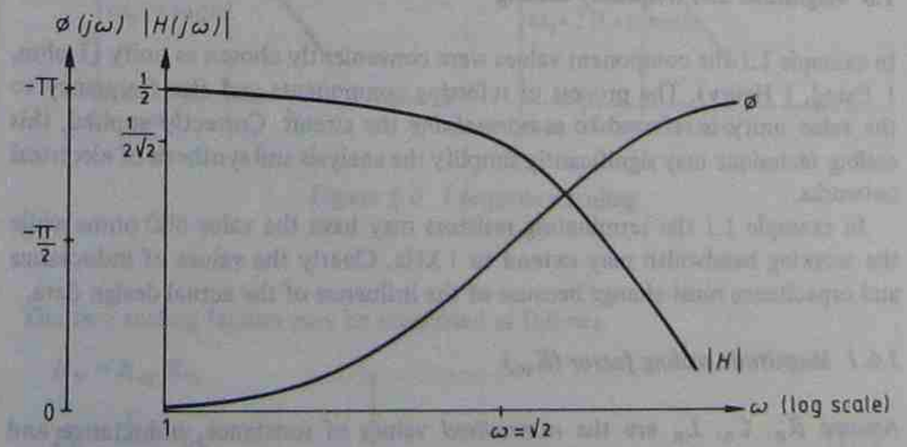


Figure 1.5 Response for example 1.1

The gain and phase responses for the circuit are shown in figure 1.5 plotted to a *logarithmic* frequency scale; a logarithmic frequency scale will be used through-

out the book unless otherwise stated, since this allows compression of the frequency scale.

The voltage gain may be expressed in decibels by using the expression:

$$\text{dB} = 20\log_{10}|H(j\omega)| \tag{1.6}$$

Applying equation (1.6) to the voltage gain of the filter circuit under consideration gives the expression:

$$\text{dB} = 20\log_{10}\frac{1}{\sqrt{\{(2-\omega^2)^2+4\omega^2\}}} = -10\log_{10}\{(2-\omega^2)^2+4\omega^2\}$$

At $\omega = 0$, $\text{dB} = -10\log_{10}2^2 = -6$ dB

$\omega = \sqrt{2}$, $\text{dB} = -10\log_{10}8 = -9$ dB

$\omega \gg \sqrt{2}$, $\text{dB} = -10\log_{10}\omega^4 = -40\log_{10}\omega$

From the results it can be seen that the 3 dB *bandwidth* of the response occurs when $\omega = \sqrt{2}$ and for the condition that ω is very large then the gain *rolls off* at 40 dB per decade change of frequency. We may summarise the performance of any order n of filter having the *maximally flat* Butterworth characteristic shown in figure 1.5, as having a roll-off given by $20n$ dB/decade and having a phase shift of $n\pi/4$ at the 3 dB frequency with a final phase shift of $n\pi/2$ at frequencies far removed from the cut-off frequency.

1.6 Magnitude and frequency scaling

In example 1.1 the component values were conveniently chosen as unity (1 ohm, 1 Farad, 1 Henry). The process of referring components and also frequency to the value unity is referred to as *normalising* the circuit. Correctly applied, this *scaling* technique may significantly simplify the analysis and synthesis of electrical networks.

In example 1.1 the terminating resistors may have the value 600 ohms while the working bandwidth may extend to 1 kHz. Clearly the values of inductance and capacitance must change because of the influence of the actual design data.

1.6.1 Magnitude scaling factor (K_m)

Assume R_n, C_n, L_n are the *normalised* values of resistance, inductance and capacitance and that R_0, L_0, C_0 are the corresponding *scaled* values. Then clearly

$$R_0 = K_m R_n$$

The criterion is that the reactance values should remain unchanged after the application of magnitude scaling, thereby preserving the shape of the response characteristic.

$$\omega L_0 = K_m(\omega L_n),\ L_0 = K_m L_n$$

$$\frac{1}{\omega C_0} = \frac{K_m}{\omega C_n},\ C_0 = \frac{C_n}{K_m}$$

1.6.2 Frequency scaling factor ($K_f = \omega_0/\omega_n$)

The criterion above must still hold for the reactance values after scaling has been applied when the frequency changes from ω_n to ω_0, again to preserve the shape of the response characteristic when moving to higher frequencies.

$$\omega_n L_n = \omega_0 L_0,\ L_0 = \left(\frac{\omega_n}{\omega_0}\right)L_n = \frac{L_n}{K_f}$$

$$\frac{1}{\omega_0 C_0} = \frac{1}{\omega_n C_n},\ C_0 = \left(\frac{\omega_n}{\omega_0}\right)C_n = \frac{C_n}{K_f}$$

Consideration of figure 1.6 shows that the frequency scaling factor shifts each frequency from the normalised circuit to the scaled circuit *in the same proportion.*

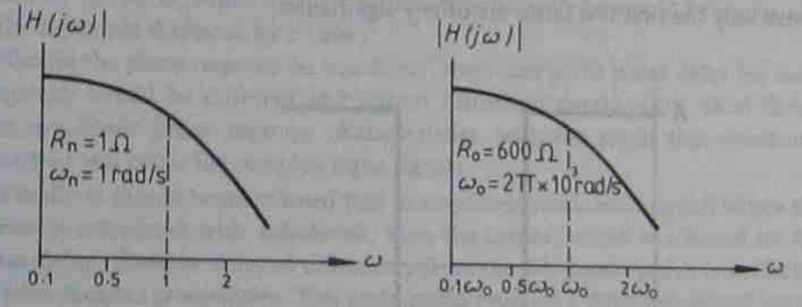


Figure 1.6 Frequency scaling

The two scaling factors may be combined as follows:

$$\left.\begin{aligned} R_0 &= K_m R_n \\ L_0 &= \frac{K_m}{K_f} L_n \\ C_0 &= \frac{C_n}{K_m K_f} \end{aligned}\right\} \tag{1.7}$$

Returning to Example 1.1, assume that the circuit is scaled for $R_L = R_S = 600\ \Omega$, $\omega_0 = 2\pi \times 10^3$ rad/s. Hence

$$L_0 = \frac{600 \times 1}{2\pi \times 10^3} = 0.095 \text{ H}$$

$$C_0 = \frac{1}{600 \times 2\pi \times 10^3} = 0.265\ \mu\text{F}$$

1.7 Effect of complex signals

We have previously considered how the magnitude and phase responses of a filter circuit are affected by a change in the frequency of the input signal where the input was a *pure* sine wave. However, in many filter applications, the input signal is a complex waveform which is made up from an infinite number of sine waves; a typical input signal may be considered to be the so-called *square* form shown in figure 1.7. The Fourier series for such a waveform may be represented by the expression

$$f(t) = A \sin \omega_0 t + \frac{A}{3} \sin 3\omega_0 t + \frac{A}{5} \sin 5\omega_0 t + \ldots \tag{1.8}$$

where only the first few terms are of any signifiance.

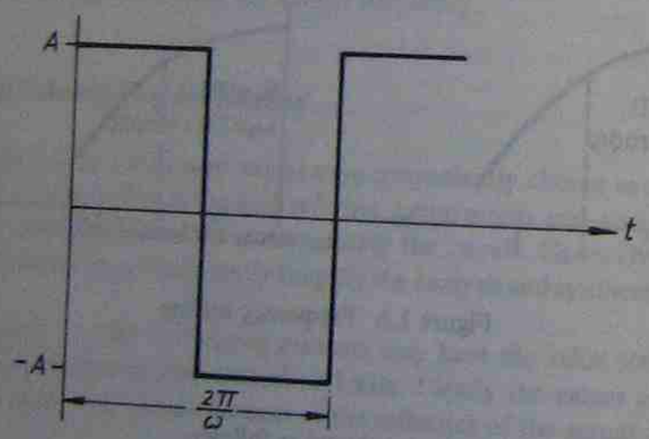


Figure 1.7 Square-wave input

The normal low-pass filter will be tuned to the fundamental component ω_0, all other components ($3\omega_0$, $5\omega_0$ etc.) being suppressed.

There is another type of filter, however, which is used to reproduce, ideally, at its output terminals, an exact replica of the input waveform such as that described by equation (1.8).

The problem arises when such a complex waveform is applied as the input signal to a filter. The fundamental along with all the significant harmonics must show the correct relative amplitude relationship at the output of the filter to maintain the original input signal. Furthermore, the various frequency components should not be displaced in time relative to one another. Filters designed around a constant delay for a given range of frequencies are called Bessel or Thomson filters and are not considered in this book.

Consider the response diagram shown in figure 1.5 and further assume that the phase characteristic is linear from dc to around the cut-off frequency (f_0). A sine-wave input signal is displaced in time at the output and may be determined by using the expression

$$\tau_p = \frac{\beta}{\omega} \tag{1.9}$$

where β is the phase shift in radians
ω is the input frequency in radians/second
τ_p is the phase delay.

For a linear phase response, the phase delay for each frequency component would be the same, which means that the output waveform would be equivalent to the input but displaced by a time τ.

Should the phase response be non-linear, then clearly the phase delay for each frequency would be different and output distortion would occur. Most filters have non-linear phase response characteristics, with the result that waveform distortion will occur for complex input signals.

Finally it should be mentioned that in amplitude modulated signals where the *carrier* is associated with *side-bands*, then the carrier, which is affected by the phase delay, could be delayed differently from the side-bands which are affected by what is called *group-delay*. This again could produce a distorted output signal.

2 Filter Types

2.1 Introduction

Modern filter design is based on the accurate selection of an appropriate input/output relationship which will satisfy the required specification of the filter. The steps involved in any successful filter design may be summarised as follows:

(a) Specification of a suitable filter characteristic - for example, attenuation, phase shift, size, weight etc.
(b) Selection of an appropriate rational function necessary to realise the specification.
(c) Analysis of the selected circuit (equivalent to the required transfer function) and the calculation of the component values of the chosen filter circuit.
(d) Preliminary computer analysis of the design using SPICE or any CAD (Computer Aided Design) package, followed by
(e) the construction and testing of the chosen circuit.

The second step usually involves what is often described as 'the approximation technique'. The filter specification, in the form of pass-band gain, transition bandwidth, stop-band attenuation and cut-off frequency, is normally given in the frequency domain. The selection of a realisable mathematical relationship which approximates to the given specification usually involves considerable algebraic manipulation. Two of the most popular relationships are considered here in a simplified form. The student is referred to the Bibliography for books covering the theory in greater detail.

From the outset it should be noted that almost all filter circuits are derived from the equivalent low-pass prototype. A typical *ideal* low-pass characteristic is shown in figure 2.1.

The requirements of the specification are that all frequencies below ω_c (cut-off) are passed with no attenuation and above ω_c are stopped with infinite attenuation. An additional requirement is that the phase $\phi(j\omega)$ should be a linear function of ω. An 'actual' magnitude characteristic is shown in figure 2.2.

It should be stressed at this point that the technique applied to filter design is one of choosing a filter response characteristic which closely approximates to the ideal and which finally results in a circuit realisation that contains a manageable number of components.

After the circuit has been realised using nominal component values to ascertain the correctness of the design, it is desirable to introduce tolerances within the

components to assess whether or not the original specification will still be met in a practical design.

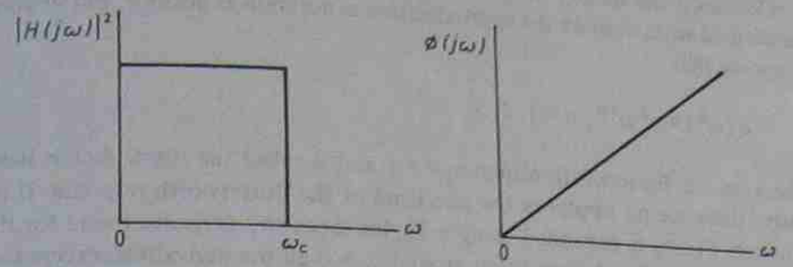


Figure 2.1 Ideal gain/phase responses

Generally, however, these parasitics, as they are often called, may be ignored in favour of completing an uncomplicated design, although sometimes they may be included within the design.

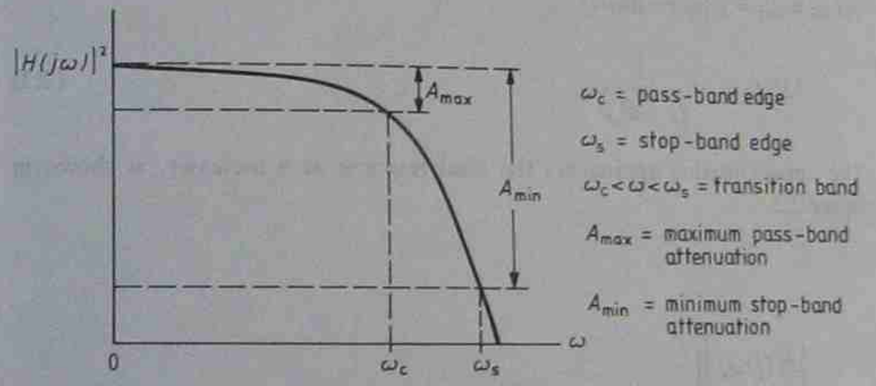


Figure 2.2 An actual gain response

2.2 The Butterworth approximation

The Butterworth characteristic takes the form described by the equation

$$|H(j\omega)|^2 = \frac{1}{1 + F(\omega^2)} \tag{2.1}$$

where $F(\omega^2) \gg 1$ for $\omega \gg \omega_c$ and $0 \leqslant F(\omega^2) \ll 1$ for $0 \leqslant \omega < 1$, so that in the pass-band $|H(j\omega)| \approx 1$ and in the stop-band $|H(j\omega)| \approx 0$.

Obviously the quality of the approximation depends on $F(\omega^2)$ which must be defined so as to make the approximation as accurate as possible. Butterworth proposed that

$$F(\omega^2) = \epsilon^2 \omega^{2n};\ n = 1, 2, 3 \ldots$$

where, in the Butterworth response, $\epsilon = 1$ and is called the *ripple factor*. Being unity, there are no ripples in the pass-band of the Butterworth response. If the function $F(\omega^2)$ is expanded using a Taylor series and derivates found for this expansion, then for the condition around $\omega = 0$ all the derivatives, except one, equal zero. This is often referred to as the *maximally flat* condition.

Insertion into equation (2.1) yields

$$|H(j\omega)| = \frac{1}{(1 + \epsilon^2 \omega^{2n})^{\frac{1}{2}}} \tag{2.2}$$

Equation (2.2) defines the AMPLITUDE response of an nth-order Butterworth filter and this approximation produces a particularly flat response close to $\omega = 0$. At $\omega = \omega_c = 1$ (normalised):

$$|H(j\omega)| = \frac{1}{(1 + \epsilon^2)^{\frac{1}{2}}} \tag{2.3}$$

The approximation approaches the ideal response as n increases, as shown in figure 2.3.

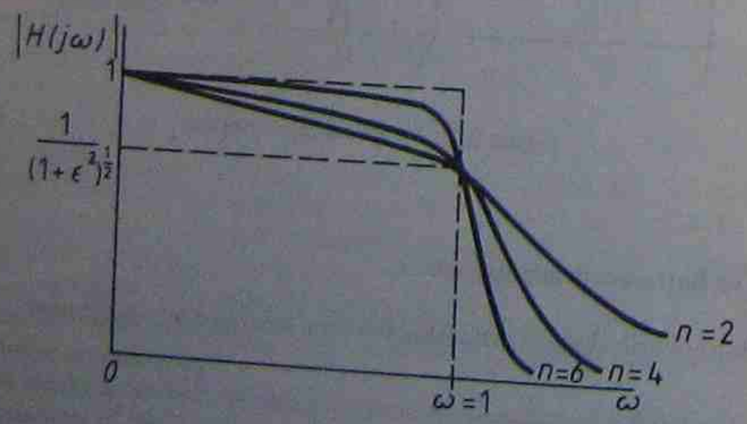


Figure 2.3 Butterworth gain response

The so-called *maximally flat* response is typical of the Butterworth approximation. The ripple factor ϵ has no effect on the pass-band response since in all of the cases the pass-band attenuation is taken to be 3 dB; this yields $\epsilon = 1$ from equation (2.6). Equation (2.2) may be written

$$|H(j\omega)| = \frac{1}{(1 + \omega^{2n})^{\frac{1}{2}}}$$

and for $\omega \gg 1$ this becomes $|H(j\omega)| \approx 1/\omega^n$ and the stop-band attenuation is then

$$\alpha_s = 20\log_{10}|H(j\omega)| = -20n\log_{10}\omega \tag{2.4}$$

Equation (2.4) shows that the attenuation 'rolls-off' at $20n$ dB/decade change in frequency (giving the slope of the response characteristic in the transition band).

The pass-band attenuation is obtained from

$$\alpha_p = 20\log_{10}|H(j\omega)| = 20\log_{10}\frac{1}{(1 + \omega^{2n})^{\frac{1}{2}}}$$

$$= -10\log_{10}(1 + \omega^{2n}) \tag{2.5}$$

It was stated that $\epsilon = 1$, but equation (2.5) can be used to obtain the ripple factor for given filter specifications; an example would be for the Chebyshev response, to be discussed later.

From equation (2.5) the pass-band attenuation may be obtained in the form $\alpha_p = -10\log_{10}(1 + \epsilon^2\omega^{2n})$ and at $\omega = 1$ this should be A_{MAX}, that is $|A_{MAX}| = 10\log_{10}(1 + \epsilon^2)$ yielding

$$\epsilon = (10^{A\,MAX/10} - 1)^{\frac{1}{2}} \tag{2.6}$$

Furthermore, it is important that the order (n) of the filter be obtained from the given specifications, and

$$|A_{MIN}| = 10\log_{10}(1 + \epsilon^2\omega^{2n}) \tag{2.7}$$

Substituting for ϵ gives

$$A_{MIN} = 10\log_{10}\,[1 + \{10^{A\,MAX/10} - 1\}\,(\omega^{2n})]$$

and

$$\omega^{2n} = \left\{\frac{10^{A\,MIN/10} - 1}{10^{A\,MAX/10} - 1}\right\} \quad \text{and} \quad \omega = (\omega_s/\omega_c)$$

Taking logarithms, we obtain

$$n = \frac{1}{2\log_{10}\left(\frac{\omega_s}{\omega_c}\right)} \times \log_{10}\left\{\frac{10^{A\,MIN/10} - 1}{10^{A\,MAX/10} - 1}\right\} \tag{2.8}$$

Using equations (2.6) and (2.8), the ripple factor (width) and the order n may be calculated. For example

for $A_{MAX} = 3$ dB in the Butterworth case

$$\epsilon = (e^{0.3} - 1)^{\frac{1}{2}} = 1$$

Since $F(\omega^2)$ is a polynomial in ω, it follows that the roots of the denominator of equation (2.2) must be obtained since they dictate the nature of the response for any order n.

Using $|H(j\omega)| = \dfrac{1}{(1 + \omega^{2n})^{\frac{1}{2}}}$ with $\epsilon = 1$

and writing $s = j\omega$, then

$$|H(s)|^2 = \frac{1}{1 - (s^2)^n}$$

Solving for the roots of $1 - (s^2)^n = 0$ yields:

for n even

$s^{2n} = -1 = \exp[j(2k - 1)\pi]$ whose $2n$ roots are:

$$p_k = \exp\left\{\frac{j(2k-1)\pi}{2n}\right\} \qquad k = 1, 2 \ldots 2n$$

or $$p_k = \cos\left[\frac{(2k-1)\pi}{2n} + j\sin\frac{(2k-1)\pi}{2n}\right]$$

for n odd

$s^{2n} = 1 = \exp(j2k\pi)$ whose $2n$ roots are:

$$p_k = \exp\left[j\frac{k}{n}\pi\right] \qquad k = 0, 1 \ldots (2n-1)$$

or $$p_k = \cos\frac{k}{n}\pi + j\sin\frac{k}{n}\pi$$

The transfer functions may be summarised:

for n even

$$|H(s)|^2 = \frac{1}{\prod_{k=1}^{n/2}(s^2 + 2\cos\theta_k s + 1)}$$

$$\theta_k = \frac{(2k-1)\pi}{2n}$$

for n odd

$$|H(s)|^2 = \frac{1}{(s+1)^{(n-1)/2}\prod_{k=1}^{(n-1)/2}(s^2 + 2\cos\theta_k s + 1)}$$

$$\theta_k = \frac{k\pi}{n}$$

For example, $n = 2$:

$$|H(s)|^2 = \frac{1}{s^2 + 2\cos\frac{\pi}{4}s + 1} = \frac{1}{s^2 + \sqrt{2}s + 1}$$

For example, $n = 3$:

$$|H(s)|^2 = \frac{1}{(s+1)\left(s^2 + 2\cos\frac{\pi s}{3} + 1\right)} = \frac{1}{(s^3 + 2s^2 + 2s + 1)}$$

A table of factors may now be constructed and is shown in table 2.1 up to and including the eighth-order polynomial.

Table 2.1 Table of factors for Butterworth polynomial

n	*Roots*	*Factors*	*Polynomial*
1	-1	$s+1$	$s+1$
2	$-\cos\frac{\pi}{4} \pm j\sin\frac{\pi}{4}$	$s^2 + \sqrt{2}s + 1$	$s^2 + \sqrt{2}s + 1$
3	$-1, -\cos\frac{\pi}{3} \pm j\sin\frac{\pi}{3}$	$(s+1), (s^2 + s + 1)$	$s^3 + 2s^2 + 2s + 1$
4	$-\cos k\frac{\pi}{8} \pm j\sin k\frac{\pi}{8}$ $k = 1, 3 \ldots$	$(s^2 + 0.765s + 1), (s^2 + 1.848s + 1)$	$s^4 + 2.613s^3 + 3.414s^2 + 2.613s + 1$
5	$-1, -\cos k\frac{\pi}{5} \pm j\sin k\frac{\pi}{5}$ $k = 1, 2$	$(s+1), (s^2 + 0.618s + 1),$ $(s^2 + 1.618s + 1)$	$s^5 + 3.236s^4 + 5.236s^3 + 5.236s^2 + 3.236s + 1$
6	$-\cos k\frac{\pi}{12} \pm j\sin k\frac{\pi}{12}$ $k = 1, 3, 5$	$(s^2 + 0.518s + 1), (s^2 + \sqrt{2}s + 1),$ $(s^2 + 1.932s + 1)$	$s^6 + 3.864s^5 + 7.464s^4 + 9.142s^3 + 7.464s^2 + 3.864s + 1$
7	$-1, -\cos k\frac{\pi}{7} \pm j\sin k\frac{\pi}{7}$ $k = 1, 2, 3$	$(s+1), (s^2 + 0.44s + 1),$ $(s^2 + 1.247s^2 + 1), (s^2 + 1.802s + 1)$	$s^7 + 4.494s^6 + 10.098s^5 + 14.592s^4 + 14.592s^3 + 10.098s^2 + 4.494s + 1$
8	$-\cos k\frac{\pi}{16} \pm j\sin k\frac{\pi}{16}$ $k = 1, 3, 5, 7$	$(s^2 + 0.399s + 1), (s^2 + 1.111s + 1),$ $(s^2 + 1.166s + 1), (s^2 + 1.962s + 1$	$s^8 + 5.126s^7 + 13.137s^6 + 21.846s^5 + 25.688s^4 + 21.846s^3 + 13.137s^2 + 5.126s + 1$

An example will now be considered with which to illustrate how such a table of factors, as shown in table 2.1, may be derived.

Example 2.1

A Butterworth low-pass filter is required to provide the following specifications: pass band up to 10 kHz with 3 dB of loss, stop band to be at least 100 dB down at 100 kHz. Calculate the order of the filter and the roots of the Butterworth polynomial.

Here $A_{MAX} = 3$ dB, $A_{MIN} = 100$ dB, $\epsilon = 1$.

$$n = \frac{1}{2\log_{10}10} \log_{10} \left\{\frac{10^{10} - 1}{10^3 - 1}\right\}$$

$$= \tfrac{1}{2} \times \log_{10} \frac{10^{10}}{0.995} = 5$$

We have an odd function, hence $\theta_n = k\pi/n$ and a root on the negative real axis.

$$\frac{k\pi}{n} = \theta_n$$

$k = 0 \quad \theta_0 = 0$

$k = 1 \quad \theta_1 = \pi/5 = 36^\circ$

$k = 2 \quad \theta_2 = 2\pi/5 = 72^\circ$

Roots are $(s + \sigma_0)(s + \sigma_1 + j\omega_1)(s + \sigma_1 - j\omega_1)(s + \sigma_2 + j\omega_2)(s + \sigma_2 - j\omega_2)$.

$\sigma_1 \pm j\omega_1 = \cos 36^\circ \pm j\sin 36^\circ = 0.809 \pm j0.587$

$\sigma_2 \pm j\omega_2 = \cos 72^\circ \pm j\sin 72^\circ = 0.309 \pm j0.951$

$\sigma_0 = \cos 0^\circ \pm j\sin 0^\circ = 1$

And

$$|H(s)|^2 = \frac{1}{(s+1)(s^2 + 1.618s + 1)(s^2 + 0.618s + 1)}$$

Example 2.2

It is required to determine the order of a Butterworth filter having the following specification: $A_{MIN} = 20$ dB, $A_{MAX} = 3$ dB, $\omega_s/\omega_c = 1.5$.

From equation (2.8):

$$n = \frac{1}{2\log_{10}1.5} \times \log_{10} \left\{\frac{10^2 - 1}{10^{0.3} - 1}\right\} = 5.67$$

Choose $n = 6$ and from equations (2.6) and (2.7) $\epsilon = (10^{0.3} - 1)^{\frac{1}{2}} = 1$ which is as expected for Butterworth.

$$A_{MIN} = 10\log_{10}(1 + 1^2(1.5)^{12}) = 10 \times 2.12 = 21.2 \text{ dB}$$

which meets the specification.

2.3 The Chebyshev approximation

It was seen that the Butterworth form gives a good approximation to the ideal around ($\omega = 0$) but is less than ideal in the vicinity of the cut-off point ($\omega = \omega_c$).

This is because $F(\omega^2) = \omega^{2n}$ is chosen such that the zeros of $F(\omega^2)$ are at one point ($\omega = 0$). An alternative approximation, called the Chebyshev approximation, spreads the zeros of $F(\omega^2)$ across the pass band and constrains $H(j\omega)$ to attain its maximum value at a number of points within the pass band. It will be seen that the ripple factor ϵ affects the way the response behaves in the pass band. Depending on the order of the filter, pass-band ripples occur having equal heights and variable frequency, the ripple height (or width) being determined by the ripple factor.

Here the Chebyshev polynomial is defined by

$$F(\omega^2) = \epsilon^2 C_n^2(\omega) \text{ and} \tag{2.9}$$

$$|H(j\omega)| = \frac{1}{(1 + \epsilon^2 C_n^2(\omega))^{\frac{1}{2}}} \text{ where} \tag{2.10}$$

$$\epsilon = (10^{A_{MAX}/10} - 1)^{\frac{1}{2}} \text{ as before}$$

The polynomial $C_n(\omega)$ is defined as

$$C_n(\omega) = \cos(n\cos^{-1}\omega) \quad \text{for } \omega \leqslant 1$$
$$C_n(\omega) = \cosh(n\cosh^{-1}\omega) \quad \text{for } \omega > 1 \tag{2.11}$$

$C_n(\omega)$ exhibits an equal ripple property in the pass band.

$n = 0;\ C_0(\omega) = 1$

$n = 1;\ C_1(\omega) = \omega$

To obtain the higher-order polynomials we use the relationship

$$C_{n+1}(\omega) = 2\omega C_n(\omega) - C_{n-1}(\omega) \tag{2.12}$$

Table 2.2 may be constructed using equations (2.11) and (2.12).

Table 2.2 A table of Chebyshev polynomials

n	$C_n(\omega)$
0	1
1	ω
2	$2\omega^2 - 1$
3	$4\omega^3 - 3\omega$
4	$8\omega^4 - 8\omega^2 + 1$
5	$16\omega^5 - 20\omega^3 + 5\omega$
6	$32\omega^6 - 48\omega^4 + 18\omega^2 - 1$
7	$64\omega^7 - 112\omega^5 + 56\omega^3 - 7\omega$
8	$128\omega^8 - 256\omega^6 + 160\omega^4 - 32\omega^2 + 1$

If these functions are plotted we see the equal ripple characteristics for $\omega \leqslant 1$, and two examples are given in figure 2.4.

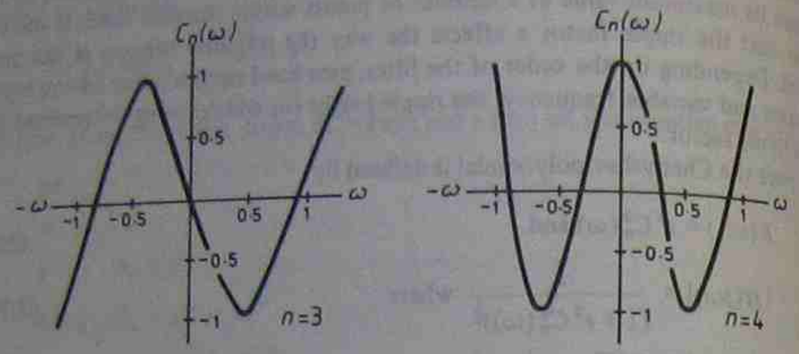


Figure 2.4 Plots of $C_n(\omega)$ for $n = 3$, $n = 4$

Consideration of equation (2.10), repeated here

$$|H(j\omega)| = \frac{1}{(1 + \epsilon^2 C_n^2(\omega))^{\frac{1}{2}}}$$

shows that within the pass band ($0 < \omega \leqslant 1$), $C_n^2(\omega) = 1$ and the function will vary between $1/(1 + \epsilon^2)^{\frac{1}{2}}$ and 1.

Within the transition band and through the stop band, however, $C_n^2(\omega)$ increases with ω such that for $\epsilon^2 C_n^2(\omega) \gg 1$, $|H(j\omega)|$ decreases rapidly. These features are shown in figure 2.5 from which the *ripple width* may be determined:

$$\text{ripple width} = 1 - \frac{1}{(1 + \epsilon^2)^{\frac{1}{2}}} \tag{2.13}$$

To determine the order n of the filter from given specifications, we note from equation (2.10)

$$A_{\text{MIN}} = 10\log_{10}[1 + \epsilon^2 C_n^2(\omega)]$$

for the edge of the stop band at $\omega = \omega_s/\omega_c$ normalised. From which

$$\epsilon^2 C_n^2(\omega) = \epsilon^2(\cosh(n\cosh^{-1}\omega))^2 = 10^{A\,\text{MIN}/10} - 1 \quad \text{giving}$$

$$|A_{\text{MIN}}| = 10\log_{10}[1 + \epsilon^2\{\cosh(n\cosh^{-1}\omega)\}^2] \tag{2.14}$$

and $\epsilon^2 = 10^{A\,\text{MAX}/10} - 1$

Which upon insertion into (2.14) gives

$$\cosh(n\cosh^{-1}\omega) = \left[\frac{10^{A\,\text{MIN}/10} - 1}{10^{A\,\text{MAX}/10} - 1}\right]^{\frac{1}{2}}$$

and which finally yields

$$n = \frac{\cosh^{-1}[(10^{A\,\text{MIN}/10} - 1)/(10^{A\,\text{MAX}/10} - 1)]^{\frac{1}{2}}}{\cosh^{-1}\left(\frac{\omega_s}{\omega_c}\right)} \tag{2.15}$$

Also since for the pass band pass $C_n^2(\omega) = 1$:

$$|A_{\text{MAX}}| = 10\log_{10}[1 + \epsilon^2] \tag{2.16}$$

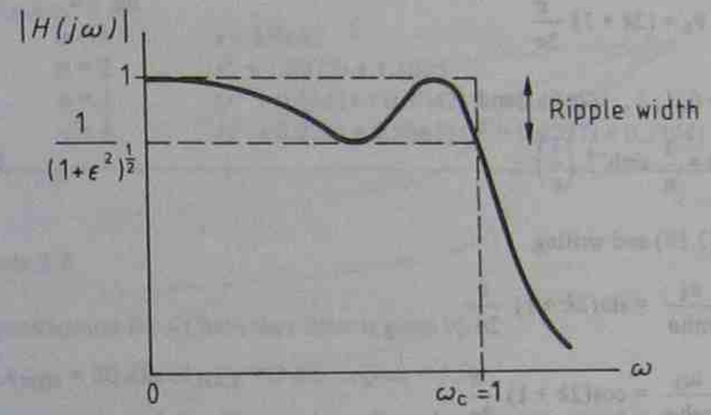


Figure 2.5 Effect of ripple width

Unlike the Butterworth approximation, it cannot be assumed that the 3 dB or *half-power* frequency is the cut-off frequency. The end of the pass band or ripple band is always at $\omega = \omega_c$. For the condition $\epsilon = 1$ gives a ripple width of 3 dB and $\omega = \omega_c$ as for the Butterworth approximation.

To determine the half-power frequency we note that $\epsilon^2 C_n^2(\omega) = 1$ which for $\omega = \omega_{3\text{dB}}$ gives

$$C_n(\omega_{3\text{dB}}) = \frac{1}{\epsilon} = \cosh(n\cosh^{-1}\omega_{3\text{dB}})$$

which yields $\cosh^{-1}\omega_{3\text{dB}} = \frac{1}{n}\cosh^{-1}\left(\frac{1}{\epsilon}\right)$ and finally

$$\omega_{3\text{dB}} = \cosh\left[\frac{1}{n}\cosh^{-1}\left(\frac{1}{\epsilon}\right)\right] \tag{2.17}$$

Finally it is necessary to determine the roots of the Chebyshev polynomial for various values of the order n of the filter.

The roots are obtained from the expression

$$1 + \epsilon^2 C_n^2(\omega) = 0 \tag{2.18}$$

and are $s_k = \sigma_k \pm j\omega_k, \quad k = 0, 1, 2 \ldots 2n-1$

where $\sigma_k = \sinh a \sin(2k+1)\frac{\pi}{2n}$

$$\omega_k = \cosh a \cos(2k+1)\frac{\pi}{2n} \tag{2.19}$$

$$\theta_k = (2k+1)\frac{\pi}{2n}$$

for $k = 0, 1, 2 \ldots 2n-1$ and

$$a = \frac{1}{n}\sinh^{-1}\left(\frac{1}{\epsilon}\right) \tag{2.20}$$

Using (2.19) and writing

$$\frac{\sigma_k}{\sinh a} = \sin(2k+1)\frac{\pi}{2n}$$

$$\frac{\omega_k}{\cosh a} = \cos(2k+1)\frac{\pi}{2n}$$

then

$$\left(\frac{\sigma_k}{\sinh a}\right)^2 + \left(\frac{\omega_k}{\cosh a}\right)^2 = 1$$

which indicates that the poles (roots) of $|H(s)|^2$ lie on an ellipse. It is interesting to note that those of the Butterworth polynomial lie on a circle. Furthermore, it should be noted that the Chebyshev poles depend on the ripple factor ϵ, and tables are constructed for various values of ripple width. These are shown in table 2.3.

It may further be shown for the special case when $\omega = \omega_c$, that is for $\epsilon = 1$, the Chebyshev response has a slope which is n times that of the Butterworth response.

If, briefly, we consider the relative phase responses shown in figure 2.6, we may conclude that the choice of response type will be governed by: high attenuation in the stop band and sharper roll-off in the vicinity of the cut-off frequency as opposed to a non-linear phase characteristic in the stop band.

Table 2.3 Poles of the Chebyshev polynomial for values of A_{MAX}

$A_{MAX} = 0.1$ dB	
$n = 1$	$s + 6.552$
$n = 2$	$s^2 + 2.3724s + 3.314$
$n = 3$	$(s^2 + 0.9694s + 1.6897)(s + 0.9694)$
$n = 4$	$(s^2 + 0.5283s + 1.33)(s^2 + 1.2754s + 0.6229)$
$A_{MAX} = 0.5$ dB	
$n = 1$	$s + 2.8627$
$n = 2$	$s^2 + 1.4256s + 1.5162$
$n = 3$	$(s^2 + 0.6264s + 1.1424)(s + 0.6264)$
$n = 4$	$(s^2 + 0.3507s + 1.0635)(s^2 + 0.8466s + 0.3564)$
$A_{MAX} = 1$ dB	
$n = 1$	$s + 1.9652$
$n = 2$	$s^2 + 1.0977s + 1.1025$
$n = 3$	$(s^2 + 0.4941s + 0.9942)(s + 0.4941)$
$n = 4$	$(s^2 + 0.279s + 0.9865)(s^2 + 0.6737s + 0.2794)$

Example 2.3

The specification for a Chebyshev filter is given by

$$A_{MIN} = 20 \text{ dB}, \; A_{MAX} = 3 \text{ dB}, \; \omega_s/\omega_c = 1.5$$

Obtain the order of the filter and confirm that the stop band requirement is fulfilled.

From equation (2.15):

$$n = \frac{\cosh^{-1}[(10^2 - 1)/(10^{0.3} - 1)]^{\frac{1}{3}}}{\cosh^{-1} 1.5} = 3.11$$

Choose $n = 4$ and from equations (2.6) and (2.14)

$$\epsilon = (10^{0.3} - 1)^{\frac{1}{2}} = 1$$

giving

$$A_{MIN} = 10\log_{10}[1 + (1)^2 \{\cosh(4\cosh^{-1}1.5)\}^2]$$
$$= 10\log_{10}[1 + \cosh(3.85)]$$
$$= \underline{27.4 \text{ dB}}$$

Note that this fourth-order Chebyshev circuit satisfies the requirements better than the previously considered sixth-order Butterworth circuit, at the expense of some pass-band ripple.

ST. GEORGE LIBRARY Technical College

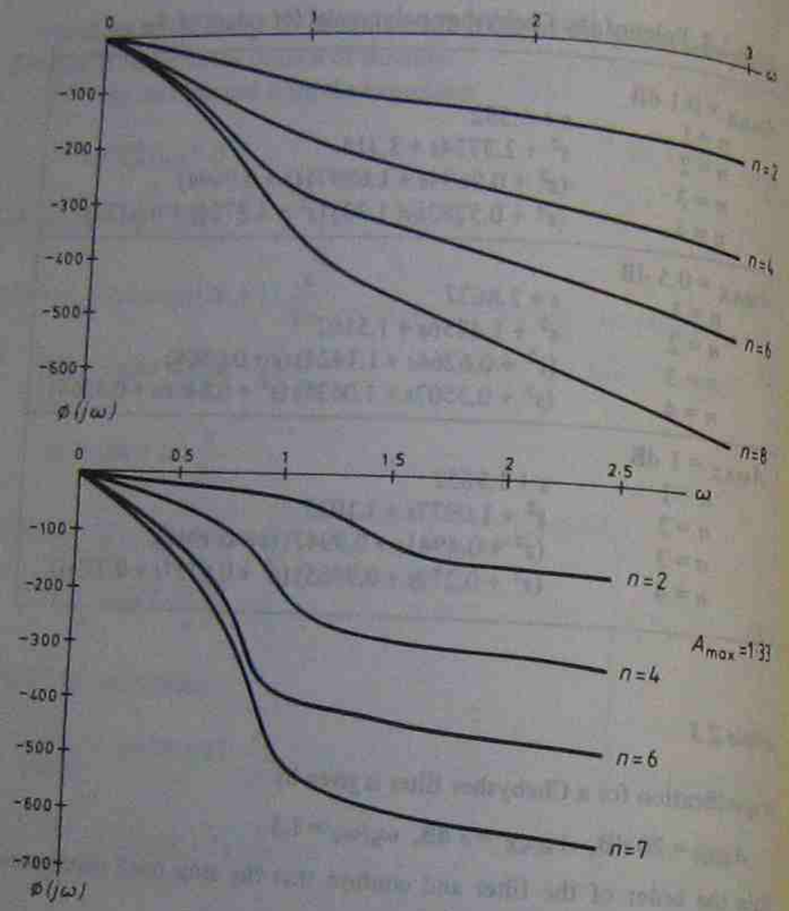


Figure 2.6 Butterworth and Chebyshev phase responses

Example 2.4

A Chebyshev filter is required to provide the following specification:

pass-band maximum ripple width 0.5 dB up to 3 kHz

stop-band minimum to be 60 dB at 30 kHz

Calculate the order of the filter and the roots of the polynomial.

Given $A_{MAX} = 0.5$ dB, $A_{MIN} > 60$ dB.

$$n = \frac{\cosh^{-1}[(10^6 - 1)/(10^{0.05} - 1)]^{\frac{1}{2}}}{\cosh^{-1} 30/3} = 2.88$$

Take $n = 3$: using equation (2.6) we find that $\epsilon = 0.35$.

$$a = \frac{1}{3} \sinh^{-1}\left(\frac{1}{0.35}\right) = 0.59$$

$$\theta_k = (2k + 1)\frac{\pi}{2n}$$

giving

$$\theta_0 = \frac{\pi}{6} = 30°$$

$$\theta_1 = 90°$$

There are poles at $s_1 = -\sinh a = -\sinh 0.59 = -0.625$

$$s_2 = -0.625 \sin 30° \pm j1.18 \cos 30°$$
$$= -0.313 \pm j1.022$$

$$s_3 = -0.625 \sin 90° \pm j1.18 \cos 90°$$
$$= -0.625 \text{ (repeated root)}$$

The transfer function takes the form

$$|H(s)|^2 = \frac{1}{(s + 0.625)(s + 0.313 + j1.022)(s + 0.313 - j1.022)}$$

$$= \frac{1}{(s + 0.625)(s^2 + 0.626s + 1.142}$$

This result should be compared with the data given in table 2.3. It should again be noted that we have also obtained important design coefficients which will later be used in the design of filters having specified ripple widths.

We may write

$$|H(s)|^2 = \frac{1}{(s + b_0)(s^2 + b_1 s + b_2)}$$

where $b_0 = 0.625$, $b_1 = 0.625$, $b_2 = 1.142$.

Table 2.4 shows such coefficients for various order n of the filter and different pass-band ripple widths.

2.4 Conclusions

We may conclude that the Chebyshev response gives a superior pass-band response, particularly if the ripple width is not too large, than does the Butterworth response. Unfortunately this superiority in amplitude response must be traded

Table 2.4 Chebyshev coefficients for various orders and maximum ripple widths

Order n	A_{MAX} (dB)	b_0	b_1
2	0.5	1.516	1.426
	1.0	1.1025	1.0977
	3.0	0.708	0.645
3	0.5	1.142	0.626
		0.626	
	1.0	0.994	0.494
		0.494	
	3.0	0.839	0.298
		0.298	
4	0.5	1.0635	0.3507
		0.3564	0.8466
	1.0	0.9865	0.279
		0.279	0.6737
	3.0	0.903	0.1703
		0.1959	0.4112

against a poorer, less linear, phase response, up to the cut-off frequency. The Butterworth phase response is far more linear, as shown in figure 2.6. A further point is that in the vicinity of the cut-off frequency, the roll-off of the response for the Chebyshev is n times steeper than that for the Butterworth.

Problems

2.1. A Butterworth low-pass filter has the following specifications: pass band up to 100 Hz with 3 dB of loss, stop band to be at least 80 dB down at 1 kHz. Calculate the order of the filter, the roots of the Butterworth polynomial, and the values of the pass-band and stop-band loss.
[Ans. 4; 0.765, 1, 1.848, 1; 3 dB, 80 dB]

2.2. Obtain the normalised Chebyshev coefficients for a fourth-order, 1 dB ripple width, low-pass filter.
[Ans. $\epsilon = 0.509$; 0.279, 0.9865; 0.6737, 0.2793]

2.3. A Chebyshev low-pass filter is required to have a pass-band ripple width of 0.5 dB from dc to 150 Hz, and to be at least 100 dB down at 1.5 kHz. Calculate the order of the filter and the Chebyshev polynomial. Check also the values of the pass-band and stop-band loss.
[Ans. 5; 0.362; 0.586, 0.476; 0.224, 1.035; 0.5 dB, 114.8 dB]

3 The Operational Amplifier

3.1 Introduction

It is useful to consider at this stage the basic theory concerning integrated circuit operational amplifiers, since they are the foundation of most of the filter circuits to be studied. The reader should refer to the more specialised texts for a fuller treatment of the subject matter.

The operational amplifier (op-amp) represents a basic integrated circuit building block and is a broad-band, high-gain, differential input device. Initially, the op-amp was applied to the solution of such mathematical functions as integration, differentiation and summing by electronic means. Nowadays, the applications of the device are legion.

At the present time, the availability of inexpensive integrated circuit amplifiers, either in single, dual or quad form, has made the packaged op-amp a useful replacement for discrete amplifiers in many active circuit realisations.

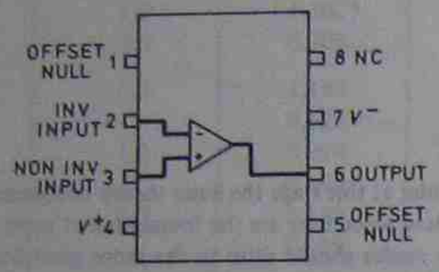
The versatility of the op-amp is improved by using negative feedback which improves stability of gain, reduces output impedance and improves linearity of operation. Positive and negative power supplies are required, both referenced to earth potential.

A popular and inexpensive op-amp is the μA741 which features offset-null capability, protection against short-circuit, and built-in frequency compensation, which results in unity gain bandwidth of 1 MHz. The μA748 is similar to the μA741 but features externally selected frequency compensation through the choice of a single capacitor.

Dual and quad op-amps have been introduced to reduce cost and space; in particular, the quad op-amp has found particular favour with filter designers. A typical specification and op-amp pin layout is shown in figure 3.1 for the popular μA741 monolithic amplifier.

3.2 The ideal op-amp

A simplified model for representing an op-amp is shown in figure 3.2 where R_i, R_o are the input and output impedances and A is the open-loop gain of the op-amp.



ELECTRICAL CHARACTERISTICS (V_S = ±15 V, T_A = 25°C unless otherwise specified)

Parameters		Conditions	Min.	Typ.	Max.
Input offset voltage		$R_S < 10\,k\Omega$		2.0	6.0
Input offset current				20	200
Input bias current				80	500
Input resistance			0.3	2.0	
Input capacitance				1.4	
Offset voltage adjustment range				±15	
Input voltage range			±12	±13	
Common mode rejection ratio		$R_S < 10\,k\Omega$	70	90	
Supply voltage rejection ratio		$R_S < 10\,k\Omega$		30	150
Large signal voltage rain		$R_L > 2\,k\Omega$, V_{OUT} = ±10 V	20 000	200 000	
Output voltage swing		$R_L > 10\,k\Omega$	±12	±14	
		$R_L > 2\,k\Omega$	±10	±13	
Output resistance				75	
Output short-circuit current				25	
Supply current				1.7	2.8
Power consumption				50	85
Transient response (unity gain)	Risetime	V_{IN} = 20 mV, R_L = 2 kΩ, C_L < 100 pF		0.3	
	Overshoot			5.0	
Slew rate		$R_L > 2\,k\Omega$		0.5	
The following specifications apply for 0°C < T_A < +70°C:					
Input offset voltage					7.5
Input offset current					300
Input bias current					800
Large signal voltage gain		$R_L > 2\,k\Omega$, V_{OUT} = ±10 V	15 000		
Output voltage swing		$R_L > 2\,k\Omega$	±10	±13	

Figure 3.1 μA741 specification

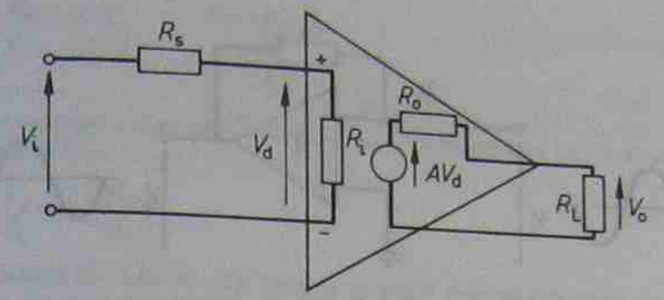


Figure 3.2 Basic voltage amplification circuit

Analysis of the circuit yields several equations:

$$V_d = \frac{R_i}{R_i + R_s} V_i \tag{3.1}$$

$$V_o = \frac{R_L}{R_o + R_L} A V_d \tag{3.2}$$

or

$$\frac{V_o}{V_i} = \frac{A R_i R_L}{(R_i + R_s)(R_o + R_L)} = \frac{A}{\left(1 + \frac{R_s}{R_i}\right)\left(1 + \frac{R_o}{R_L}\right)}$$

Normally $R_i \gg R_s$, $R_L \gg R_o$ and we obtain

$$\frac{V_o}{V_i} \approx A \tag{3.3}$$

This latter result shows that under conditions approaching ideal, the circuit behaves as an ideal *voltage-controlled-voltage-source* (VCVS). We conclude, therefore, that an ideal op-amp has the following: infinite input impedance, zero output impedance, infinite open-loop gain and infinite bandwidth.

3.3 Two important op-amp configurations

We now consider two basic circuits which will be used in subsequent chapters.

3.3.1 The inverting circuit

The basic connection is shown in figure 3.3 where the input is connected via resistor R_1 to the inverting terminal, the non-inverting terminal being connected to earth. Feedback is applied via resistor R_2.

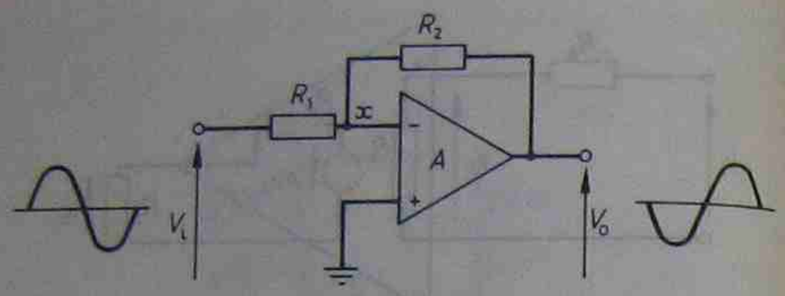


Figure 3.3 Inverting amplifier

Nodal analysis at point x yields

$$\frac{V_x - V_i}{R_1} + \frac{V_x - V_o}{R_2} = 0 \quad \text{also} \quad V_o = -AV_x$$

Eliminating V_x we obtain after a little manipulation, the expression

$$\frac{V_o}{V_i} = -\frac{A R_2}{(1+A)\left(R_1 + \frac{R_2}{1+A}\right)} \tag{3.4}$$

Normally the open-loop gain A is large so that the closed-loop gain becomes to a good approximation

$$\frac{V_o}{V_i} = -\frac{R_2}{R_1} \tag{3.5}$$

The negative sign implies inversion of the input signal (a 180° phase shift between input and output).

3.3.2 *The non-inverting circuit*

In this configuration, the signal is applied to the non-inverting terminal and the inverting terminal is connected to earth via resistor R_1. Feedback is again supplied through resistor R_2 as shown in figure 3.4.

Applying nodal analysis at point x yields

$$\frac{V_x}{R_1} + \frac{V_x - V_o}{R_2} = 0 \quad \text{also} \quad A(V_i - V_x) = V_o \text{ or}$$

$$V_x = V_i - \frac{V_o}{A}$$

Eliminating V_x we obtain after manipulation

$$\frac{V_o}{V_i} = \frac{A}{(1+A)} \times \frac{R_1 + R_2}{R_1 + \frac{R_2}{1+A}} \tag{3.6}$$

If we assume a large open-loop gain A, then the expression for the closed-loop gain reduces to

$$\frac{V_o}{V_i} = 1 + \frac{R_2}{R_1} \tag{3.7}$$

Equation (3.7) shows that there is no phase difference between input and output. We find that the inverting circuit has a gain with feedback, fixed by the ratio of two resistors, so that in theory, closed-loop gains less than unity may be achieved, although we shall see that this condition has little practical significance for filter circuits.

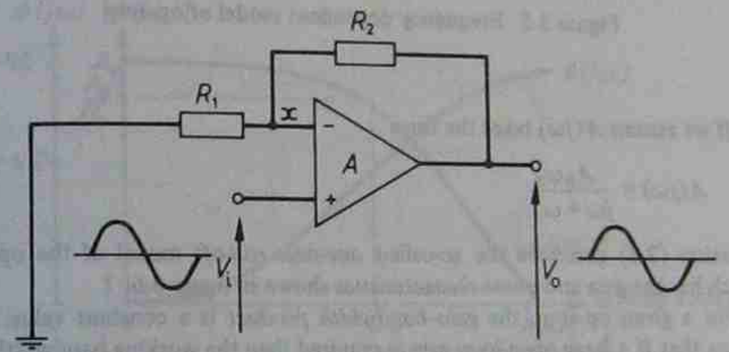


Figure 3.4 Non-inverting amplifier

The open-loop gain may be assumed to be high and the output impedance low. The input impedance, however, is fixed by the value of the resistor R_1 and consequently care must be exercised for some applications.

The non-inverting circuit has a minimum gain of unity when feedback is applied. The open-loop gain is also high and the circuit exhibits a low output impedance. Unlike the inverting circuit, the non-inverting configuration produces a high input impedance, AR_i (notice that now node x is not at virtual earth). An interesting modification to the circuit may be made by making R_2 a short-circuit and R_1 an open-circuit; the gain becomes unity $[V_o/V_i = 1 + (0/\infty)]$ and the circuit is then described as a *unity gain buffer* or *voltage follower* circuit.

3.4 The non-ideal op-amp

It is possible to construct op-amps to give a good approximation to the ideal within prescribed frequency limits. However, the gain of the op-amp is a function of the frequency of the applied signal and the actual behaviour can only approach the ideal over a specified range of frequency. Consider the simple op-amp model shown in figure 3.5.

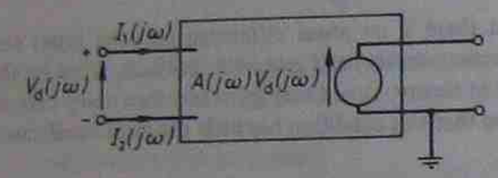


Figure 3.5 Frequency dependent model of op-amp

If we assume $A(j\omega)$ takes the form

$$A(j\omega) = \frac{A_0\omega_a}{j\omega + \omega_a} \tag{3.8}$$

equation (3.8) describes the so-called *one-pole roll-off* model of the op-amp which has the gain and phase characteristics shown in figure 3.6.

For a given op-amp, the *gain-bandwidth product* is a constant value. This means that if a large open-loop gain is required then the working bandwidth will be narrow. Conversely, if the open-loop gain requirement is to be low then a much wider bandwidth will be obtained.

For example:

$A_0 = 10^5, \omega_a = 2\pi \times 10$

$GB = 2\pi \times 10^6$

For $A_0 = 10^3$, bandwidth $= \dfrac{2\pi \times 10^6}{10^3} = 2\pi \times 10^3$ rad/s

For $A_0 = 10$, bandwidth $= \dfrac{2\pi \times 10^6}{10} = 2\pi \times 10^5$ rad/s

The reader is directed to problem 3.4 where a different view of the bandwidth of an inverting circuit is proposed.

We may conclude that the gain-bandwidth product is an important item of data when selecting an op-amp for a particular design:

$$GB = A_0\omega_a \tag{3.9}$$

As we have seen, typical values would be $A_0 = 10^5$, $\omega_a = 2\pi \times 10$ rad/s, which when inserted into the magnitude expression obtained from equation (3.8) yields

$$|A(jA_0\omega_a)| = \frac{A_0\omega_a}{((A_0\omega_a)^2 + \omega_a^2)^{0.5}} \approx 1,\ \omega = A_0\omega_a$$

At very low frequencies ($\omega \to 0$) the gain becomes

$$|A(j0)| = \frac{A_0\omega_a}{\omega_a} = A_0$$

and for the values of A_0, ω_a quoted, it is seen that the gain is

$$|A(j0)| = 20\log_{10}10^5 = 100 \text{ dB}$$

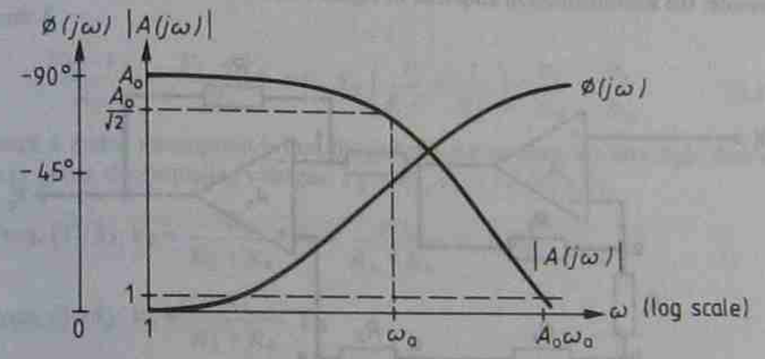


Figure 3.6 Response of non-ideal op-amp

We see that the gain is 100 dB down when the frequency is equal to the gain-bandwidth product. Note also that there is an additional 90° phase shift through the op-amp over the working frequency range.

It should finally be mentioned here that the quality of the output signal depends on the op-amp's ability to cope with the rate at which the output signal varies with time. This capability is referred to as the *slew rate* of the op-amp and is the maximum rate-of-change of the output voltage which the device can handle.

$$\text{slew rate} = \left.\frac{dV_o}{dt}\right|_{max} \tag{3.10}$$

Selection of op-amps with superior GB and SR is left to the designer, who must choose between cost and suitability of a device for his particular require-

ment. It should be noted that non-ideal op-amps have finite, (usually) high values of input resistance and non-zero values of output resistance. Other quantities of interest are: common-mode rejection, input offset voltages and currents etc.; this latter condition is of importance to the filter designer and will be briefly investigated here.

In the case of an ideal op-amp, the output voltage will be zero if both input voltages are zero. For a real op-amp however, there may be an output voltage even if both input voltages are zero. This output *offset* voltage is obtained by grounding each input terminal and measuring the output voltage. As an example, consider $A_o = 10^4$ and that a V_o of 5 V is measured; then

$$V_{offset} = \frac{5}{10^4} = 0.5 \text{ mV}$$

In order to illustrate the important analytical method of nodal analysis, consider the instrumentation amplifier of figure 3.7.

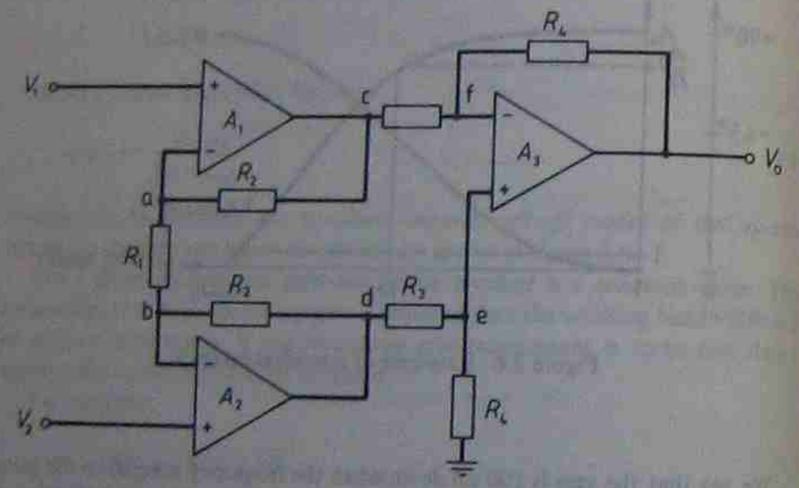


Figure 3.7 Instrumentation amplifier circuit

Example 3.1

The circuit shown in figure 3.7 is that of an instrumentation amplifier. Obtain the relationship between the output and the two inputs. Assume that the op-amps used have high open-loop gains and input impedance values.

Using nodal analysis at the labelled nodes, we obtain the following equations:

node a

$$\frac{V_a - V_c}{R_2} + \frac{V_a - V_b}{R_1} = 0, V_a\left(\frac{1}{R_1} + \frac{1}{R_2}\right) = \frac{V_c}{R_2} + \frac{V_b}{R_1} \quad (3.11)$$

node b

$$\frac{V_b - V_d}{R_2} + \frac{V_b - V_a}{R_1} = 0, \; V_b\left(\frac{1}{R_1} + \frac{1}{R_2}\right) = \frac{V_a}{R_1} + \frac{V_d}{R_2} \quad (3.12)$$

node c

$$\frac{V_c - V_f}{R_3} + \frac{V_c - V_a}{R_2} = 0, \; V_c\left(\frac{1}{R_2} + \frac{1}{R_3}\right) = \frac{V_a}{R_2} + \frac{V_f}{R_3} \quad (3.13)$$

node e

$$\frac{V_e}{R_4} + \frac{V_e - V_d}{R_3} = 0, \; V_e\left(\frac{1}{R_3} + \frac{1}{R_4}\right) = \frac{V_d}{R_3} \quad (3.14)$$

node f

$$\frac{V_f - V_o}{R_4} + \frac{V_f - V_c}{R_3} = 0, \; V_f\left(\frac{1}{R_3} + \frac{1}{R_4}\right) = \frac{V_o}{R_4} + \frac{V_c}{R_3} \quad (3.15)$$

Since a major assumption is that the gains of the op-amps are very high, then we may quote the following voltages: $V_b = V_2$, $V_a = V_1$, $V_e = V_f$.

From (3.15): $V_f = \dfrac{R_3}{R_3 + R_4} V_o + \dfrac{R_4}{R_3 + R_4} V_c = V_e$

From (3.14): $V_e = \dfrac{R_4}{R_3 + R_4} V_d$

Equating these two expressions yields:

$$V_d = \frac{R_3}{R_4} V_o + V_c \quad (3.16)$$

From (3.12): $V_c = \dfrac{R_1 + R_2}{R_1} V_2 - \dfrac{R_2}{R_1} V_1 - \dfrac{R_3}{R_4} V_o$ using (3.16)

From (3.11): $V_c = \dfrac{R_1 + R_2}{R_1} V_1 - \dfrac{R_2}{R_1} V_2$

Equating these two expressions finally yields:

$$V_o = \left(1 + \frac{2R_2}{R_1}\right)\frac{R_4}{R_3}(V_2 - V_1) \quad (3.17)$$

The resistor R_1 is usually made variable in order to adjust the gain. A typical instrumentation op-amp would have $A_o = 2 \times 10^5$, $R_i = 10^{12}$ ohms.

This circuit is an example of a differential-in and differential-out (A_1, A_2 amplifier with negative feedback and equalised amplification.

Problems

3.1. For the circuit shown in figure 3.8 it is required that $V_o = \frac{V_2}{3} - 2V_1$.

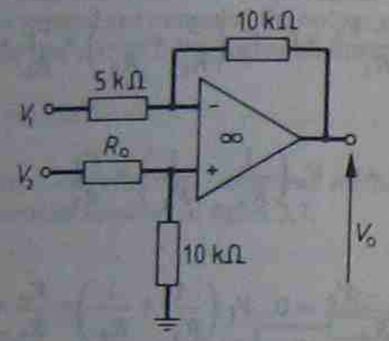


Figure 3.8

(a) Determine the value of R_a necessary to give the stated relationship
(b) If $V_2 = 10$ V and $V_1 = -10$ V, calculate the current in each resistor
[*Ans.* 80 kΩ, 1.78 mA, 1.22 mA, 0.11 mA]

3.2. Obtain the expression $\frac{V_o}{V_i} = \frac{ab}{1 + b\left\{\frac{(1+a)}{(1+c)}\right\}}$ for the circuit shown in figure 3.9.

3.3. Derive the expression $\frac{V_o}{V_i} = \frac{G_3(G_1 + G_2)}{G_2 G_4\left\{\frac{G_3 + G_5}{G_4} - \frac{G_1}{G_2}\right\}}$ for the circuit shown in figure 3.10 where $G_1 = \frac{1}{R_1}$ etc.

If now all of the components, with the exception of G_4, are fixed resistors of equal value, sketch the variation of V_o/V_i with variation of G_4 assuming that it is a pure resistor.

3.4. The circuit shown in figure 3.11 is that of a summing amplifier with inverting gain. The op-amp has an open-loop gain given by $A(j\omega) = A_o\omega_a/j\omega + \omega_a$. Derive the expression

$$V_o = -\frac{A_o\omega_a(V_1 + V_2)}{(1+2k)\left(j\omega + \omega_a + \frac{A_o\omega_a}{1+2k}\right)}$$

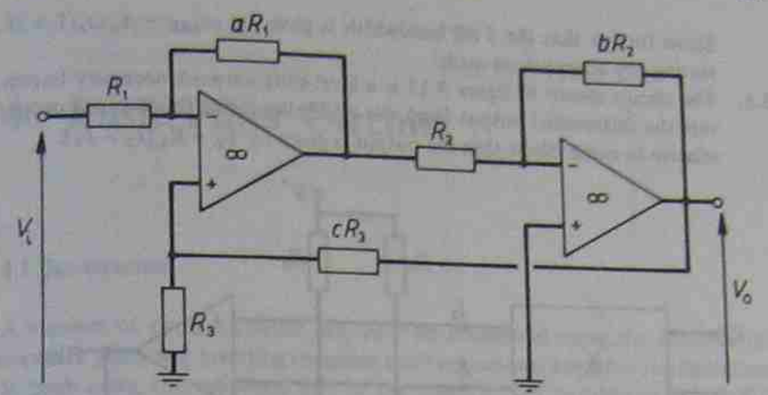


Figure 3.9

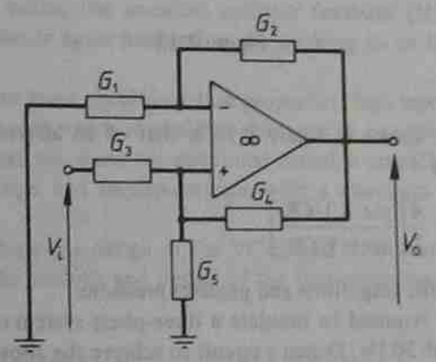


Figure 3.10

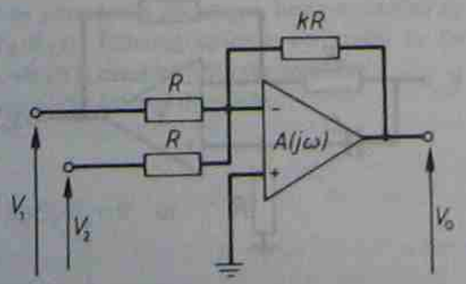


Figure 3.11

Show further that the 3 dB bandwidth is given by $\omega_{3dB} = A_0\omega_a/1 + 2$ stating any assumptions made.

3.5. The circuit shown in figure 3.12 is a level-shift network necessary to convert the differential output from the multiplier into a single-ended output relative to earth. Show that the output is given by $V_o = R_L(I_2 - I_1)$.

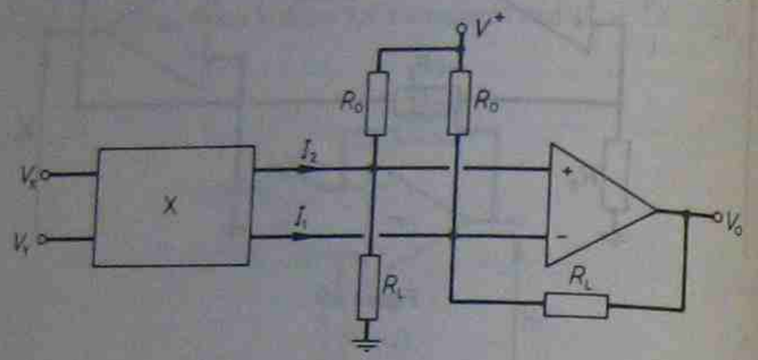


Figure 3.12

3.6. The circuit shown in figure 3.13 is that of an all-pass filter. Derive the relationship

$$\frac{V_o}{V_i} = \frac{A}{2}\left[\frac{j\omega - 1/CR}{j\omega + 1/CR}\right]$$

and obtain the magnitude and phase expressions.

It is now required to simulate a three-phase system of voltages having a frequency of 50 Hz. Design a circuit to achieve the above using the all-pass design assuming ideal op-amps and 1 μF capacitors.

[*Ans.* Two cascaded circuits: $R = 1.84\ k\Omega$, $C = 1\ \mu F$]

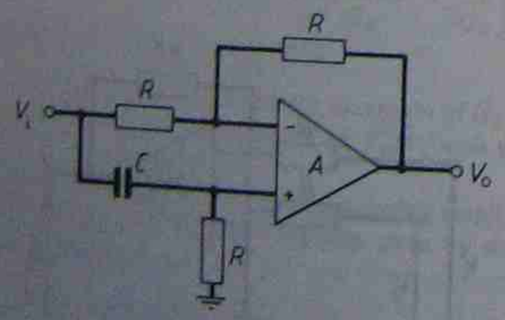


Figure 3.13

4 Basic Filter Circuits

4.1 Introduction

A number of popular circuits will now be considered using the non-inverting (positive gain) and inverting (negative gain) operational amplifier configurations. In both cases, the open-loop gain of the amplifier will be taken to be high (10^5 or greater) and in the first class of circuits to be considered, the so-called voltage-controlled-voltage-source (VCVS), it will be seen that the circuit gain may be adjusted by means of a resistor ratio. In the second class of circuits where there are two feedback paths, the so-called multiple feedback (MFB) circuits, the operational amplifier is again assumed to be working as an ideal *infinite* gain device.

Both circuits have good circuit isolation properties (high input impedance and low output impedance) which means that they may be cascaded to form higher-order filters without the need for additional isolation amplifiers. Both circuits allow for easy design and implementation with a minimum number of components.

Before undertaking the design of the VCVS and MFB second-order circuits, we will consider the analysis and design of the first-order low-pass VCVS filter.

4.2 The voltage-controlled-voltage-source first-order filter

The simplest active filter employing the non-inverting (positive gain) feedback amplifier is shown in figure 4.1. The amplifier is assumed to be a VCVS having a very high open-loop gain A, the circuit gain being controlled by resistors R_3 and R_2 $[K = 1 + (R_3/R_2)]$. Filtering action is performed by the R-C frequency selective network which is cascaded with the amplifier.

Nodal analysis at nodes (a) and (b) yields:

Node (a)

$$\frac{V_a - V_i}{R_1} + j\omega C_1 V_a = 0 \quad \text{or}$$

$$V_a\left(\frac{1}{R_1} + j\omega C_1\right) = V_i/R_1 \tag{4.1}$$

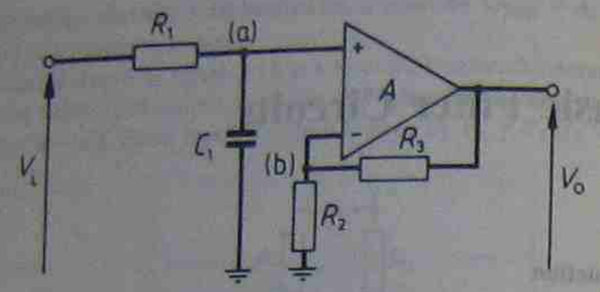


Figure 4.1 First-order low-pass VCVS filter

Node (b)

$$\frac{V_b - V_o}{R_3} + \frac{V_b}{R_2} = 0 \quad \text{or}$$

$$V_b\left(\frac{1}{R_2} + \frac{1}{R_3}\right) = \frac{V_o}{R_3} \text{ giving}$$

$$V_b = \left(\frac{R_2}{R_2 + R_3}\right) V_o \tag{4.2}$$

and $V_o = A(V_a - V_b)$, $(V_a - V_b) = \frac{V_o}{A} \to 0$ for A large.

Therefore $V_a \approx V_b$ and combining equations (4.1) and (4.2) gives

$$\frac{V_o}{V_i}(j\omega) = \frac{K(1/C_1R_1)}{j\omega + 1/C_1R_1} \tag{4.3}$$

where $K = 1 + R_3/R_2$.

This may be compared with the form for a low-pass circuit:

$$\frac{V_o}{V_i}(j\omega) = \frac{Kb_0\omega_0}{j\omega + b_0\omega_0} \tag{4.4}$$

Comparing equations (4.3) and (4.4) provides $b_0\omega_0 = 1/C_1R_1$, $K = 1 + R_3/R_2$. We see that we have two equations and four unknowns. We therefore select the capacitor value (we have a more restricted choice in capacitors than with resistors).

Example 4.1

A circuit is required to have a voltage gain of 4 in the pass band, to be 3 dB down up to the cut-off frequency of 100 Hz and to be at least 20 dB down at 1 kHz. The design is to be Butterworth using a VCVS circuit.

The performance requirement is shown in figure 4.2.

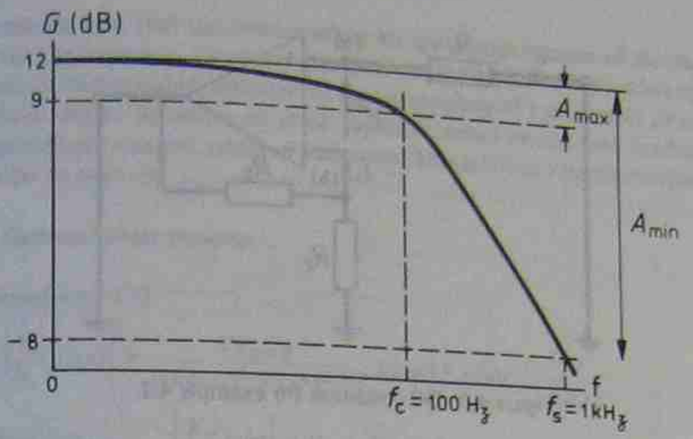


Figure 4.2 Response for example 4.1

To find the order n of the circuit we use equation (2.8):

$$n = \frac{1}{2\log_{10}\left(\frac{\omega_s}{\omega_c}\right)} \log_{10}\left\{\frac{10^{A\,\text{min}/10} - 1}{10^{A\,\text{max}/10} - 1}\right\}$$

$$n = \frac{1}{2\log_{10}10} \log_{10}\left\{\frac{10^2 - 1}{10^{0.3} - 1}\right\}$$

$$= 0.5\log_{10}\frac{99}{0.995} = 0.999 \quad \text{or} \quad n = 1$$

Practical experience shows that capacitor values, appropriate to the frequency range under consideration, may be selected by using the following expression:

$$C_1 = \frac{10}{f_c}\,\mu\text{F} = \frac{10}{100} \times 10^{-6} = 0.1\,\mu\text{F}$$

For a Butterworth response, $b_0 = 1$, therefore

$$R_1 = \frac{1}{b_0\omega_0 C_1} = \frac{1}{2\pi \times 100 \times 10^{-7}} = 15.9\,\text{k}\Omega$$

To work out resistors R_2 and R_3, good circuit design suggests that the condition for minimum dc offset should be met. Re-draw the circuit for the dc condition as shown in figure 4.3.

For $V_a = V_b$, $V' = 0$, then $V_{R_1} = V(R_2//R_3)$ or

$$R_1 = \frac{R_2R_3}{R_2 + R_3} = \frac{R_3}{1 + R_3/R_2} = \frac{R_3}{K}$$

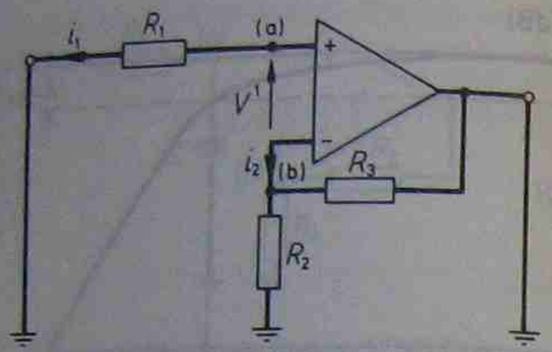


Figure 4.3 DC condition for example 4.1

giving $R_3 = KR_1$.

$$\frac{R_3}{R_2} = K - 1, \text{ hence } R_2 = \left(\frac{K}{K-1}\right) R_1$$

Summary of design equations

$$C_1 = \frac{10}{f_c}\ \mu\text{F}, \quad R_2 = \left(\frac{K}{K-1}\right) R_1$$

$$R_1 = \frac{1}{b_0 \omega_0 C_1} \quad R_3 = KR_1$$

To complete the design we find that

$R_2 = 21.2\ \text{k}\Omega$

$R_3 = 63.7\ \text{k}\Omega$

The circuit and response are shown in figure 4.4.

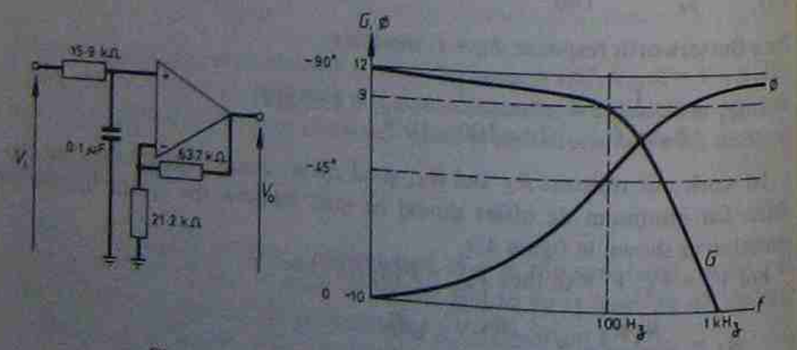


Figure 4.4 Circuit and response for example 4.1

It can be seen that the resistor values are quite large because of the chosen value for the capacitor. Of course the choice of the capacitor value is left to the designer; in the example, selection of capacitor values of 1 μF or 0.47 μF could have been made. Selection of these capacitor values would have resulted in correspondingly reduced values of resistance. This is left as a simple exercise for the reader to perform.

4.2.1 Gain and phase response

From equation (4.4):

$$\left|\frac{V_o}{V_i}(j\omega)\right| = \frac{Kb_0\omega_0}{((b_0\omega_0)^2 + \omega^2)^{0.5}}; \quad \phi(j\omega) = -\tan^{-1}\frac{\omega}{b_0\omega_0}$$

For $\omega = \omega_0$ $\left|\frac{V_o}{V_i}(j0)\right| = K$; $\phi(j0) = 0°$

For $\omega = b_0\omega_0$ $\left|\frac{V_o}{V_i}(jb_0\omega_0)\right| = \frac{K}{b_0}$; $\phi(jb_0\omega_0) = -45°$

For $\omega \to \infty$ $\left|\frac{V_o}{V_i}(j\infty)\right| = 0$; $\phi(j\infty) = -90°$

For $\omega \gg \omega_0$, roll-off (dB) $= 20\log_{10}\left|\frac{V_o}{V_i}(j\omega)\right| = -20\log_{10}\omega$

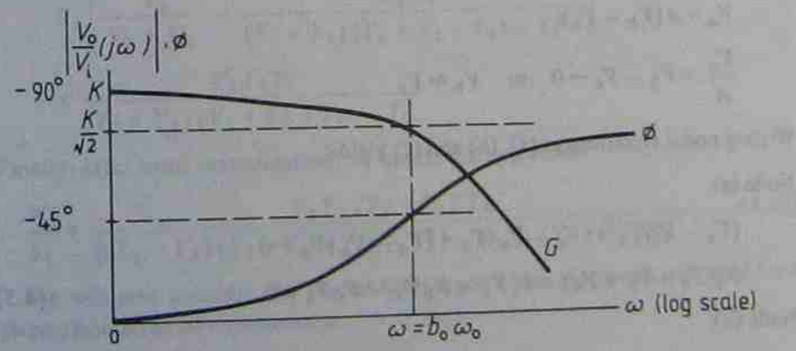


Figure 4.5 Gain/phase responses for first-order VCVS filter

4.3 The voltage-controlled-voltage-source second-order filter

This type of circuit was first proposed by Sallen and Key and is relatively easy to design and implement. It employs the infinite gain VCVS concept associated

with the operational amplifier and the circuit offers external gain adjustment via components Y_5 and Y_6. It will be seen in chapter 8 that this type of filter structure, employing the non-inverting (positive gain) operational amplifier, exhibits a high Q-factor sensitivity with the circuit gain $[K = 1 + (R_6/R_5)]$. Because of this high sensitivity, the configuration is restricted to low Q and low gain values with a trade-off being a wide frequency bandwidth.

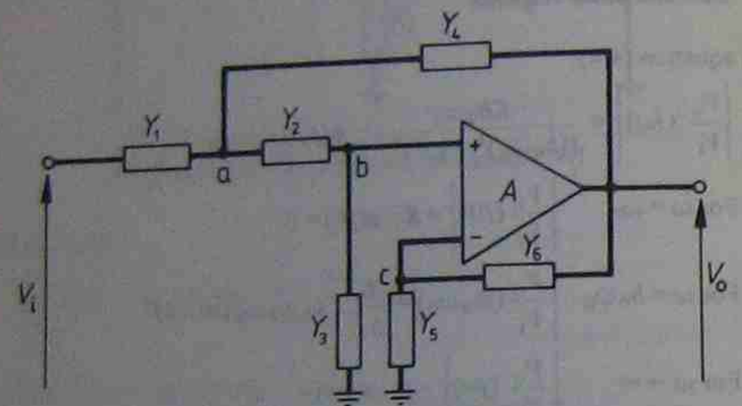


Figure 4.6 Basic second-order VCVS filter

A is assumed very large which implies that $V_b \approx V_c$.

$$V_o = A(V_b - V_c)$$

$$\frac{V_o}{A} = V_b - V_c \to 0 \quad \text{or} \quad V_b \approx V_c$$

Writing nodal equations at (a), (b) and (c) yields:

Node (a)

$$(V_a - V_i)Y_1 + (V_a - V_b)Y_2 + (V_a - V_o)Y_4 = 0$$

$$V_a(Y_1 + Y_2 + Y_4) - V_iY_1 - V_bY_2 - V_oY_4 = 0 \qquad (4.5)$$

Node (b)

$$V_bY_3 + (V_b - V_a)Y_2 = 0$$

$$V_b(Y_3 + Y_2) - V_aY_2 = 0 \qquad (4.6)$$

Node (c)

$$V_cY_5 + (V_c - V_o)Y_6 = 0$$

$$V_c(Y_5 + Y_6) - V_oY_6 = 0 \qquad (4.7)$$

From which

$$V_c = \frac{Y_6}{Y_5 + Y_6} V_o \qquad (4.8)$$

$$V_b = \frac{Y_2}{Y_2 + Y_3} V_a \qquad (4.9)$$

$$V_a = \frac{Y_1}{Y_1 + Y_2 + Y_4} V_i + \frac{Y_2}{Y_1 + Y_2 + Y_4} V_b + \frac{Y_4}{Y_1 + Y_2 + Y_4} V_o$$

Substituting this into equation (4.9) yields

$$V_b = \left(\frac{Y_2}{Y_2 + Y_3}\right) \frac{1}{Y_1 + Y_2 + Y_4} [Y_1V_i + Y_2V_b + Y_4V_o] \qquad (4.10)$$

from which

$$(Y_1 + Y_2 + Y_4)\left(\frac{Y_2 + Y_3}{Y_2}\right) V_b - Y_2V_b = Y_1V_i + Y_4V_o$$

that is

$$V_b[(Y_2 + Y_3)(Y_1 + Y_2 + Y_4) - Y_2^2] = Y_1Y_2V_i + Y_2Y_4V_o \qquad (4.11)$$

Equating equations (4.8) and (4.11) yields

$$\frac{Y_6}{Y_5 + Y_6} V_o = \frac{Y_1Y_2V_i + Y_2Y_4V_o}{(Y_2 + Y_3)(Y_1 + Y_2 + Y_4) - Y_2^2} \text{ and}$$

$$V_o\left[\frac{Y_6}{Y_5 + Y_6} - \frac{Y_2Y_4}{(Y_2 + Y_3)(Y_1 + Y_2 + Y_4) - Y_2^2}\right]$$

$$= \frac{Y_1Y_2V_i}{(Y_2 + Y_3)(Y_1 + Y_2 + Y_4) - Y_2^2}$$

Finally, after some manipulation, we obtain the expression

$$\frac{V_o}{V_i} = \frac{Y_1Y_2(Y_5 + Y_6)/Y_6}{[(Y_2 + Y_3)(Y_1 + Y_2 + Y_4) - Y_2^2] - Y_2Y_4(Y_5 + Y_6)/Y_6} \qquad (4.12)$$

We will now consider the several filter types which the circuit will yield for a given association of components.

4.3.1 The low-pass filter

Here $Y_1 = \frac{1}{R_1}$, $Y_2 = \frac{1}{R_2}$, $Y_3 = j\omega C_1$, $Y_4 = j\omega C_2$, $Y_5 = \frac{1}{R_3}$, $Y_6 = \frac{1}{R_4}$.

Equation (4.12) yields the following expression upon insertion of the components:

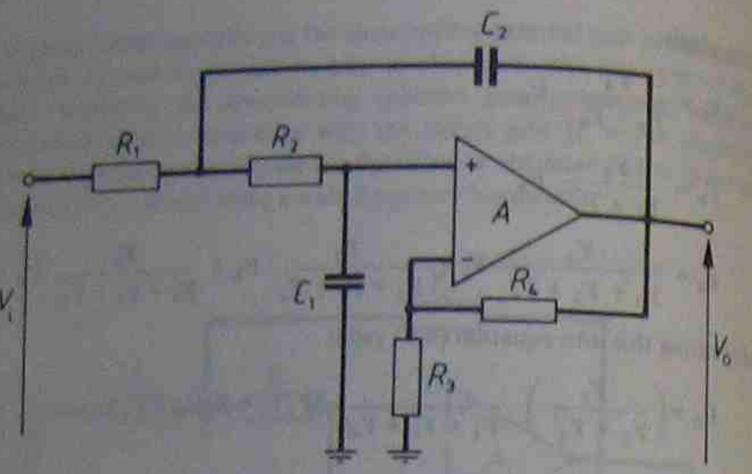


Figure 4.7 Second-order low-pass VCVS filter

$$\frac{V_o}{V_i}(j\omega) = \frac{\frac{1}{R_1R_2}\left(\frac{1}{R_3}+\frac{1}{R_4}\right)\Big/\frac{1}{R_4}}{\left(\frac{1}{R_2}+j\omega C_1\right)\left(\frac{1}{R_1}+\frac{1}{R_2}+j\omega C_2\right)-\frac{1}{R_2^2}-\frac{j\omega C_2}{R_2}\left(\frac{1}{R_3}+\frac{1}{R_4}\right)\Big/\frac{1}{R_4}}$$

Manipulation of the expression yields

$$\frac{V_o}{V_i}(j\omega) = \frac{\frac{1}{R_1R_2}\left(1+\frac{R_4}{R_3}\right)}{-\omega^2C_1C_2+\frac{j\omega C_1}{R_1}+\frac{j\omega C_1}{R_2}+\frac{j\omega C_2}{R_2}-\frac{j\omega C_2}{R_2}\left(1+\frac{R_4}{R_3}\right)+\frac{1}{R_1R_2}}$$

Now $K = 1 + R_4/R_3$ which is the dc gain of the stage and the expression may finally be written as

$$\frac{V_o}{V_i}(j\omega) = \frac{\frac{1}{C_1C_2R_1R_2}K}{-\omega^2+j\omega\left\{\frac{1}{C_2R_1}+\frac{1}{C_2R_2}+\frac{1}{C_1R_2}-\frac{K}{C_1R_2}\right\}+\frac{1}{C_1C_2R_1R_2}} \tag{4.13}$$

If we note further that $\omega_0^2 = \dfrac{1}{C_1C_2R_1R_2}$ then we may compare equation (4.13) with the standard form for a second-order low-pass filter:

$$\frac{V_o}{V_i}(j\omega) = \frac{Kb_0}{-\omega^2+jb_1\omega\omega_0+b_0\omega_0^2} \tag{4.14}$$

where ω_0 is the 3 dB frequency for the Butterworth case and $b_0\omega_0$ is the frequency at the end of the ripple band for the Chebyshev filter. The coefficients b_0, b_1 etc. are the appropriate coefficients for the prescribed filter response.

Comparison between equations (4.13) and (4.14) yields

$$\left.\begin{aligned} b_0\omega_0^2 &= \frac{1}{C_1C_2R_1R_2} \\ b_1\omega_0 &= \frac{1}{C_2R_1}+\frac{1}{C_2R_2}+\frac{1}{C_1R_2}+\frac{K}{C_1R_2} \end{aligned}\right\} \tag{4.15}$$

We must now introduce normalisation into the components and design parameters in order to produce a rationalised design procedure. Let $\omega_0^2 = 1$ rad/s, $C_1 = C_2 = 1$ F. Then equation (4.15) becomes

$$\left.\begin{aligned} b_0 &= \frac{1}{R_1R_2} \\ b_1 &= \frac{1}{R_1}+\frac{1}{R_2}(2-K) \end{aligned}\right\} \tag{4.16}$$

Notice that there are six unknowns and therefore a certain degree of component selectivity will have to be made since a unique solution is not possible.

From equation (4.16), $R_1 = 1/b_0R_2$ and on substituting into the second equation (4.16) we obtain

$$b_1 = b_0R_2+\frac{(2-K)}{R_2}$$

which may be written

$$b_0R_2^2 - b_1R_2 + (2-K) = 0$$

and which has a solution

$$R_2 = (b_1 \pm \sqrt{(b_1^2 - 4b_0(2-K))})/2b_0 \tag{4.17}$$

So far we have normalised the capacitors C_1, C_2 and obtained the normalised values for the resistors R_1, R_2. We now have to obtain values for the gain resistors R_3, R_4 for the good circuit design condition of minimum dc offset.

Figure 4.8 shows the re-drawn circuit for the dc condition where the capacitors are now represented by open circuits.

For minimum dc offset, $V_+ = V_-$ or

$$V(R_1+R_2) = V(R_3//R_4)$$

hence

$$R_1+R_2 = \frac{R_3R_4}{R_3+R_4} = \frac{R_4/R_3}{1+R_4/R_3}$$

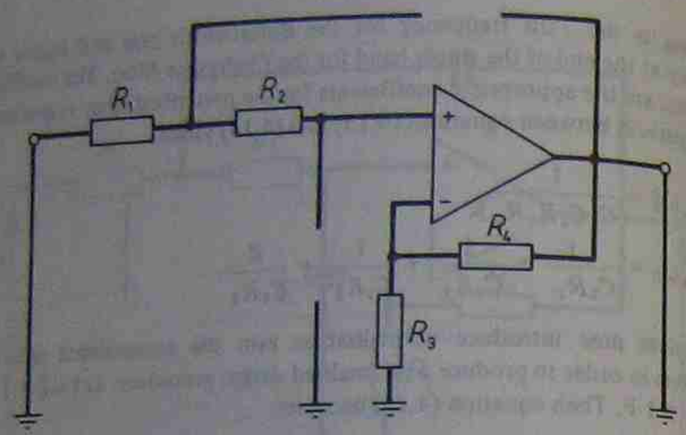


Figure 4.8 DC condition for second-order low-pass filter

giving

$$\frac{R_4}{R_3} = K - 1 = K(R_1 + R_2)$$

hence

$$R_3 = \left(\frac{K}{K-1}\right)(R_1 + R_2);\ R_4 = K(R_1 + R_2)$$

We now have all the necessary information to enable a complete design to be achieved.

Example 4.2

A low-pass filter is required to provide the specifications given in figure 4.9. The design is to be a Butterworth maximally flat realisation having a pass-band gain of 20 dB.

The requirement is for the output to be attenuated by 3 dB at 2 kHz and to have at least 40 dB of attenuation at 20 kHz.

Using equations from chapter 2 we obtain the order n of the filter and its performance in the pass and stop bands:

$$n = \frac{1}{2\log_{10}\left(\frac{\omega_s}{\omega_c}\right)} \log_{10}\left\{\frac{10^{A\,\mathrm{min}/10} - 1}{10^{A\,\mathrm{max}/10} - 1}\right\}$$

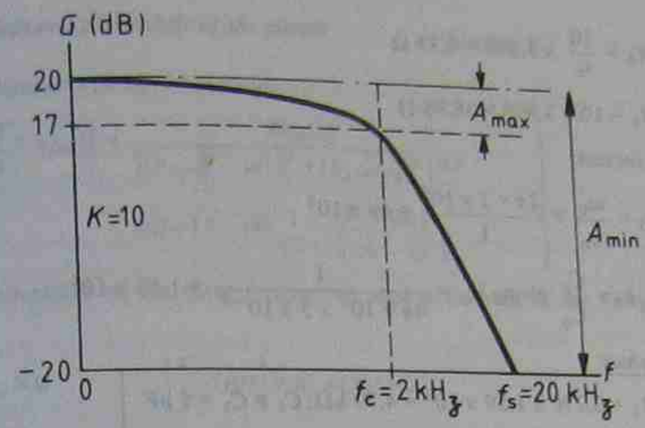


Figure 4.9 Specification for example 4.2

Insert values:

$$n = \frac{1}{2\log_{10}10} \log_{10}\left\{\frac{10^4 - 1}{10^{0.3} - 1}\right\}$$

producing $n = 2$ (a second-order circuit). At 2 kHz

$$A_{\max} = 10\log_{10}\left(1 + \left(\frac{\omega}{\omega_c}\right)^{2n}\right)$$

$$= 10\log_{10}(1 + 1^4)$$

$$= 3 \text{ dB for } \omega = \omega_c$$

$$A_{\min} = 10\log_{10}\left(1 + \left(\frac{20}{2}\right)^4\right) = 40 \text{ dB for } \omega = \omega_s$$

The filter will therefore perform exactly as required by the design.

We now choose values for the capacitors and a good choice is made using the formula $C = (10/f_c)\ \mu$F. Here $f_c = 2$ kHz giving $C_1 = C_2 = 10/(2 \times 10^3) = 5$ nF. The Butterworth coefficients for $n = 2$ are found from the table 2.1: $b_0 = 1$, $b_1 = 1.414$, hence

$$R_2 = \frac{1.414 \pm \sqrt{[2 - 4(2 - 10)]}}{2} = \frac{1.414 \pm 5.83}{2}$$

$= 3.622\ \Omega$ (the negative value is inadmissible) and

$$R_1 = \frac{1}{b_0 R_2} = 0.276\ \Omega$$

$$R_3 = \frac{10}{9} \times 3.898 = 4.33\ \Omega$$

$$R_4 = 10 \times 3.898 = 38.98\ \Omega$$

Scaling factors:

$$K_f = \frac{\omega_o}{\omega_n} = \frac{2\pi \times 2 \times 10^3}{1} = 4\pi \times 10^3$$

$$k_m k_f = \frac{C_n}{C_o} \text{ giving } k_m = \frac{1}{4\pi \times 10^3 \times 5 \times 10^{-9}} = 1.59 \times 10^4$$

Scaled values:

$R_1 = 0.276 \times 1.59 \times 10^3 = 4.39\ \text{k}\Omega$; $C_1 = C_2 = 5\ \text{nF}$
$R_2 = 3.622 \times 1.59 \times 10^3 = 57.5\ \text{k}\Omega$
$R_3 = 4.33 \times 1.59 \times 10^3 = 68.9\ \text{k}\Omega$
$R_4 = 38.98 \times 1.59 \times 10^3 = 620.3\ \text{k}\Omega$

The circuit is shown in figure 4.10 with its response.

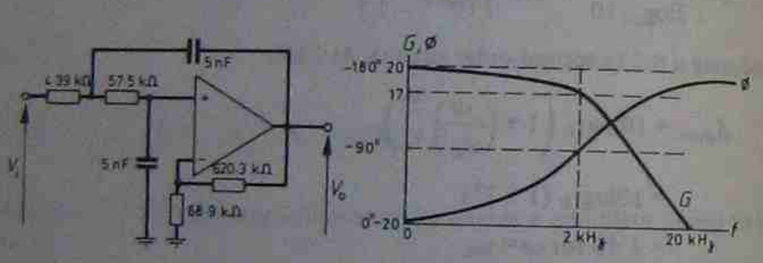


Figure 4.10 Final circuit for example 4.2

It should be pointed out that an adjustment to the scaling may be made in order that realistic values of resistors and capacitors may be achieved. Consider that we select $C_1 = C_2 = 0.1\ \mu\text{F}$, then

$$k_m = \frac{C_o}{k_f C_o} = \frac{1}{4\pi \times 10^3 \times 10^{-7}} = 795$$

The resistor values would be

$R_1 = 219.6\ \Omega$
$R_2 = 2.88\ \text{k}\Omega$
$R_3 = 3.45\ \text{k}\Omega$
$R_4 = 31\ \text{k}\Omega$
$C_1 = C_2 = 0.1\ \mu\text{F}$

It can be seen that with a larger value of capacitor, the resistor values are more realistic.

4.3.2 Gain and phase shift of the circuit

Re-call equation (4.14) from which

$$\left.\begin{aligned} \left|\frac{V_o}{V_i}(j\omega)\right| &= \frac{Kb_0\omega_0^2}{[(b_0\omega_0^2 - \omega^2)^2 + (b_1\omega\omega_0)^2]^{0.5}} \\ \phi(j\omega) &= -\tan^{-1}\frac{(b_1\omega\omega_0)}{(b_0\omega_0^2 - \omega^2)} \end{aligned}\right\} \quad (4.18)$$

A consideration of the three frequencies used in the case of the VCVS circuit yields

$\omega = 0$ $\quad \left|\frac{V_o}{V_i}(j0)\right| = K,\ \phi(j0) = 0°$

$\omega = \infty$ $\quad \left|\frac{V_o}{V_i}(j\infty)\right| \to 0,\ \phi(j\infty) \to -180°$

$\omega' = \sqrt{b_0}\omega_0$ $\quad \left|\frac{V_o}{V_i}(j\omega')\right| = \frac{K\sqrt{b_0}}{b_1},\ \phi(j\omega') = -90°$

These are shown in figure 4.11 for $b_0 = 1$, $b_1 = \sqrt{2}$

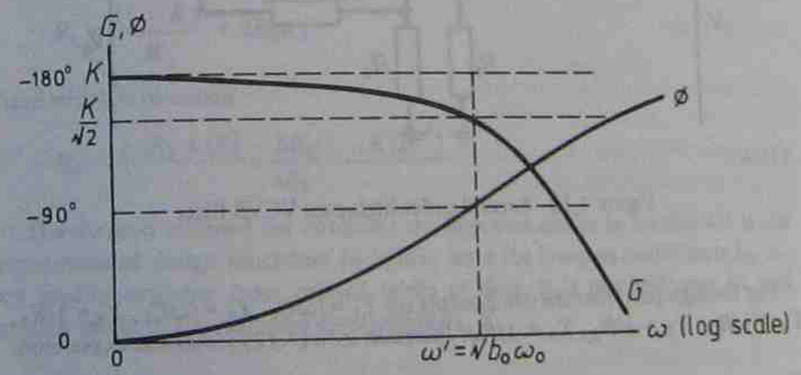


Figure 4.11 Gain/phase responses for second-order VCVS filter

An insight into the action of the filter may be obtained by noting that at low frequencies the reactances of C_1 and C_2 are high and the signal is applied via resistors R_1 and R_2 with negligible attenuation. At higher frequencies the reactances of C_1 and C_2 are lowered significantly. Capacitor C_1 shunts a proportion of the input signal to ground, thereby reducing the potential at node b, while C_2 feeds back a proportion of the output to node a with a corresponding rise in nodal potential. There are two roll-off paths, via (R_1 C_2) and (R_2 C_1) which combine to produce the second-order response.

4.4 The VCVS high-pass filter

This circuit is shown in figure 4.12 from which note that the frequency selective components (R_1, R_2, C_1, C_2) are interchanged with those of the low-pass filter shown in figure 4.7. Notice that the resistors *become* capacitors and capacitors *become* resistors. This concept is associated with what is referred to as *frequency transformation* and is useful in translating a high-pass filter specification into an equivalent low-pass specification. The low-pass design (prototype) is realised as in example 4.2, after which frequency transformation is applied to realise the prototype high-pass filter. Finally, frequency and magnitude scaling produces the final circuit design.

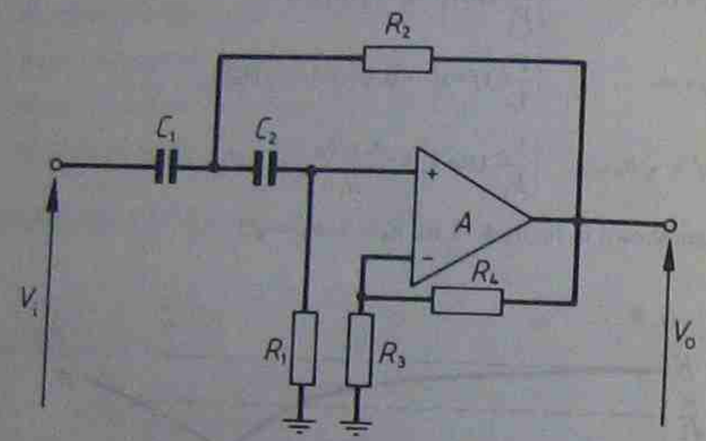


Figure 4.12 Second-order high-pass VCVS filter

For the high-pass filter the components are $Y_1 = j\omega C_1$, $Y_2 = j\omega C_2$, $Y_3 = 1/R_1$, $Y_4 = 1/R_2$, $Y_5 = 1/R_3$, $Y_6 = 1/R_4$. Insertion into (4.12) yields the expression:

$$\frac{V_o}{V_i}(j\omega) = \frac{-\omega^2 C_1 C_2 (1 + R_4/R_3)}{\left(j\omega C_2 + \frac{1}{R_1}\right)\left(j\omega C_1 + j\omega C_2 + \frac{1}{R_2}\right) + \omega^2 C_2^2 - \frac{j\omega C_2}{R_2}\left(1 + \frac{R_4}{R_3}\right)}$$

and

$$\frac{V_o}{V_i}(j\omega) = \frac{-\omega^2 C_1 C_2 K}{-\omega^2 C_1 C_2 - \omega^2 C_2^2 + \frac{j\omega C_2}{R_2} + \omega^2 C_2^2 + \frac{j\omega C_1}{R_1} + \frac{j\omega C_2}{R_1} + \frac{1}{R_1 R_2} - \frac{j\omega C_1 K}{R_1}}$$

which on simplification becomes

$$\frac{V_o}{V_i}(j\omega) = \frac{-\omega^2 K}{-\omega^2 + j\omega\left(\frac{1}{C_1 R_1} + \frac{1}{C_2 R_1} + \frac{(1-K)}{C_1 R_2}\right) + \frac{1}{C_1 C_2 R_1 R_2}} \tag{4.19}$$

Equation (4.19) should be compared with the standard form for a high-pass filter given by

$$\frac{V_o}{V_i}(j\omega) = \frac{-\omega^2 K}{-\omega^2 + j\omega\omega_0 B_1 + B_0\omega_0^2} \tag{4.20}$$

from which we obtain, noting that $B_0 = 1/b_0$, $B_1 = b_1/b_0$:

$$B_0\omega_0^2 = \frac{1}{C_1 C_2 R_1 R_2}$$

$$B_1\omega_0 = \frac{(1-K)}{C_1 R_2} + \frac{1}{C_1 R_1} + \frac{1}{C_2 R_1}$$

Again normalising ω_0, C_1, C_2 as before, $\omega_0 = 1$ rad/s, $C_1 = C_2 = 1$ F. Therefore

$$B_0 = \frac{1}{R_1 R_2} \quad \text{and} \quad B_0 R_2 = \frac{1}{R_1}$$

$$B_1 = \frac{(1-K)}{R_2} + 2B_0 R_2$$

from which is obtained

$$R_2 = \frac{[-B_1 \pm (B_1^2 - 8B_0(1-K))^{0.5}]}{4B_0} \tag{4.21}$$

The method outlined for obtaining the high-pass design in section 4.4 is the recommended design procedure to follow, since the low-pass coefficients b_0, b_1 are readily available from printed tables of data. It is relatively rare to find printed tables of the high-pass coefficients B_0, B_1.

Example 4.3

A high-pass filter is required to meet the specifications shown in figure 4.13. Obtain the component values necessary to satisfy the requirement.

The response will be Chebyshev having a ripple width of 3 dB and a high-frequency gain of 20 dB. The low-pass coefficients for $A_{max} = 3$ dB are from table 2.4: $b_1 = 0.65$, $b_0 = 0.708$. Using equation (2.15) to determine the order n of the filter and noting that for the identical 3 dB frequency the prototype low-pass filter will have the inverse of the ratio of frequencies as is the case for the high-pass filter:

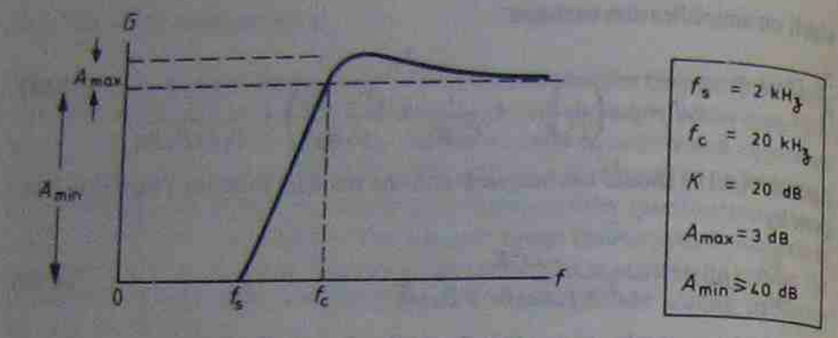


Figure 4.13 Response characteristic for example 4.3

$$\left.\frac{\omega_s}{\omega_c}\right|_{LP} = \left.\frac{\Omega_c}{\Omega_s}\right|_{HP} \tag{4.22}$$

Therefore

$$\frac{\omega_s}{\omega_c} = \frac{20}{2} = 10 \quad \text{and} \quad \omega_s = 10 \times \omega_c = 200\ \text{kHz}$$

also the order

$$n = \cosh^{-1}\frac{[(10^4 - 1)/(10^{0.3} - 1)]^{0.5}}{\cosh^{-1}10} = 1.77 \text{ (round up to 2)}$$

Figure 4.14 shows the high-pass characteristic frequency-translated into the equivalent low-pass specification.

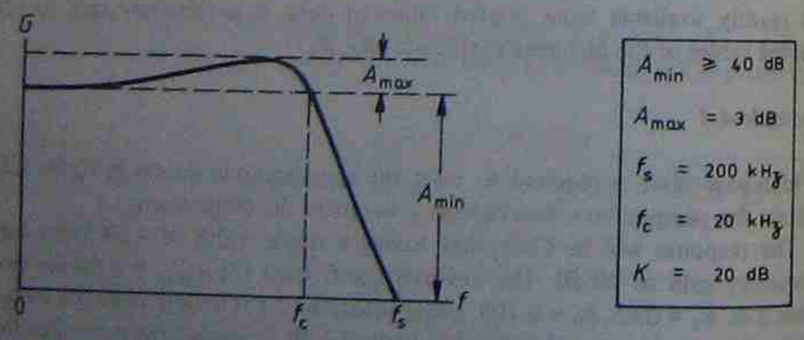


Figure 4.14 Low-pass response characteristic for example 4.3

The value for n suggests that the attenuation in the stop band will be greater than that for the identical order Butterworth filter. Using equation (2.14) and noting that from equation (2.6) $\epsilon = (10^{0.3} - 1)^{0.5} = 1$:

$$A_{min} = 10\log_{10}[1 + (\cosh 2\cosh^{-1}10)^2] = 45.9\ \text{dB}$$

The prototype *normalised* low-pass circuit may now be designed.
Let $C_1 = C_2 = 1$ F, $\omega_0 = 1$ rad/s.
Using equations (4.16) and (4.17)

$$R_1 = \frac{1}{b_0 R_2}, \quad R_2 = \frac{[b_1 \pm (b_1^2 - 4b_0(2 - K))^{0.5}]}{2b_0}$$

Also noting for $\epsilon = 1$, $A_{max} = 3$ dB and from table 2.4 $b_0 = 0.708$, $b_1 = 0.65$:

$$R_2 = \frac{0.65 \pm (0.65^2 - 4 \times 0.708\,(2 - 10))^{0.5}}{2 \times 0.708} = 3.85\ \Omega$$

$$R_1 = \frac{1}{0.708 \times 3.85} = 0.37\ \Omega$$

$$R_3 = \frac{10}{9 \times 4.22} = 4.7\ \Omega$$

$$R_4 = 10 \times 4.22 = 42.2\ \Omega$$

$$C_1 = C_2 = 1\ \text{F}$$

We must now apply the frequency transformation to the low-pass components in order to obtain the equivalent high-pass components. A brief outline of the concept is now included. We have seen that there is an inverse relationship between the ratio of the stop-band and pass-band frequencies for the low- and high-pass circuits. In general this may be written (for a normalised cut-off frequency)

$$\Omega\,(\text{HP}) = \frac{1}{\omega}\ (\text{LP})$$

which for the VCVS low-pass $\omega_0 = \dfrac{1}{(b_0 C_1 C_2 R_1 R_2)^{0.5}}$ and

$$\Omega_0 = (b_0 C_1 C_2 R_1 R_2)^{0.5} = \left(\frac{b_0}{R_1' R_2' C_1' C_2'}\right)^{0.5}$$

where $R_1' = \dfrac{1}{C_1}$

$$R_2' = \frac{1}{C_2}$$

$$C_1' = \frac{1}{R_1}$$

$$C_2' = \frac{1}{R_2}$$

where it can be seen that the capacitor (LP) becomes a resistor (HP) and a resistor (LP) becomes a capacitor (HP).

Therefore, for the prototype high-pass the normalised components may be calculated:

$$C_1 = \frac{1}{0.37} = 2.702\ \text{F},\ C_2 = \frac{1}{3.85} = 0.26\ \text{F}$$

$$R_1 = \frac{1}{1} = 1\ \Omega,\ R_2 = \frac{1}{1} = 1\ \Omega,\ R_3 = 4.7\ \Omega,\ R_4 = 42.4\ \Omega$$

The gain resistors are unaffected by the transformation. For scaling, select $C = 0.01\ \mu\text{F}$ which yields

$$k_f = 2\pi \times 2 \times 10^4 = 4\pi \times 10^4 \qquad k_m = \frac{10^8}{4\pi \times 10^4} = 796$$

The components are

$R_1 = R_2 = 796\ \Omega$ $R_3 = 3.74\ \text{k}\Omega$ $R_4 = 33.6\ \text{k}\Omega$

$C_1 = 2.702/k_m k_f = 0.027\ \mu\text{F}$

$C_2 = 0.26/k_m k_f = 0.0026\ \mu\text{F}$

The final circuit is shown in figure 4.15 with a CAD prediction of the response. Note that the program gives $\phi(j\omega)$ from $-180°$ to $-360°$.

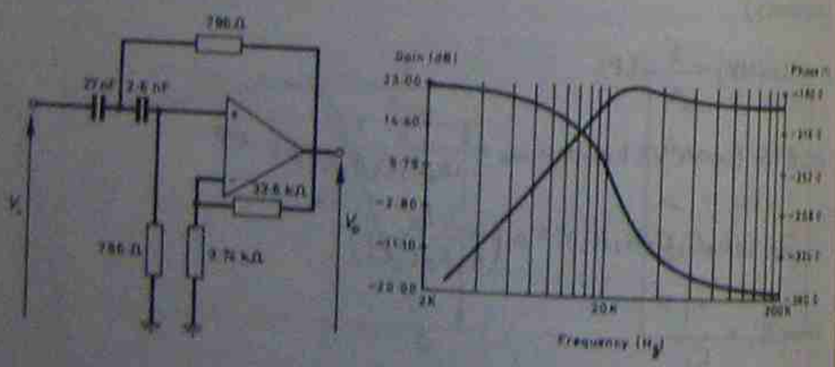


Figure 4.15 Circuit and gain/phase responses for VCVS high-pass filter

4.4.1 Gain and phase response

From equation (4.20):

$$\left|\frac{V_o}{V_i}(j\omega)\right| = \frac{\omega^2 K}{\{(b_0\omega_0^2 - \omega^2) + (\omega\omega_0 b_1)^2\}^{0.5}}$$

$$\phi(j\omega) = 180° - \tan^{-1}\frac{\omega\omega_0 b_1}{b_0\omega_0^2 - \omega^2}$$

Insertion of $\omega = 0, \omega = \sqrt{b_0}\omega_0, \omega = \infty$ yields

$$\left|\frac{V_o}{V_i}(j\omega)\right| = 0,\ \frac{\sqrt{b_0}K}{b_1},\ K;\ \phi(j\omega) = 180°, 90°, 0°$$

The predicted shape of the response is seen from figure 4.15.

Circuit operation is clearly the converse of that for the low-pass filter, as explained earlier. The reactances of C_1 and C_2 are very high at low frequencies, thereby preventing the signal from reaching the op-amp. At higher frequencies these reactances are very low and the signal may pass with negligible attenuation. Feedback increases via resistor R_3 and there are again two roll-off paths.

4.5 The VCVS band-pass filter

The basic structure of the circuit is that shown in figure 4.16 where the components are

$$Y_1 = \frac{1}{R_1},\ Y_2 = j\omega C_1,\ Y_3 = \frac{1}{R_3},\ Y_4 = \frac{1}{R_2},\ Y_5 = \frac{1}{R_5}$$

$$Y_6 = \frac{1}{R_4},\ Y_7 = j\omega C_2$$

Notice that a capacitor C_2 is connected between node a and ground.

Nodal analysis yields the three equations

$$V_a\left(\frac{1}{R_1} + \frac{1}{R_2} + j\omega(C_1 + C_2)\right) = \frac{V_i}{R_1} + j\omega C_1 V_b + \frac{V_o}{R_2}$$

$$V_b\left(\frac{1}{R_3} + j\omega C_1\right) = j\omega C_1 V_a \qquad (4.23)$$

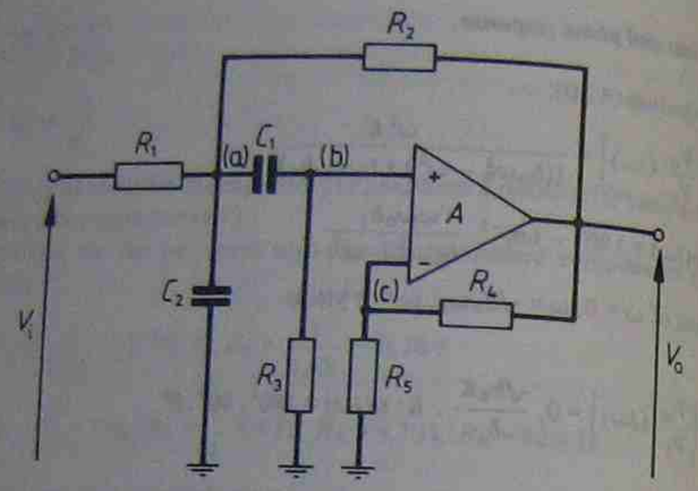


Figure 4.16 VCVS band-pass circuit

$V_c = \dfrac{V_o}{K}$, where

$$K = 1 + \frac{R_4}{R_5}$$

Also for A large, $V_b \approx V_c$ and

$$\frac{V_o}{V_i}(j\omega) = \frac{j\omega K/C_2R_1}{-\omega^2 + \dfrac{j\omega}{C_2}\left[\dfrac{1}{R_1} + \dfrac{1}{R_3} + \dfrac{1}{R_2}(1-K) + \dfrac{C_2}{C_1R_3}\right] + \dfrac{(R_1+R_2)}{C_1C_2R_1R_2R_3}} \tag{4.24}$$

The general form for a band-pass function having a positive gain is given by the expression

$$\frac{V_o}{V_i}(j\omega) = \frac{j\omega\omega_0 G/Q}{-\omega^2 + j\omega\omega_0/Q + \omega_0^2} \tag{4.25}$$

In the design of band-pass filters, three design parameters must be specified. They are centre frequency gain (G), centre frequency (f_0) and the Q-factor of the circuit (Q). The relationship between the Q-factor and the centre frequency also provides the 3 dB bandwidth of the circuit:

$$\Delta f = f_2 - f_1 = f_0/Q \tag{4.26}$$

These relationships are shown in figure 4.17.

Comparing equations (4.24) and (4.25) yields for normalised values of $C_1 = C_2 = 1$ F, $\omega_0 = 1$ rad/s:

$$G/Q = \frac{K}{R_1} = KG$$

$$1/Q = G_1 + G_2(1-K) + 2G_3$$

$$G_3 = \frac{1}{G_1 + G_2} \tag{4.27}$$

where $G_1 = \dfrac{1}{R_1}$ etc.

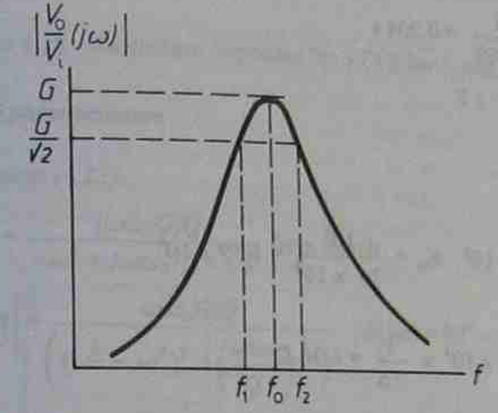


Figure 4.17 Gain response of a band-pass filter

At this point it is expedient to choose the condition that $K = 2$, which implies that $R_4 = R_5$. The reader may pursue conditions other than those for $K = 2$ but will find that the design equations become difficult to manipulate. The equations (4.27) now become

$$G_2 = \left\{\frac{-1 \pm ((G-1)^2 + 8Q^2)^{0.5}}{2Q}\right\}$$

$$G_1 = \frac{G}{2Q}$$

$$G_3 = \frac{1}{G_1 + G_2}$$

Example 4.4

A band-pass filter is required to have a centre frequency gain of 20 dB at 1 kHz and a Q-factor of 2. Obtain the design using 0.01 μF capacitors.

Using equation (4.26) the bandwidth is found to be $\Delta f = 1000/2 = 500$ Hz, giving $f_0 = 1$ kHz, $f_2 = 1.25$ kHz, $f_1 = 750$ Hz.
Choose $K = 2$; $R_4 = R_5$:

$$G_2 = \left\{ \frac{-1 \pm ((10-1)^2 + 8 \times 4)^{0.5}}{4} \right\} = 2.408 \text{ s}$$

$$G_1 = \frac{10}{4} = 2.5 \text{ s}$$

$$G_3 = \frac{1}{4.908} = 0.204 \text{ s}$$

$$C_1 = C_2 = 1 \text{ F}$$

Scaling

$$k_f = 2\pi \times 10^3, \; k_m = \frac{10^5}{2\pi \times 10^3} = 1.59 \times 10^4$$

$$R_1 = 1.59 \times 10^4 \times \frac{1}{2.5} = 6.36 \; \Omega$$

$$R_2 = 1.59 \times 10^4 \times \frac{1}{2.408} = 6.603 \text{ k}\Omega$$

$$R_3 = 1.59 \times 10^4 \times \frac{1}{0.204} = 77.9 \text{ k}\Omega$$

$$C_1 = C_2 = 0.01 \; \mu\text{F}$$

$$R_4 = KR_3 = 155.8 \text{ k}\Omega$$

$$R_5 = \left(\frac{K}{K-1} \right) R_3 = 155.8 \text{ k}\Omega$$

The circuit with its response is shown in figure 4.18.

The operation of the circuit may be summarised briefly as follows. Capacitors C_1 and C_2 have high reactance values at low frequencies and maximum attenuation of the input signal occurs. At high frequencies these reactances are very low and most of the input signal and the feedback signal through R_2 are by-passed to earth. Over a narrow range of frequencies around the centre frequency, frequency selection takes place, involving R_1, R_2, R_3, C_1 and C_2. Note that at low and high frequencies the circuit approximates to a low-pass and then a high-pass filter.

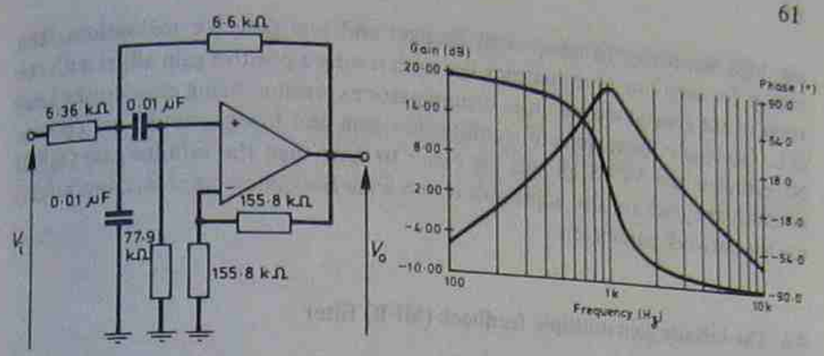


Figure 4.18 Gain/phase responses for VCVS band-pass filter

4.5.1 Gain and phase response

Re-stating equation (4.25):

$$\frac{V_o}{V_i}(j\omega) = \frac{j\omega\omega_0 G/Q}{-\omega^2 + j\omega\omega_0/Q + \omega_0^2} \quad \text{from which}$$

$$\left| \frac{V_o}{V_i}(j\omega) \right| = \frac{\omega\omega_0 G/Q}{\left((\omega_0^2 - \omega^2)^2 + \left(\frac{\omega\omega_0}{Q} \right)^2 \right)}, \; \phi(j\omega) = 90^\circ - \tan^{-1} \frac{\omega\omega_0}{\omega_0^2 - \omega^2}$$

Insertion of $\omega = 0$, $\omega = \omega_0$, $\omega \to \infty$ yields the response form shown in figure 4.19.

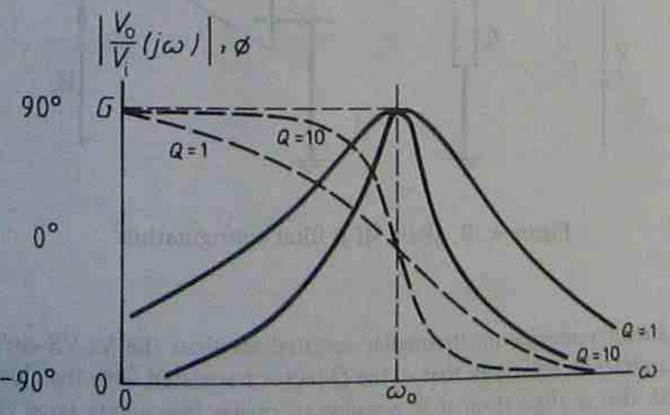


Figure 4.19 Response of VCVS band-pass filter for various values of Q

The general advantages of all the structures considered so far are that they are relatively easy to design and implement. The designer has control over the element values and the *spread* of these values. The disadvantages are those involv-

ing high sensitivity to component changes and low Q-factor realisations. The reason for such low Q realisations lies with having a positive gain allied with the roots of the general second-order denominator expression being close to the $(j\omega)$ axis. Obviously for reasons of stability, low-gain and low-Q circuits are chosen. Nevertheless the VCVS circuits are easier to tune than the infinite gain (MFB) counterparts, and are also adjustable over a wide frequency range without affecting the network parameters.

4.6 The infinite gain multiple feedback (MFB) filter

This circuit is an alternative to the previously considered VCVS circuit which had a relatively low closed-loop gain, insofar as the operational amplifier is deemed to be working as an *ideal* VCVS with an infinite gain. The overall circuit gain is of course finite and fixed by the circuit components. The expression *multiple feedback* refers to the two feedback paths through admittances Y_3 and Y_4 as can be seen in figure 4.20, from which it will be seen that the output has a 180° phase shift in addition to the phase shift through the network.

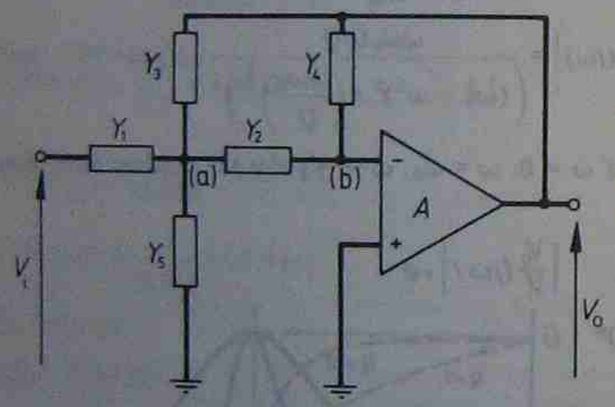


Figure 4.20 Basic MFB filter configuration

This circuit possesses much smaller sensitivities than the VCVS circuit previously considered, especially that of the Q-factor associated with the closed-loop gain. With this configuration it is possible to realise reasonably large values of Q-factor at low frequencies, although the trade-off is a narrower bandwidth and a much higher open-loop gain requirement. Although we are assuming that the gain is infinite when conducting the analysis, it is in reality far from being the case.

Assuming that $A \to \infty$ and $V_b \to 0$, nodal analysis at nodes (a) and (b) gives:

Node (a)

$$(V_a - V_i)Y_1 + V_aY_5 + (V_a - V_o)Y_3 + (V_a - 0)Y_2 = 0$$

$$V_a(Y_1 + Y_2 + Y_3 + Y_5) - V_iY_1 - V_oY_3 = 0 \qquad (4.28)$$

Node (b)

$$(0 - V_a)Y_2 + (0 - V_o)Y_4 = 0$$

$$V_a = -\frac{Y_4}{Y_2}V_o \qquad (4.29)$$

which produces, when inserted into equation (4.28)

$$\frac{V_o}{V_i} = -\frac{Y_1Y_2}{Y_4(Y_1 + Y_2 + Y_3 + Y_5) + Y_2Y_3} \qquad (4.30)$$

4.6.1 The multiple feedback low-pass filter

The circuit configuration is shown in figure 4.21 where $Y_1 = 1/R_1$, $Y_2 = 1/R_3$, $Y_3 = 1/R_2$, $Y_4 = j\omega C_2$, $Y_5 = j\omega C_1$.

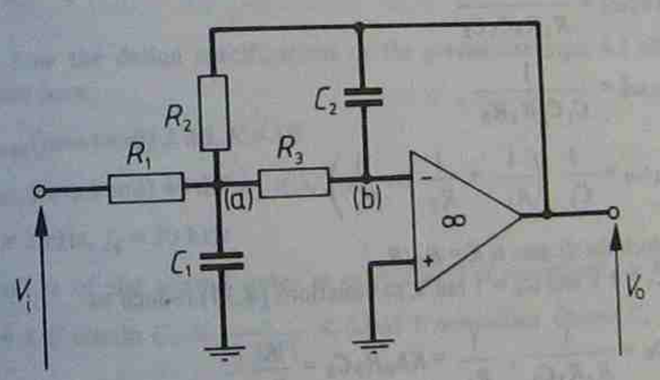


Figure 4.21 MFB low-pass filter

Insertion of admittances into equation (4.30) gives the expression

$$\frac{V_o}{V_i}(j\omega) = \frac{-\frac{1}{R_1R_3}}{j\omega C_2\left(\frac{1}{R_1} + \frac{1}{R_3} + \frac{1}{R_2} + j\omega C_1\right) + \frac{1}{R_1R_2}}$$

which after manipulation becomes

$$\frac{V_o}{V_i}(j\omega) = -\frac{\dfrac{1}{R_1R_3C_1C_2}}{-\omega^2 + \dfrac{j\omega}{C_1}\left(\dfrac{1}{R_1}+\dfrac{1}{R_2}+\dfrac{1}{R_3}\right) + \dfrac{1}{C_1C_2R_2R_3}} \tag{4.31}$$

This expression may be compared with that for a typical low-pass response, namely

$$\frac{V_o}{V_i}(j\omega) = \frac{-Kb_0\omega_0^2}{-\omega^2 + jb_1\omega\omega_0 + b_0\omega_0^2} \tag{4.32}$$

The gain and phase responses have the same shape as for the VCVS filter but it should be noted that the MFB filter provides an *extra* phase shift of 180° because of the input signal being applied to the inverting terminal. Therefore

$$\phi(j\omega) = 180^\circ - \tan^{-1}\frac{b_1\omega\omega_0}{b_0\omega_0^2 - \omega^2} \tag{4.33}$$

Comparing equation (4.31) and equation (4.32) yields

$$\begin{aligned} Kb_0\omega_0^2 &= \frac{1}{R_1R_3C_1C_2} \\ b_0\omega_0^2 &= \frac{1}{C_1C_2R_2R_3} \\ b_1\omega_0 &= \frac{1}{C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}\right) \end{aligned} \tag{4.34}$$

From which the dc gain is $K = R_2/R_1$.

Let $C_1 = 1$ F and $\omega_0 = 1$ rad/s, so equations (4.34) reduce to

$$\begin{aligned} Kb_0 &= \frac{1}{R_1R_3C_2}\cdot\frac{1}{R_1} = Kb_0R_3C_2 = \frac{K}{R_2} \\ b_0 &= \frac{1}{R_2R_3C_2}\cdot\frac{1}{R_3} = b_0R_2C_2 \\ b_1 &= \left(\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}\right) \end{aligned}$$

which on substituting for $1/R_1$ and $1/R_3$ gives

$$b_1 = \frac{K}{R_2} + \frac{1}{R_2} + b_0R_2C_2$$

which yields the quadratic expression

$$b_0R_2^2 = \frac{b_1R_2}{C_2} + \frac{(1+K)}{C_2} = 0 \tag{4.35}$$

which has a solution

$$R_2 = \frac{\dfrac{b_1}{C_2} \pm \left(\left(\dfrac{b_1}{C_2}\right)^2 - \dfrac{4(1+K)b_0}{C_2}\right)^{0.5}}{2b_0} \tag{4.36}$$

For R_2 to be real and positive, the following condition for C_2 must prevail:

$$\left(\frac{b_1}{C_2}\right)^2 \geqslant \frac{4(1+K)b_0}{C_2} \quad \text{or}$$

$$b_1^2 \geqslant 4(1+K)C_2b_0 \quad \text{and}$$

$$C_2 \leqslant \frac{b_1^2}{4(1+K)b_0} \tag{4.37}$$

Note here that there are five unknowns and three basic equations, and note again that a unique solution is not possible.

Example 4.5

Consider now the design specifications of the previous example 4.2 which are summarised here:

A_{max}(pass-band) 3 dB, $K = 10$

A_{min}(stop-band) 40 dB, $b_1 = \sqrt{2}$, $b_0 = 1$

$f_c = 2$ kHz, $f_s = 20$ kHz

The circuit is of the second order as calculated. For normalised $\omega_0 = 1$ rad/s and $C_1 = 1$ F obtain $C_2 \leqslant \dfrac{2}{4 \times 11} \leqslant 0.045$ F *normalised*. Choose $C_2 = 0.01$ F and

$$R_2 = \frac{\dfrac{\sqrt{2}}{0.01} \pm \left(\left(\dfrac{\sqrt{2}}{0.01}\right)^2 - \dfrac{4(11)}{0.01}\right)^{0.5}}{2}$$

This yields $R_2 = 133\ \Omega$ or $8.3\ \Omega$, giving

$$R_1 = \frac{R_2}{K} = 13.3\ \Omega \text{ or } 0.83\ \Omega$$

$$R_3 = \frac{1}{b_0R_2C_2} = 0.752\ \Omega \text{ or } 12.12\ \Omega$$

Select $C_1 = \dfrac{10}{f_c}\ \mu\text{F} = \dfrac{10}{2 \times 10^3} = 5\ \text{nF}.$

Scaling:

$$k_f = 2\pi \times 2 \times 10^3 = 4\pi \times 10^3$$

$$k_m = \frac{C_n}{k_f C_0} = \frac{1}{4\pi \times 10^3 \times 5 \times 10^{-9}} = 1.59 \times 10^4$$

Then

$R_1 = 2.117\ \text{k}\Omega$ or $13.2\ \text{k}\Omega$

$R_2 = 2.11\ \text{M}\Omega$ or $132\ \text{k}\Omega$

$R_3 = 11.96\ \text{k}\Omega$ or $192\ \text{k}\Omega$

$C_1 = 5\ \text{nF}$

$$C_2 = \frac{0.01}{4\pi \times 10^3 \times 1.59 \times 10^4} = 50\ \text{pF}$$

It can be seen here that the resistor values are very large, and although a MICROCAP or SPICE simulation may give acceptable results, in practice the circuit may fail to meet the specification.

We may effect a better design by changing the scaling factors. Choose $C_1 = 0.1\ \mu\text{F}$ then

$$k_m = \frac{1}{4\pi \times 10^3 \times 10^{-7}} = 796$$

and the new values are

$R_1 = 10.6\ \text{k}\Omega$ or $656\ \Omega$

$R_2 = 105\ \text{k}\Omega$ or $6.56\ \text{k}\Omega$

$R_3 = 597\ \Omega$ or $9.6\ \text{k}\Omega$

$C_1 = 0.1\ \mu\text{F}$

$C_2 = 1\ \text{nF}$

Note the 100:1 capacitor spread and 10:1 spread of resistor values.

The circuit with the response is shown in Figure 4.22.

4.6.2 Gain and phase shift of the circuit

From equation (4.32) we may write

$$\left|\frac{V_o}{V_i}(j\omega)\right| = \frac{Kb_0\omega_0^2}{[(b_0\omega_0^2 - \omega^2) + (b_1\omega\omega_0)^2]^{0.5}}$$

$$\phi(j\omega) = 180^\circ - \tan^{-1}\frac{b_1\omega\omega_0}{b_0\omega_0^2 - \omega^2}$$

By considering the three conditions $\omega = 0$, $\omega = \omega'$, $\omega = \infty$, a reasonable idea of the variation of gain and phase with variation of ω may be obtained and the response is shown in figure 4.23. Note that the MICROCAP simulation gives $\phi(j\omega)$ from -180° to -360°; this is because of the formulation of its programs and is consistent with the result shown in figure 4.23.

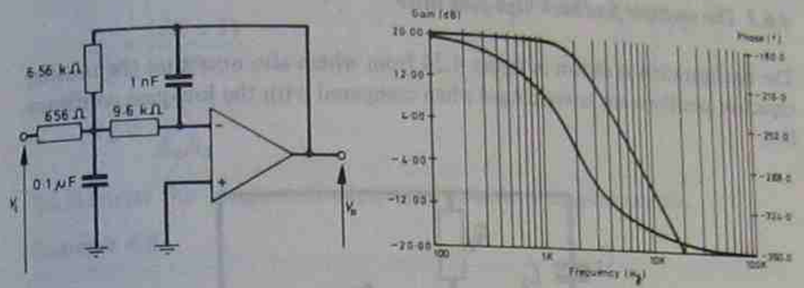


Figure 4.22

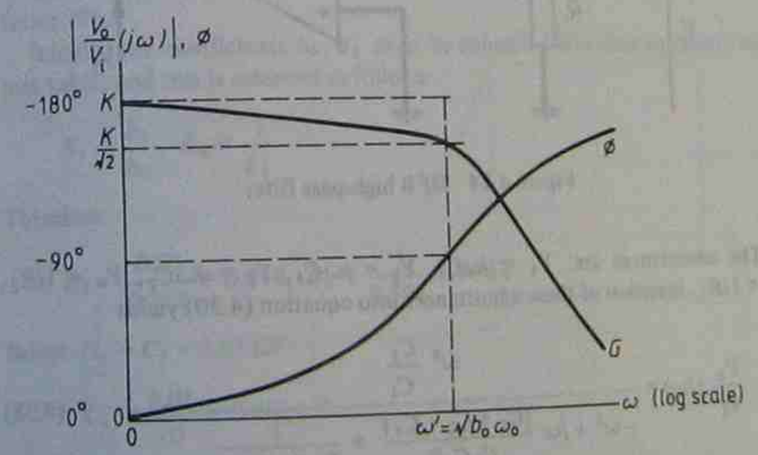


Figure 4.23 Gain/phase response for low-pass filter

$$\omega = 0 \quad \left|\frac{V_o}{V_i}(j0)\right| = K \text{ (the dc gain)};\ \phi(j0) = 180^\circ$$

$$\omega = \infty \quad \left|\frac{V_o}{V_i}(j\infty)\right| \to 0;\ \phi(j\infty) \to 0^\circ$$

$$\omega = \omega' \left| \frac{V_o}{V_i}(j\omega') \right| = K \Big/ \frac{b_0}{b_1} \; ; \; \phi(j\omega') = 90^\circ \quad \text{where } \omega' = \sqrt{b_0}\omega_0$$

Referring to figure 4.21, it can be seen that at low frequencies capacitors C_1 and C_2 present high reactance values and the signal passes with very little attenuation, the gain being given by $H(j\omega) = -R_2/R_1$. At high frequencies the reactances of C_1 and C_2 are low, the signal being by-passed to ground via C_1. There are again two roll-off circuits to give the prescribed second-order response.

4.6.3 *The multiple feedback high-pass filter*

The configuration is shown in figure 4.24 from which also note that the resistor/capacitor positions are interchanged when compared with the low-pass configuration.

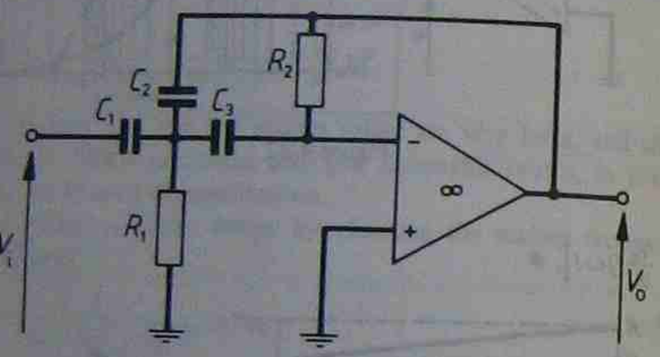


Figure 4.24 MFB high-pass filter

The admittances are: $Y_1 = j\omega C_1$, $Y_2 = j\omega C_3$, $Y_3 = j\omega C_2$, $Y_4 = 1/R_1$, $Y_5 = 1/R_1$. Insertion of these admittances into equation (4.30) yields

$$\frac{V_o}{V_i}(j\omega) = \frac{\omega^2 \dfrac{C_1}{C_2}}{-\omega^2 + j\omega \dfrac{(C_1 + C_2 + C_3)}{C_2 C_3 R_2} + \dfrac{1}{R_1 R_2 C_2 C_3}} \tag{4.38}$$

which may be compared with the standard form

$$\frac{V_o}{V_i}(j\omega) = \frac{K\omega^2}{-\omega^2 + jB_1\omega\omega_0 + B_0\omega_0^2} \tag{4.39}$$

From which we obtain the following relationships:

$$K = \frac{C_1}{C_2}$$

$$B_1\omega_0 = \frac{(C_1 + C_2 + C_3)}{C_2 C_3 R_2}$$

$$B_0\omega_0^2 = \frac{1}{R_1 R_2 C_2 C_3} \tag{4.40}$$

Consider normalised values of $C_1 = C_3$ and ω_0, then equations (4.40) reduce to

$$C_2 = \frac{1}{K}$$

$$R_2 = \frac{(2K + 1)}{B_1}$$

$$R_1 = \frac{K}{B_0 R_2}$$

To illustrate the design method, we will consider a simple example.

Example 4.6

Consider the previous specifications for the VCVS filter (example 4.4) which is summarised: $K = 10$, $A_{max} = 3$ dB, $A_{min} \geqslant 40$ dB, $f_c = 20$ kHz, $f_s = 2$ kHz.

The low-pass Chebyshev coefficients are: $b_1 = 0.65$, $b_0 = 0.708$, scaling factor 796.

Initially the coefficients b_0, b_1 must be converted into their equivalent high-pass values and this is achieved as follows:

$$B_1 = \frac{b_1}{b_0}, \; B_0 = \frac{1}{b_1}$$

Therefore

$$B_1 = \frac{0.65}{0.708} = 0.918, \; B_0 = \frac{1}{0.708} = 1.412$$

Select $C_1 = C_3 = 0.01\ \mu$F

$$C_2 = \frac{0.01}{10} = 0.001\ \mu\text{F}$$

$$R_2 = \frac{21}{0.918} = 22.88\ \Omega \qquad \text{normalised}$$

$$R_1 = \frac{10 \times 0.708}{22.88} = 0.309\ \Omega$$

Applying the scaling factor to the normalised resistor values produces $R_1 = 246\ \Omega$, $R_2 = 18.2$ kΩ. The circuit with its response is shown in figure 4.25.

An alternative approach is to convert the high-pass specification into the equivalent low-pass specification and then design the prototype low-pass. Finally,

GEORGE LIBRARY Technical College

application of the $CR:RC$ transformation produces the prototype high-pass. It is recommended that the reader adopts this approach when designing high-pass filter circuits.

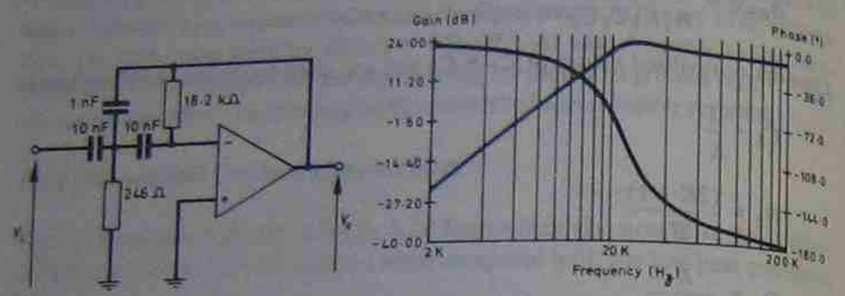
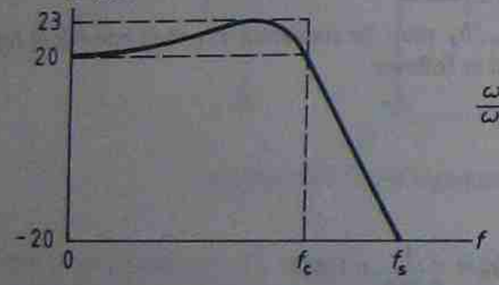


Figure 4.25 Circuit and response for MFB high-pass filter

Consider the specification 'translated' to the equivalent low-pass specification in figure 4.26.



$$\frac{\omega_s}{\omega_c}\bigg|_{LP} = \frac{\Omega_c}{\Omega_s}\bigg|_{HP} = \frac{200}{20} = 10$$

Figure 4.26 Specification for example 4.6

Using the low-pass design with $b_1 = 0.65$, $b_0 = 0.708$:

$$C_2 \leqslant \frac{0.65^2}{4 \times 11 \times 0.708} < 0.0135 \text{ F}$$

Select $C_2 = 0.01$ F:

$$R_2 = \frac{\dfrac{0.65}{0.01} \pm \sqrt{\left[\left(\dfrac{0.65}{0.01}\right)^2 - \dfrac{4 \times 11 \times 0.708}{0.01}\right]}}{2 \times 0.708} = 69.4\ \Omega \text{ or } 22.38\ \Omega$$

giving $R_1 = 6.94\ \Omega$ or $2.238\ \Omega$.

Consideration of component magnitudes suggests we select $R_1 = 2.238\ \Omega$, $R_2 = 22.38\ \Omega$ and choose $C = 0.01\ \mu$F.

Scaling factors are

$$k_f = 2\pi \times 2 \times 10^4 = 4\pi \times 10^4$$

$$k_m = \frac{1}{4\pi \times 10^4 \times 10^{-8}} = 796$$

$$R_3 = \frac{1}{b_0 R_2 C_2} = \frac{1}{0.708 \times 22.38 \times 0.01} = 6.31\ \Omega$$

We now convert these components into the equivalent high-pass components.

Noting that $C_{LP} \rightarrow R_{HP} = \dfrac{1}{C_{LP}}$ and

$$R_{LP} \rightarrow C_{HP} = \frac{1}{R_{LP}}$$

$$R_1 = \frac{1}{1} = 1\ \Omega,\ R_2 = \frac{1}{0.01} = 100\ \Omega$$

$$C_3 = \frac{1}{6.31} \text{ F}$$

$$C_1 = \frac{1}{2.24} \text{ F},\ C_2 = \frac{1}{22.4} \text{ F}$$

The scaled values are

$R_1 = 796\ \Omega$, $R_2 = 79.6\ \text{k}\Omega$, $C_3 = 1.58$ nF, $C_1 = 4.46$ nF, $C_2 = 446$ pF

and the final circuit with its response is shown in figure 4.27.

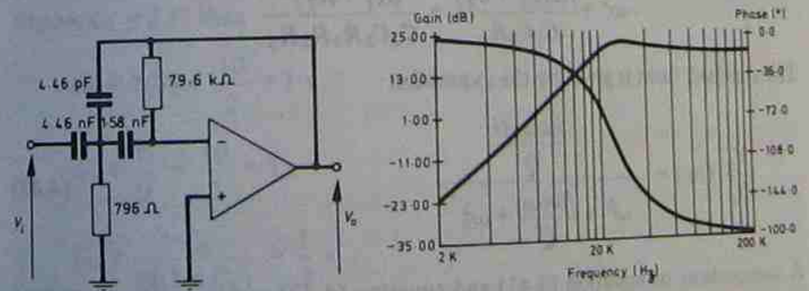


Figure 4.27 Circuit and responses for example 4.6

Note that the capacitor values are rather small and the spread between the two resistor values is 100:1 and between the capacitor values 3.5:1. The circuit

operation is the inverse of the low-pass filter. At low frequencies, capacitors C_1 and C_3 block the signal and at high frequencies allow the signal through with negligible signal loss.

4.6.4 The multiple feedback band-pass filter

The concept of an infinite gain VCVS circuit is particularly useful when applied to the MFB band-pass circuit in figure 4.28.

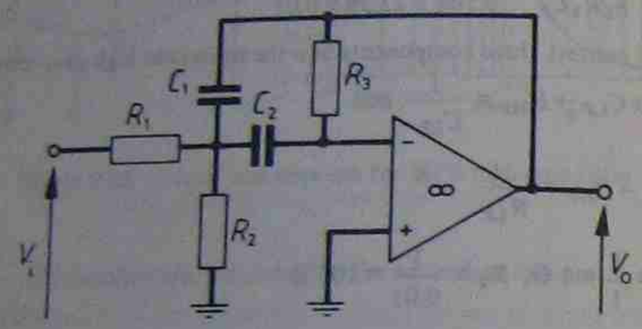


Figure 4.28 MFB band-pass filter

The conditions imposed on the admittances are: $Y_1 = 1/R_1$, $Y_2 = j\omega C_2$, $Y_3 = j\omega C_1$, $Y_4 = 1/R_3$, $Y_5 = 1/R_2$ which when inserted into equation (4.30) gives, after manipulation, the expression

$$\frac{V_o}{V_i}(j\omega) = \frac{-j\dfrac{\omega}{C_1R_1}}{-\omega^2 + \dfrac{j\omega(C_1+C_2)}{C_1C_2R_3} + \dfrac{(R_1+R_2)}{C_1C_2R_1R_2R_3}} \tag{4.41}$$

The standard form is given by the expression

$$\frac{V_o}{V_i}(j\omega) = \frac{-\dfrac{j\omega\omega_0 G}{Q}}{-\omega^2 + \dfrac{j\omega\omega_0}{Q} + \omega_0^2} \tag{4.42}$$

A comparison of equation (4.41) and equation (4.42) produces for normalised C_1 and ω_0:

$$Q = \frac{R_3C_2}{1+C_2}$$

$$\frac{1}{R_1} + \frac{1}{R_2} = C_2R_3 = Q(1+C_2) \tag{4.43}$$

$$\frac{1}{R_1} = \frac{G}{Q}$$

Noting that $G_1 = \dfrac{1}{R_1}$ etc., subsequent manipulation of equation (4.43) gives

$$G_2 = Q(1+C_2) - G_1 = Q(1+C_2) - \frac{G}{Q}$$

For G_2 to be real and positive, then conditions must be imposed in the inequality

$$C_2 \geqslant \frac{G}{Q^2} - 1 \tag{4.44}$$

Example 4.7

A band-pass filter is required to have a centre frequency gain of 20 dB, a bandwidth of 500 Hz and a centre frequency of 1 kHz. The design is to use readily available capacitors.

$$G = \text{Antilog}\frac{20}{20} = 10$$

$$Q = \frac{f_0}{\Delta f} = \frac{1000}{500} = 2$$

Select $C_1 = 1$ F, giving $C_2 \geqslant \dfrac{10}{4} - 1 \geqslant 1.5$ F.

Choose $C_2 = 2$ F, then

$$G_2 = 2(3) - \frac{10}{2} = 1 \text{ s}$$

$$G_1 = \frac{G}{Q} = \frac{10}{2} = 5 \text{ s}$$

$$G_3 = \frac{C_2}{Q(1+C_2)} = \frac{1}{3} \text{ s}$$

Scaling: $k_f = 2\pi \times 10^3$, $k_m = 1.59 \times 10^4$ for $C_1 = 0.01\ \mu$F.

$$R_1 = 1.59 \times 10^4 \times \frac{1}{5} = 3.14\ \text{k}\Omega$$

$R_2 = 1.59 \times 10^4 \times 1 = 15.9\ \text{k}\Omega$

$R_3 = 1.59 \times 10^4 \times 3 = 47.7\ \text{k}\Omega$

$C_2 = 0.02\ \mu\text{F}$

$C_1 = 0.01\ \mu\text{F}$

The circuit with its response is shown in figure 4.29.

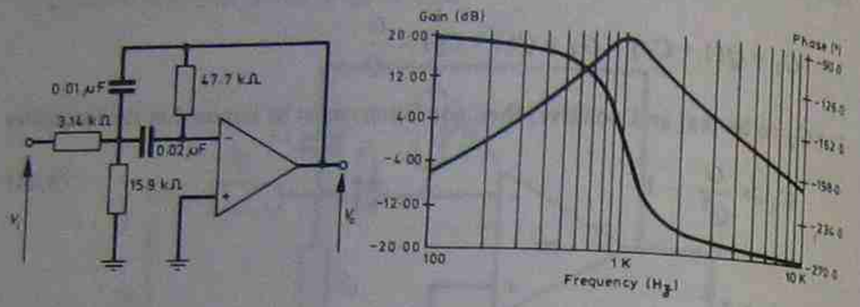


Figure 4.29 Circuit and responses for MFB band-pass filter

It is left as an exercise for the reader to show that the gain and phase response take the forms as for the VCVS band-pass circuit with the exception that the phase response extends from $-90°$ down to $-270°$. The filtering action is similar to the previously considered VCVS circuit but note that at low frequencies C_2 blocks the signal and at high frequencies C_1 is a short-circuit, producing a unity gain circuit.

The principal advantages of the MFB circuits over the VCVS is that they are less sensitive to component variations (see chapter 8) and have one component fewer within the design. A disadvantage is that they produce an additional phase shift of 180°.

4.7 Higher-order filter circuits

It was mentioned earlier in chapter 3 that a principal advantage of a VCVS circuit was the good circuit isolation properties it gave. These properties are usefully employed when cascading sections to produce higher-order filters, the basic principle being shown in figure 4.30.

Here

$H_1(j\omega) = |H_1(j\omega)|,\ \phi_1(j\omega)$ etc. from which

$$\frac{V_{o_3}}{V_i} = H(j\omega) = H_1(j\omega) \times H_2(j\omega) \times H_3(j\omega) \quad \text{where}$$

$$|H(j\omega)| = |H_1| \times |H_2| \times |H_3| \qquad (4.45)$$

$$\phi(j\omega) = \phi_1 + \phi_2 + \phi_3$$

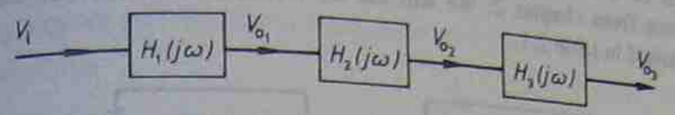


Figure 4.30 Cascaded filter sections

For example, a unity-gain third-order filter constructed from cascaded first-order and second-order circuits would provide $|H| = 1, \phi = -270°$.

If an *even*-order function is required, then there will be one or more even-order stages, whereas if the order is *odd*, one first-order stage in combination with one or more even-order stage(s) will be required. Although only active circuit realisations are to be considered, it should be noted that a combination of active and passive circuits may be employed. A third-order single amplifier filter circuit is shown in figure 4.31, which illustrates an alternative technique.

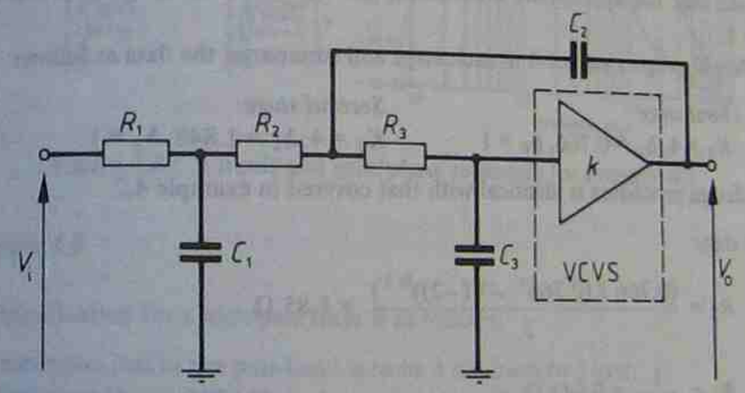


Figure 4.31 Third-order low-pass circuit

An advantage of the cascaded section technique is that each circuit (or active resonator) is tuned to the same frequency. A distinct disadvantage is their rather poor sensitivity response to small changes in circuit parameters and amplifier gain; this imposes a limit on the number of circuits which may be cascaded.

Example 4.8

A fourth-order, low-pass, Butterworth filter is required to have a gain of 16 and a cut-off frequency of 1 kHz. VCVS circuits are to be used with 0.01 μF capacitors.

We may represent the filter as two dissimilar sections as shown in figure 4.32. The operator s is used in order to correspond with tables in chapter 2, which is seen to be a fourth-order function whose roots may be obtained using the equations from chapter 2. We will use the results from that section which are summarised in table 2.1.

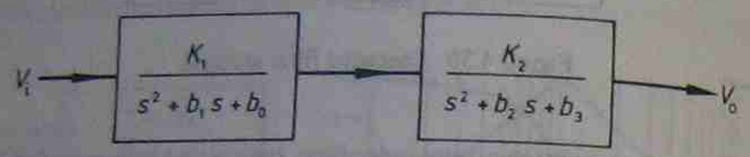


Figure 4.32 Block-diagram representation of example 4.7

Here

$$\frac{V_o}{V_i}(s) = \frac{K_1 K_2}{(s^2 + b_1 s + b_0)(s^2 + b_2 s + b_3)}$$

We see that the appropriate coefficients are $b_1 = 0.766$, $b_0 = 1$; $b_2 = 1.848$, $b_3 = 1$.

We will assign a gain of 4 to each stage and summarise the data as follows:

First stage	*Second stage*
$K_1 = 4, b_1 = 0.766, b_0 = 1$	$K_2 = 4, b_2 = 1.848, b_3 = 1$

The design procedure is identical with that covered in example 4.2.

First stage

$$R_2 = \frac{(0.766 \pm (0.766^2 - 4(-2))^{0.5})}{2} = 1.85\ \Omega$$

$$R_1 = \frac{1}{1.85} = 0.541\ \Omega$$

$$C_1 = C_2 = 1\ \text{F}$$

$$R_3 = \frac{4}{3} \times 2.39 = 3.187\ \Omega,\quad R_4 = 3 \times 3.187 = 9.56\ \Omega$$

Second stage

$$R_2 = \frac{(1.848 \pm (1.848^2 - 4(-2))^{0.5})}{2} = 2.613\ \Omega$$

$$R_1 = \frac{1}{2.613} = 0.383\ \Omega$$

$$R_3 = \frac{4}{3} \times 2.996 = 3.99\ \Omega$$

$$R_4 = 3 \times 3.99 = 11.97\ \Omega$$

$$C_1 = C_2 = 1\ \text{F}$$

Scaling:

$$k_f = 2\pi \times 10^3$$

$$k_m = \frac{10^8}{2\pi \times 10^3} = 1.59 \times 10^4$$

The final scaled circuit with its response is shown in figure 4.33. Note well from the design that the two sections which comprise the filter are not identical because of the coefficients (b_1, b_2) being unequal.

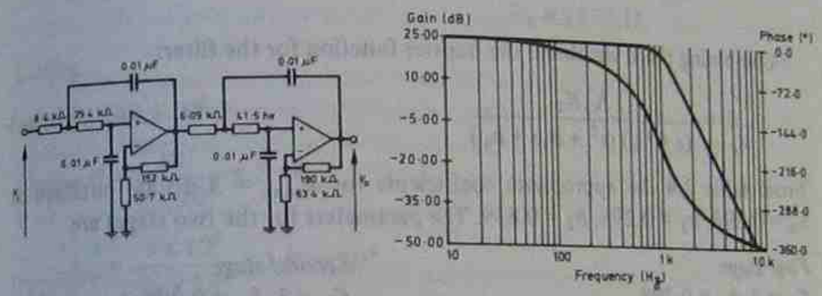


Figure 4.33 Circuit and gain/phase responses for example 4.8

Example 4.9

The specification for a high-pass filter is as follows:

(a) maximum loss in the pass-band is to be 3 dB down to 2 kHz;
(b) minimum loss at 250 kHz to be greater than 60 dB;
(c) the high-frequency gain is to be 20 dB.

The data suggests a Chebyshev response. From equation (2.15) the order n is obtained and note that $\Omega_c/\Omega_s = 8 = \omega_s/\omega_c$:

$$n = \frac{\cosh^{-1}\left[\frac{(10^6 - 1)}{(10^{0.3} - 1)}\right]^{0.5}}{\cosh^{-1} 8} = 2.74$$

Round up to 3.

Using equation (2.14):

$$\text{dB} = 10\log_{10}\,[1 + \{\cosh 3\cosh^{-1} 8\}^2] = 66\ \text{dB}$$

The design data are now translated into the equivalent low-pass and the steps followed as shown in example 4.3.

We note that $\Omega_c/\Omega_s = 8 = \omega_s/\omega_c$, and for $\omega_c = 1$ rad/s normalised $\omega_s = 8$ (16 kHz).

We shall require a first-order section and a second-order section cascaded as shown in figure 4.34.

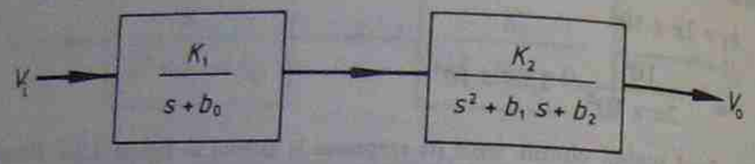


Figure 4.34 Cascaded low-pass equivalent for example 4.9

Combining these we obtain the transfer function for the filter:

$$\frac{V_o}{V_i} = \frac{K_1 K_2}{(s+b_0)(s^2+b_1 s+b_2)} \tag{4.46}$$

From table 2.4 the appropriate coefficients for $A_{max} = 3$ dB are obtained as $b_0 = 0.298, b_1 = 0.298, b_2 = 0.839$. The parameters for the two stages are

First stage
$K_1 = 2, b_0 = 0.298$

Second stage
$K_2 = 5, b_1 = 0.298, b_2 = 0.839$

First stage
This employs the VCVS circuit of example 4.1.

$$R_1 = \frac{1}{b_0} = \frac{1}{0.298} = 3.36\ \Omega$$

$$C = 1\ \text{F}$$

$$R_2 = R_3 = KR_1 = 6.72\ \Omega$$

Second stage
This uses the circuit of example 4.2.

$$R_2 = \frac{(0.298 \pm (0.298^2 + 4(0.839 \times 3))^{0.5})}{2 \times 0.839} = 2.07\ \Omega$$

$$R_1 = \frac{1}{2.07 \times 0.839} = 0.576\ \Omega$$

$$R_3 = \frac{5}{4} \times 2.646 = 3.31\ \Omega$$

$$R_4 = 5 \times 2.646 = 13.23\ \Omega$$

$$C_1 = C_2 = 1\ \text{F}$$

Use the $RC{:}CR$ transformation outlined in example 4.3 to obtain the low-pass to high-pass transformation of components as follows:

First stage
$R_1 = 1\ \Omega$

$$C = \frac{1}{3.36} = 0.298\ \Omega$$

$$R_2 = R_3 = 6.72\ \Omega$$

Second stage
$R_1 = R_2 = 1\ \Omega$

$$C_1 = \frac{1}{0.576} = 1.736\ \text{F}$$

$$C_2 = \frac{1}{2.07} = 0.483\ \text{F}$$

$$R_3 = 3.31\ \Omega$$

$$R_4 = 13.23\ \Omega$$

Scaling:

$$k_f = 4\pi \times 10^3$$

Choose $C = \frac{10}{2 \times 10^3} = 5\ \text{nF}$

$$k_m = \frac{5 \times 10^9}{4\pi \times 10^3} = 1.59 \times 10^4$$

The resistor values are multiplied by 1.59×10^4 and the capacitor values divided by 2×10^8.

The final circuit with its response is shown in figure 4.35.

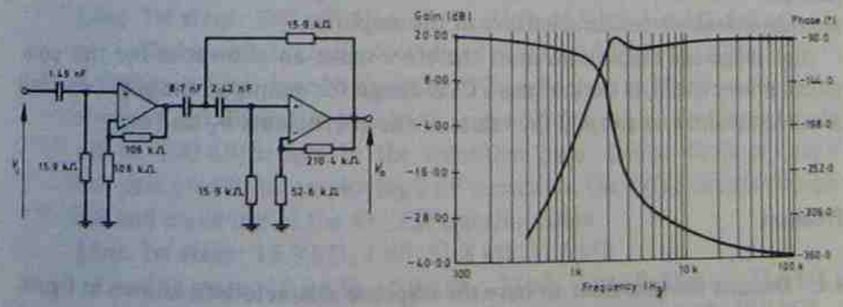


Figure 4.35 Circuit and responses for third-order high-pass filter

We noted earlier that cascaded filter sections possess poor sensitivities to component variations. Another design problem results in difficulties in the tuning of

these circuits because of the large number of components used to realise the particular design.

4.8 Conclusions

To conclude this chapter, we note that a number of popular circuits have been studied using the positive gain VCVS and negative gain (infinite gain) MFB concepts. We noted the merits of the VCVS in that they have an easier closed-loop gain adjustment, a higher input impedance and a low output impedance, which makes for easier cascading of sections without the need for isolation transformers. The MFB circuit has one fewer component, perhaps two in the band-pass configuration, and is also easy to tune. Because of the lower input impedance, this circuit may suffer when used in cascade to produce a higher-order filter.

Later, in chapter 8, it will be seen that the VCVS filter has poor Q and ω_0 sensitivities with gain, which imposes restrictions on the Q-values that may be realised.

A brief mention should be made here concerning the variation of input impedance with frequency throughout the frequency range of the filter. Since we are using reactive elements and an op-amp whose gain varies with frequency, it should come as no surprise to find that the input impedance may swing from being inductive reactive to capacitive reactive for a wide variation in the input signal frequency.

A final point to consider concerns the accuracy of the design when variation in the filter parameters, caused by a finite and perhaps low value of A_0, is taken into account. We saw in chapter 2 that for so long as A_0 is large then the behaviour of the circuit is independent of the characteristics of the amplifier and dependent only on the external circuit components. When A_0 becomes small (less than 500 for a 741), then the situation changes and the circuit becomes more dependent on the characteristics of the amplifier.

Any design and calculations must therefore make an allowance for this condition. Briefly stated, in the low-pass VCVS design for example, the gain constant K will be modified, as also will the values of the coefficients b_1 and b_0.

Problems

4.1. Design a low-pass filter to have the response characteristic shown in figure 4.36. Use the design equations to confirm (or otherwise) the pass-band and stop-band attenuations. Design both VCVS and MFB filters using 0.1 μF capacitors.
[*Ans.* VCVS: 2.93 kΩ, 38.4 kΩ, 0.1 μF, 414 kΩ, 45.6 kΩ.
MFB: 65.6 kΩ, 656 kΩ, 8.49 kΩ, 0.1 μF, 2 nF]

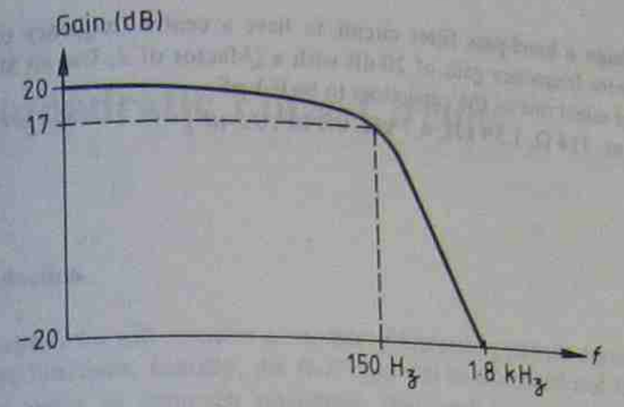


Figure 4.36

4.2. A filter is to have a gain of 5 and a permitted ripple width of 1 dB, the end of the ripple occurring at 2.4 kHz. The gain is to be at least −26 dB down at 200 Hz. Design the prototype low-pass and high-pass circuits and obtain the scaled circuit using 0.01 μF capacitors. Check out the pass-band and stop-band attenuations for the order of the filter chosen and sketch the expected frequency response. Design both VCVS and MFB circuits.
[*Ans.* VCVS: 22.9 nF, 4.26 nF, 6.7 kΩ, 6.7 kΩ, 21.7 kΩ, 87 kΩ.
MFB: 10 nF, 2 nF, 1.1 nF, 6.67 kΩ, 333 kΩ]

4.3. A filter is required to have a gain of 18 dB with a permitted pass-band ripple of 3 dB up to 6.8 kHz. The output is to be −80 dB down at 1 kHz. Design the filter detailing *all* the stages in the design including the confirmation of the actual pass-band and stop-band attenuations. Use VCVS circuits with gains of 2 and 4 respectively, using 2 nF capacitors throughout.
[*Ans.* 1st stage: 340 pF, 10.6 nF, 11.7 kΩ, 11.7 kΩ, 142 kΩ, 142 kΩ.
2nd stage: 1.72 nF, 450 pF, 11.7 kΩ, 11.7 kΩ, 87 kΩ, 261 kΩ]

4.4. A high-pass filter is to be designed to be maximally flat Butterworth having a gain of 20 dB in the pass band, a cut-off frequency of 5 kHz and a roll-off of 100 dB/decade in the transition band. Design the filter using a low-pass prototype employing 2 nF capacitors. Use VCVS circuits throughout and make use of the RC:CR transformation.
[*Ans.* 1st stage: 15.9 kΩ, 2 nF, 31.8 kΩ, 31.8 kΩ.
2nd stage: 15.9 kΩ, 15.9 kΩ, 1.78 nF, 2.25 nF, 57.9 kΩ, 71.6 kΩ.
3rd stage: 15.9 kΩ, 15.9 kΩ, 3.5 nF, 1.14 nF, 66.9 kΩ, 82.7 kΩ]

4.5. A band-pass filter is required having a centre frequency gain of 20 dB, a centre frequency of 500 Hz and a pass band of 100 Hz. Design the VCVS circuit using $K = 2$, and capacitors selected from $C = 10/f_0$ μF.
[*Ans.* 15.9 kΩ, 10.09 kΩ, 41 kΩ, 82 kΩ, 82 kΩ, 0.02 μF, 0.02 μF]

4.6. Design a band-pass filter circuit to have a centre frequency of 1 kHz, a centre frequency gain of 20 dB with a Q-factor of 2. Use an MFB circuit and select one of the capacitors to be 0.1 μF.
[*Ans.* 314 Ω, 1.59 kΩ, 4.7 kΩ, 0.1 μF, 0.2 μF]

5 Biquadratic Filter Circuits

5.1 Introduction

In this chapter, we will consider a number of interesting circuits associated with biquadratic functions. Initially, the R–C type will be discussed and the analysis conducted using an approach resembling that used for analogue computing circuitry. It should be pointed out here that some authors prefer to use state-space methods of analysis for these circuits, and for this the reader is referred to the relevant literature.

Next we consider an increasingly popular type of filter based on the fabrication of op-amps and capacitors on the same chip, the capacitors being switched to simulate resistor values. Such *switched capacitor* circuits are becoming increasingly popular in communications, spectrum analysers and medical electronics where a degree of programmability may be achieved.

Finally, brief consideration is given to the concept of an *all-resistor* biquad which uses the roll-off of the op-amp at high frequencies which was considered in chapter 3.

5.2 The biquadratic circuit

This concept provides for stable and easily tuned filter circuits. They require more components than the MFB and VCVS circuits previously considered, but in its bandpass form it is capable of producing much greater Q-factors than the VCVS and MFB circuits. Furthermore, with the use of the *quad* op-amp chip a *universal biquad* circuit is realisable which means that the five basic filter forms may be obtained using appropriately selected component values. The name *biquad* is coined because the general circuit can be shown to be defined (mathematically) as the ratio of two quadratic expressions.

A starting point for the analysis will use the transfer function for the high-pass filter:

$$H(j\omega) = \frac{K\omega^2}{-\omega^2 + \left(\frac{\omega_0}{Q}\right) j\omega + \omega_0^2} \tag{5.1}$$

and if we write $s = j\omega$ this becomes

$$H(s) = \frac{-Ks^2}{s^2 + \left(\frac{\omega_0}{Q}\right) s + \omega_0^2}$$

and to simplify the manipulation we will normalise the angular frequency to $\omega_0 = 1$ rad/s. Therefore

$$H(s) = \frac{-Ks^2}{s^2 + \frac{s}{Q} + 1} \tag{5.2}$$

This expression may be re-written as

$$H(s) = \frac{V_o}{V_i}(s) = \frac{-K}{1 + \frac{1}{Qs} + \frac{1}{s^2}} \times \frac{\left(\frac{P}{K}\right)}{\left(\frac{P}{K}\right)}$$

where, note that the expression, although modified, is mathematically unaltered and which may be considered to be equivalent to

$$V_o = -P$$

$$V_i = \frac{P}{K} + \frac{P}{K}\left(\frac{1}{Qs}\right) + \frac{P}{K}\left(\frac{1}{s^2}\right) \tag{5.3}$$

or

$$P = KV_i - \frac{P}{Qs} - \frac{P}{s^2} = KV_i + \frac{V_o}{s} + \frac{V_o}{s^2}$$

This equation states that an output $P = -V_o$ is the sum of three input voltages and this specification may be realised by the circuit shown in figure 5.1.

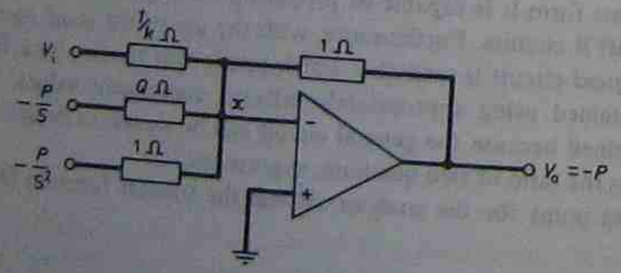


Figure 5.1 Simple summing amplifier

Nodal analysis at x yields, assuming $V_x = 0$:

$$(0 - V_i)K + \left(0 + \frac{P}{s}\right)\frac{1}{Q} + \left(0 + \frac{P}{s^2}\right) + (0 - V_o) = 0$$

or

$$-KV_i = V_o + \frac{V_o}{Qs} + \frac{V_o}{s^2} = V_o\left(1 + \frac{1}{Qs} + \frac{1}{s^2}\right)$$

$$\frac{V_o}{V_i} = \frac{-Ks^2}{s^2 + \frac{s}{Q} + 1}$$

From equation (5.3) it will be seen that we require to produce voltages $(-P/s = V_o/s)$ and $(-P/s^2 = V_o/s^2)$, and it should be noted that $(1/s)$ implies an integrator which has both R and C set to unit value of 1 ohm and 1 Farad. Similarly, $(1/s^2 = 1/s \times 1/s)$ implies a cascading of two identical integrator circuits which combine to produce the second-order effect.

It was mentioned earlier that the circuit synthesis would be similar to the method applied to the analysis of analogue computing circuits and with this in mind a convenient starting point will be a consideration of the circuit shown in figure 5.2.

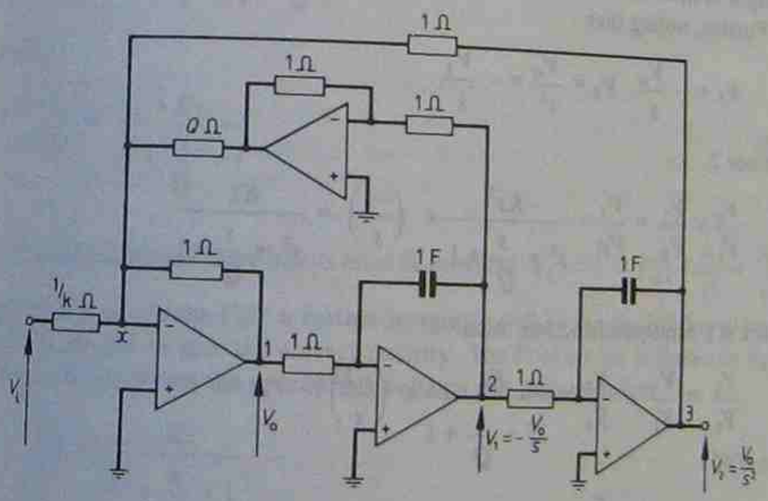


Figure 5.2 Circuit realisation for the basic biquad

Obviously, the transfer function $H(s)$ demands that we produce a second-order function which we see is satisfied by cascading two identical integrator sections. A further requirement is the summation of the three quantities given by equation (5.3), which are mixed at node x to the input of the first op-amp. Note

ST. GEORGE LIBRARY

further that this circuit, and indeed most biquad circuits, employ op-amps in the inverting mode and therefore an *odd* number of stages must be used to avoid the possibility of positive feedback arising with subsequent instability occurring. Note also that the inner-loop op-amp also produces negative feedback, the signal then being applied to node x through resistor Q ohms.

Nodal equation at x yields

$$(0 - V_i)K + (0 - V_o) + \left(0 - \frac{V_o}{s}\right)\frac{1}{Q} + \left(0 - \frac{V_o}{s^2}\right) = 0$$

or

$$-KV_i = V_o\left(1 + \frac{1}{Qs} + \frac{1}{s^2}\right)$$

and

$$\frac{V_o}{V_i} = \frac{-Ks^2}{s^2 + \frac{s}{Q} + 1}$$

The circuit shown in figure 5.2 will, therefore, represent the high-pass function (output at node 1).

Further, noting that

$$V_1 = -\frac{V_o}{s},\ V_2 = \frac{V_o}{s^2} = -\frac{V_1}{s}$$

at node 2

$$\frac{V_o}{V_i} \times \frac{V_1}{V_o} = \frac{V_1}{V_i} = \frac{-Ks^2}{s^2 + \frac{s}{Q} + 1} \times \left(\frac{-1}{s}\right) = \frac{Ks}{s^2 + \frac{s}{Q} + 1}$$

which is a band-pass function. Also

$$\frac{V_2}{V_i} = \frac{V_1}{V_i} \times \frac{V_2}{V_1} = \frac{Ks}{s^2 + \frac{s}{Q} + 1} \times \left(\frac{-1}{s}\right)$$

$$= -\frac{K}{s^2 + \frac{s}{Q} + 1}$$

which is a low-pass function.

If only low-pass and band-pass configurations are required, one of the op-amps may be eliminated by combining the operations of summation and integration as shown in figure 5.3.

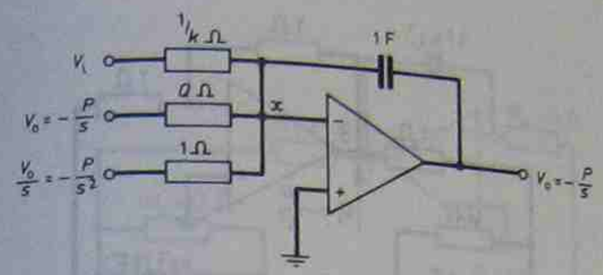


Figure 5.3 Summation and integration

At node x, nodal analysis yields

$$(0 - V_i)K + (0 - V_o)\frac{1}{Q} + \left(0 - \frac{V_o}{s}\right) + (0 - V_o)s = 0$$

or

$$-KV_i = V_o\left(s + \frac{1}{s} + \frac{1}{Q}\right)$$

giving

$$\frac{V_o}{V_i} = \frac{-Ks}{s^2 + \frac{s}{Q} + 1} \qquad (5.4)$$

It can be seen that three inputs must be provided $V_i; -\frac{P}{s} = V_o; -\frac{P}{s^2} = \frac{V_o}{s}$ and clearly to obtain V_o/s a further integrator will be required followed by an inverter circuit to give the correct polarity. The final circuit is shown in figure 5.4 which also shows the appropriate voltages and their polarities.

$$\frac{V_1}{V_i} = \frac{-Ks}{s^2 + \frac{s}{Q} + 1} \qquad \text{band-pass}$$

$$\frac{V_2}{V_i} = \frac{V_1}{V_i} \times \frac{V_2}{V_1} = \frac{-Ks}{s^2 + \frac{s}{Q} + 1} \times \left(-\frac{1}{s}\right) = \frac{K}{s^2 + \frac{s}{Q} + 1} \qquad \text{low-pass}$$

We will now consider in detail two useful circuits employing the biquad concept.

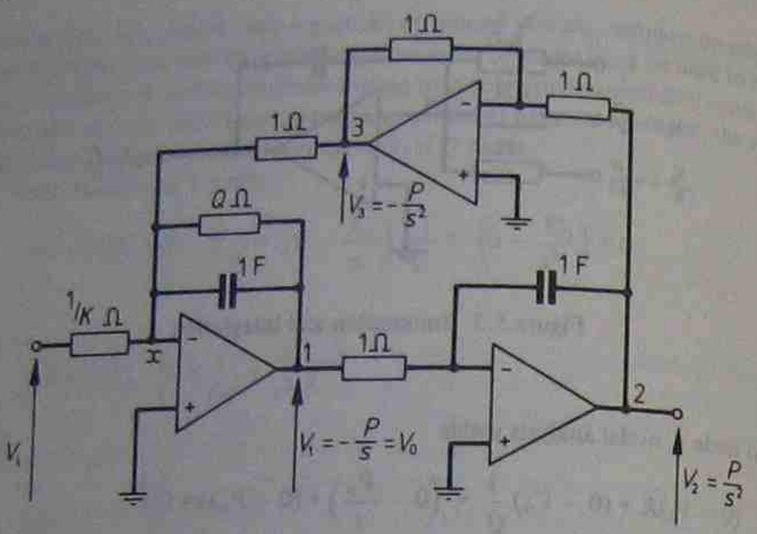


Figure 5.4 Circuit realisation of low-pass and band-pass cases

5.3 The band-pass circuit

If we are only interested in the band-pass circuit, figure 5.4 may be re-drawn as shown in figure 5.5, which includes the modification shown in figure 5.3. Note that one of the separate integrators has been removed (since we are combining summation and integration at the input) and that this particular configuration has the advantage of providing a positive-going output.

It is left as an exercise for the reader to show that consideration of conditions at node x and $\frac{V_o}{V_i} = \frac{V_y}{V_i} \times \frac{V_o}{V_y}$ produces the relationship

$$\frac{V_o}{V_i} = \frac{Ks}{s^2 + \frac{s}{Q} + 1} \tag{5.5}$$

Consider now the circuit with actual component values and consider further the conditions $C_1 = C_2 = C$ and $R_4 = R_5 = R_6$. From figure 5.5, nodal analysis at x yields

$$\frac{V_i}{R_1} + \left(sC + \frac{1}{R_2}\right) V_y + \frac{V_\omega}{R_3} = 0$$

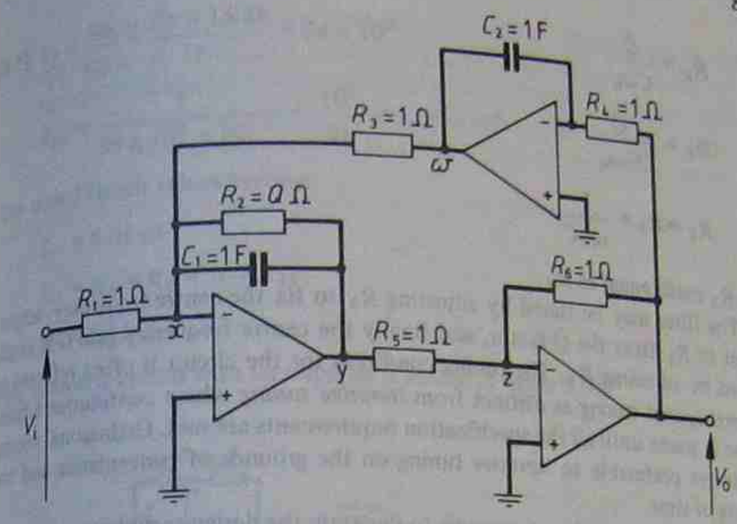


Figure 5.5 Biquad band-pass circuit

also

$$V_z = -V_y, \; V_\omega = -\frac{V_z}{sCR_4} = \frac{V_y}{sCR_4}$$

which finally gives

$$\frac{V_o}{V_i} = \frac{\frac{s}{CR_1}}{s^2 + \frac{s}{CR_2} + \frac{1}{C^2 R_3 R_4}} \tag{5.6}$$

Writing $s = j\omega$, equation (5.6) may be re-written in the standard form

$$\frac{V_o}{V_i}(j\omega) = \frac{j\omega G \frac{\omega_0}{Q}}{-\omega^2 + j\omega \frac{\omega_0}{Q} + \omega_0^2} \tag{5.7}$$

where

$$\omega_0 = \frac{1}{(C^2 R_3 R_4)^{0.5}}$$

$$G = \frac{R_2}{R_1}$$

$$R_2 = \frac{Q}{C\omega_0}$$

$$R_1 = \frac{Q}{CG\omega_0} \qquad (5.8)$$

$$R_3 = R_4 = \frac{1}{\omega_0 C}$$

for R_3 made equal to R_4.

The filter may be tuned by adjusting R_3 to fix the centre frequency, adjustment of R_2 fixes the Q-factor, and finally the centre frequency gain G is established by adjusting R_1. The tuning condition for the circuit is often referred to as *orthogonal tuning* as distinct from *iterative tuning* where continuous adjustment is made until all the specification requirements are met. Orthogonal tuning is always preferable to iterative tuning on the grounds of convenience and the saving of time.

We will now consider an example to illustrate the design technique.

Example 5.1

A band-pass filter is required having a centre frequency of 1500 Hz and a bandwidth of 300 Hz, the centre frequency gain being 5. Obtain the scaled circuit design.

Initially, normalise the centre frequency and capacitor values, that is $\omega_0 = 2\pi f_0 = 1$ rad/s, $C = 1$ F. Equations (5.8) reduce to

$$R_3 = R_4 = \frac{1}{\omega_0} = 1,\ R_3 = \frac{1}{R_4},\ R_1 = \frac{Q}{G},\ R_2 = Q$$

From chapter 4, equation (4.26):

$$Q = \frac{f_0}{\Delta f} = \frac{1500}{300} = 5$$

$$G = 5$$

$$R_1 = \frac{5}{5} = 1\ \Omega$$

$$R_2 = 5\ \Omega$$

$$R_3 = R_4 = 1\ \Omega$$

We must now choose C, and we use $C = (10/f_c)\ \mu\text{F} = 10/1500 = 0.0067\ \mu\text{F}$ which is difficult to achieve, so we choose $C = 0.01\ \mu\text{F}$ from which the scaling factors are

$$k_f = \frac{\omega_o}{\omega_n} = \frac{2\pi \times 1500}{1} = 3\pi \times 10^3$$

$$k_m = \frac{1}{3\pi \times 10^3 \times 10^{-8}} = \frac{10^5}{3\pi} = 1.06 \times 10^4$$

The scaled circuit values become:

$$C = 0.01\ \mu\text{F}$$

$$R_1 = R_3 = R_4 = 10.6\ \text{k}\Omega$$

$$R_2 = 53\ \text{k}\Omega$$

The complete circuit with its response is shown in figure 5.6.

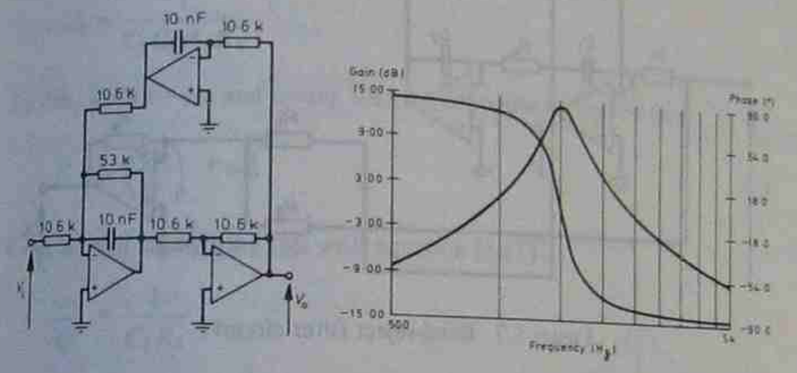


Figure 5.6 Band-pass biquad

It will be observed from the transfer function given by equation (5.7) that the phase shift will vary between +90° ($\omega = 0$) through 0° ($\omega = 3000\pi$) to −90° ($\omega \to \infty$).

5.4 The band-reject, notch or band-elimination circuit

It was mentioned earlier that with the introduction of the four op-amp (quad) chip, all of the five basic filters could be realised within the framework of a so-called *universal* filter. Consideration is now given to the band-reject circuit and figure 5.4 is re-drawn with the inclusion of an additional summing amplifier, as shown in figure 5.7.

The band-reject circuit, of which there are several forms (see figures 5.22 and 5.23), is particularly useful in applications where a specific frequency is to be

removed. An example would be the removal of the hum associated with the 50 Hz power frequency in electronic medical equipment or instrumentation systems. Such a band-reject filter, as shown in figure 5.7, uses the output from the previously considered band-pass circuit, which is added to the input signal via the summing amplifier which then inverts the combined signal.

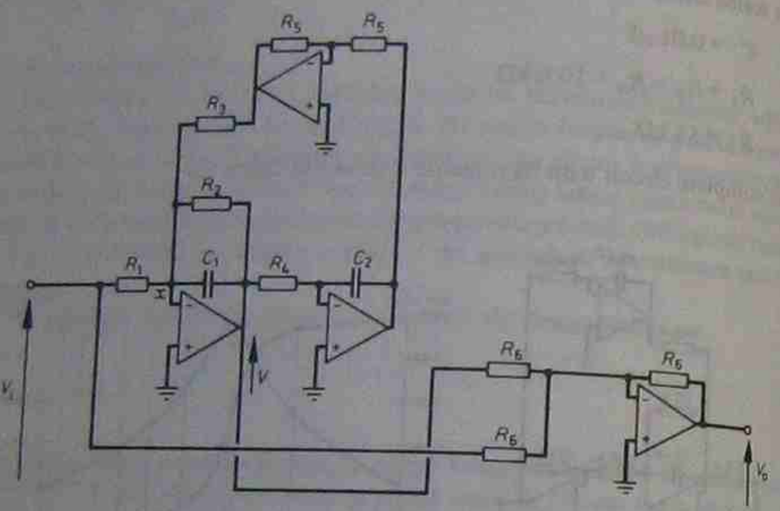


Figure 5.7 Band-reject filter circuit

It is left as an exercise for the reader to show that

$$\frac{V_1}{V_i}(s) = \frac{-\dfrac{s}{C_1R_1}}{s^2 + \dfrac{s}{C_1R_2} + \dfrac{1}{C_1C_2R_3R_4}} \tag{5.9}$$

and

$$V_o = -(V_i + V_1)$$

or

$$\frac{V_o}{V_i}(s) = -\left(1 + \frac{V_1}{V_i}\right) = -\left(1 - \frac{\dfrac{s}{C_1R_1}}{s^2 + \dfrac{s}{C_1R_2} + \dfrac{1}{C_1C_2R_3R_4}}\right)$$

$$\frac{V_o}{V_i}(s) = -\left(\frac{s^2 + \dfrac{s}{C_1R_2} - \dfrac{s}{C_1R_1} + \dfrac{1}{C_1C_2R_3R_4}}{s^2 + \dfrac{s}{C_1R_2} + \dfrac{1}{C_1C_2R_3R_4}}\right) \tag{5.10}$$

If we impose the condition $R_1 = R_2$ ($G = 1$) then the expression reduces to

$$\frac{V_o}{V_i}(s) = \frac{-(s^2 + \omega_0^2)}{s^2 + \dfrac{s}{C_1R_2} + \omega_0^2} \tag{5.11}$$

where

$$\omega_0^2 = \frac{1}{C_1C_2R_3R_4}$$

Further, write $s = j\omega$ and noting that the denominator has the form

$$s^2 + s\left(\frac{\omega_0}{Q}\right) + \omega_0^2$$

it can be seen, comparing this with equation (5.11)

$$\frac{\omega_0}{Q} = \frac{1}{C_1R_2}$$

producing

$$Q = \frac{C_1R_2}{(C_1C_2R_3R_4)^{0.5}} = \left(\frac{R_2^2C_1}{C_2R_3R_4}\right)^{0.5}$$

Normally $C_1 = C_2 = C$ and the expressions become

$$\omega_0 = \frac{1}{C(R_3R_4)^{0.5}}, \quad Q = \frac{R_2}{(R_3R_4)^{0.5}}$$

Furthermore, if we make $R_3 = R_4$ then

$$\omega_0 = \frac{1}{CR_3}, \quad Q = \frac{R_2}{R_3}$$

and by fixing C, ω_0 is adjusted by R_3 and Q by R_2. Now

$$\left|\frac{V_o}{V_i}(j\omega)\right| = \frac{\omega_0^2 - \omega^2}{\left\{(\omega_0^2 - \omega^2)^2 + \left(\dfrac{\omega}{C_1R_2}\right)^2\right\}^{0.5}}$$

and it can be seen that at selected frequencies

$$\left|\frac{V_o}{V_i}(j0)\right| = 1, \quad \left|\frac{V_o}{V_i}(j\infty)\right| = 1, \quad \left|\frac{V_o}{V_i}(j\omega_0)\right| = 0 \quad \text{or}$$

$$0\ \text{dB}, \qquad 0\ \text{dB}, \qquad -\infty\ \text{dB}$$

The suggestion here is that for the condition $\omega = \omega_0$, the circuit will provide infinite attenuation. In practice, this is not the case because of the imperfections within the components. It is interesting to consider the phase produced by the circuit at the selected frequencies:

$$\phi(j\omega) = \phi_1 - \phi_2 \text{ where } \phi_1 = \angle\omega_0^2 - \omega^2$$

$$\phi_2 = \tan^{-1}\left(\frac{\dfrac{\omega}{C_1R_2}}{\omega_0^2 - \omega^2}\right)$$

It is useful to draw these as independent angles and then to combine them as shown in figure 5.8, which also shows the gain response.

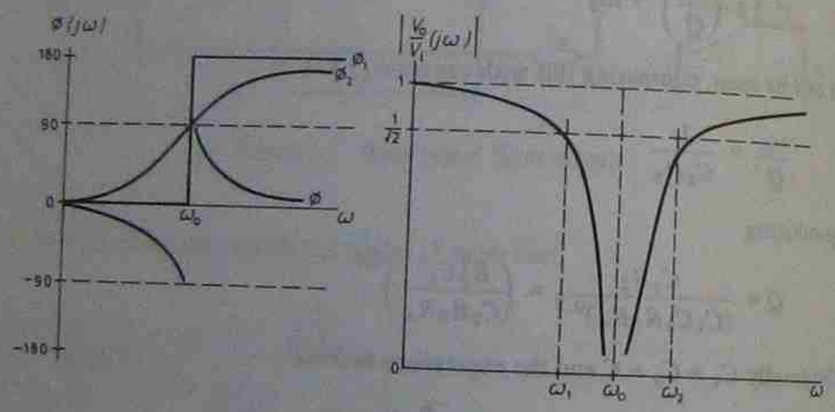


Figure 5.8 Gain/phase responses of the biquad band-reject

At the 3 dB points, bandwidth $= \dfrac{\omega_0}{Q} = \omega_2 - \omega_1$.

Example 5.2

An undesirable 50 Hz signal is required to be removed from the output of an electrocardiogram. The filter must pass frequencies above 55 Hz and below 45 Hz with no more than 3 dB of loss. The low- and high-frequency loss must be 0 dB.

From the specifications we see, using equation (4.26) from chapter 4, that

$$Q = \frac{50}{(55-45)} = 5 = \frac{R_2}{R_3}$$

$$\omega_0 = 2\pi \times 50 = \frac{1}{CR_3}$$

Normalise $\omega_0 = 1$ rad/s, $C = 1$ F, and using the previously derived equations:

$$R_3 = R_4 = 1\ \Omega$$

$$\frac{R_2}{R_1} = G$$

and

$$G = 1$$

Therefore $R_2 = Q\ \Omega = R_1$.

Let $R_5 = R_6 = 1\ \Omega \quad R_3 = R_4 = 1\ \Omega \quad R_1 = R_2 = Q\ \Omega \quad C_1 = C_2 = 1$ F.

Scaling: $k_f = \dfrac{2\pi \times 50}{1} = 100\,\pi$

Choose $C = 1\ \mu$F: $k_m = \dfrac{1}{100\pi \times 10^{-6}} = 3180$

Then $R_3 = R_4 = R_5 = R_6 = 3.18\ \text{k}\Omega$
$R_1 = R_2 = 15.9\ \text{k}\Omega$
$C_1 = C_2 = 1\ \mu$F.

The circuit with its response is shown in figure 5.9.

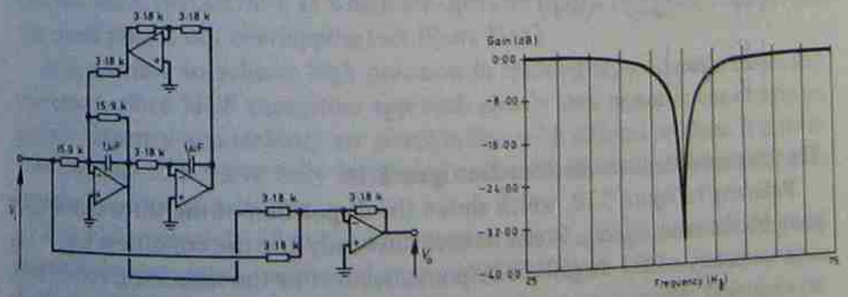


Figure 5.9 Circuit and response for example 5.2

A further note concerning the band-reject circuit

Re-call equation (5.11) which may be re-written

$$\frac{V_o}{V_i}(s) = \frac{-(s^2 + \omega_p^2)}{s^2 + \left(\frac{\omega_0}{Q}\right)s + \omega_0^2} \qquad (5.12)$$

where ω_p is the frequency at which the maximum attenuation occurs. It should be noted that $\omega_0 < \omega_p < \omega_0$ and three principal conditions arise:

$\frac{\omega_p}{\omega_0} < 1$ which defines a *high-pass* notch

$\frac{\omega_p}{\omega_0} = 1$ which defines a *regular* notch

$\frac{\omega_p}{\omega_0} > 1$ which defines a *low-pass* notch

If we consider the magnitude characteristic of equation (5.12):

$$\left|\frac{V_o}{V_i}(j\omega)\right| = \frac{\omega_p^2 - \omega^2}{\left\{(\omega_0^2 - \omega^2)^2 + \omega^2\left(\frac{\omega_0}{Q}\right)^2\right\}^{0.5}}$$

It can be seen that three principal points arise:

$$\left|\frac{V_o}{V_i}(j\omega_p)\right|_{\omega=\omega_p} = 0$$

$$\left|\frac{V_o}{V_i}(j0)\right|_{\omega=0} = \left(\frac{\omega_p}{\omega_0}\right)^2$$

$$\left|\frac{V_o}{V_i}(j\infty)\right|_{\omega\to\infty} = 1$$

The three conditions are illustrated in figure 5.10.

Referring to figure 5.10, which shows the disposition of the three conditions given by the ratio ω_p/ω_0, it can be seen that only for the condition $\omega_p = \omega_0$ does symmetry of the magnitude response occur. For the remaining conditions, asymmetrical responses are produced with the low-pass notch exhibiting gain up to the rejection frequency range and the high-pass notch exhibiting attenuation. The reader is directed to example 5.6, figure 5.23, for an illustration of the three notch conditions.

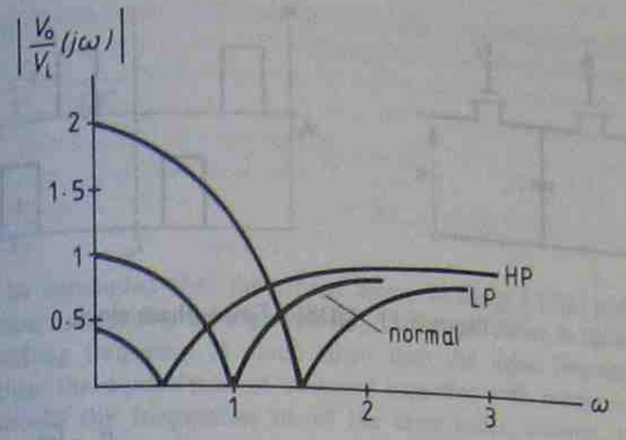


Figure 5.10 Normal, high- and low-pass notch responses

5.5 Switched capacitor filters

These are sometimes referred to as *analogue sampled data filters* and use MOS technology. Using current IC technology it is quite difficult to achieve accurate values of resistance and capacitance which exhibit good linearity and stability with variation of temperature. Most of the active *R–C* filters which have been completely integrated require careful trimming and consequently are too expensive for commercial mass production.

One approach which makes active filters more sympathetic towards IC technology is the active *R* circuit, which although it dispenses with the need for capacitors, still depends on resistor tolerances and op-amp gain-bandwidth products. A big disadvantage is that resistors require more chip area than is the case with capacitors.

The true switched capacitor filter in many ways overcomes most of the aforementioned problems. In MOS technology the switched capacitor can be implemented using two MOSFETs which are operated using a two-phase clock system, the clock phases not overlapping (see figure 5.11).

It is possible to achieve high precision in realising capacitor ratio values and, moreover, since MOS capacitors approach closely ideal capacitor characteristics, greater linearity and stability are possible than with diffused resistors. It is therefore possible to achieve fully integrated and stable active filter circuits using the switched capacitor principle.

The basic principle of switched capacitor action may be understood by considering the two-phase, non-overlapping MOSFET clock system shown in figure 5.11.

With ϕ_1 high, C is connected to V_i. When ϕ_2 is high, C is connected to V_o, otherwise C is connected to neither V_i nor V_o. The principle of operation is now discussed with reference to figure 5.12.

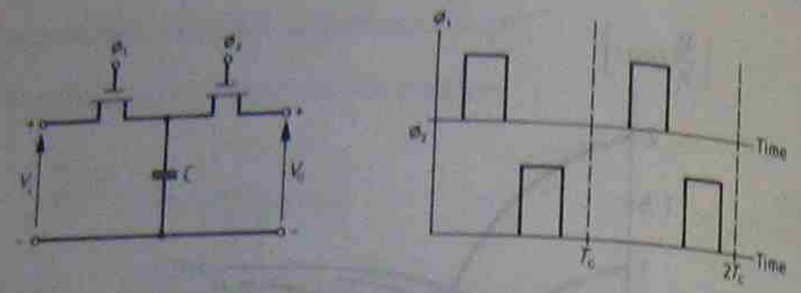


Figure 5.11 MOSFET two-phase clock

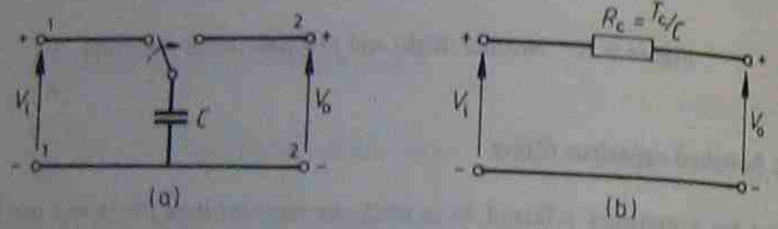


Figure 5.12 Switched capacitor with circuit equivalent

The single-pole double-throw (SPDT) switch is assumed to be driven using ϕ_1, ϕ_2 in figure 5.11 and it will be assumed that the switch is initially connected to position 1. We will also assume that V_i is a function of time and that the periodic time T of V_i is much greater than the switching frequency T_c of the SPDT switch.

In position 1, capacitor C charges to V_i and with the switch made to position 2, the capacitor discharges (or charges) to V_o.

The charge transferred in T_c seconds is given by

$$q(t) = C(V_i - V_o)$$

and

$$i(t) = \frac{\Delta q}{\Delta t} = \frac{C}{T_c}(V_i - V_o) \tag{5.13}$$

A resistor required to give the same magnitude of current will have a value given by the expression

$$R_c = \frac{V_i - V_o}{i} = \frac{T_c}{C(V_i - V_o)}$$

where

$$R_c = \frac{T_c}{C} = \frac{1}{f_c C} \tag{5.14}$$

and

$$f_c = \frac{1}{T_c}$$

It may be concluded that the circuit shown in figure 5.12(a) performs the same function as a resistor having a value T_c/C ohms, as shown in figure 5.12(b). If the switching frequency is much larger than the signal frequencies being handled, then the equivalence of switched capacitor with resistor is accurate enough. Should the frequencies be of the same order, however, then data-sampling techniques must be employed.

A simple circuit with which to illustrate the procedure is the inverting integrating amplifier shown in figure 5.13.

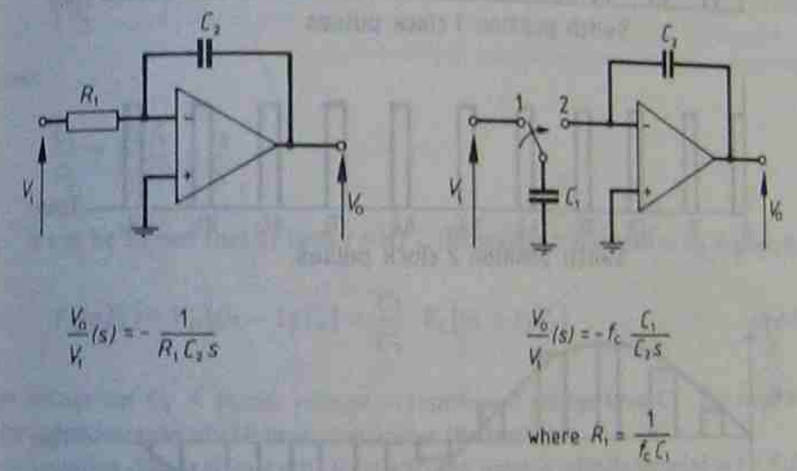


Figure 5.13 Integrator circuit with switched capacitor equivalent

With the switch in position 1, C_1 charges to V_i and when the switch is moved to position 2, this charge is removed from C_1 and transferred to C_2. It should be noted that all of the charge is transferred because the virtual-earth condition of the op-amp constrains the voltage across C_1 to be zero. The resultant charge on C_2 is the charge previous to making the switch *less* the charge transferred from C_1. This charge is subtracted because of the inverting operation of the amplifier.

It can be seen that the time constant of the integrator may easily be adjusted by means of the clock frequency.

The nature of the operation of a sampled data integrator circuit may be further illustrated by reference to figure 5.14.

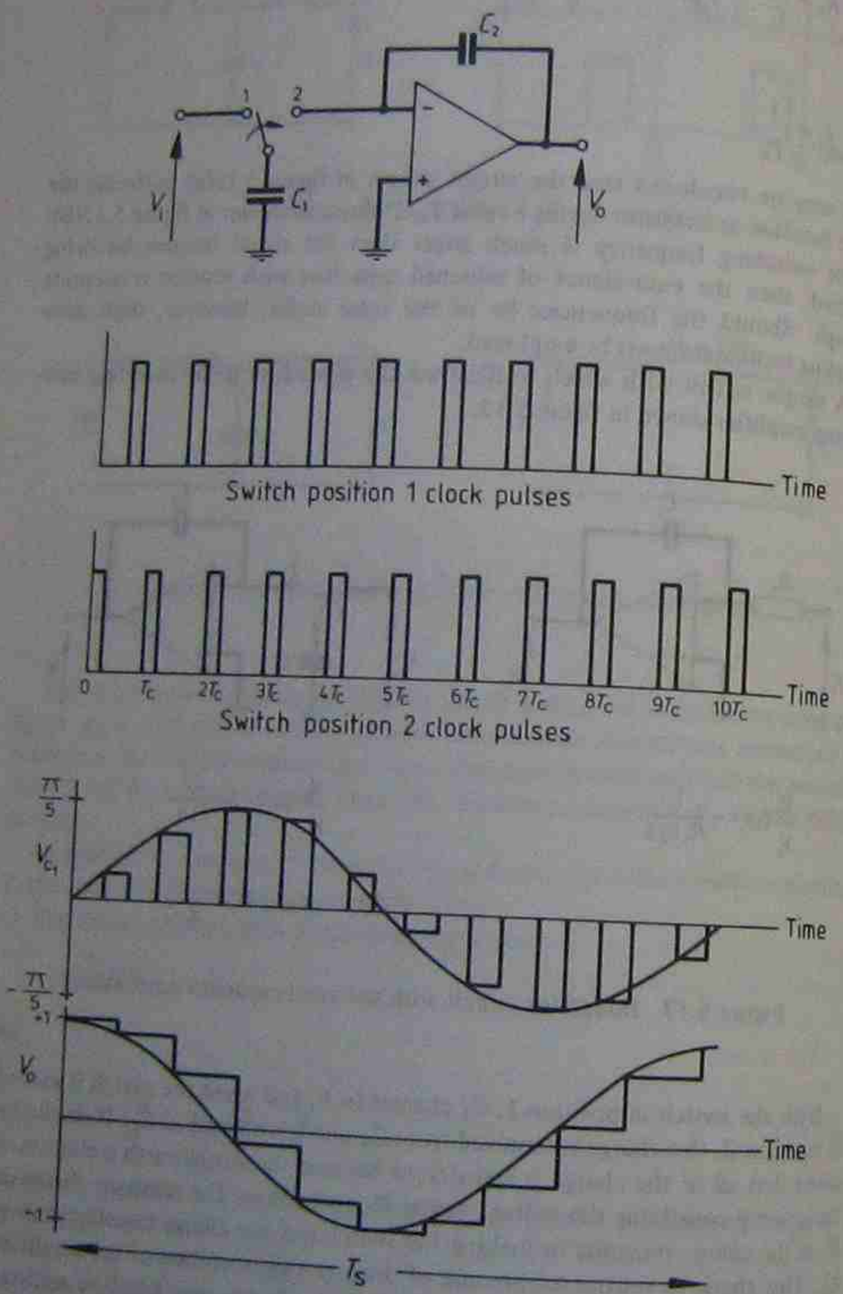


Figure 5.14 Sampled data concept

The input is assumed to be sinusoidal having a frequency f_s and a peak amplitude of 1 V. The input is sampled in the centre of each clock cycle and this value (V_i) is held on the capacitor C_1 for half a clock period and then is transferred to C_2 when the switch is made to position 2. The output is held on C_2 until a new sample of V_i is transferred to C_2 and so on. The clock frequency is made ten times larger than the signal frequency ($f_c = 10\ f_s$); also, using the previous expression with $s = j\omega$ in order to consider circuit operation with frequency:

$$\frac{V_o}{V_i}(j\omega) = -f_c \frac{C_1}{j\omega C_2}$$

$$\left|\frac{V_o}{V_i}(j\omega)\right| = \frac{C_1}{C_2} \times \frac{f_c}{2\pi f_s}$$

when

$$\left|\frac{V_o}{V_i}(j\omega)\right| = 1$$

then

$$\frac{C_1}{C_2} = \frac{2\pi f_s}{f_c} = \frac{\pi}{5}$$

It can be shown that at time $t = nT_c$, the voltage transferred to C_2 is given by

$$V_o(nT_c) = V_o[(n-1)T_c] - \frac{C_1}{C_2} V_i[(n-1)T_c] \qquad (5.15)$$

or voltage on C_2 = initial voltage − transferred voltage from C_1. The negative sign arises because of the inversion across the amplifier.

Assuming that at time $t = 0$, $V_o = 1$ V, the input is initially sampled at $t = T_c/2$ and is held on C_1 for a further period of time $t = T_c/2$ before the voltage on C_1 is transferred to C_2 at $t = T_c$.

From figure 5.14 it can be seen that

$$T_c = \frac{T_s}{10} = \frac{360^\circ}{10} \quad \text{or} \quad \frac{T_c}{2} \equiv 18^\circ$$

Using equation (5.15) we may compute the voltage V_o for any switching time, for example

$$t = \frac{T_c}{2} \equiv 18^\circ$$

The sampled voltage on C_1 is $(\pi/5)$ sin 18° and the voltage on C_2 for $T_c \leqslant t \leqslant 2T_c$ is given by

$$V_o(2T_c) = V_o(t = T_c) - \frac{\pi}{5} V_i \left(t = \frac{T_c}{2}\right)$$

$$= 1 - \frac{\pi}{5} \sin 18° = 0.806 \text{ V}$$

This voltage is held on C_2 until $t = 2T_c$ while the input voltage is being sampled at the new time $t = (3T_c)/2$ giving a new voltage on C_1 of $(\pi/5)$ sin 54°. The voltage on C_2 for $2T_c \leqslant t \leqslant 3T_c$ is given by

$$V_o(3T_c) = V_o(2T_c) - \frac{\pi}{5} V_i(2T_c)$$

$$= 0.806 - 0.508 = 0.298 \text{ V}$$

This process may be repeated for all the clock intervals to obtain the waveform shown in figure 5.14.

Consider now the circuits shown in figure 5.15, the second circuit being the switched capacitor equivalent of the first circuit.

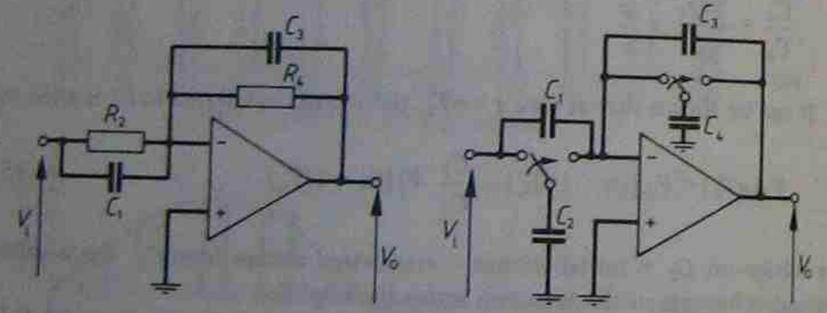


Figure 5.15 Further circuit with its switched capacitor equivalent

Here

$$C_2 = \frac{1}{f_c R_2}, \quad C_4 = \frac{1}{f_c R_4}$$

also

$$\frac{V_o}{V_i}(s) = -\frac{C_1}{C_3} \frac{\left(s + \frac{1}{R_2 C_1}\right)}{\left(s + \frac{1}{R_4 C_3}\right)}$$

and on substituting for C_2, C_4:

$$\frac{V_o}{V_i}(s) = -\frac{C_1}{C_3} \frac{\left(s + f_c \frac{C_2}{C_1}\right)}{\left(s + f_c \frac{C_4}{C_3}\right)} \tag{5.16}$$

To illustrate only, consider now an example of the technique outlined above.

Example 5.3

We wish to implement a switched capacitor design to give the specification shown in figure 5.16.

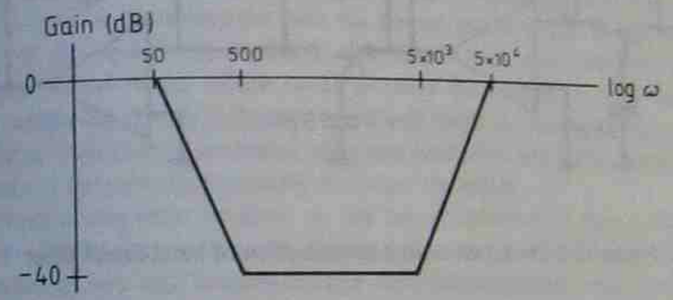


Figure 5.16 Band-reject specification

The transfer function for this band-reject response takes the form

$$\frac{V_o}{V_i}(s) = \frac{(s + 500)(s + 5 \times 10^3)}{(s + 50)(s + 5 \times 10^4)}$$

which may be broken down into the form

$$\frac{V_o}{V_i}(s) = \frac{(s + 500)}{(s + 50)} \times \frac{(s + 5 \times 10^3)}{(s + 5 \times 10^4)} \tag{5.17}$$

Select $C_1 = C_3$ for convenience and also assign a value of 10 pF to these components. Further, select the clock frequency f_c to be 10 kHz.

Equating the coefficients of equations (5.16) and (5.17) gives for the first section:

$$\frac{C_2}{C_1} = 0.05 \quad \text{or} \quad C_2 = 0.5 \text{ pF}$$

$$\frac{C_4}{C_3} = 0.005 \quad \text{or} \quad C_4 = 0.05 \text{ pF}$$

ST. GEORGE LIBRARY

Repeating for the second section gives the capacitor values

$C_3 = 5$ pF, $C_4 = 50$ pF

It will be observed that there is a large *spread* of capacitor values and that the capacitor values used are relatively small. The final circuit is shown in figure 5.17.

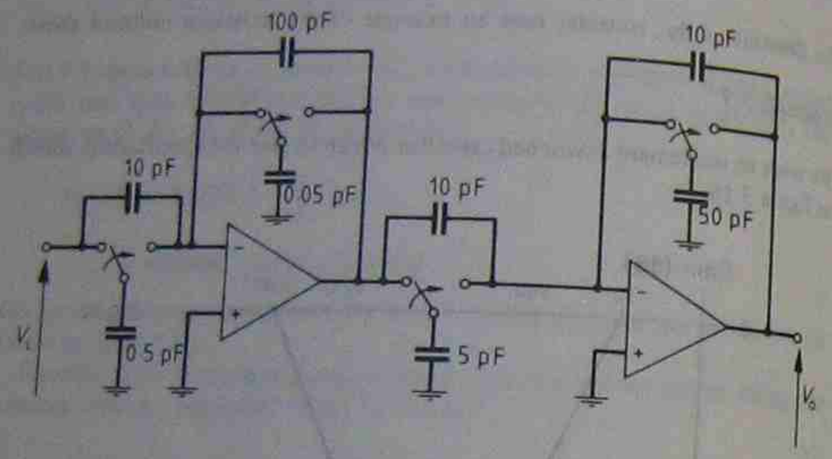


Figure 5.17 Switched capacitor realisation of band-reject filter

Conclusions

It is perhaps timely here to remind ourselves of some of the problems associated with integrated circuit switched-capacitor filters, of which the following are typical:

1. the MOS switch used in the implementation should have high off-resistance and low on-resistance, as well as having low parasitic capacitance values from the source and drain to the substrate;
2. should the clocking rate be of the order of signal frequency, then sampled data techniques must be used for the analysis, which means that the input signal should be band limited;
3. consideration of the circuits shows that the switched-capacitor resistor does not provide a continuous path for op-amp leakage currents, which are essential to minimise offsets at the input to the op-amp. This may complicate the design of circuits that require a precise value of closed-loop gain, such as is often required in VCVS filter circuits.

On the plus side we find that many modern communication systems can readily take advantage of switched-capacitor filters because the filter parameters

of gain, Q-factor, centre frequency and bandwidth, are accurately controlled by capacitor ratios and clock frequency.

Typical devices manufactured currently are dual-in-line, that is, two separate filters which may be used singly or cascaded to produce all the five filter functions; the programming requirements are the selection of a clock frequency, power supplies and three or four resistors.

5.6 All-resistor active biquad circuits

It has been previously mentioned that at high frequencies the op-amp derived circuits cease to have benefits over the normal passive circuit design. The main cause of the op-amp limitation is due to the finite gain-bandwidth product imposed on the device which tends to move the location of the poles in the filter, which for certain Q-factors could well result in circuit instability. Also, as the signal frequency is increased, slew-rate limitation may occur because of the inability of the op-amp accurately to output the signal.

Consideration must be given to the use of inductors at these considerably higher working frequencies. Of course, inductors are a much more viable consideration at very high frequencies and are indeed extensively used. Nevertheless, integrated circuit technology has the advantage of compatibility between devices, and inductor technology, although advancing, is still not yet fully compatible with silicon chip technology. A starting point will be the one-pole roll-off model of the op-amp discussed in chapter 3.

$$A(j\omega) = \frac{A_0\omega_a}{j\omega + \omega_a} = \frac{A_0}{1 + \dfrac{j\omega}{\omega_a}} \tag{5.18}$$

It may be observed that the gain/frequency relationship has a similar relationship to that of an elementary R-C network. The conclusion may therefore be drawn that biquad filters may be designed using only operational amplifiers and resistors with the exclusion of capacitors. A number of circuits have been designed by many workers around the model of the op-amp specified in equation (5.18) to work at relatively high frequencies in the range 20–300 kHz.

The circuit in figure 5.18 shows an active-R biquad circuit.
The constraints on R_4 and R_5 are given by the expressions

$$R_4 = (1 - \beta)R$$
$$R_5 = \beta R \tag{5.19}$$
$$\text{for } 0 < \beta < 1$$

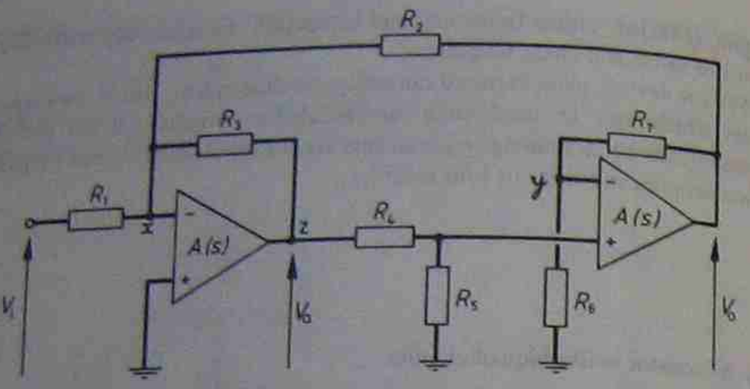


Figure 5.18 Active-*R* biquad circuit

Applying nodal analysis at points x, y and z within the circuit in figure 5.18 yields

$$\left.\begin{aligned}
&\frac{(V_1 - V_x)}{R_1} + \frac{(V_z - V_x)}{R_2} + \frac{(V_z - V_x)}{R_3} = 0 \\
&\frac{(V_o - V_y)}{R_7} = \frac{V_y}{R_6} \\
&V_z = -A(s)V_x = -\frac{A_0\omega_a}{s + \omega_a} V_x \\
&V_o = \frac{A_0\omega_a}{s + \omega_a}(\beta V_a - V_y)
\end{aligned}\right\} \quad (5.20)$$

After considerable manipulation of equation (5.20), and by making the valid assumption that the open-loop gain (A_0) is much larger than the ratio of the resistor values, we finally obtain the desired transfer function:

$$\frac{V_z}{V_1}(s) = \frac{a_1 s + a_0}{s^2 + b_1 s + b_0} \quad (5.21)$$

where

$$\left.\begin{aligned}
a_1 &= \frac{A_0\omega_a}{1 + \frac{R_1}{R_2} + \frac{R_1}{R_3}} \\
a_0 &= \frac{(A_0\omega_a)^2}{\left(1 + \frac{R_1}{R_2} + \frac{R_1}{R_3}\right)\left(1 + \frac{R_7}{R_6}\right)} \\
b_1 &= A_0\omega_a\left(\frac{1}{1 + \frac{R_7}{R_6}} - \frac{1}{1 + \frac{R_3}{R_2} + \frac{R_3}{R_1}}\right) \\
b_0 &= \frac{\frac{R_3}{R_2}(\beta A_0^2\omega_a^2)}{1 + \frac{R_3}{R_2} + \frac{R_3}{R_1}}
\end{aligned}\right\} \quad (5.22)$$

After re-arranging equations (5.22) and introducing a coefficient y, where $y > 1$ to ensure positive resistor ratios, we obtain the expressions

$$\left.\begin{aligned}
&1 + \frac{R_1}{R_2} + \frac{R_1}{R_3} = \frac{A_0\omega_a}{a_1} \\
&1 + \frac{R_3}{R_2} + \frac{R_3}{R_1} = \frac{A_0\omega_a R_3}{a_1 R_1} \\
&\frac{R_7}{R_6} = \frac{A_0\omega_a}{b_1\left(1 - \frac{1}{y}\right)} \\
&R_4 = R(1 - \beta) \\
&R_5 = R\beta \\
&\beta = \frac{b_0}{A_0\omega_a\left(A_0\omega_a - \frac{b_1}{y} - a_1\right)} \\
&\frac{R_1}{R_2} = \frac{b_0}{A_0\omega_a\beta a_1}
\end{aligned}\right\} \quad (5.23)$$

Consider now the design of an all op-amp band-pass filter.

Example 5.4

An all-resistor band-pass biquad circuit is to be designed to have the following specification: $Q = 5$, $G = 10$, $f_0 = 80$ kHz. Obtain the realisation of the filter

assuming that the op-amp employed has the following parameters: $A_0 = 8 \times 10^4$, $\omega_a = 10\pi$ rad/s. The second pole of the op-amp is taken to be at $\omega = 8\pi \times 10^5$ rad/s.

Using the equation for a band-pass response:

$$\frac{V_o}{V_i}(j\omega) = \frac{j\omega\omega_0 \dfrac{G}{Q}}{-\omega^2 + \dfrac{j\omega\omega_0}{Q} + \omega_0^2}$$

Comparing with equation (5.21) with $a_0 = 0$, $s = j\omega$, the equation becomes, after insertion of the given data

$$\frac{V_o}{V_i}(j\omega) = \frac{j10^6\omega}{-\omega^2 + j10^5\omega + 2.53 \times 10^{11}}$$

where $a_1 = 10^6$, $b_1 = 10^5$, $b_0 = 2.53 \times 10^{11}$. Selecting $R = 2$ kΩ, equations (5.23) produce

$\beta = 0.07$, $R_1 = 145\ \Omega$ (selected), $R_2 = 100\ \Omega$

$R_3 = 2.07$ kΩ, $R_4 = 1.86$ kΩ, $R_5 = 140\ \Omega$, $R_6 = 120\ \Omega$ (selected)

$R_7 = 33$ kΩ

The final circuit with its response is shown in figure 5.19.

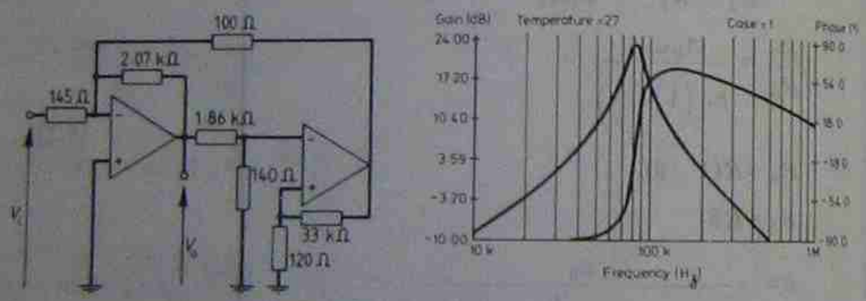


Figure 5.19 All-resistor band-reject circuit

5.7 Conclusions

It can readily be seen from the foregoing analysis that active-R circuits are more difficult to design and implement than the usual R–C biquad circuits previously considered. One might ask whether such circuits are really desirable, when at frequencies in excess of 100 kHz inductors become far more attractive components to use in terms of size, cost and ease of manufacture.

It was mentioned earlier that until inductors become more integrated, circuit-compatible active realisations will still be sought after. Nevertheless, active-R filters are almost impossible to produce on a production-line basis, mainly because of the sensitivities associated with the gain-bandwidth product ($A_0\omega_a$) being very dependent on power supply and temperature variations. Also, since perfect matching of op-amps of the same type is virtually impossible, individual circuit tuning must always be applied. Slew-rate limiting of the op-amp also causes problems at high frequencies where even low signals could cause slewing problems.

Some workers in this field have used CCVS (current-controlled-voltage-source) and CCCS (current-controlled-current-source) op-amp circuits instead of the more normal VCVS circuit. These circuits have the benefit of low input impedance levels which can lead to filters having a much higher frequency range.

Problems

5.1. Using the circuit shown in figure 5.5, obtain the design of a band-pass filter having the following specification: centre frequency gain 10 dB, centre frequency 10 kHz, with a bandwidth of 250 Hz. Use 10 nF capacitors in the design.
[*Ans.* $R_1 = 2$ kΩ, $R_2 = 6.4$ kΩ, $R_3 = R_4 = 1.6$ kΩ]

5.2. Using the circuit shown in figure 5.7, design a band-reject filter having the following specification: centre frequency 1 kHz, three dB bandwidth 100 Hz, low- and high-frequency loss 0 dB. Use 0.1 μF capacitors in the design.
[*Ans.* $R_1 = R_2 = 15.9$ kΩ, $R_3 = R_4 = R_5 = R_6 = 1.59$ kΩ]

5.3. The circuit shown in figure 5.20 is that of a band-pass filter having adjustable *ganged* capacitors and a variable resistor R_1. Obtain the transfer function of the filter in the form:

$$\frac{V_o}{V_i}(j\omega) = \frac{j2\omega\omega_0/Q}{-\omega^2 + j\omega\dfrac{\omega_0}{Q} + \omega_0^2}$$

where $\omega_0^2 = \dfrac{1}{C^2R^2}$, $Q = K$.

Design a filter to have a centre frequency of 10 kHz and a bandwidth of 1 kHz. Use 0.1 μF capacitors in the design.
[*Ans.* $R_1 = 1.59$ kΩ, $R = 159\ \Omega$, $C = 0.1$ μF]

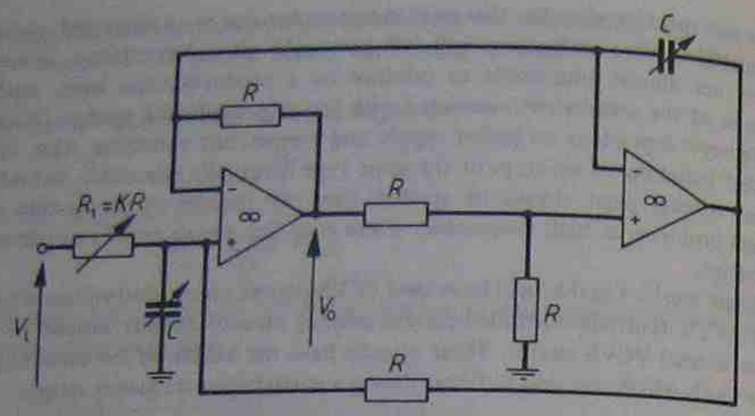


Figure 5.20

5.4. Obtain the transfer function $\frac{V_o}{V_i}(j\omega)$ for the circuit shown in figure 5.21 in the form:

$$\frac{V_o}{V_i}(j\omega) = \frac{\omega^2 + j\frac{\omega}{C_2R_2}\left(\frac{R_2}{R_3} - 1\right) - \frac{1}{C_1C_2R_1R_2}}{-\omega^2 + j\frac{\omega}{C_2R_2} + \frac{1}{C_1C_2R_1R_2}}$$

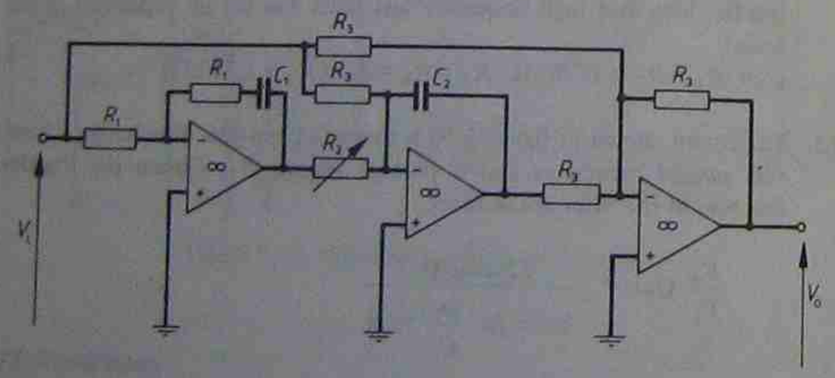


Figure 5.21

Examine the conditions when: (a) $R_2 = R_3$, (b) $R_2 = 2R_3$. Which responses do these conditions represent?

5.5. The circuit shown in figure 5.22 is that of a twin-T band-reject filter circuit employing two unity-gain op-amps. Write the nodal equations at the nodes 2, 3 and 4 and obtain the standard transfer function for the circuit, showthat $\omega_o = \frac{1}{CR}$, $Q = \frac{1}{4(1-K)}$. Design a band-reject filter to have $Q = 10$, $f_o = 1$ kHz and using 0.1 μF capacitors.
[*Ans.* $R = 1.59$ kΩ]

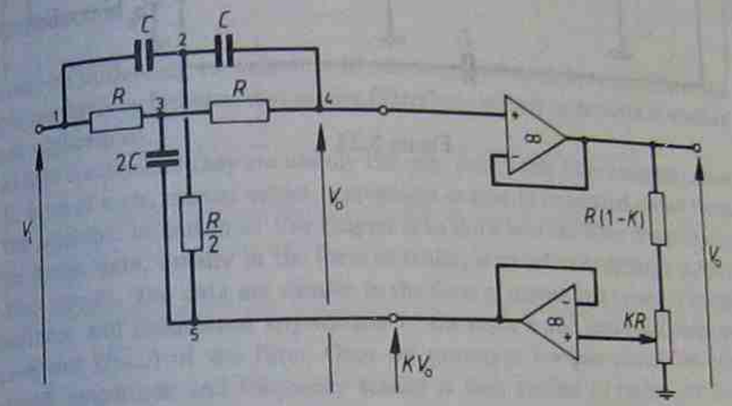


Figure 5.22

5.6. Obtain the transfer function for the circuit shown in figure 5.23 in the form:

$$\frac{V_o}{V_i}(j\omega) = \frac{\omega_z^2 - \omega^2}{-\omega^2 + j\omega\frac{\omega_0}{Q} + \omega_0^2}$$

where

$$\omega_z^2 = \frac{R_2}{R_1R_3R_5C_1C_2}, \quad \omega_0^2 = \frac{1}{R_4R_5C_1C_2}$$

$$Q = \frac{R_5R_6}{R_5 + R_6}\left(\frac{C_2}{R_4R_5C_1}\right)^{\frac{1}{2}}$$

Which filter response does the transfer function represent?

If the values of ω_0, R_1, C_1 and C_2 are normalised and the conditions $R_5 = R_6$, $R_2 = KR_1$ imposed, obtain the component values in terms of K, Q and ω_z. Explain the function of each of the variable components in the *tuning* of the circuit.

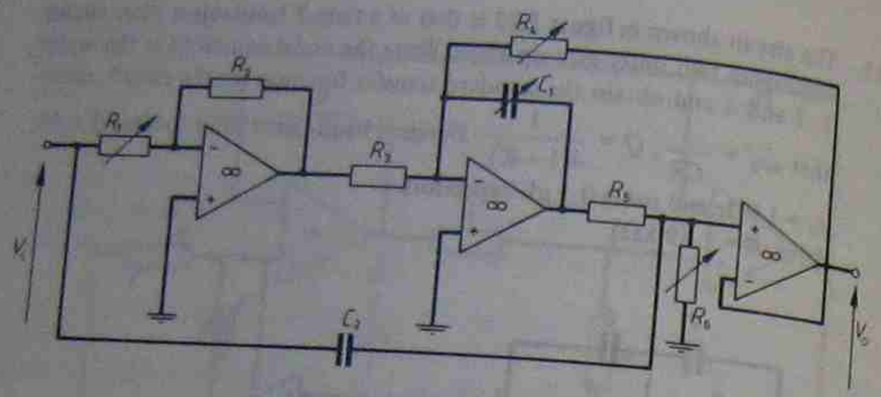


Figure 5.23

6 Passive Filter Circuit Design

6.1 Introduction

It would be misleading to suggest, with the advent of operational amplifiers and digital synthesis techniques, that passive filters have no role to perform in modern circuit applications.

At high frequencies they are usually the only practicable filter available, albeit in the form of coils, coaxial cables, waveguides or even in integrated circuit form.

The principal intention of this chapter is to show how the filter designer may utilise design data, usually in the form of tables, as an aid to designing a particular filter circuit. The data are usually in the form of *normalised* values of circuit capacitance and inductance appropriate to the order n and particular response requirement $H(j\omega)$ of the filter. Once the prototype low-pass circuit has been obtained, magnitude and frequency scaling is then applied to realise the final circuit. Another, important, reason for introducing passive circuit realisations is that from these circuits the gyrator-based equivalents considered in chapter 7 are derived.

Finally, a method of frequency transformation from low-pass to high-pass and band-pass is outlined which shows how prototype low-pass design data may be used to implement other responses.

6.2 The coefficient matching technique

This is a simple and effective means of obtaining a circuit transfer function, but is limited to simple structures. The method will be outlined by considering the circuit shown in figure 6.1. The circuit is of the second order, has two energy storage elements, and is terminated at both ends in resistors R_s (source) and R_L (load).

Analysis of the circuit yields:

$$\frac{V_o}{V_i}(s) = \frac{1}{R_s + sL + \dfrac{R_L}{1 + sCR_L}} \left(\frac{\dfrac{1}{Cs}}{R_L + \dfrac{1}{Cs}} \right) R_L$$

$$\frac{V_o}{V_i}(s) = \frac{\frac{1}{LC}}{s^2 + s\frac{(L + CR_L R_s)}{LCR_L} + \frac{(R_L + R_s)}{LCR_L}} \tag{6.1}$$

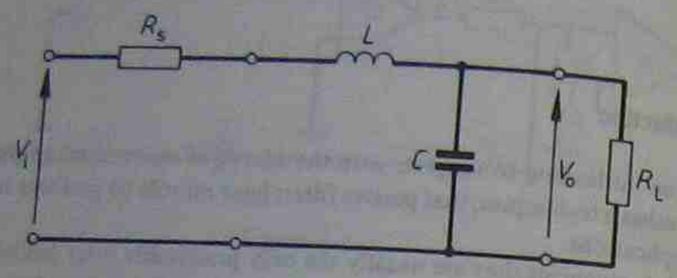


Figure 6.1 *L–C* section with resistor terminations

If now we assume the specifications of a second-order Butterworth response, then

$$\frac{V_o}{V_i}(s) = \frac{K}{s^2 + \sqrt{2}s + 1} \tag{6.2}$$

Treating these expressions as being identical, we may therefore compare coefficients. Furthermore, it is usual to work with normalised terminating resistors ($R_s = R_L = 1\ \Omega$):

$$\frac{L + C}{LC} = \sqrt{2},\ \frac{2}{LC} = 1,\ K = \frac{1}{LC}$$

From which we obtain $L = \sqrt{2}$ H, $C = \sqrt{2}$ F.

The method appears elegant and straightforward, but becomes increasingly difficult to apply as the order n of the circuit increases.

There now follows the outline of a technique which may be used to obtain the normalised values of inductance and capacitance for a given network and a given response. The main aim here is to demonstrate how tables of design data are established and through the use of these tables passive filter realisations may be achieved.

Readers who require a deeper and more detailed account of the technique are referred to the appropriate texts listed in the Bibliography.

6.3 The design of doubly terminated *L–C* filters

The method is based on the extensive use of modern network theory and was originally outlined by Darlington and other workers. The design strategy involves

the integration of the following concepts: insertion loss, reflection coefficient, network input impedance and the appropriate response requirement of the filter.

Consider the loss-less network with resistive terminations shown in figure 6.2.

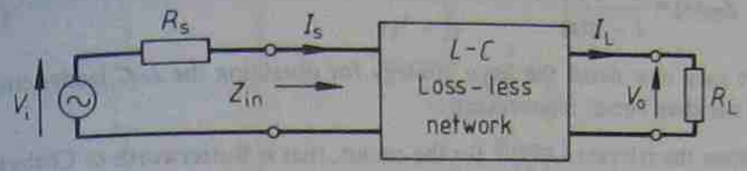


Figure 6.2 Loss-less network, doubly terminated

Any *L–C* network may be characterised at the source end of the network by a reflection coefficient ρ expressed as:

$$\rho(s) = \frac{Z_{IN}(s) - R_s}{Z_{IN}(s) + R_s} \tag{6.3}$$

Assuming a loss-less network and unequal terminations, then the power in the load may be expressed by:

$$P_L = \left|\frac{V_i R_L}{R_s + R_L}\right|^2 \Big/ R_L \tag{6.4}$$

Maximum power occurs when the terminating resistors are equal ($R_L = R_s$):

$$P_{max} = \frac{V_o^2}{R_L} = \left(\frac{V_i}{2}\right)^2 \Big/ R_s \tag{6.5}$$

Using these equations we obtain:

$$\frac{P_L}{P_{max}} = \frac{4R_L R_s}{(R_s + R_L)^2} \tag{6.6}$$

Using definitions of insertion loss, reflection coefficient and response, it can be shown for an input impedance specified as

$$Z_{IN}(j\omega) = R + jX$$

associated with a loss-less circuit working under maximum power conditions ($R_L = R_s$):

$$\rho(s)\,\rho(-s) = 1 - \frac{4R_L R_s}{(R_L + R_s)^2}\,|H(s)|^2 \tag{6.7}$$

where $H(s)$ is the appropriate circuit response.

Furthermore, by making $R_L = R_s = 1\ \Omega$, equations (6.3) and (6.7) yield

$$\rho(s)\,\rho(-s) = 1 - |H(s)|^2 = \left|\frac{Z_{IN}(s) - 1}{Z_{IN}(s) + 1}\right|^2$$

from which is finally obtained:

$$Z_{IN}(s) = \frac{1 + \rho(s)}{1 - \rho(s)} \tag{6.8}$$

We may now detail the basic strategy for obtaining the L–C ladder circuits from the given circuit requirements:

1. obtain the relevant $|H(s)|^2$ for the circuit, that is Butterworth or Chebyshev;
2. obtain $\rho(s)\rho(-s) = 1 - |H(s)|^2$;
3. deduce $Z_{IN}(s)$;
4. expand $Z_{IN}(s)$ into a continued fraction and obtain the ladder configuration of the normalised circuit;
5. scale the circuit to obtain the specified realisation.

To illustrate the use of the technique outlined above, we will now consider a number of examples.

Example 6.1

A third-order Butterworth filter is required normalised at 1 rad/s and terminated in equal 1 Ω resistors. Obtain the normalised low-pass design.

Using equation (2.2) from chapter 2 with $\epsilon = 1$, $s = j\omega$, $n = 3$:

$$|H(s)|^2 = \frac{1}{1 + (-js)^6} = \frac{1}{1 - s^6}$$

$$\rho(s)\,\rho(-s) = 1 - \frac{1}{1 - s^6} = \frac{s^6}{s^6 - 1}$$

$$= \left(\frac{s^3}{s^3 - 1}\right)\left(\frac{s^3}{s^3 + 1}\right)$$

from which

$$\rho(s) = \frac{s^3}{(s - P_1)(s - P_2)(s - P_3)} = \frac{s^3}{s^3 + 2s^2 + 2s + 1} \quad \text{(see table 2.1)}$$

where P_1, P_2, P_3 are the poles of the denominator, the denominator of this expression being obtained from table 2.1.

Inserting the expression for $\rho(s)$ into equation (6.8) gives, after rationalisation:

$$Z_{IN}(s) = \frac{2s^3 + 2s^2 + 2s + 1}{2s^2 + 2s + 1}$$

Finally, the continued division process is applied as follows:

$$2s^2 + 2s + 1 \;\overline{)\; 2s^3 + 2s^2 + 2s + 1 \;}\; s$$

$$2s^3 + 2s^2 + s$$

$$s + 1 \;\overline{)\; 2s^2 + 2s + 1 \;}\; 2s$$

$$2s^2 + 2s$$

$$1 \;\overline{)\; s + 1 \;}\; s$$

$$s$$

$$1$$

The two circuits which will satisfy the requirement are shown in figure 6.3.

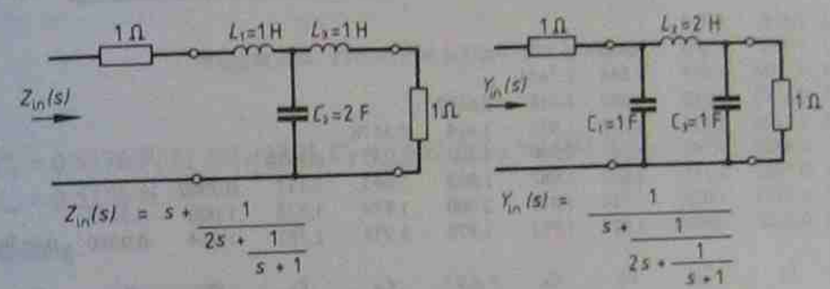


Figure 6.3 Prototype T and π circuits

By continuing to increase the order n of the response, a table of component values may be drawn up as shown in table 6.1. Of course, the procedure outlined above may be applied to deduce any loss-less configuration, but this is unnecessary once the table of component values has been drawn up. All that will be required is to scale the normalised values appropriate to the quoted filter specifications.

The method is also valid for unequal terminations and the reader is directed towards the more detailed literature on the topic.

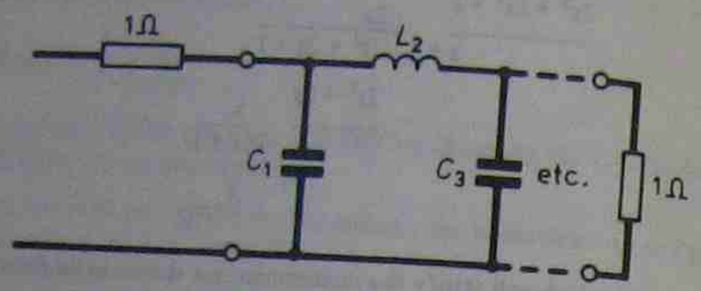
Finally, by substituting

$$|H(s)|^2 = \frac{1}{1 + \epsilon^2 C_n^2(s)}$$

for the Chebyshev response, we may use the technique to determine an appropriate table for the inductors and capacitors for the Chebyshev circuit. It should be noted that there will be several such tables corresponding to the particular ripple value ϵ chosen, or more usually, the value of A_{MAX} permitted in the pass band.

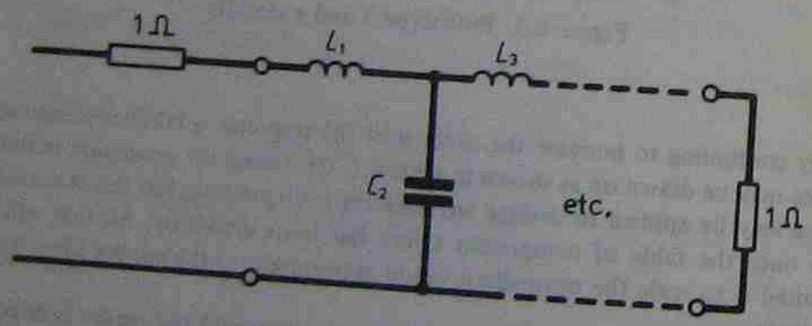
ST GEORGE'S LIBRARY Technical College

Table 6.1 Element values for Butterworth filters normalised for 1 radian/s, 1 ohm, 1 ohm



n	C_1	L_2	C_3	L_4	C_5	L_6	C_7	L_8	C_9	L_{10}
2	1.414	1.414								
3	1.000	2.000	1.000							
4	0.7654	1.848	1.848	0.7654						
5	0.6180	1.618	2.000	1.618	0.6180					
6	0.5176	1.414	1.932	1.932	1.414	0.5176				
7	0.4450	1.247	1.802	2.000	1.802	1.247	0.4450			
8	0.3902	1.111	1.663	1.962	1.962	1.663	1.111	0.3902		
9	0.3473	1.000	1.532	1.879	2.000	1.879	1.532	1.000	0.3473	
10	0.3129	0.9080	1.414	1.782	1.975	1.975	1.782	1.414	0.9080	0.3129

n	L_1	C_2	L_3	C_4	L_5	C_6	L_7	C_8	L_9	C_{10}



Example 6.2

Using the data in table 6.1 obtain the prototype low-pass circuit for a sixth-order filter. Scale the circuit to have a cut-off frequency of 100 Hz and to have equal 600 Ω terminations.

We will use the π-circuit; it is left as an exercise for the reader to take the T-circuit equivalent. This example will serve to illustrate the use of tables of data in designing passive, ladder-type filters.

Referring to table 6.1 we may obtain the normalised values of capacitance and inductance. For example, using the π-circuit equivalent the first component is a capacitor $C_1 = 0.5176$ F followed by a series inductor $L_2 = 1.414$ H which is then followed by a shunt capacitor $C_3 = 1.932$ F and so on. The values are listed below and are shown in figure 6.4:

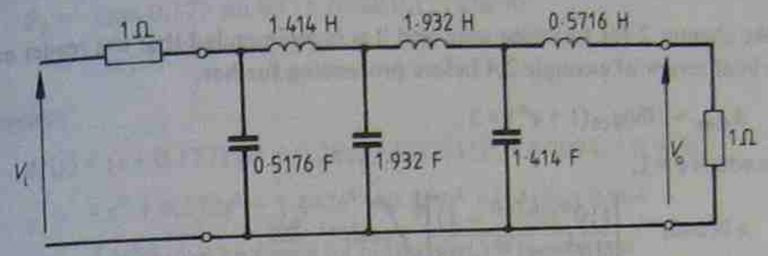


Figure 6.4 Prototype sixth-order low-pass

$C_1 = 0.5176$ F, $L_2 = 1.414$ H, $C_3 = 1.932$ F, $L_4 = 1.932$ H, $C_5 = 1.414$ F, $L_6 = 0.5176$ H

Scaling:

$$k_f = \frac{2\pi \times 100}{1} = 200\pi \qquad L_o = L_n \left(\frac{k_m}{k_f}\right)$$

$$k_m = 600 \qquad C_o = \frac{C_n}{k_m k_f}$$

The final scaled circuit is shown in figure 6.5.

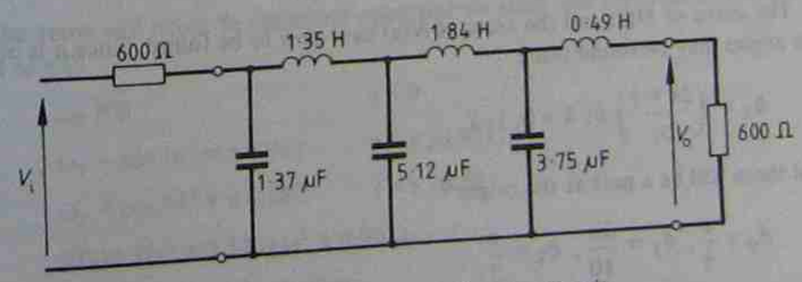


Figure 6.5 Scaled sixth-order circuit

An example will now be given to show how the method may be used to design passive Chebyshev filter circuits.

Example 6.3

A filter is to be designed around the following specifications:

Pass band up to 5 kHz with a permitted ripple of 3 dB.
Stop band should have a minimum loss greater than 100 dB at 30 kHz. The terminations are to be 600 Ω pure resistors.

See chapter 2 for formulae used and it is recommended that the reader make a brief review of example 2.4 before proceeding further:

$$A_{max} = 10\log_{10}(1+\epsilon^2) = 3$$

from which $\epsilon = 1$:

$$n \geqslant \cosh^{-1}\left[\frac{(10^{A\,min/10}-1)}{(10^{A\,max/10}-1)}\right]^{\frac{1}{2}} \Big/ \cosh^{-1}\frac{\omega_s}{\omega_p}$$

$$\geqslant \cosh^{-1} 10^5/\cosh^{-1} 6 = 4.92$$

We take $n = 5$:

$$C_5^2(s) = -(16s^5 + 20s^3 + 5s)^2 \text{ (see table 2.2 where } s = j\omega)$$

$$|H(s)|^2 = \frac{1}{1+\epsilon^2 C_5^2(s)}$$

$$\rho(s)\,\rho(-s) = 1 - |H(s)|^2 = \frac{\epsilon^2 C_5^2(s)}{1+\epsilon^2 C_5^2(s)}$$

which yields on substituting $C_5^2(s)$:

$$\rho(s)\,\rho(-s) = \frac{(16s^5+20s^3+5s)^2}{(16s^5+20s^3+5s)^2 - 1} = \frac{N(s)}{D(s)}$$

The zeros of $D(s)$ and the zeros of $N(s)$ have now to be found. Since n is odd the angles may be found using:

$$\theta_k = \left(\frac{2k+1}{2n}\right)\pi;\ k = 0, 1, 2$$

and there will be a pole at the origin:

$$\theta_0 = \frac{\pi}{5},\ \theta_1 = \frac{3\pi}{10},\ \theta_2 = \frac{\pi}{2}$$

There will be poles of $\rho(s)\rho(-s)$ in the form $(s-P_0)(s-P_1)(s-P_2)$, obtained as follows:

$$a = \frac{1}{n}\sinh^{-1}\left(\frac{1}{\epsilon}\right) = 0.177$$

$$P_0 = -\sinh 0.177 \sin 18^\circ \pm j\cosh 0.177 \cos 18^\circ$$
$$= -0.0547 \pm j0.9652$$

$$P_1 = -\sinh 0.177 \sin 54^\circ \pm j\cosh 0.177 \cos 54^\circ$$
$$= -0.143 \pm j0.597$$

$$P_2 = -\sinh 0.177 \sin 90^\circ \pm j\cosh 0.177 \cos 90^\circ$$
$$= -0.177$$

Therefore

$$D(s) = (s+0.177)(s^2+0.286s+0.375)(s^2+0.1094s+0.966)$$
$$= s^5 + 0.572s^4 + 1.442s^3 + 0.559s^2 + 0.419s + 0.064$$

The zeros of $N(s)$ may be found by noting that $s = j\omega$ and that

$$N(j\omega) = |C_5^2(\omega)|^2$$

The zeros occur for the condition $(16\omega^5 - 20\omega^3 + 5\omega)^2 = 0$, and note that we may write, using equation (2.11) with $C_n(\omega) = 0$:

$$C_5^2(\omega) = 0 = \cos(5\cos^{-1}\omega)$$

and

$$5\cos^{-1}\omega = \frac{k\pi}{2};\ k = 1, 3, 5, 7$$

from which

$$\omega = \cos\frac{k\pi}{10}$$

The zeros will occur in conjugate pairs and we shall also find that a zero occurs at $\omega = 0$:

$\omega_0 = 0$ $\qquad s = 0$

$\omega_1 = \cos 18^\circ = \pm 0.951$ $\qquad s = \pm j0.951$

$\omega_3 = \cos 54^\circ = \pm 0.58$ $\qquad s = \pm j0.58$

$$N(s) = s(s^2+0.336)(s^2+0.904)$$

$$N(s) = s^5 + 1.24s^3 + 0.304s$$

From which

$$\rho(s) = \frac{N(s)}{D(s)} = \frac{s^5 + 1.24s^3 + 0.304s}{s^5 + 0.572s^4 + 1.442s^3 + 0.559s^2 + 0.419s + 0.064}$$

Using $Z_{IN}(s) = \dfrac{1 + \rho(s)}{1 - \rho(s)}$, we obtain after routine manipulation:

$$Z_{IN}(s) = \frac{2s^5 + 0.522s^4 + 2.682s^3 + 0.56s^2 + 0.723s + 0.064}{0.572s^4 + 0.202s^3 + 0.56s^2 + 0.115s + 0.064}$$

Employing the synthetic division process, $Z_{IN}(s)$ reduces to:

$$Z_{IN}(s) = 3.496s + \cfrac{1}{0.79s + \cfrac{1}{4.39s + \cfrac{1}{0.734s + \cfrac{1}{3.406s + 1}}}}$$

The normalised circuit is shown in figure 6.6.

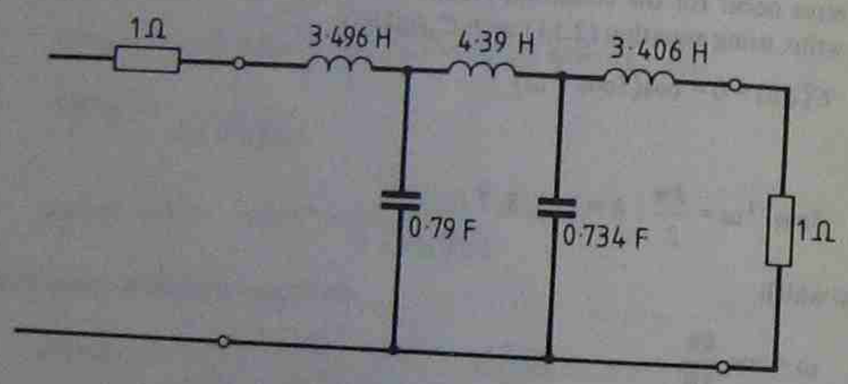


Figure 6.6 Prototype fifth-order Chebyshev filter

Scaling factors:

$k_f = 2\pi \times 5 \times 10^3 = \pi \times 10^4$

$k_m = 600$

$C_o = \dfrac{C_n}{k_m k_f}$

$R_o = 600\ \Omega$

$L_o = L_n \left(\dfrac{k_m}{k_f}\right)$

The final, scaled circuit is shown in figure 6.7 which also shows the frequency response.

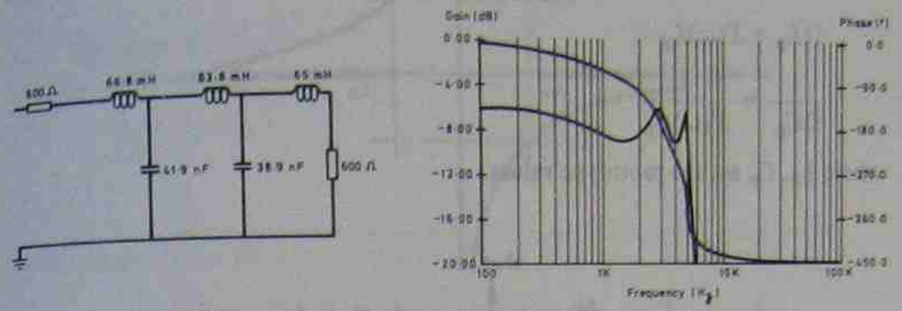


Figure 6.7 Scaled circuit with response for example 6.3

The foregoing example clearly indicates that the design of passive filters having either a Butterworth or Chebyshev response is a laborious and time-consuming exercise. Fortunately, the designer has recourse to tables wherein the normalised inductor and capacitor component values are computed for any order n of filter. The only hardship facing the designer is that of scaling the circuit to the filter specification. Of course if the reader objects to his work being done by others, then by all means reject the idea of tables!

6.4 Frequency transformations

Filter design is usually based on consideration of the prototype low-pass circuit. Other types of response are often required and these may be achieved using an appropriate frequency transformation which means that the specification of a low-pass filter may be transformed into those of any other filter using a *reactance transformation*. By means of such a transformation, the low-pass frequency variable Ω is replaced by a realisable function $X(\omega)$ associated with the variable ω. Other filter responses may be achieved by the appropriate choice of $X(\omega)$, it being noted that such a function is physically realised by using pure inductors and capacitors.

Let $\Omega = X(\omega)$ be the transformation, and

$$X(\omega) = H \frac{(\omega_2^2 - \omega^2)(\omega_4^2 - \omega^2)\ldots}{(\omega_1^2 - \omega^2)(\omega_3^2 - \omega^2)\ldots} \qquad (6.9)$$

The reactance function has the form shown in figure 6.8.

If we define the low-pass cut-off frequency to be $-\Omega c$ and Ωc, then the pass-band extends from $-\Omega c \leqslant \Omega \leqslant \Omega c$ which is itself transformed into a number of

pass bands with centre frequencies at 0, $\pm\omega_0$, $\pm 2\omega_0$. The components of the transformed filters are obtained by applying the transformation to the reactance of the elements of the prototype. For example

$$j\Omega L_p = jH\omega L_p$$

$$\frac{1}{j\Omega C_p} = \frac{1}{jH\omega C_p}$$

where L_p, C_p are the prototype values.

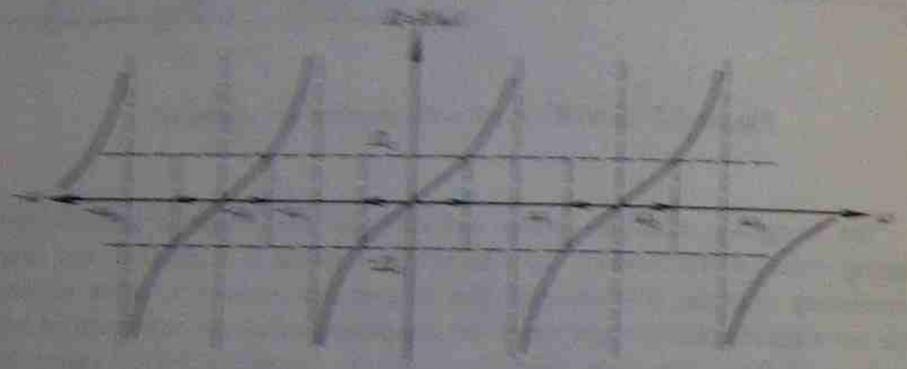


Figure 6.8 Reactance function representing $X(\omega)$

6.4.3 Low-pass to high-pass

Here we select the reactance function:

$$\Omega = -\frac{H}{\omega} \tag{6.10}$$

This function is shown in figure 6.9.

It can be seen that using this transformation, the low-pass pass-band is transformed into two pass-bands ($-\omega_c$ to $-\infty$) and (ω_c to ∞).

$$j\Omega L_p = -j\frac{H}{\omega}L_p = \frac{1}{j\omega C_p} \quad \text{where } C_p = \frac{1}{HL_p}$$

$$\frac{1}{j\Omega C_p} = -\frac{\omega}{jHC_p} = \frac{j\omega}{HC_p} = j\omega L_p \quad \text{where } L_p = \frac{1}{HC_p}$$

The circuits are shown in figure 6.10.

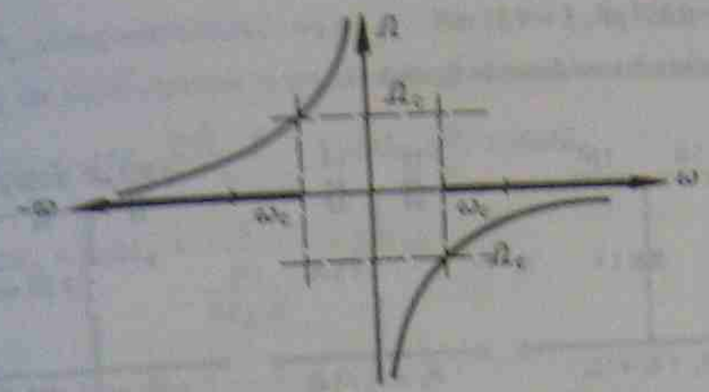


Figure 6.9 Reactance function $\Omega = -(H/\omega)$

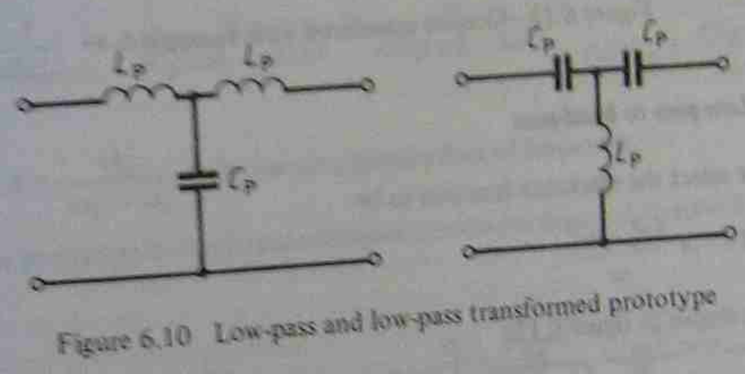


Figure 6.10 Low-pass and low-pass transformed prototype

Example 6.4

Design a high-pass network having a cut-off frequency of 5 kHz and being terminated in 600 Ω resistors.

From table 6.1 we obtain, for a normalised low pass having equal terminations:

$$L_p = 1\text{ H},\ C_p = 2\text{ F},\ \Omega_c = 1,\ H = 1$$

Using the relationships obtained:

$$C_p = \frac{1}{L_p} = 1\text{ F},\ L_p = \frac{1}{C_p} = 0.5\text{ H}$$

Scaling values are

$$k_m = 600,\ k_f = \frac{2\pi \times 5 \times 10^3}{1} = 3.14 \times 10^4$$

giving for the scaled circuit:

$C = 0.053\ \mu F,\ L = 9.55\ mH$

The three circuits are shown in figure 6.11.

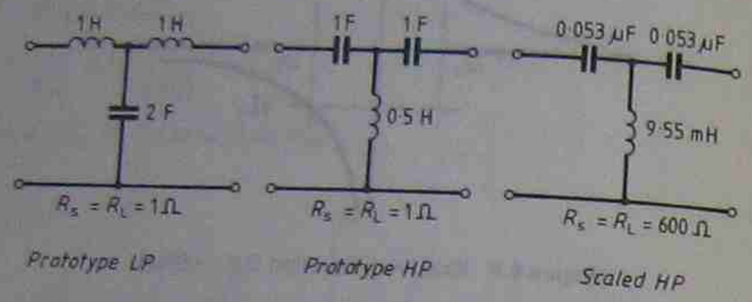


Figure 6.11 Circuits associated with example 6.4

6.4.2 *Low-pass to band-pass*

Here we select the reactance function to be:

$$\Omega = -A\,\frac{\omega_n^2 - \omega^2}{\omega} \tag{6.11}$$

which is shown in figure 6.12.

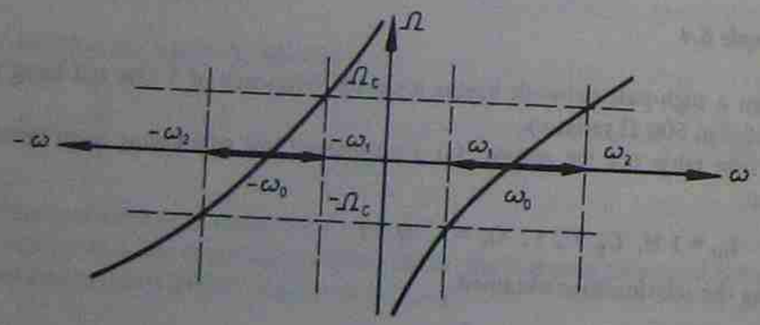


Figure 6.12 Reactance function $\Omega = -A\,((\omega_n^2 - \omega^2)/\omega)$

The function is seen to transform the low-pass pass-band $(-\Omega_c, \Omega_c)$ into two band-pass pass-bands $(\omega_1, -\omega_2)$. The transformed frequencies are

Ω_c corresponds to $(-\omega_1, \omega_2)$

$-\Omega_c$ corresponds to $(\omega_1, -\omega_2)$

Applying the transformation to the low-pass circuit we obtain for $\omega_0^2 = \omega_1\omega_2$:

$$j\Omega L_p = -jAL_p\,\frac{(\omega_0^2 - \omega^2)}{\omega} = -jAL_p\,\frac{\omega_0^2}{\omega} + j\omega AL_p$$

$$j\Omega L_p = j\omega AL_p + \frac{1}{\dfrac{j\omega}{AL_p\omega_0^2}} = j\omega L_{BP} + \frac{1}{j\omega C_{BP}}$$

where $L_{BP} = AL_p$, $C_{BP} = \dfrac{1}{AL_p\omega_0^2}$

Similarly:

$$\frac{1}{j\Omega C_p} = -\frac{\omega}{jC_p(\omega_0^2 - \omega^2)A} = \frac{j\omega}{AC_p(\omega_0^2 - \omega^2)} = \frac{1}{j\omega AC_p + \dfrac{AC_p\omega_0^2}{j\omega}}$$

where $A = \dfrac{\Omega_c}{\omega_2 - \omega_1} = Q$, for normalised values of frequency.

The prototype low-pass and band-pass circuits are shown in figure 6.13.

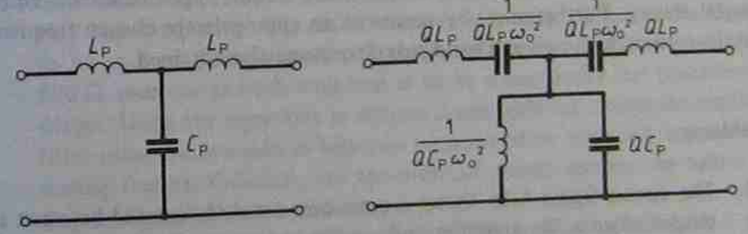


Figure 6.13 Low-pass and band-pass transformed prototypes

Example 6.5

A band-pass filter is required to have a centre frequency of 500 Hz, a bandwidth of 100 Hz and to be terminated at both ends with 1 kΩ resistors.

The normalised circuits will be those previously shown in figure 6.13 but with $\omega_0 = 1$ rad/s:

$$A = Q = \frac{f_0}{\Delta f} = \frac{500}{100} = 5$$

The scaling factors are:

$$k_m = 1000,\ k_f = 2\pi \times 500$$

$$QL_p = 5\ \text{H};\ \frac{1}{QL_p} = 0.2\ \text{F};\ QC_p = 10\ \text{F};\ \frac{1}{QC_p} = 0.1\ \text{H}$$

The scaled circuit with its response is shown in figure 6.14.

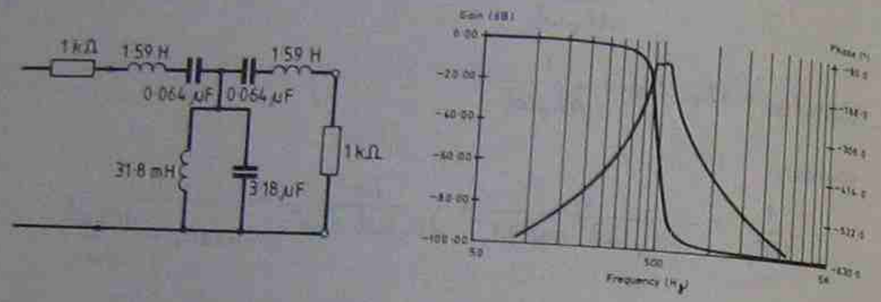


Figure 6.14 Scaled band-pass circuit and responses

We will end this chapter by noting that by means of tables of data, prototype passive filters may be designed around appropriately quoted specifications, namely Butterworth, Chebyshev, or the more exotic types based on elliptic integral theory. Furthermore, by means of an appropriately chosen frequency transformation, high-pass and band-pass circuits may be obtained.

Problems

6.1. The circuit figure 6.15 shows a cross-over network inserted between the output of an audio amplifier and woofer and tweeter loudspeakers. Using the data in table 6.1, design the prototype low- and high-pass filters and scale the circuit to give a cross-over frequency of 4 kHz with equal terminations of 8 Ω.
[*Ans.* 0.477 mH, 0.159 mH, 6.6 μF; 3.33 μF, 0.238 mH, 9.95 μF]

6.2. A Chevyshev filter is to have the following specification: pass band from 1 kHz to infinity with a permitted 1 dB of ripple; stop-band attenuation to be at least 80 dB down at 100 Hz. The circuit is to be terminated in 1 kΩ resistors at both ends.
Using example 6.3 as a guide, obtain the complete scaled circuit, working throughout from first principles.
[*Ans.* $n = 4$, 76 nF, 0.15 H, 56 nF, 0.2 H]

6.3. A band-pass filter is required to have a centre frequency of 500 Hz, a bandwidth of 100 Hz and is to be terminated in 1 kΩ pure resistors. Design the filter using a third-order low-pass prototype and the low-pass to band-pass transformation.
[*Ans.* 1.59 H, 64 nF, 31.8 mH, 3.18 μF, 64 nF, 1.59 H]

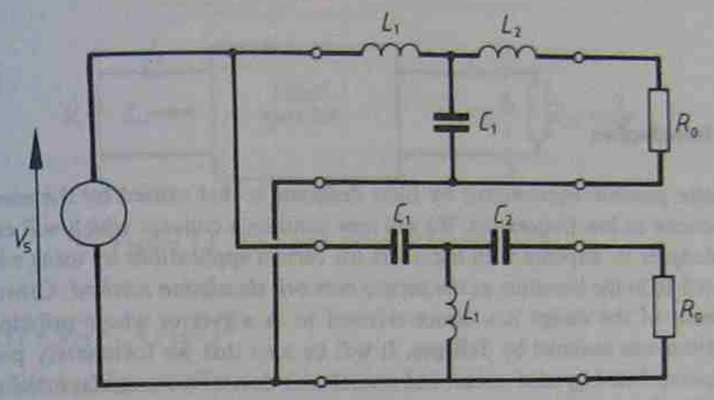


Figure 6.15

6.4. A filter circuit is required to have the following specifications: stop band to extend from dc up to 100 Hz; pass band to extend from 1 kHz up to a very high frequency. The cut-off frequency is to be 1 kHz and the filter is to be at least 200 dB down at 100 Hz; the circuit is to be terminated in 800 Ω resistors at both ends and is to be a maximally flat Butterworth design. Using the equations in chapter 2 and table 6.1, design the required filter using the low-pass to high-pass transformation and apply appropriate scaling factors. Calculate, for the order of circuit chosen, the values of stop- and pass-band attenuations.
[*Ans.* $n = 10$: 0.405 H, 0.22 μF, 0.09 H, 0.11 μF, 0.07 H, 0.1 μF, 0.713 μF, 0.14 μF, 0.14 H, 0.633 μF]

7 Gyrator-based Filter Circuits

7.1 Introduction

A major problem encountered by filter designers is that caused by the need for inductance at low frequencies. We will now consider a concept which will enable the designer to dispense with inductors for certain applications by using what is referred to in the literature as *the passive network simulation method*. Central to the basis of the design is a device referred to as a gyrator whose principle of operation was outlined by Tellegan. It will be seen that we fortunately possess an op-amp based gyrator circuit and several such devices are manufactured using IC technology. A distinct advantage of the active simulation approach is that the circuits produced have low sensitivities which are basically those of the passive counterparts. We shall see that a distinct disadvantage of producing *synthetic* inductance, as it is called, is that being an active device, one of its terminals must be connected to earth. Furthermore, if matched op-amps are used, then the effects of non-ideal amplifier parameters on the inductance realisation are minimised.

7.2 Properties of an ideal gyrator

The *ideal* gyrator is a four-terminal (two-port) network which presents, at either port, an input impedance which is proportional to the admittance connected across the other port. With reference to figure 7.1 it is considered that

$$Z_{11} = r^2 Y_{22} \tag{7.1}$$

likewise

$$Y_{11} = g^2 Z_{22} \tag{7.2}$$

where g is defined as the *gyration conductance*:

$$g = \frac{1}{r} \tag{7.3}$$

Observation of equations (7.1) and (7.2) suggests that if the termination is a pure resistor then the input impedance is also a pure resistor. If the terminating component is a pure capacitor, however, then the input impedance is a pure inductor and vice versa. We may conclude, therefore, that the device *inverts* the

impedance connected to it. Furthermore, the *ideal* gyrator is assumed to dissipate or store no power, and is referred to as being a passive, loss-less element.

To study further the behaviour of the gyrator, we will characterise the network of figure 7.1 by its appropriate y parameters.

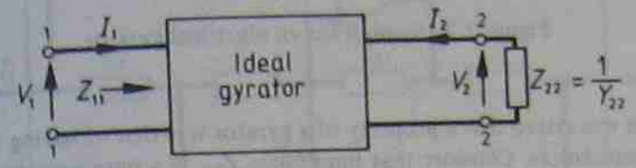


Figure 7.1 Two-port network terminated in Z_{22}

$$I_1 = y_{11} V_1 + y_{12} V_2 \tag{7.4}$$
$$I_2 = y_{21} V_1 + y_{22} V_2$$

where y_{11} = short-circuit input admittance
y_{12} = reverse short-circuit transfer admittance
y_{21} = forward short-circuit transfer admittance
y_{22} = short-circuit output admittance.

From figure 7.1 we see that using the given polarities for currents and voltages

$$I_2 = -Y_{22} V_2 \tag{7.5}$$

Manipulation of equations (7.4) and (7.5) yields

$$Y_{11} = \frac{I_1}{V_1} = y_{11} - \frac{y_{12} y_{21}}{y_{22} + Y_{22}} \tag{7.6}$$

The conditions which must be imposed to enable the two-port network to act as an ideal gyrator are

$$y_{11} = 0 = y_{22}$$
$$y_{12} y_{21} = -g^2 \tag{7.7}$$
$$Y_{11} = g^2 Z_{22}$$

Choosing $y_{12} = g$ and $y_{21} = -g$ we obtain from equation (7.4) the relationships which exist between the current at one port and the voltage at the other port in the form

$$I_1 = g V_2 \tag{7.8}$$
$$I_2 = -g V_1$$

The symbol for an ideal gyrator is shown in figure 7.2.

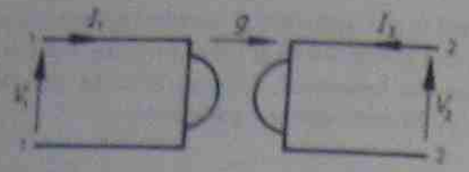


Figure 7.2 Symbol for an electrical gyrator

Earlier it was stated that a property of a gyrator was that of having the ability to invert impedances. Consider that impedance Z_{22} is a pure capacitor $Z(j\omega) = 1/j\omega C$. Using equation (7.1) produces

$$Z_{11} = r^2/j\omega C = j\omega(r^2 C)$$

Here it can be seen that the capacitor C F has been *inverted* into an inductor $L = r^2 C$ H.

A circuit which may be considered as a practical realisation of an ideal gyrator is shown in figure 7.3.

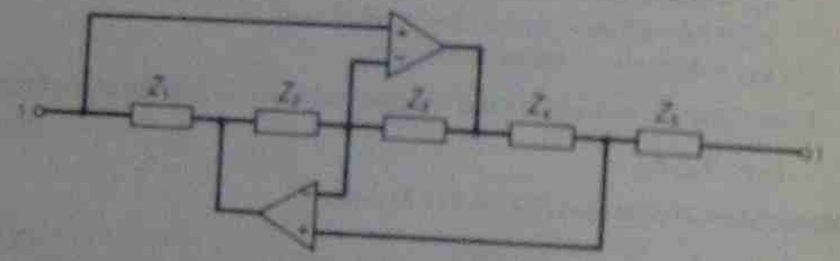


Figure 7.3 Two op-amp realisation of a gyrator

7.3 The generalised impedance converter – GIC

The circuit which is shown in figure 7.3 may be re-drawn as shown in figure 7.4 where the impedance Z_5 has been removed from outside the circuit enclosed within the box and which is referred to as a GIC.

If the circuit of figure 7.4 is analysed for its input impedance it can be shown that

$$Z_{11} = \frac{V_1}{I_1} = \frac{Z_1 Z_3 Z_5}{Z_2 Z_4} \tag{7.9}$$

Further, if the element Z_5 is removed and terminals 2–2 created in its place, and an element Z_0 connected across terminals 1–1, it can also be shown that

$$Z_{22} = \frac{V_2}{I_2} = \frac{Z_2 Z_4 Z_0}{Z_1 Z_3} \tag{7.10}$$

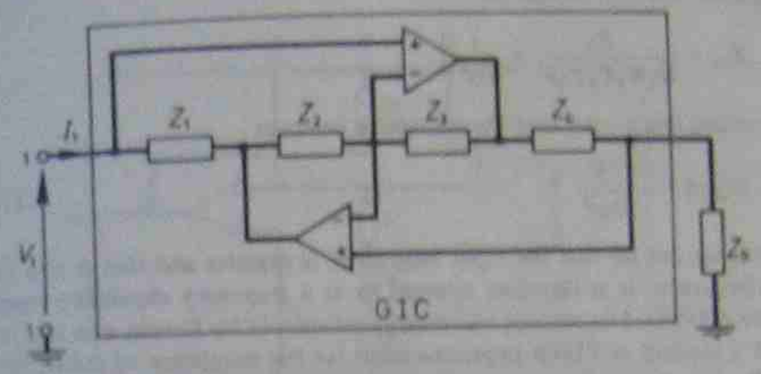


Figure 7.4 Circuit drawn as a GIC

The new circuit is shown in figure 7.5.

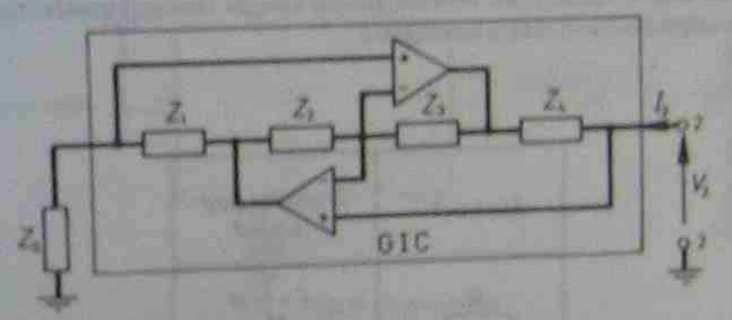


Figure 7.5 Circuit drawn as a GIC – termination Z_0

The GIC circuits of figures 7.4 and 7.5 are mainly due to Antoniou, and we shall see that both circuits have an important role to play in the simulation of so-called *synthetic inductors* associated with active realisations of passive circuits.

Consider the circuit shown in figure 7.4 and let element Z_4 be a capacitor with the remaining elements being pure resistors. Using equation (7.9) we find that

$$Z_{11} = \frac{R_1 R_3 R_5 C_4 s}{R_2} = L_{eq} s \tag{7.11}$$

It is seen that the GIC *converts* a resistor R_s into an equivalent inductor having the value given by equation (7.11). Similarly, using the circuit of figure 7.5 and equation (7.10) with Z_0 and Z_4 capacitors, the remaining elements being pure resistors, we find that

$$Z_{22} = \frac{R_2}{R_1 R_3 C_4 C_0 s^2} = \frac{1}{Ds^2}$$

Also noting that $s = j\omega$, then the expression becomes

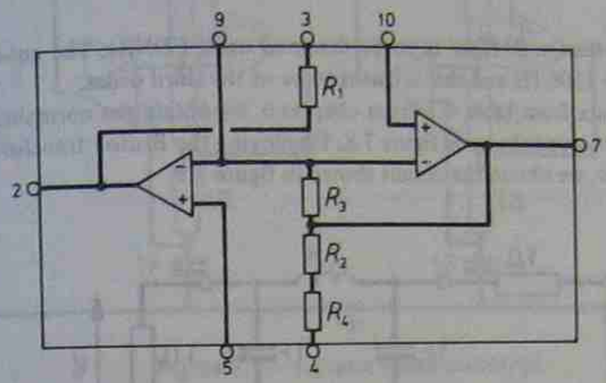
$$Z(j\omega) = -\frac{1}{D\omega^2} \tag{7.12}$$

Here we can see that the input impedance is negative and that it also varies with frequency. It is therefore referred to as a *frequency dependent negative resistor* (FDNR). The concept was initially introduced by Bruton who also introduced a method of FDNR implementation for the simulation of ladder filters. He also introduced the notion of scaling elements by a factor $(1/s)$ whereby resistors became inductors, inductors became resistors and capacitors became FDNRs. Using this idea it can be seen that by using such a factor, inductors are no longer described. The principle is outlined below and summarised in figure 7.6 where the symbol for an FDNR should be noted. It should also be noted that the Bruton transformed network should have an identical transfer function to the original network being investigated.

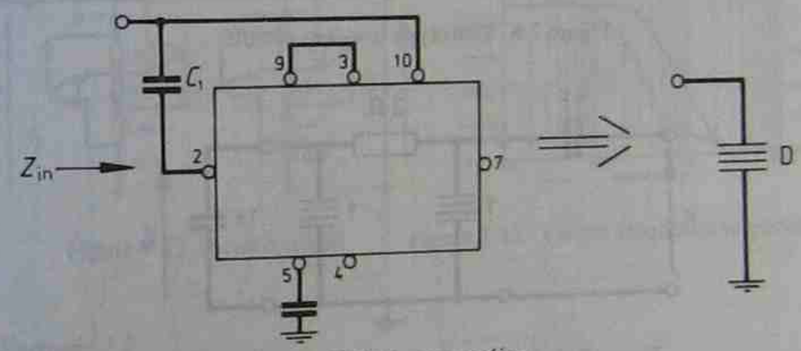
Element	Transformed element
R $Z = R$	$C = 1/R$ $Z' = R/s$
L $Z = sL$	$R = L$ $Z' = sL/s = L$
C $Z = 1/Cs$	$D = C$ $Z' = 1/Cs^2$

Figure 7.6 Transformed elements

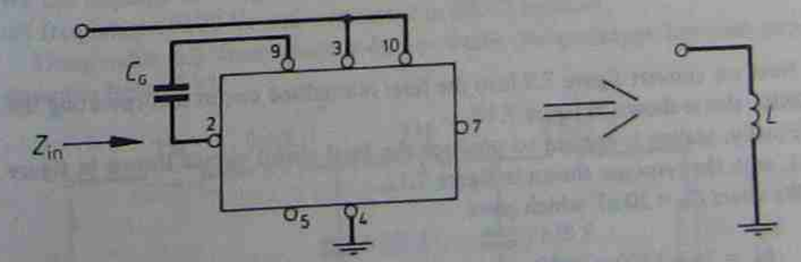
Gyrator module integrated circuits are produced; one such being manufactured by National Semiconductors is shown in figure 7.7.



Gyrator module



FDNR connection



Grounded inductor connection

Figure 7.7 Gyrator module schematic

Two examples will now be considered with which to illustrate the principles outlined in this chapter.

Example 7.1

A low pass Butterworth filter is to be designed using FDNRs. The cut-off frequency is to be 1500 Hz and the circuit is to be of the third order.

Using the data from table 6.1 from chapter 6, we obtain the normalised low-pass passive prototype shown in figure 7.8. Employing the Bruton transformation to each element, we obtain the circuit shown in figure 7.9.

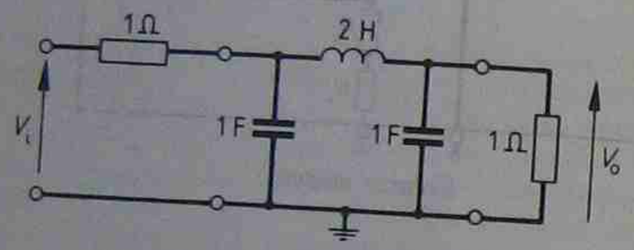


Figure 7.8 Prototype low-pass circuit

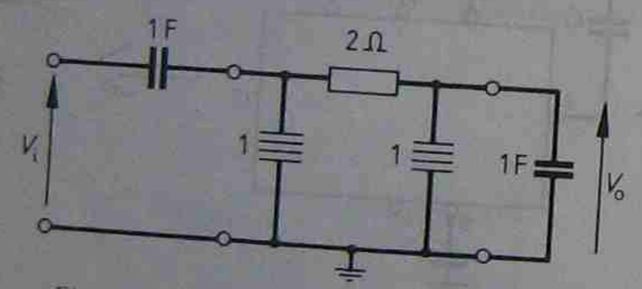


Figure 7.9 Bruton transformation of figure 7.8

Next we convert figure 7.9 into the final normalised circuit incorporating the FDNRs; this is shown in figure 7.10.

Finally, scaling is applied to produce the final scaled circuit shown in figure 7.11, with the response shown in figure 7.12.

We select $C_0 = 20$ nF which gives

$$k_f = 2\pi \times 1500 = 3000\pi$$

$$k_m = \frac{1}{3000\pi \times 20 \times 10^{-9}} = 5305$$

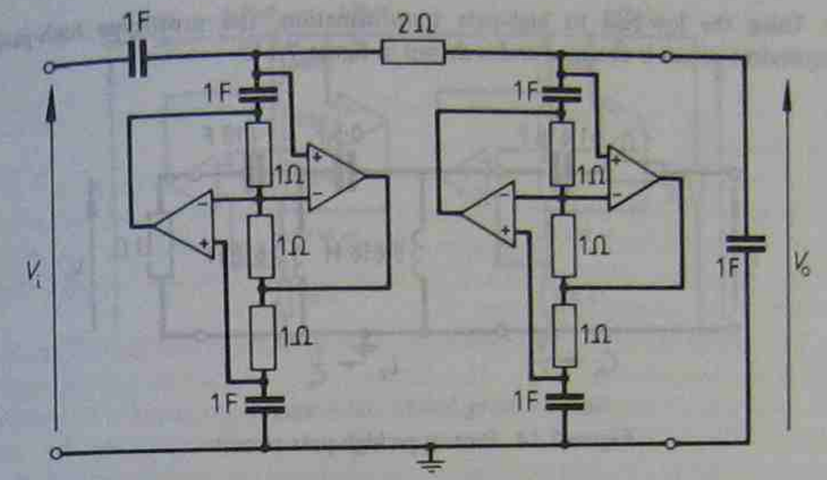


Figure 7.10 Gyrator-based prototype

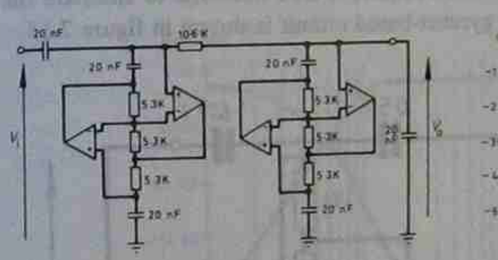


Figure 7.11 Scaled circuit

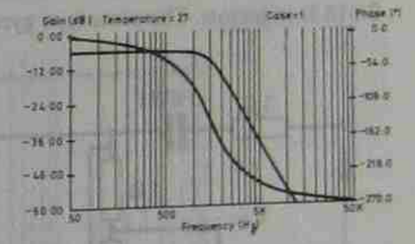


Figure 7.12 Circuit frequency responses

Example 7.2

We are required to design a fifth-order Butterworth high-pass filter having a cut-off frequency of 100 Hz and terminated in 600 Ω resistors.

Using table 6.1 from chapter 6, we obtain the prototype low-pass circuit shown in figure 7.13.

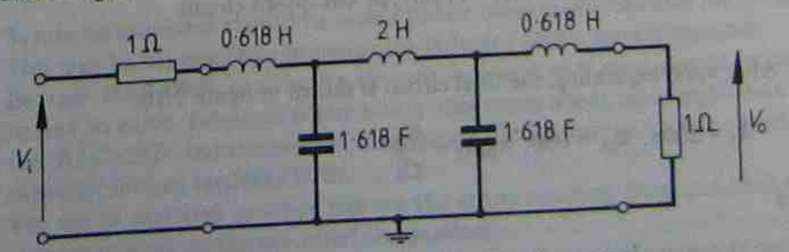


Figure 7.13 Prototype T-section low-pass

ST. GEORGE LIBRARY

Using the low-pass to high-pass transformation, the prototype high-pass equivalent circuit is obtained and is shown in figure 7.14.

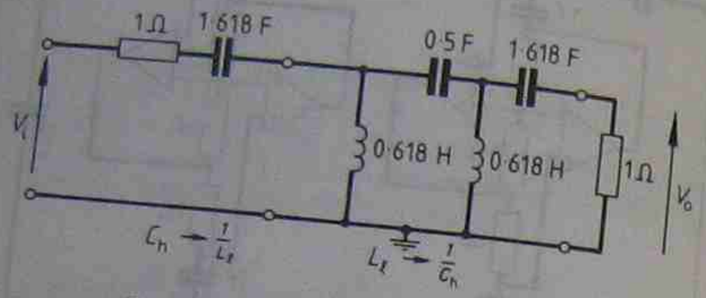


Figure 7.14 Prototype high-pass circuit

Clearly, it can be seen that we shall require a GIC network to simulate the 0.618 H inductors. The prototype gyrator-based circuit is shown in figure 7.15.

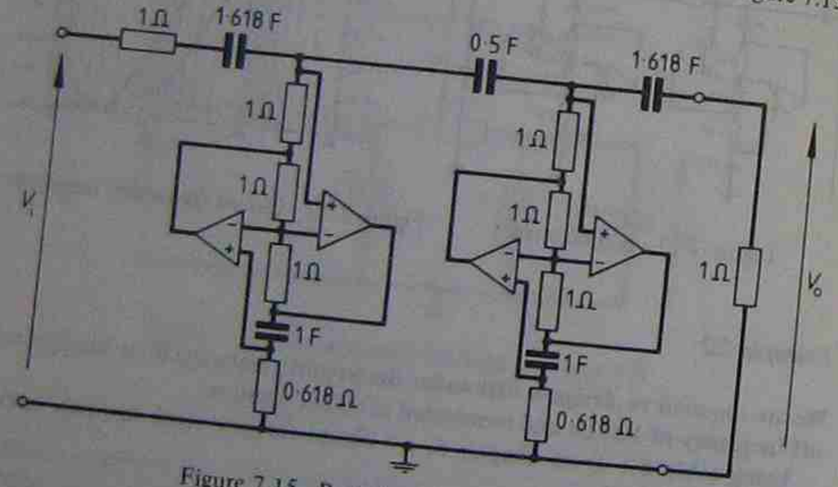


Figure 7.15 Prototype GIC-based circuit

After applying scaling, the final circuit is shown in figure 7.16.

$$k_f = 200\pi,\ k_m = 600,\ k_m k_f = \frac{C_n}{C_0}$$

and

$$C_0 = \frac{1}{200\pi \times 600} = 2.65\ \mu F$$

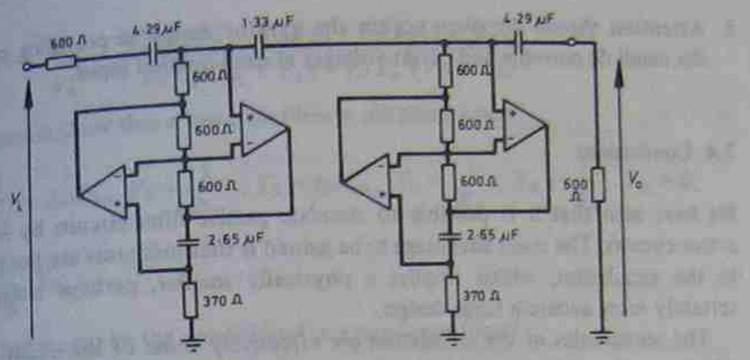


Figure 7.16 Scaled gyrator circuit

The frequency response of the circuit is shown in figure 7.17.

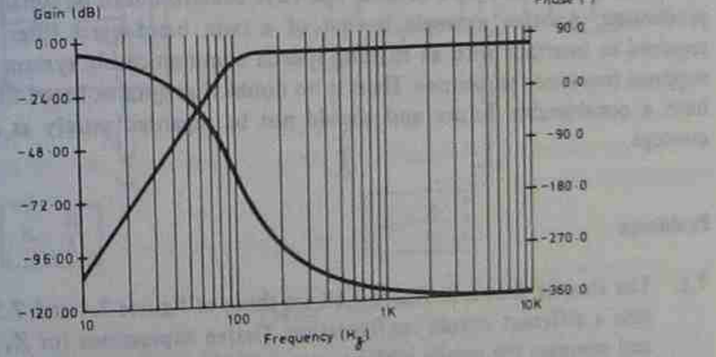


Figure 7.17

From a consideration of the two examples discussed, a number of points emerge which are relevant to the correct realisation of gyrator-based circuits.

1. Synthetic inductors should be tested before being connected into the circuit. This may be achieved by resonating the inductor with a known capacitor.
2. Because the FDNR is an active element, one end of it must always be connected to earth. Low-pass filters satisfy this requirement since application of the *RLC*:*CRD* transformation transforms the prototype passive, earthed capacitor into an earthed FDNR.
3. The use of matched op-amps reduces the errors resulting from non-ideal op-amp parameters on the simulated components.
4. Floating synthetic inductors can be made, but their performance is not as good as those which may be grounded.

5. Attention should be given within the gyrator design to providing paths for the small dc currents and offset voltages at each op-amp input.

7.4 Conclusions

We have seen that it is possible to simulate passive filter circuits by means of active circuits. The main advantage to be gained is that inductors are not involved in the simulation, which implies a physically smaller, perhaps cheaper and certainly more accurate filter design.

The sensitivities of the simulation are effectively those of the passive circuit which are themselves quite low. Also, band-pass circuits may be realised using the basic GIC circuits, but having modified terminating components. The technique is beyond the intended scope of this book, but suffice it to say that we may realise certain complex impedances by considering the terminations to be a combination of resistors and capacitors.

For the present, gyrator filters have been integrated on to a single chip for use as a video filter within a TV receiver and have been found to be suitable for mass production. Another example is that of a twin band-reject filter which was required to interface with an existing speech communication system in order to suppress two-tone frequencies. There is no doubt that gyrator-based filter circuits have a considerable future and should not be regarded purely as a 'novelty' concept.

Problems

7.1. The circuits shown in figure 7.18 are those of figures 7.4 and 7.5 re-drawn into a different circuit configuration. Derive expressions for Z_{11} and Z_{22} and compare the results with equations (7.9) and (7.10) in the text.

7.2. Using the data for a Butterworth filter from chapter 2, design a fourth-order filter using the techniques outlined in chapter 7 to have the following specification. Pass band to extend from dc up to 1 kHz with a maximum attenuation of 3 dB. Stop-band attenuation to be 80 dB down at 10 kHz. Use 0.01 μF capacitors in the design.
[*Ans.* The two FDNRs will contain 16 kΩ, 29 kΩ and 0.01 μF]

7.3. Using the prototype circuits obtained in example 6.3, obtain the gyrator circuit form to give the following specification: pass-band 1 kHz up to infinity with 1 dB of ripple; stop-band attenuation to be better than 80 dB at 100 Hz. Use a magnitude scaling factor of 1000.
[*Ans.* FDNR1: 1 kΩ, 1 kΩ, 1 kΩ, 159 nF, 940 Ω.
FDNR2: 1 kΩ, 1 kΩ, 1 kΩ, 159 nF, 1.27 kΩ]

7.4. The circuit shown in figure 7.19 is often attributed to Fliege and possesses low sensitivity. Using nodal analysis and the assumption that the op-amps are ideal, show that

$$\frac{V_o}{V_i} = \frac{Y_A(Y_1Y_3 - Y_0Y_2) + Y_BY_2(Y_4 + Y_5)}{Y_1Y_3(Y_A + Y_5) + Y_2Y_4(Y_B + Y_0)}$$

Further show that a low-pass filter is obtained when

$$Y_1 = \frac{1}{R_1} + j\omega C_1,\ Y_2 = \frac{1}{R_2},\ Y_3 = j\omega C_3,\ Y_4 = \frac{1}{R_4},\ Y_5 = \frac{1}{R_5},\ Y_A = 0,$$

$$Y_B = \frac{1}{R_B}\ Y_0$$

What would be the conditions for a band-pass filter?

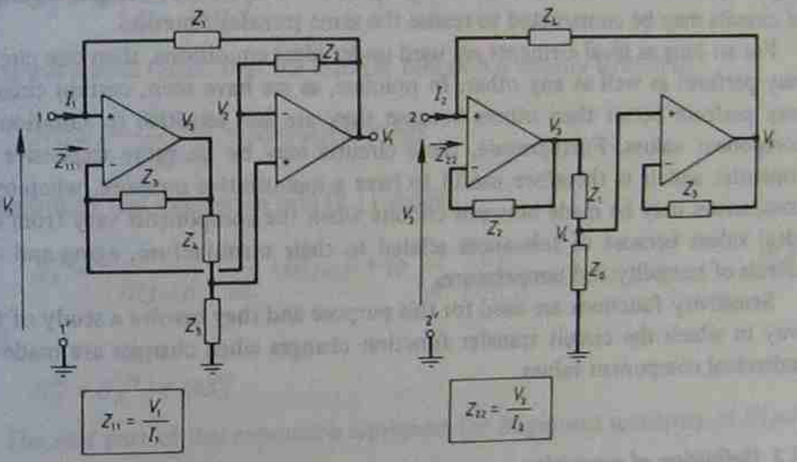


Figure 7.18

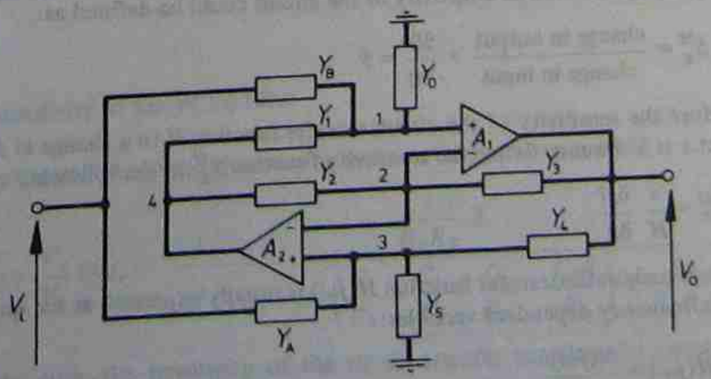


Figure 7.19

8 Sensitivity

8.1 Introduction

Circuits or networks are constructed to produce a specified input/output relationship and since the solution to the design problem may not be unique, a variety of circuits may be constructed to realise the same transfer function.

For so long as ideal elements are used under ideal conditions, then one circuit may perform as well as any other. In practice, as we have seen, certain circuits may perform better than others because they are less sensitive to variation in component values. Furthermore, these circuits may be no more expensive to construct and it is therefore useful to have a quantitative measure, whereby a comparison may be made between circuits when the components vary from the ideal values because of tolerances related to their manufacture, aging and the effects of humidity and temperature.

Sensitivity functions are used for this purpose and they involve a study of the way in which the circuit transfer function changes when changes are made to individual component values.

8.2 Definition of sensitivity

Consider a circuit in which a change of 10 per cent is made in one of the circuit components and assume that this change produces a change in the output of the circuit of 50 per cent. The sensitivity of the circuit could be defined as:

$$S_x^H = \frac{\text{change in output}}{\text{change in input}} = \frac{50}{10} = 5$$

Therefore the sensitivity of the circuit transfer function H to a change in a component x is 5. We may define the sensitivity function S_x^H in the following way:

$$S_x^H = \frac{x}{H}\frac{\partial H}{\partial x} \tag{8.1}$$

In circuit analysis the transfer function $H(j\omega)$ is usually expressed as the quotient of two frequency dependent variables:

$$H(j\omega) = \frac{N(j\omega)}{D(j\omega)}$$

which on differentiation yields:

$$\frac{\partial H}{\partial x} = \frac{N'D - D'N}{D^2} \tag{8.2}$$

where N', D' are the partial derivatives of N and D with respect to x. By combining equations (8.1) and (8.2) we obtain:

$$S_x^H = \frac{x}{H}\frac{\partial H}{\partial x} = \frac{x}{H}\left[\frac{N'D - D'N}{D^2}\right] \quad \text{and} \quad H = \frac{N}{D}$$

which finally produces the expression:

$$S_x^H = x\left[\frac{N'}{N} - \frac{D'}{D}\right] \tag{8.3}$$

It was stated earlier that the transfer function is usually written as

$$H(j\omega) = \frac{N(j\omega)}{D(j\omega)} = |H(j\omega)|\exp(j\phi)$$

Substituting this expression into (8.1) yields:

$$S_x^H = \frac{x}{|H(j\omega)|}\frac{\partial}{\partial x}|H(j\omega)| + j\phi\left(\frac{x}{\phi}\frac{\partial \phi}{\partial x}\right) \tag{8.4}$$

and

$$S_x^H = S_x^{|H|} + j\phi S_x^{\phi}$$

The *real* part of this expression represents the magnitude sensitivity of $H(j\omega)$ with respect to $x(S_x^{|H(j\omega)|})$. The *imaginary* part represents the phase sensitivity with respect to $x(S_x^{\phi})$. It should be noted that angle ϕ must be in radians. Also note that to obtain S_x^{ϕ}, the imaginary part of the expression must be divided by ϕ radians.

8.3 Sensitivity of the VCVS filter

From chapter 4 we re-call the transfer function for the filter, namely

$$H(j\omega) = \frac{V_o}{V_i}(j\omega) = \frac{\dfrac{1}{R_1R_2}K}{-\omega^2C_1C_2 + j\omega\left(\dfrac{C_1}{R_1} + \dfrac{C_1}{R_2} + \dfrac{C_2}{R_2} - \dfrac{C_2K}{R_2}\right) + \dfrac{1}{R_1R_2}}$$

Consider now the sensitivity of the circuit transfer function for variation of capacitor C_1; using equation (8.3) we obtain

$$S_{C_1}^H = C_1\left[\frac{N'}{N} - \frac{D'}{D}\right], \text{ where } N' = \frac{\partial N}{\partial C_1}, D' = \frac{\partial D}{\partial C_1}$$

$$S_{C_1}^H = C_1\left[\frac{0}{N} - \frac{\left\{-\omega^2 C_2 + j\omega\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\right\}}{D}\right]$$

$$S_{C_1}^H = \frac{\omega^2 C_1 C_2 - j\omega C_1\left(\frac{1}{R_1} + \frac{1}{R_2}\right)}{D}$$

Similarly for the remaining components:

$$S_{C_2}^H = \frac{\omega^2 C_1 C_2 - j\omega C_2 \frac{(1-K)}{R_2}}{D} \tag{8.5}$$

$$S_{R_1}^H = \frac{\omega^2 C_1 C_2 - j\omega\left\{\frac{C_1}{R_2} + \frac{C_2}{R_2}(1-K)\right\}}{D}$$

$$S_{R_2}^H = \frac{\omega^2 C_1 C_2 - j\omega\frac{C_1}{R_1}}{D}$$

where

$$D = -\omega^2 C_1 C_2 + j\omega\left\{\frac{C_1}{R_1} + \frac{C_1}{R_2} + \frac{C_2}{R_2}(1-K)\right\} + \frac{1}{R_1 R_2}$$

It can be seen from the expressions that the circuit sensitivity for each component varies with frequency and is not constant. To illustrate how sensitivity analysis may be used to predict changes in gain and phase caused by changes in component values, a practical example will now be considered.

Example 8.1

A VCVS filter design is shown in figure 8.1 which is to have a pass-band gain of 20 dB and a cut-off frequency of 10 kHz. The gain and phase responses are shown in figure 8.2 with the response to a −10 per cent change in C_1 also shown.

Finally, the magnitude and phase sensitivities are shown in figure 8.3 and values calculated using equations (8.5).

We wish to calculate:

(a) the change in gain at 6 kHz and
(b) the change in phase at 3 kHz when the capacitor value is decreased by 10 per cent.

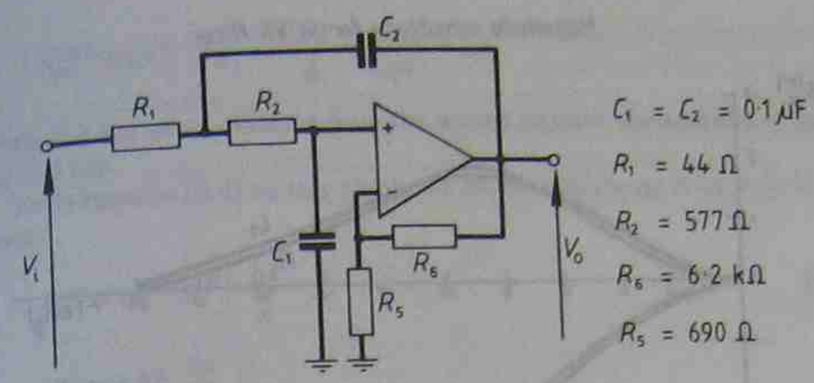


Figure 8.1 VCVS low-pass filter

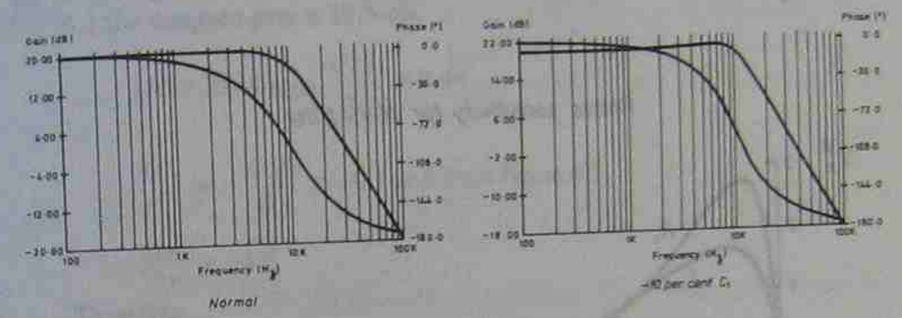


Figure 8.2 Gain/phase

Insertion of the component values into equation (8.5) gives the following sensitivities:

$$S_{R_1}^H = \frac{\omega^2 + j1.4 \times 10^5 \omega}{D}$$

$$S_{R_2}^H = \frac{\omega^2 - j2.23 \times 10^5 \omega}{D}$$

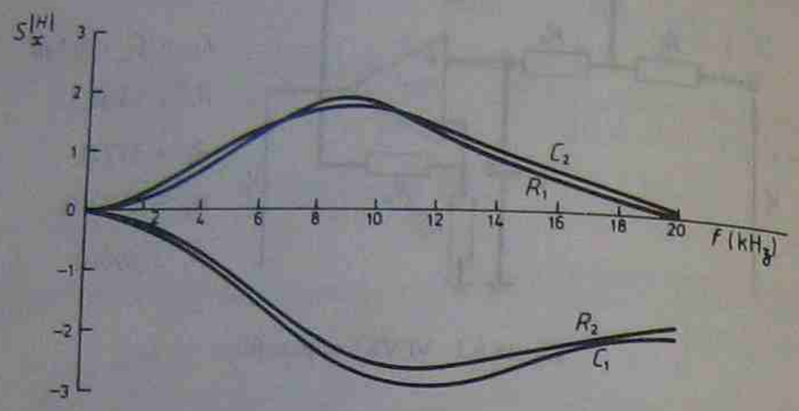
$$S_{C_1}^H = \frac{\omega^2 - j2.45 \times 10^5 \omega}{D}$$

$$S_{C_2}^H = \frac{\omega^2 + j1.56 \times 10^5 \omega}{D}$$

where

$$D = 3.94 \times 10^9 - \omega^2 + j8.86 \times 10^4 \omega$$

Magnitude sensitivity for VCVS filter



Phase sensitivity for VCVS filter

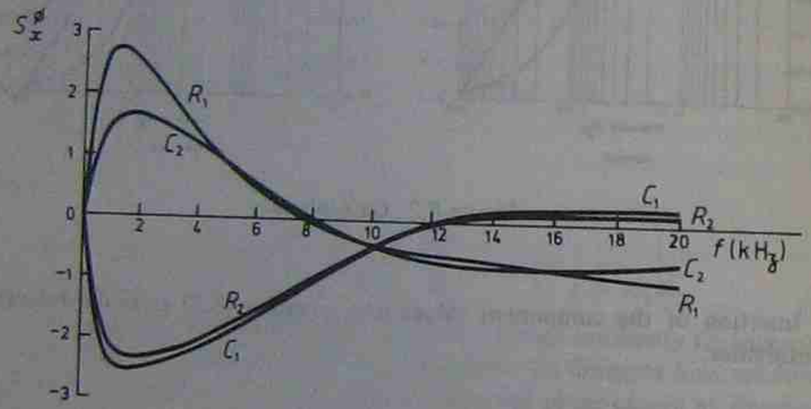


Figure 8.3 Magnitude/phase sensitivities

The next step is to consider a range of values for ω and obtain the component sensitivities for each value of ω. Finally, magnitude and phase sensitivity graphs are plotted as shown in figure 8.3. To illustrate the method of obtaining the sensitivities, we will consider the conditions for a frequency of 5 kHz.

$$S_{R_1}^{H} = \frac{(2\pi \times 5 \times 10^3)^2 + j1.4 \times 2\pi \times 5 \times 10^8}{(3.94 \times 10^9 - (2\pi \times 5 \times 10^3)^2) + j8.86 \times 2\pi \times 5 \times 10^7} = 1.11\angle 34.2^\circ$$

$S_{R_1}^{H} = 0.92 + j0.62$ which on using (8.4) gives

$$S_{R_1}^{|H|} = 0.92 \quad S_{R_1}^{\phi} = \frac{0.62}{\phi} = \frac{0.62}{\pi/4} = 0.8$$

where ϕ is the phase obtained from the *normal* response characteristic of figure 8.2 at 5 kHz.

From equation (8.4) we may obtain the incremental change in response in the form:

$$\partial H \approx |H|\, S_x^{|H|}\, \frac{\partial x}{x} \tag{8.6}$$

$$\partial \phi \approx \phi\, S_x^{\phi}\, \frac{\partial x}{x}$$

(a) From figure 8.2 we see that the gain at 6 kHz is 21 dB and also from figure 8.2 the designed gain is 19.5 dB.

$$|H| = \text{Antilog}_{10}\, \frac{19.5}{20} = 9.44$$

$$\frac{\partial C_1}{C_1} = -\frac{0.01}{0.1} = -0.1 \text{ and from figure 8.3}$$

$$S_{C_1}^{|H|} = -1.5$$

Therefore

$$\partial H \approx 9.44 \times -1.5 \times -0.1 = 1.416 \quad \text{giving}$$

$$|\hat{H}| = 9.44 + 1.416 = 10.86 \quad \text{or} \quad 20.8 \text{ dB}$$

This is good agreement with the value of 21 dB.

(b) Also from figure 8.2 we see that at $f = 3$ kHz, $\phi = 26^\circ$ and from figure 8.3 $S_{C_1}^{\phi} = -2.35$.

$$\partial\phi = 26 \times -2.35 \times -0.1 = 6^\circ$$

giving $\hat{\phi} = -26^\circ + 6^\circ = -20^\circ$ which gives good agreement with -21° from figure 8.2.

The two sensitivity plots show that the magnitudes $S_{C_2}^{|H|}$ and $S_{R_2}^{|H|}$ are both positive and have their largest values over a frequency range 5–14 kHz. Over this range of frequency we would expect the most marked effect on the gain response and would anticipate an increase in the gain. The converse is true for $S_{C_1}^{|H|}$ and $S_{R_1}^{|H|}$ where a reduction in gain is anticipated.

The phase sensitivity characteristics show a low negative value around 10 kHz; this suggests that within this frequency range the phase response should show very small deviation from the normal.

A further point to note from the two graphs is that the resistors (R_1, R_2) and capacitors (C_1, C_2) should be chosen to have identical temperature coefficients.

8.4 Sensitivity of K, Q and ω_0

Sensitivity analysis may also be applied to ascertain the variation of Q and ω_0 due to component variation within the circuit.

To illustrate the technique we will use the VCVS circuit of chapter 4, in which was stated the voltage transfer function:

$$\frac{V_o}{V_i}(j\omega) = \frac{\dfrac{1}{C_1C_2R_1R_2}K}{-\omega^2 + j\omega\left\{\dfrac{1}{C_2R_1} + \dfrac{1}{C_2R_2} + \dfrac{1}{C_1R_2} - \dfrac{K}{C_1R_2}\right\} + \dfrac{1}{C_1C_2R_1R_2}}$$

which may be conveniently re-written as

$$\frac{V_o}{V_i}(j\omega) = \frac{K\omega_0^2}{-\omega^2 + j\omega\left(\dfrac{\omega_0}{Q}\right) + \omega_0^2} \tag{8.7}$$

where

$$K = 1 + \frac{R_6}{R_5}, \quad \omega_0 = \frac{1}{\sqrt{(C_1C_2R_1R_2)}}, \quad \frac{\omega_0}{Q} = \frac{1}{C_2R_1} + \frac{1}{C_2R_2} + \frac{1}{C_1R_2}(1-K)$$

For convenience, let $P = \dfrac{\omega_0}{Q}$ and from equation (8.1):

$$S_x^{\omega_0} = \frac{x}{\omega_0}\frac{\partial\omega_0}{\partial x}, \quad S_x^Q = S_x^{\omega_0} - S_x^P \tag{8.8}$$

If we now apply equations (8.8) for each component of the filter, we obtain

$$S_{R_1}^{\omega_0} = \frac{R_1}{\omega_0}\frac{\partial\omega_0}{\partial R_1} = -\frac{R_1R_1^{-1}(C_1C_2R_1R_2)^{-\frac{1}{2}}(C_1C_2R_1R_2)^{\frac{1}{2}}}{2} = -\frac{1}{2}$$

Similarly for resistor R_2 and capacitors C_1 and C_2. Consider now variation of Q with each component.

$$S_{R_1}^P = \frac{R_1}{P}\frac{\partial P}{\partial R_1} = \frac{R_1}{P}\left(-\frac{R_1^{-2}}{C_2}\right) = -\frac{1}{C_2R_1P}, \text{ also}$$

$$P = \frac{\omega_0}{Q}, \text{ giving } S_{R_1}^P = -\frac{Q}{\omega_0C_2R_1} = \sqrt{\left(\frac{C_1R_2}{C_2R_1}\right)}\,Q$$

Using now the previous result for $S_{R_1}^{\omega_0}$ and combining with the result for $S_{R_1}^P$:

$$S_{R_1}^Q = S_{R_1}^{\omega_0} - S_{R_1}^P = -\frac{1}{2} + \sqrt{\left(\frac{C_1R_2}{C_2R_1}\right)}\,Q \tag{8.9}$$

A similar analysis will yield the sensitivities due to the remaining components.

Sensitivity values for the gain resistors may be found as follows by re-calling the expressions:

$$P = \frac{1}{C_2R_1} + \frac{1}{C_2R_2} + \frac{1}{C_1R_2}(1-K), \quad (1-K) = -\frac{R_6}{R_5}$$

$$P = \frac{1}{C_2R_1} + \frac{1}{C_2R_2} - \frac{R_6}{C_1C_2R_5}, \text{ also}$$

$$S_K^Q = S_K^{\omega_0} - S_K^P \text{ and } S_K^{\omega_0} = 0, \text{ therefore}$$

$$S_K^P = \frac{K}{P}\frac{\partial P}{\partial K} = -\frac{K}{P} = -\frac{KQ}{\omega_0} \text{ and}$$

$$S_K^Q = \frac{QK}{C_1R_2}\sqrt{(C_1C_2R_1R_2)} = KQ\sqrt{\left(\frac{C_2R_1}{C_1R_2}\right)} \tag{8.10}$$

We may also obtain the sensitivities of gain and Q-factor for variation of the gain adjustment resistors:

$$S_{R_6}^Q = S_{R_6}^{\omega_0} - S_{R_6}^P = 0 - \frac{R_6}{C_1R_2R_5P} = -Q\frac{R_6}{R_5}\sqrt{\left(\frac{C_2R_1}{C_1R_2}\right)}$$

$$= (K-1)Q\sqrt{\left(\frac{C_2R_1}{C_1R_2}\right)} \tag{8.11}$$

Similarly it can be shown that for variation of R_5 and R_6:

$$\left.\begin{aligned} S_{R_5}^Q &= -S_{R_6}^Q = -(K-1)Q\sqrt{\left(\frac{C_2R_1}{C_1R_2}\right)} \\ S_{R_6}^K &= \frac{K-1}{K} = -S_{R_5}^K \end{aligned}\right\} \tag{8.12}$$

It is convenient to take as many of the components equal to unity as is possible. Let us make $R_1 = R_2 = 1\ \Omega$, $C = C_2 = 1$ F and resistor $R_5 = 1\ \Omega$. To obtain the value for R_6 we use equation (8.7) and with $\omega_0 = 1$ rad/s to yield the following relationship:

$$3 - K = \frac{1}{Q} \quad \text{and } K = 1 + \frac{R_6}{R_5}$$

from which we obtain the value for R_6:

$$R_6 = 2 - \frac{1}{Q}$$

Imposing these conditions keeps the spread of component values as small as possible.

The sensitivity values shown in table 8.1 may now be constructed showing the sensitivity for each component for the condition $R_1 = R_2 = 1\ \Omega$, $C_1 = C_2 = 1$ F, $R_5 = 1\ \Omega$, $R_6 = 2 - \frac{1}{Q}\ \Omega$.

Table 8.1

Component	Sensitivity
R_1	$S^Q_{R_1} = -\frac{1}{2} + Q$; $S^{\omega_0}_{R_1} = -\frac{1}{2}$
R_2	$S^Q_{R_2} = \frac{1}{2} - Q$; $S^{\omega_0}_{R_2} = -\frac{1}{2}$
C_1	$S^Q_{C_1} = \frac{1}{2} - 2Q$; $S^{\omega_0}_{C_1} = -\frac{1}{2}$
C_2	$S^Q_{C_2} = -\frac{1}{2} + 2Q$; $S^{\omega_0}_{C_2} = -\frac{1}{2}$
R_5	$S^Q_{R_5} = 1 - 2Q$
R_6	$S^Q_{R_6} = 2Q - 1$
K	$S^K_{R_5} = -\frac{2Q-1}{3Q-1}$
	$S^Q_K = 3Q - 1$
	$S^K_{R_6} = \frac{2Q-1}{3Q-1}$

To illustrate only the usefulness of sensitivity values, consider for the circuit conditions outlined above that a certain design requires $Q = 10$.

Insertion of this value yields the following sensitivity values:

$$S^Q_{R_1} = 9.5,\ S^Q_{R_2} = -9.5,\ S^Q_{C_1} = -19.5,\ S^Q_{C_2} = 19.5,\ S^Q_K = 29,$$

$$S^Q_{R_5} = -19,\ S^Q_{R_6} = 19,\ S^K_{R_5} = -\frac{19}{29} = -S^K_{R_6}$$

This circuit realisation has the advantage that all the frequency selective components are easily chosen; the resistors are equal, as are also the capacitors. This makes the scaling of the circuit much easier to implement. The price to be paid is that all of the sensitivities are high, especially ($S^Q_K = 29$). The implication is that care must be taken in adjusting the gain K to obtain the specified Q-value. The high Q sensitivity with respect to the gain is certainly a disadvantage of the VCVS circuit as we saw in chapter 4. A simple example will suffice to highlight this problem.

Example 8.2

Using table 8.1, calculate the percentage change in Q and K for a ±5 per cent change in the values of the gain adjustment resistors. Consider the circuit has $Q = 10$.

The basic sensitivity expression is obtained from equation (8.1):

$$S^Q_{R_5} = \frac{Q}{R_5}\ \frac{\partial Q}{\partial R_5} \quad \text{etc.} \quad \text{from which}$$

$$\partial Q = S^Q_{R_5} \times \frac{R_5}{Q}\ \partial R_5$$

$$= -19 \times \frac{1}{10} \times \pm 5 = \pm 9.5 \text{ per cent} \quad \text{similarly}$$

$$\partial K = S^K_{R_5} \times \frac{R_5}{K} \times \partial R_5 = -\frac{19}{29} \times \frac{1}{\left(3 - \frac{1}{10}\right)} \times \pm 5 \text{ per cent} = 1.13 \text{ per cent}$$

$$\partial Q = \frac{29 \times \left(3 - \frac{1}{10}\right)}{10} \times \pm 1.13 = \pm 9.5 \text{ per cent}$$

The calculations show that for a ±5 per cent change in tolerance value of the gain adjustment resistors there is a ±1.13 per cent change in the gain and a ±9.5 per cent change in the Q-value.

8.5 Conclusions

To conclude this chapter we should remind ourselves that sensitivity analysis of circuits is an integral part of any analogue design, be it a filter or any other circuit. Moreover, in the final analysis we are interested in minimising the deviation of the filter response due to the incremental variation of some parameter within the circuit. For example, in a highly selective (high Q) circuit the sensitivity of the poles is an important factor to consider when studying the stability of the circuit. It is only by considering the relative sensitivities of the filter circuits available that a correct choice may then be made for a particular application.

The deviation of ω_0 and Q for a particular filter, for example, may be reduced by using an op-amp having a superior GB product. This may be fine for a single design, but commercial success depends on mass-production techniques and therefore we find the op-amp derivatives of the 741 still being used. We have also

discussed in chapter 1 the need to have quality components and it was mentioned that precision components could be manufactured using thin film techniques. It is also common practice to minimise the spread of component values, especially capacitance, since the value of the total capacitance determines the area of substrate used in hybrid integrated-circuit form electrical filters.

Problems

8.1. Using the general circuit for the MFB filter from figure 4.21 and the sensitivity technique outlined in chapter 8, obtain the component sensitivities for the circuit and compare with those for the VCVS circuit.

8.2. Obtain the sensitivities for the two circuits shown using the sensitivity equations derived in this section and compare with the values given in table 8.1.

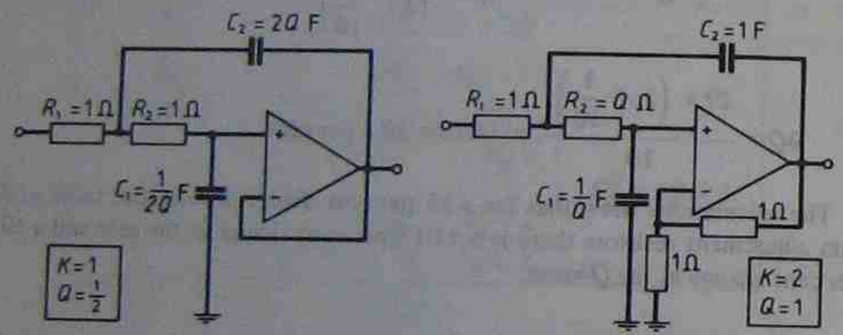


Figure 8.4

8.3. Using the graphs of figures 8.2, 8.3 etc. obtain the changes in gain and phase for the circuit of figure 8.1 for the following component variations:

(a) +10 per cent increase in R_2 at 6 kHz
(b) −10 per cent decrease in R_2 at 5 kHz
(c) ±10 per cent change in R_1 at 9 kHz
(d) +10 per cent increase in C_1 at 1 kHz
(e) −10 per cent decrease in C_1 at 10 kHz
(f) +10 per cent increase in C_2 at 5 kHz
(g) −10 per cent decrease in C_2 at 7 kHz.

Appendix: The Use of Computer Aided Design Methods

A.1 Introduction

This brief appendix deals with the formulations and methods that can be implemented for an ever-increasing number of design situations where the classical synthesis method (analytical or numerical) is considered to be inappropriate.

The modern approach to the problem is to use efficient, iterative, automatic optimisation methods to achieve a design which meets or may even exceed certain requirements. As soon as such realities as active devices, parasitic effects, non-linearities, high-frequency operation etc. are to be considered, then the classical method can, at best, only give an approximate solution to the design problem.

Generally, we could argue against the derivation of a computer algorithm for a specific filter design, since criteria such as how often the algorithm may be used and how much time is taken in its development must be overriding factors.

There are a number of CAD packages available to the designer; two of the most popular with relative novices are SPICE and MICROCAP. Quite a number of the filter circuits contained within the book were initially checked for correctness of response by means of the CAD facility. It must be stressed here that the formalities of the design were performed as in the preceding chapters and not by a computer algorithm.

A.2 The CAD approach – the MICROCAP way

The author is indebted to Spectrum Software (Sunnyvale, California, USA) for permission to include this brief overview of the MICROCAP approach to analogue circuit design.

MICROCAP is an interactive analogue circuit design and analysis tool which functions as shown in figure A.1. It allows the user to draw a circuit diagram on a CRT and automatically creates a netlist which is suitable for simulation or analysis directly from the diagram that has been drawn.

The diagrams are drawn using familiar components that are stored within a library of device models; for example, op-amps, diodes, resistors, transformers, switches, transistors (MOSFET and bipolar), etc. Running the program allows a choice from four analytical modes – ac, dc, transient and Fourier.

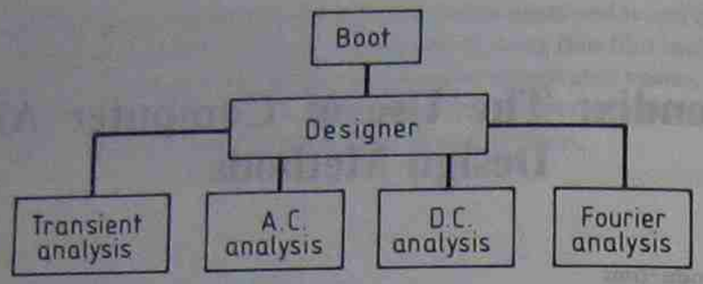


Figure A.1

Finally, the circuit diagram and the analysis may be dumped to a printer in graphic or tabular form. Alternatively, a plotter may be used to display the recorded and stored data.

Simply in order to illustrate the use of a CAD tool such as MICROCAP, we will consider a simple design example. In section A.3 we will further consider a laboratory exercise to obtain the response of the chosen design.

The design specification is that a band-pass filter is required, having a bandwidth of 500 Hz, a centre frequency of 1 kHz and a centre frequency gain of 20 dB. The actual filter design is covered in chapter 4, example 4.7, and the component values are listed as:

$$R_1 = 3.14\ \text{k}\Omega,\ R_2 = 15.9\ \text{k}\Omega,\ R_3 = 47.7\ \text{k}\Omega,\ C_1 = 0.01\ \mu\text{F},\ C_2 = 0.02\ \mu\text{F}$$

The circuit is drawn as shown in figure A.2, and the ac analysis mode is selected. This initially attributes node numbers to nodal points within the circuit. A prologue of condition is then presented to the user, as shown in table A.1; it is at this point that the user must insert the appropriate data relevant to the design, in order for a correct analysis of the circuit to be achieved. For example, in the design under consideration, the input node is 1 and the output node is 5. Notice also that the number of cases taken is 3. This facility permits the use of Worst-Case Analysis procedures based on the component tolerances. On acknowledging 'Yes' to the final prompt, the program will be run by connecting a constant-amplitude, variable-frequency source to the input node, the response being displayed on the CRT screen and shown in figure A.3.

A further prompt will ask whether another RUN is required, whether to DUMP the results to a printer, or if the user wants to return to the circuit in order to make possible alterations. Clearly, it can be seen that even by using a relatively unsophisticated CAD technique, a considerable time-saving facility is available to the designer.

A typical library of devices is shown in figure A.4, along with a specification for the op-amp used in the analysis.

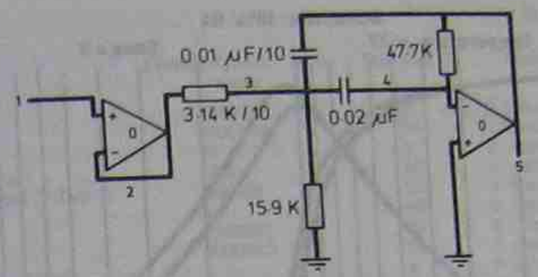


Figure A.2

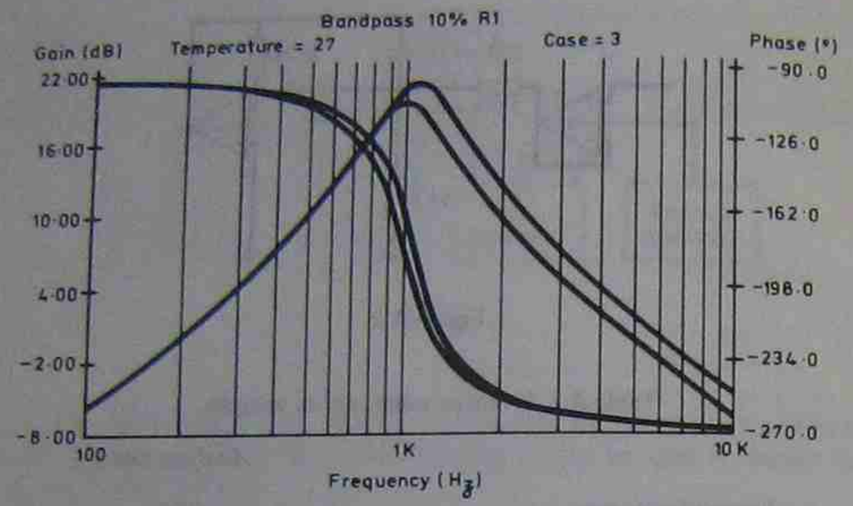
Table A.1 Prologue values per ac analysis

	Analysis limits
Lowest frequency	100
Highest frequency	1E4
Lowest gain (dB)	–8
Highest gain (dB)	22
Lowest phase shift	–270
Highest phase shift	–90
Lowest group delay	1E-9
Highest group delay	1E-4
Input node number	1
Output node number	5
Minimum accuracy (%)	5
Auto or Fixed frequency step (A, F)	A
Temperature (Low/High/Step)	27
Number of cases	3
Output: Disk, Printer, None (D, P, N)	N
Save, Retrieve, 'Normal run (S, R, N)'	N
Default plotting parameters (Y, N)	Y
Are these correct (Y, N)	Y

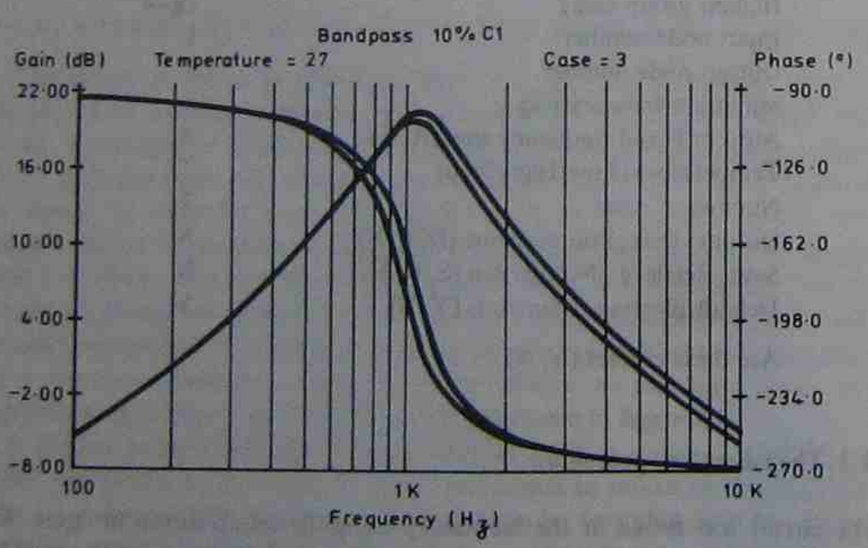
A.3 The laboratory procedure

The circuit was tested in the laboratory using the set-up shown in figure A.5. Because of the variation of input impedance of the filter over the working frequency range, a buffer amplifier was used between the source and the filter.

The results are presented in figure A.6 and may be compared favourably with those from the CAD prediction.



Frequency = 100·00000 D+02 Hz Gain = −4·925 dB
Phase angle = −267·678 Degrees Group delay = 994·43710 D-09 Sec
Gain slope = -613·66835 E-02 dB/OCT Peak gain = 21·023 dB/F
= 104·5+01 Hz



Frequency = 100·00000 D+02 Hz Gain = −6·718 dB
Phase angle = −268·992 Degress Group delay = 124·88288 D-88 Sec
Gain slope = -610·56437 E-02 dB/OCT Peak gain = 19·817 dB/F
= 965·0+01 Hz

Figure A.3 Band-pass responses for changes in R_1 and C_1

	Opamps Type 0.. Alias Value	Tolerance (%)
0:Input resistance	1000000	0
1:Open loop gain	2000000	50
2:Output resistance	75	0
3:Offset voltage (Voffset)	.005	50
4:Temp coeff. of Voffset (V/Deg C)	.000005	50
5:First pole (HZ)	5	40
6:Second pole (HZ)	1000000	60
7:Slew rate (V/Sec)	500000.2	30
8:Input offset current (Ioffset)	0	0
9:Input bias current	0	0
10:Current doubling interval (Deg. C)	0	0

E:Edit J:Jump N:Next L:Last C:Copy A:Alter alias Q:Quit

Standard Components

Standard Components Library PDC

0 :Opamps
1 :Diodes
2 :Bipolar transistors
3 :MOS Transistors
4 :Programmable waveforms
5 :Sinusoidal sources
6 :Transformers
7 :Polynomial sources
8 :Printer copy of library
9 :Passive component labels
10:Retrieve a library
11:Save library
12:Rename current library
13:Quit
14:Change colors

Your choice?

Figure A.4 Op-amp specification and component library

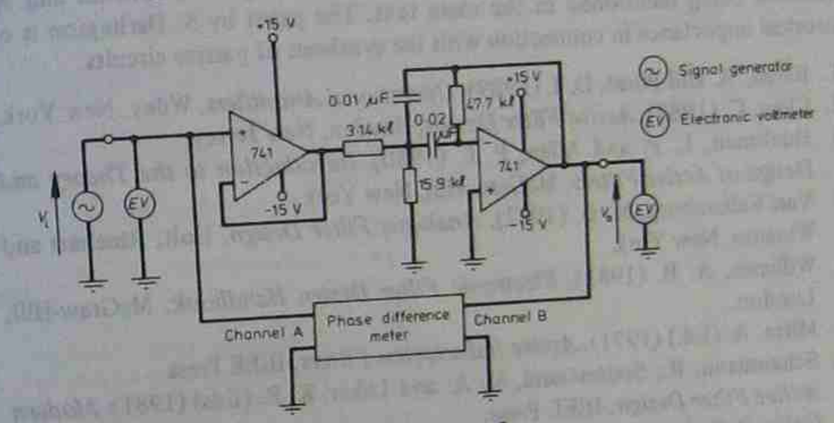


Figure A.5

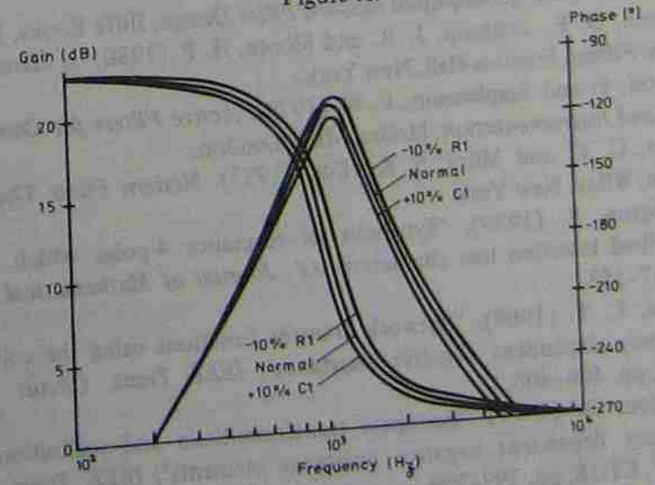


Figure A.6

Bibliography

Included is a short list for further reading. The first eleven entries are books and they are all excellent in their coverage of material. The book by A. B. Williams is very comprehensive and that by M. E. Van Valkenburg highly readable. The book by D. E. Johnson, J. R. Johnson and H.P. Moore contains charts and tables and the book by P. R. Geffe, although rather venerable, still contains good practical advice on passive circuits.

The remaining entries are technical papers, those by L. T. Bruton and A. Antoniou being mentioned in the main text. The paper by S. Darlington is of historical importance in connection with the synthesis of passive circuits.

1. Barna, A. and Porat, D. I. (1989). *Operational Amplifiers*, Wiley, New York.
2. Chen, C. (1982). *Active Filter Design*, Hayden, New Jersey.
3. Huelsman, L. P. and Allen, P. E. (1980). *Introduction to the Theory and Design of Active Filters*, McGraw-Hill, New York.
4. Van Valkenburg, M. E. (1982). *Analogue Filter Design*, Holt, Rinehart and Winston, New York.
5. Williams, A. B. (1981). *Electronic Filter Design Handbook*, McGraw-Hill, London.
6. Mitra, A. (Ed.) (1971). *Active Inductorless Filters*, IEEE Press.
7. Schaumann, R., Soderstrand, M. A. and Laker, K. R. (Eds) (1981). *Modern Active Filter Design*, IEEE Press.
8. Geffe, P. R. (1964). *Simplified Modern Filter Design*, Iliffe Books, London.
9. Johnson, D. E., Johnson, J. R. and Moore, H. P. (1980). *A Handbook of Active Filters*, Prentice-Hall, New York.
10. Bowron, P. and Stephenson, F. W. (1979). *Active Filters for Communications and Instrumentation*, McGraw-Hill, London.
11. Temes, G. C. and Mitra, S. K. (Eds) (1973). *Modern Filter Theory and Design*, Wiley, New York.
12. Darlington, S. (1939). 'Synthesis of reactance 4-poles which produce prescribed insertion loss characteristics', *Journal of Mathematical Physics*, pp. 257–353.
13. Bruton, L. T. (1969). 'Network transfer functions using the concept of frequency dependent negative resistance', *IEEE Trans. Circuit Theory*, **CT-16**, pp. 406–408.
14. Antoniou, A. (1971). 'Bandpass transformations and realisations using frequency dependent negative resistance elements', *IEEE Trans. Circuit Theory*, **CT-18**, pp. 297–299.

15. Temes, G. C., Orchard, H. J. and Jahanbegloo, M. (1978). 'Switched-capacitor filter design using the bilinear z-transform', *IEEE Trans. Circuit Syst.*, **CAS-25**, pp. 1039–1044.
16. Fidler, J. K. and Nightingale, C. (1979). 'Shape normalised sensitivity: a new sensitivity measure', *Electronics Letters*, **15** (2).
17. Waters, A. and Newsome, J. P. (1985). 'The input impedance of voltage controlled voltage source active filters', *International Journal Elec. Eng. Educ.*, **22**, pp. 63–68 (Manchester University Press).
18. Schaumann, R. (1975). 'Low-sensitivity high frequency tunable active filters without external capacitors', *IEEE Trans. Circuits and Systems*, **CAS-22** (1).

Index

algorithm
  CAD 153
  mathematical 2
amplifier
  buffer 31
  input impedance of 31
  instrumentation 34, 35
  inverting 29
  non-inverting 30
  output impedance of 31
analogue 1
  CAD tool 153
  computing 83, 85
  sampled data 97
analysis
  ac 153
  dc 153
  transient 153
  worst-case 154
Antoniou 133
attenuation 12
  infinite 94
  pass-band 15, 19
  stop-band 15, 20
  transition band 20

band pass 5
  all-resistor 107
  biquad 86, 88
  MFB 72
  passive 126, 127
  VCVS 57
band reject 5
  biquad 91
  high-pass notch 96
  low-pass notch 96
  normal notch 96
bandwidth
  3 dB 8, 94
  of MFB filter 62
  of op-amp 29, 32
biquadratic, definition 83
Bruton transformation 134
Butterworth approximation 12
  amplitude response 14
  maximally flat condition 8, 14
  phase response 24
  table of polynomials 17

capacitance, parasitic 104
capacitors
  capacitor ratio values 97
  ceramic 4
  film (mica, tantalum, aluminium) 3
carrier 11
cascaded filter sections 75, 80
CCCS (current controlled current source) 109
CCVS (current controlled voltage source) 109
Chebyshev approximation 18, 19
  amplitude response 20
  phase response 24
  table of coefficients 26
  table of polynomials 23
clock
  cycle 101
  frequency 99, 101, 105
  rate 104
  two-phase MOSFET 97
complex variable 4

Darlington 1, 114
decade 8
decibels 8
delay, group, phase 11
digital filters 1–2

FDNR (frequency dependent negative resistor) 134–6, 139

gain
  bandwidth product 32–3, 105, 109
  centre frequency 58
  of op-amp 27

GIC (general impedance converter) 132–4, 138, 140
gyration conductance 130
gyrator
  definition 130
  general 130–2
  module 135

half-power frequency 21
high-pass
  response of second-order MFB 70–1
  response of second-order VCVS 54–7

maximum power 115
mechanical filter 2
MICROCAP 66, 153–4
MOS switch 104
MOSFET 97
  two-phase clock 98

network, loss-less 115
nodal analysis 34, *see also* chapters 3, 4 and 5
normalising 8

offset
  dc offset 34
  of first-order VCVS 41–2
  of second-order VCVS 47–8
op-amp (operational amplifier) 2, 3, 27
  ideal 27
  non-ideal 32
  one-pole roll-off model 32

piezo-crystal filter 2
poles
  Butterworth 22
  Chebyshev 22
polynomial
  Butterworth 16
  Chebyshev 19
prologue, MICROCAP 154–5
prototype *see under each filter heading*

quad op-amp 27, 83, 91
quality-factor (Q-factor) 4, 44, 58, 90, 105

reactance transformation 123
reflection coefficient 115
resistor, carbon composition, carbon film, metal film 3
ripple
  factor 14–15, 19
  width 16, 20–1, 25

Sallen and Key 43
SAW (surface acoustic wave) 2
scaling
  frequency factor 8–9
  magnitude factor 8–9
sensitivities
  definition 142
  to Q, K, ω_0 148
short-circuit
  gyrator y characteristics 131
  protection against 27
  unity gain buffer 31
simulation 130, 153
slew rate 33, 109
SPICE 66, 153
switch
  off/on resistance 104
  SPDT 98

Tellegan 130
terminations of passive filter 117
transfer function 4, 6
transformation
  frequency 123
  low-pass to high-pass and band-pass 124–7
  reactance 123
tuning
  iterative 90
  orthogonal 90

universal filter *see* quad op-amp 91

voltage controlled voltage source (VCVS), ideal 29, 39, 62, 80, 143, 148

This book will appeal to students studying active filter design, probably as part of an advanced course in signal processing. It will also provide a useful introduction for postgraduate MSc courses.

The author's approach is based on analysis, followed by synthesis, the aim being to introduce the concept of filter response through traditional circuits and then progress to the increasingly popular gyrator and switched capacitor concepts.

Recognising the need to include an appraisal of circuit behaviour when subjected to non-ideal components, the coverage concludes with a chapter on sensitivity analysis.

All chapters contain worked examples; also included are graded problems (with answers).

As the subject matter lends itself to CAD simulation, an appendix outlines a technique that can assist in filter design.

After serving an apprenticeship in the steel industry, **Allan Waters** worked with Brush Electrical Engineering before embarking on an academic career. He is now a senior lecturer in the Department of Electrical and Electronic Engineering at Nottingham Polytechnic.

ISBN 0-333-48862-8

9 780333 488621