NEAR EAST UNIVERSITY

Faculty of Engineering

Department of Computer Engineering

ERROR DETECTION CORRECTION ALGORITHMS

Graduation Project
COM- 400

Student: Mohammed N. JARRAR
(992124)

Supervisor: Mr. Jamal FATHI

Nicosia - 2003

Dedicated to My Mom

ACKNOWLEDGMENTS

I could not have prepared this assignment without the generous help of my supervisor, colleaques, friends, and family.

First, I would like to thank my supervisor Mr. Jamal Fathi for his invaluable advice and belief in my work and myself. He provided valuable advice at each stage of the preparation of this assignment.

I also want to thank my friends in NEU being with them make my 4 years in NEU full of fun.

Finally, I want to thank my family, especially my parents, brother and sister. Without their endless support and love for me, I would never achieve my current position. I wish my mother lives happily always, and my father to be proud of me.

ABSTRACT

As a signal propagates along a transmission medium its amplitude decreases. This is known as signal attenuation. To compensate the attenuation, amplifiers are inserted at intervals along the cable to restore the received signal to its original level. Signal attenuation increases as a function of frequency. To overcome this problem, the amplifiers are designed to amplify different frequency by varying gains of amplifications. These devices are known as equaliser. For guided media (Twisted wires, Coaxial cables and Fiber optic cables) attenuation, is generally logarithmic and it is typically expressed as a constant number of decibels per unit distance.

Whenever an electromagnetic signal flows from one point to another, it is subject to unpredictable interference from heat, magnetism, and other forms of electricity. This interference can change the shape or timing of the signal. If the signal is carrying encoded binary data, such changes can alter the meaning of the data, changing 0 to 1 or 1 to 0, bits can be changed singly or in clumps.

CONTENTS

1. SIGNALS

1.1 Analog and Digital

Both data and the signals that represent them can take either analog or digital form. Analog refers to something that is continuous- a set of specific points of data and all possible points between. Digital refers to something that is discrete. Time is an analog quantity. It is a continuous stream that can be divided up into quarters, hundredths, thousandths, and so on. The measurement of time, however, can be either analog or digital. The hands of a traditional, or analog, clock do not jump from minute to minute or hour to hour; they move smoothly through all possible intermediate subdivisions of a 12-hour period.

Information can be analog or digital. Analog information is continuous. Digital information is discrete.

Digital and analog information can be distinguished by how we think about and refer to them. Analog quantities are generally described using various units of measure, while digital quantities are counted.

We use measuring units for analog quantities; for example, the length of a room can be 12 feet. We count digital quantities; for example, the number of students in a class can be 56.

Like the information they represent, signals can be either analog or digital. An analog signal is a continuous wave form that changes smoothly over time. As the wave moves from value A to value B, it passes through and includes an infinite number of values along its path. A digital signal, on the other hand, is discrete. It can have only a limited number of defined values, often as simple as 1 and 0. The transition of a digital signal from value is instantaneous, like a light being switched on and off.

We usually illustrate signals by plotting them on a pair of perpendicular axes. The vertical axis represents the value or strength of a signal. The horizontal axis represents the passage of time. Figure 1.1 illustrates an analog and a digital signal. The curve representing the analog signal is smooth and continuous, passing through an infinite number of points. The vertical lines of the digital signal, however, demonstrate the sudden jump the signal makes from value to value; and its flat highs and lows indicate that those values are fixed. Another way to express the difference is that the analog signal changes continuously with respect to time, while the digital signal changes instantaneously.

Signals can be analog or digital. Analog signals can have any value in a range; digital signals can have only a limited number of values.

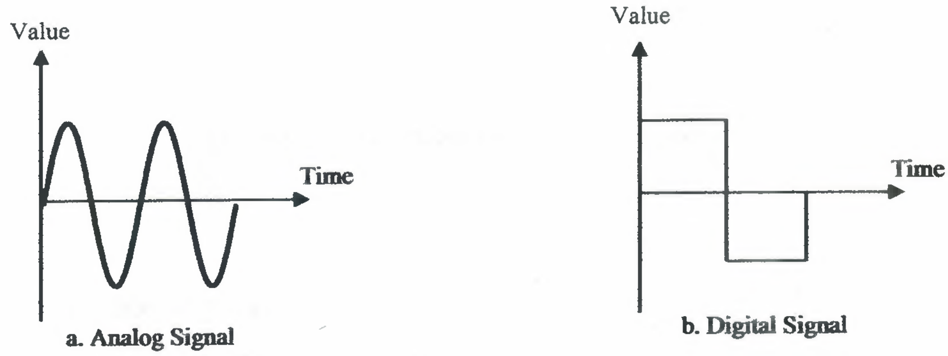


Figure 1.1 Comparison of Analog and Digital Signals

1.2 Aperiodic and Periodic Signals

Both analog and digital signals can be of two forms: periodic and aperiodic.

1.2.1 Periodic Signals

A signal is periodic if it completes a pattern within a measurable time frame, called a period, and repeats that pattern over identical subsequent periods. The completion of one full pattern is called a cycle. A period is defined as the amount of time (expressed in seconds) required to complete one full cycle. The duration of a period, represented by T, may be different for each signal, but is constant for any given periodic signal. Figure 1.2 illustrates hypothetical periodic signal.

A periodic signal consists of a continuously repeated pattern. The period of a signal (T) is expressed in seconds.

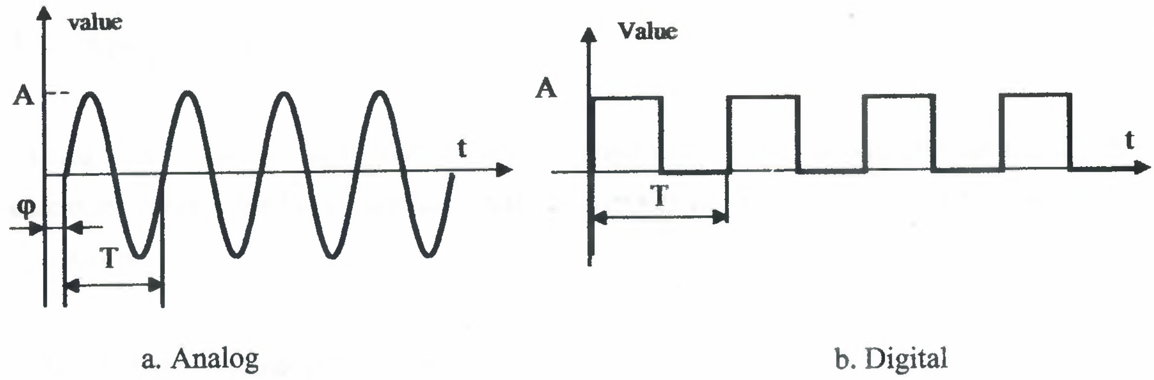


Figure 1.2 Examples of Periodic Signals

1.2.2 Aperiodic Signals

An aperiodic, or nonperiodic, signal changes constantly without exhibiting a pattern or cycle that repeats over time. Figure 1.3 shows examples of aperiodic signals.

An aperiodic, or nonperiodic, signal has no repetitive pattern.

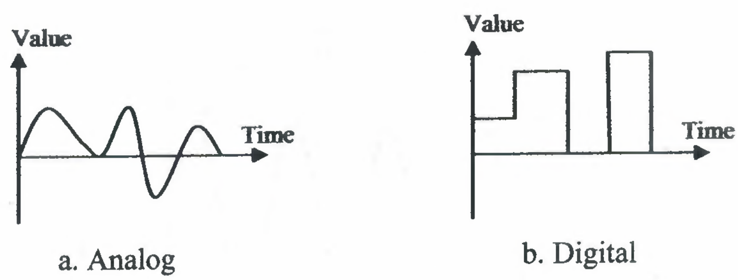


Figure 1.3 Examples of Aperiodic Signals

An aperiodic signal can be decomposed into an infinite number of periodic signals. A sine wave is the simplest periodic signal.

1.3 Analog Signals

Analog signals can be classified as simple or complex. A simple analog signal, or a sine wave, cannot be decomposed into simpler signals. A complex analog signal is composed of multiple sine waves.

1.3.1 Simple Analog Signals

The sine wave is the most fundamental form of a periodic analog signal. Visualized as a simple oscillating curve, its change over the course of a cycle is smooth and consistent, a continuous, rolling flow. Figure 1.4 shows a sine wave. Each cycle consists of a single arc above the time axis followed by a single arc below it. Sine waves can be fully described by three characteristics:

1. Amplitude
2. Period or frequency
3. Phase

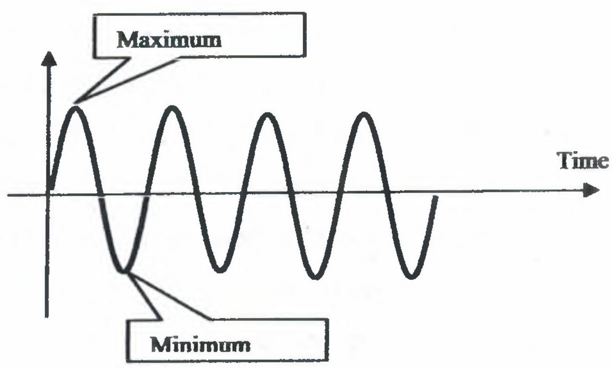


Figure 1.3 A Sine Wave

1. Amplitude

On a graph, the amplitude of a signal is the value of the signal at any point on the wave. It is equal to the vertical distance from a given point on the wave form to the horizontal axis. The maximum amplitude of a sine wave is equal to the highest value it reaches on the vertical axis.

Amplitude is measured in either volts, amperes, or watts, depending on the type of signal. Volts refers to voltage; amperes refers to current; and watts refers to power.

Amplitude refers to the height of the signal. The unit for amplitude depends on the type of the signal.

2. Period and frequency

Period refers to the amount of time, a signal needs to complete one cycle. Frequency refers to the number of periods a signal makes over the course of one second. The frequency of a signal is its number of cycle per second. Mathematically, the relationship between frequency and period is that they are the inverse of each other, if one is given, the other can be derived.

Frequency = 1/period   Period = 1/frequency

Period is the amount of time it makes a signal to complete one cycle; frequency is the number of cycle per second. Frequency and period are inverse of each other: $f = 1/T$ *and* $T = 1/f$.

Unit of Frequency Frequency is expressed in Hertz (Hz), after the German physicist Heinrich Rudolf Hertz. The communication industry uses five units to measure frequency: Hertz (Hz), Kilohertz (KHz = 10^3 Hz), Megahertz (MHz = 10^6 Hz), Gigahertz (GHz = 10^9 Hz), and Treahertz (THz = 10^{12} Hz). See Table 1.1.

Unit of period Period is expressed in second. The communication industry uses five units to measure period: second (s), millisecond (ms = 10^{-3} s), microsecond (m = 10^{-6} s), nanosecond (ns = 10^{-9} s), and picosecond (ps = 10^{-12} s). See table 1.1.

Table 1.1 Unit of Frequency and Period

Frequency		Period	
Unit	Equivalent	Unit	Equivalent
Hertz (Hz)	1 Hz	Second (s)	1s
Kilohetrz (KHz)	10^3 Hz	Millisecond (ms)	10^{-3} s
Megahertz (MHz)`	10^6 Hz	Microsecond (ms)	10^{-6} s
Gigahertz (GHz)	10^9 Hz	Nanosecond (ns)	10^{-9} s
Terahertz (THz)	10^{12} Hz	Picosecond (Ps)	10^{-12} s

Example 1.1

A sine wave has a frequency of 8 KHz. What is its period?

Solution

Let *T* be period and *f* be the frequency. Then,

$$T = 1/f = 1/8{,}000 = 0.000125 = 125 \text{ m}$$

Example 1.1

A sine wave complete one cycle in 25 m. What is its frequency?

Solution

Let T be the period and *f* be the frequency?

$$f = 1/T = 1/(25¥\ 10^{-6}) = 40{,}000 = 40 \text{ KHz}$$

3. Phase

The term phase describes the position of the waveform relative to time zero. If we think of the wave as something that can be shifted backward or forward along the time axis, phase describes the amount of that shift. It indicates the status of the first cycle.

Phase is measure in degree or radians (360 degree is 2**p** radians). A phase shift of 360 degree corresponds to a shift of complete period; a phase shift of 180 degree corresponds to a shift of a half a period; and a phase shift of 90 degree corresponds to a shift of a quarter of a period (see figure 1.4).

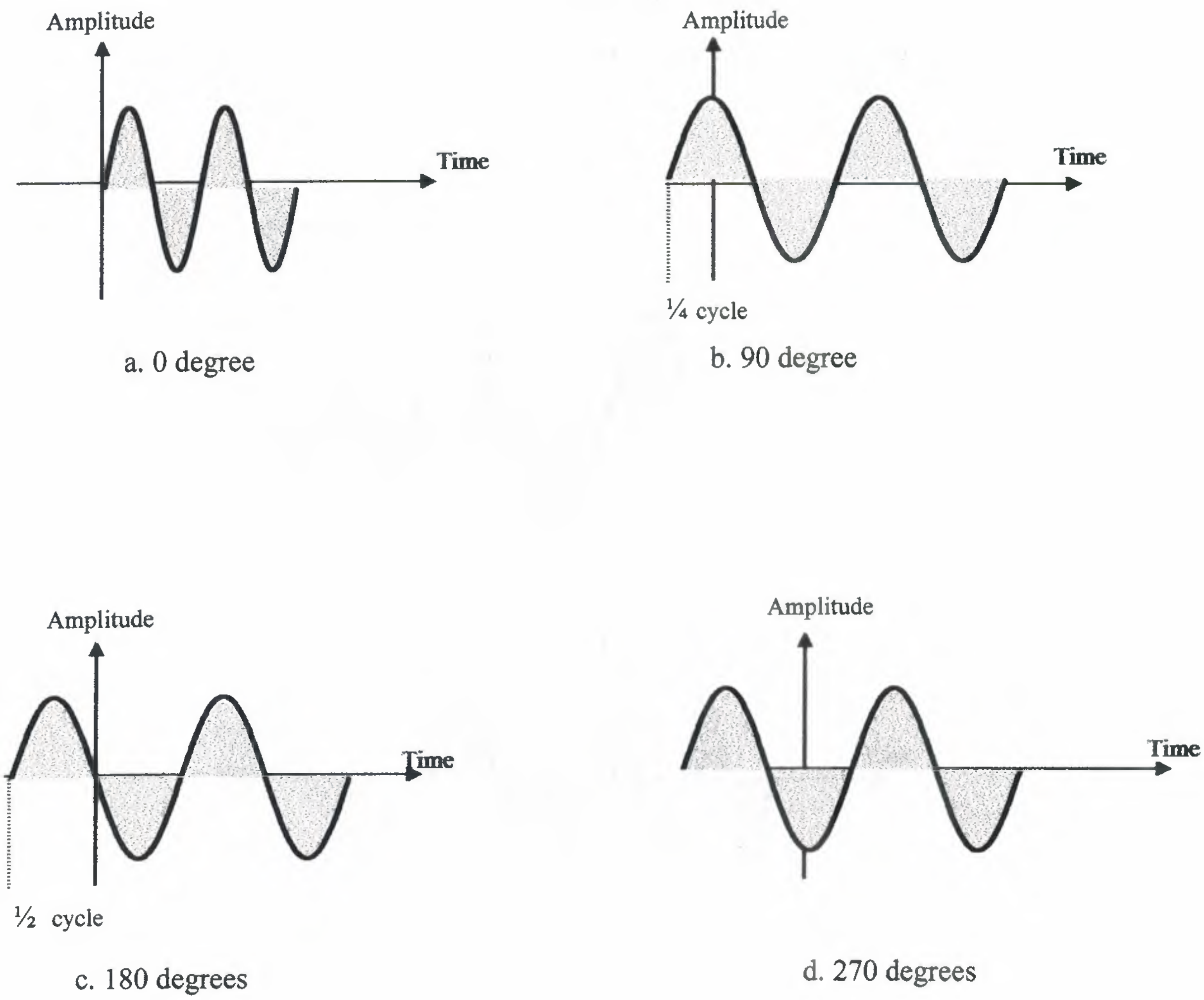


Figure 1.4 Relationship Between Different Phases

A visual comparison of amplitude, frequency, and phase provides a reference useful for understanding their function. Change in all three attributes can be introduced into a signal and controlled electronically.

More about Frequency

We know already that frequency is the relationship of a signal to time, and that the frequency of a waveform is the number of cycle it completes per second. But another way to look at frequency is as a measurement of the rate of change. Electromagnetic signal are oscillating waveforms; that is, they fluctuate continuously and predictably above and below a mean energy level. The rate at which a sine wave moves from its lowest to its highest level is its frequency. A 40 Hz signal has half the frequency of an 80 Hz signal: it completes one cycle in twice the time of the 80 Hz signal, so each cycle also takes twice as long to change from its lowest to its highest voltage levels.

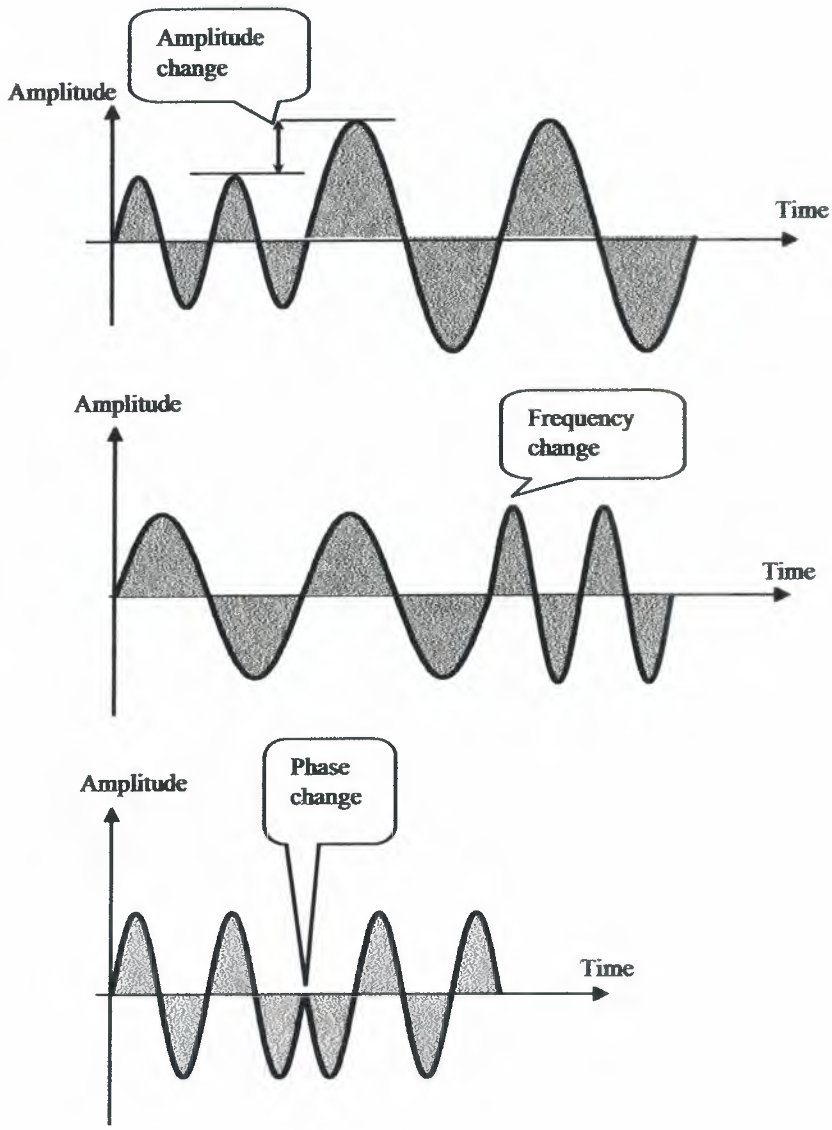


Figure 1.5 Amplitude, Frequency, and Phase Changes

Frequency, therefore, though described in cycles per second (HZ), is a general measurement of change of a signal with respect to time.

Frequency is rate of change with respect to time. Change in a short span of time means high frequency. Change in a long span of time means low frequency.

If the value of a signal changes over a very short span of time, its frequency is high. If it changes over a long span of time, its frequency is low.

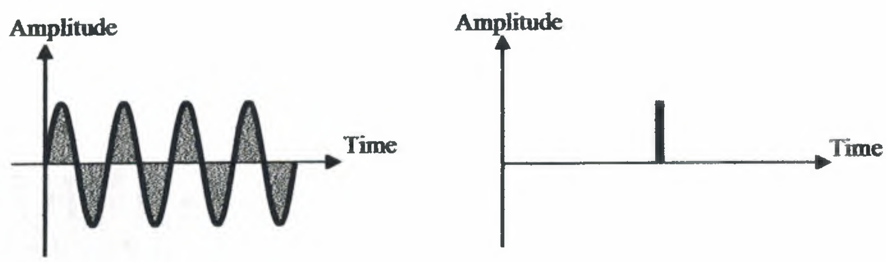
1.3.2 Time versus Frequency Domain

A sine wave is comprehensively defined by its amplitude, frequency, and phase. To show the relationship between the three characteristics (amplitude, frequency, and phase), we can use what is called a frequency-domain plot.

There are two types of frequency-domain plots:

1. Maximum amplitude versus frequency
2. Phase versus frequency.

The first type of frequency-domain plot (maximum amplitude versus frequency) is more common in data communications than the second (phase versus frequency). Figure 1.6 compares the time domain (instantaneous amplitude with respect to time) and the frequency domain (maximum amplitude with respect to frequency).



a. Time domain

b. Frequency domain

Figure 1.6 Time and Frequency Domains

Figure 1.7 gives examples of both time-domain and frequency-domain plots of three signals with varying frequencies and amplitudes. Compare the models within each pair to see which sort of information is best suited to convey.

A low-frequency signal in the frequency domain corresponds to a signal with a long period in the time domain and vice versa. A signal that changes rapidly in the time domain corresponds to high frequencies in the frequency domain.

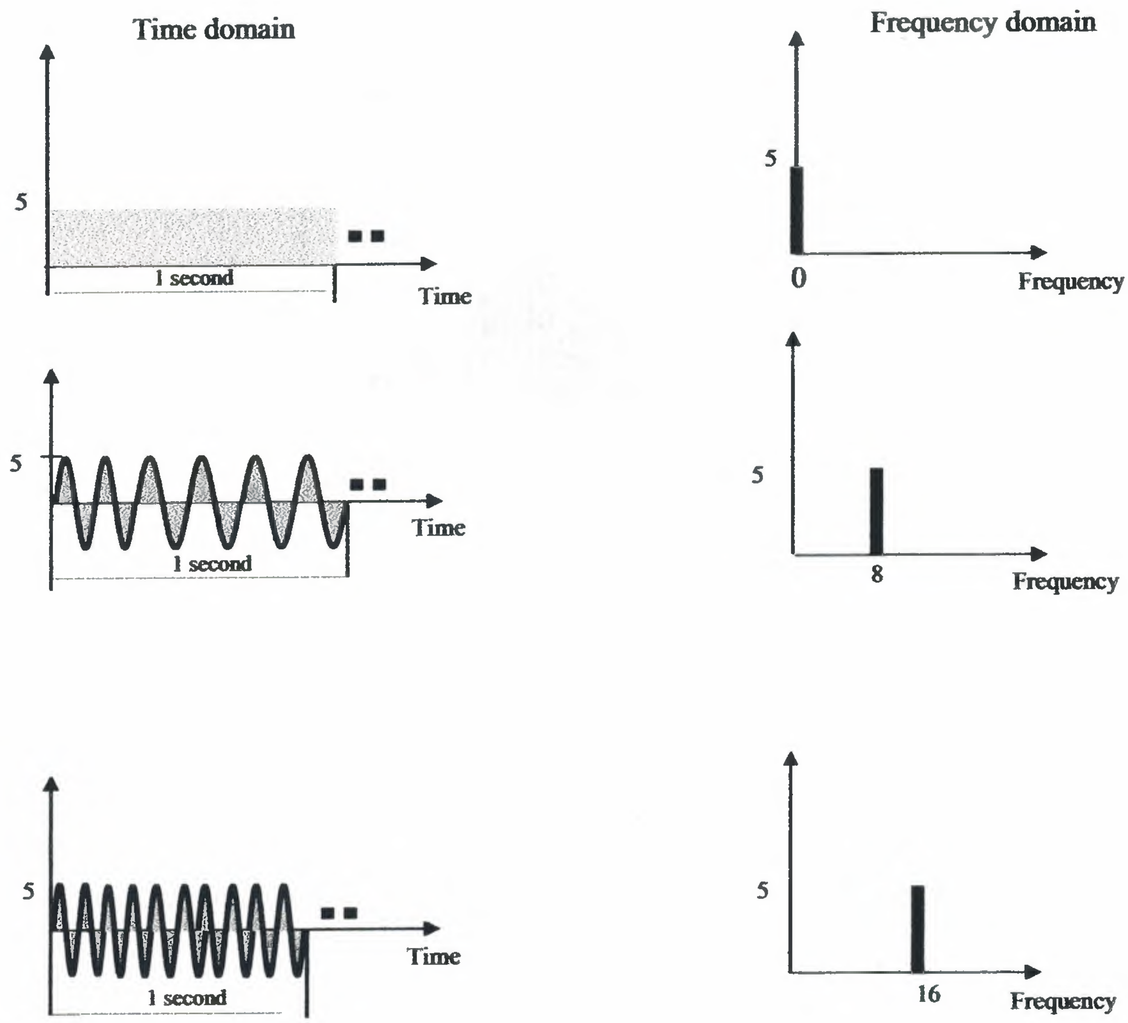


Figure 1.7 Time and Frequency Domains for Different Signals

1.3.3 Frequency Spectrum and Bandwidth

Two terms need mentioning here: spectrum and bandwidth. The frequency spectrum of a signal is the collection of all the component frequencies it contains and is shown using a frequency domain graph. The bandwidth of a signal is the width of the frequency spectrum (see figure 1.8). In other words, bandwidth refers to the range of component frequencies, and frequency spectrum refers to the elements within that range. To calculate the bandwidth, subtract the lowest frequency from the highest frequency of the range.

The frequency spectrum of a signal is the combination of all sine wave signals that make that signal.

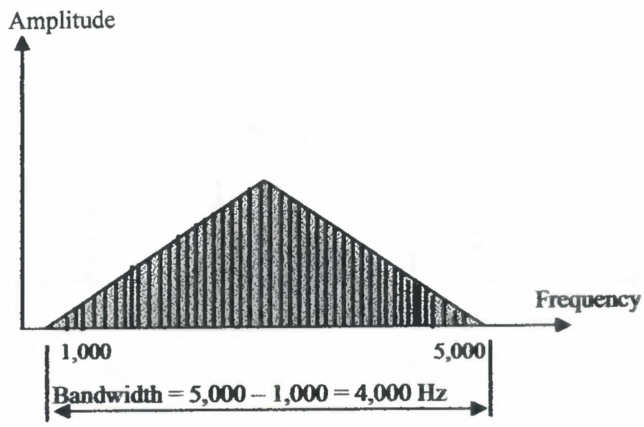


Figure 1.8 Bandwidth

Example 1.3

If a periodic signal is decomposed into five sine waves with frequencies of 100, 300, 500, 700, and 900 Hz, what is the bandwidth?

Solution

Let f_h be the highest frequency, f_l be the lowest frequency, and B be the bandwidth. Then,

$$B = f_h - f_l = 900 - 100 = 800\text{Hz}$$

Example 1.4

A signal has a bandwidth of 30Hz. The highest frequency is 60 KHz. What is the lowest frequency?

Solution

Let f_h be the highest frequency, f_l be the lowest frequency, and B be the bandwidth. Then,

$$B = f_h - f_l$$

$$20 = 60 - f_l$$

$$f_l = 60 - 20 = 40 \text{ KHz}$$

1.4 Digital Signals

In addition to being represented by an analog signal, data can also be represented by a digital signal. See Figure 1.9.

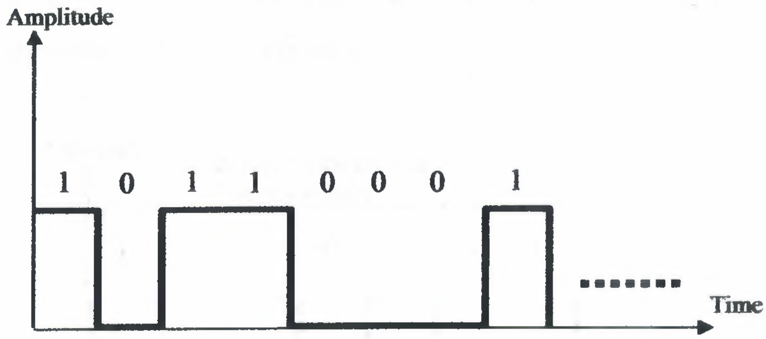


Figure 1.9 A digital Signal

1.4.1 Amplitude, Period, and Phase

The three characteristics of periodic analog signals (amplitude, period, and phase) can be redefined for a periodic digital signal. (see Figure 1.10).

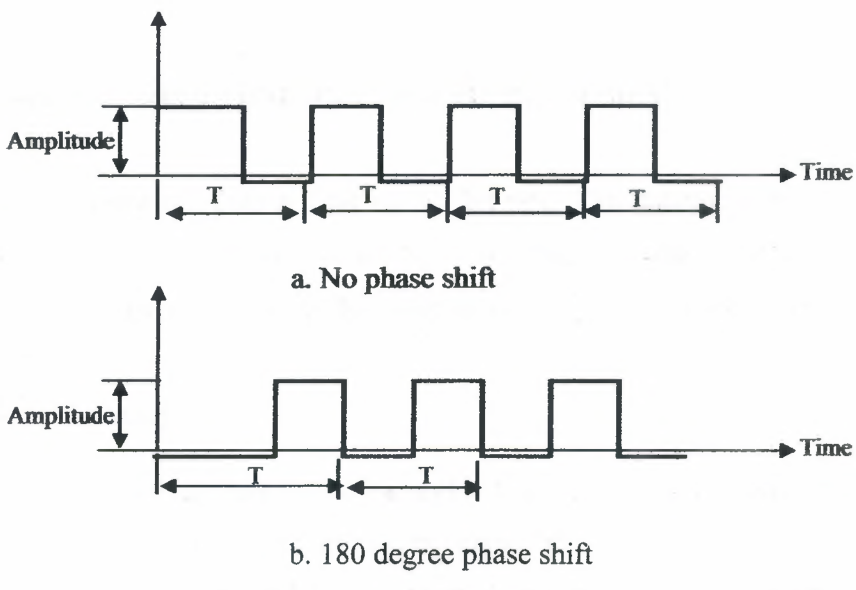


Figure 1.10 Amplitude, Period, and Phase for a Periodic Digital Signal

1.4.2 Bit Interval and Bit Rate

Most digital signals are aperiodic and thus period or frequency is not appripriate. Two new terms, bit interval (instead of period) and bit rate (instead of frequency) are used to describe digital signals. The bit interval is the time required to send one single bit. The bit rate is the number of bit interval per second. This means that the bit rate is the number of bits sent in one second, usually expressed in bps. See Figure 1.11.

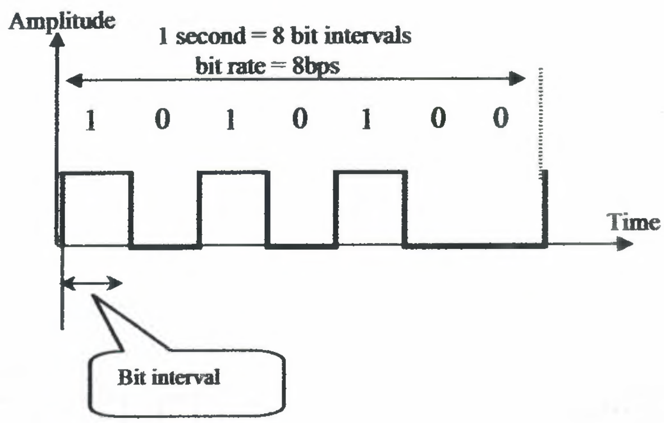


Figure 1.11 Bit Rate and Bit Interval

1.4.3 Medium Bandwidth and Data Rate: Channel Capacity

The medium bandwidth puts a limit on the bit rate. The maximum bit rate a transmission medium can transfer is called channel capacity of the medium. The capacity of a channel depends on the type of encoding technique and the signal-to-noise ratio of the system (see Figure 1.12).

Channel Capacity

The rate at which data can be transmitted over a given communication channel, under given conditions, is referred to as the channel capacity.

There are four concepts here that we are trying to relate to one another.

- Data rate: This is the rate, in bits per second (bps), at which data can be transmitted.
- Bandwidth: This is the bandwidth of the transmitted signal as constrains by the transmitter and the nature of the transmission medium, expressed by Hertz.
- Noise: The average level of noise over the communications path.

- Error rate: The rate at which errors occur, where an error is the reception of a 1 when a 0 was transmitted or the reception of a 0 when a 1 was transmitted.

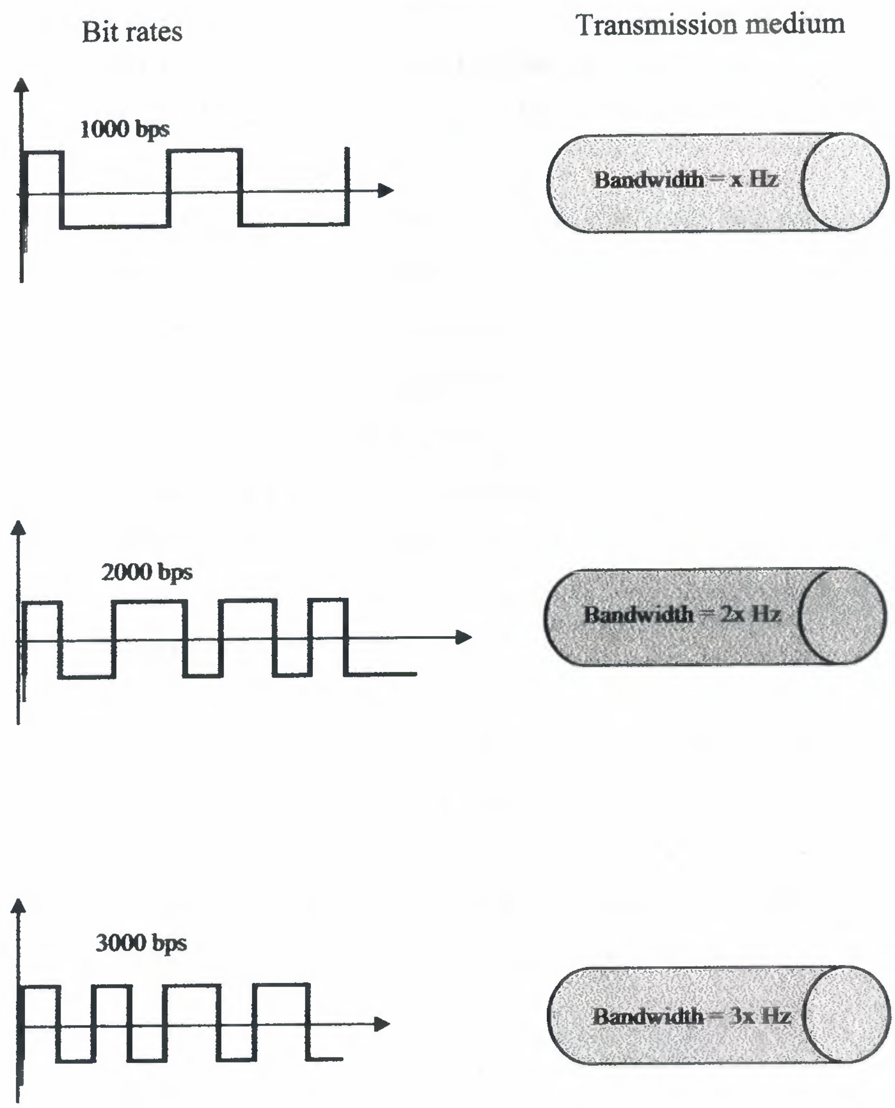


Figure 1.12 Medium Bandwidth and Data Rate

Communication facilities are expensive and, in general, the greater the bandwidth of a facility the greater the cost. Furthermore, all transmission channels of any practical interest are of limited bandwidth. The limitations arise from the physical properties of the transmission medium or from deliberate limitations at the transmitter on the bandwidth to prevent interference from other sources. Accordingly, we would like to make as efficient use as possible of a given bandwidth.

Let us consider the case of a channel that is noise-free. In this environment, the limitation on data rate is simply the bandwidth of the signal. A formulation of this limitation, due to Nyquist, states that if the rate of signal transmission is 2W, then a signal with frequencies no greater than W is sufficient to carry the data rate. The conserve is also true: Given a bandwidth of W, the highest signal rate that can be carried is 2W.

Example 2.2. A voice channel bandwidth is of W = 3100 Hz. Find the channel capacity. Solution: C = 2 W = 6200 bps.

However, as we shall see in chapter 3, signals with more than two levels can be used; that is each signal element can represent more than one bit. For example; if M possible voltage levels are used, then each signal element can be represented by $n = \log_2 M$ numbers of bits. With multilevel signaling, the Nyquist formulation becomes.

$$C = 2\,W \log_2 M$$

Thus, for M = 8, a value used with some modems, C becomes 18600 bps.

An important parameter associated with a channel is a signal-to-noise ratio (SNR) expressed as

$$SNR = 10\log_{10} (S/N)\ dB$$

Where S/N – signal –to- noise powers ratio. Clearly a high S/N will mean a high - quality signal and a low number of required intermediate repeaters.

The signal - to noise ratio is important in the transmission of digital data because it sets the upper bound on the achievable data rate. The maximum channel capacity, in bits per second, obeys the equation attributed as the Shannon – Hartley law

$$C = W \log_2 (1+ S/N)$$

Example 1.5

Consider a voice channel with bandwidth of 3000 Hz. A typical value of S/N for a telephone line is 20 dB.

Solution

$S/N = W \log_{10}(1+S/N) = 3{,}32\ W\log_{10}(1+S/N)$, $S/N = 100$

$C = 3000 \log_2 (1 + 100) = 19963$ bps

2. ENCODING

2.1 Overview

As we discussed before we must encode data into signals to send them from one place to another.

How information is encoded depends on its original format and on the format used by the communication hardware. If you want to send a love letter by smoke signal. you need to know which smoke patterns match which words in your message before you actually build your fire. Words are digital information and puffs of smoke are a digital representation of information, so defining the smoke patterns would be a form of digital-to-digital encoding. Communication technology has fundamentally the same requirements with a few additional options.

A simple signal by itself does not carry information any more than a straight line conveys words. The signal must be manipulated so that it contains identifiable changes that are recognizable to the sender and receiver as representing the information intended. First the information must be translated into agreed-upon patterns of 0s and 1s. In the case of textual data, these patterns can belong to either of two conventions:

ASCII or EBCDIC.

As we saw before, information can be of two types, digital or analog, and signals can be of two types, also digtal or analog. Therefore, four types of encoding are possible: digital-to-digital, analog-to-digital, digital-to-analog, and analog-to-analog (see Figure 2.1)

2.2 Digital-To-Digital Encoding

Digital-to-digital encoding is the representation of digital information by a digital signal. For example, when you transmit data from your computer to your printer, both the original data and the transmitted data are digital. In this type of encoding, the binary 1s and 0s generated by a computer are translated into a sequence of voltage pulses that can be propagated over a wire. Figure 2.2 shows the relationship between the digital information, the digital-to-digital encoding hardware, and the resultant digital signal.

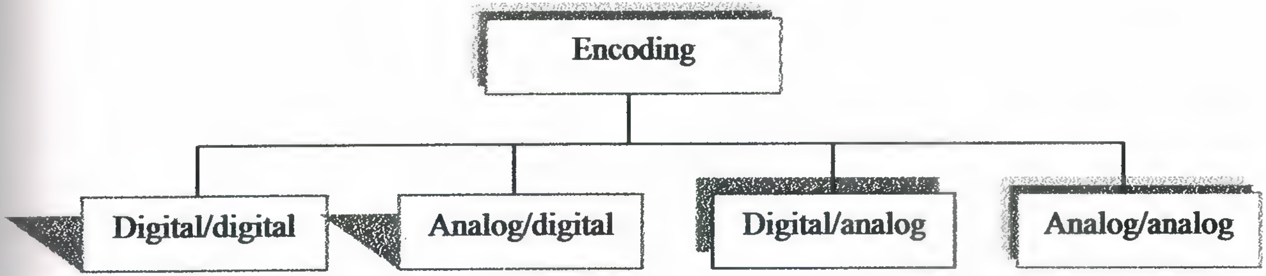


Figure 2.1 Different Encoding Schemes

Figure 2.2 Digital-to-Digital Encoding

Of the many mechanisms for digital-to-digital encoding, we will discuss only those most useful for data communication. These fall into three broad categories: unipolar, polar, and bipolar (see Figure 2.3).

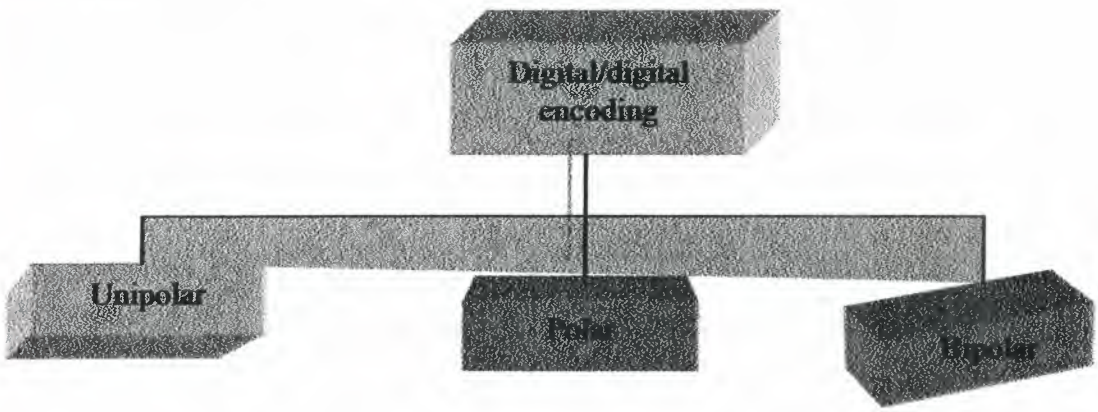


Figure 2.3 Types of Digital-to-Digital Encoding

Unipolar encoding is simple, with only one technique in use. Polar encoding ha three subcategories, NRZ, RZ and biphase, two of which have multiple variations of the own. The third option, bipolar encoding, has three variations: AMI, B8ZS, and HDB3.

2.2.1 Unipolar

Unipolar encoding is very simple and very primitive. Although it is almost obsolete today, its simplicity provides an easy introduction to the concepts developed with more complex encoding systems and allows us to examine the kinds of problems that any digital transmission system must overcome.

Digital transmission systems work by sending voltage pulses along a media link, usually a wire or cable. In most types of encoding, one voltage level stands for binary 0 and another level stands for binary 1. The polarity of a pulse refers to whether it is positive or negative. Unipolar encoding is so named because it uses only one polarity. Therefore, only one of the two binary states is encoded. usually the 1. The other state, usually the 0, is represented by zero voltage, or an idle line.

Unipolar encoding uses only one level of value

Figure 2.4 shows the idea of unipolar encoding. In this example, the is are encoded as a positive value and the 0s are idle. In addition to being straightforward. unipolar encoding is inexpensive to implement.

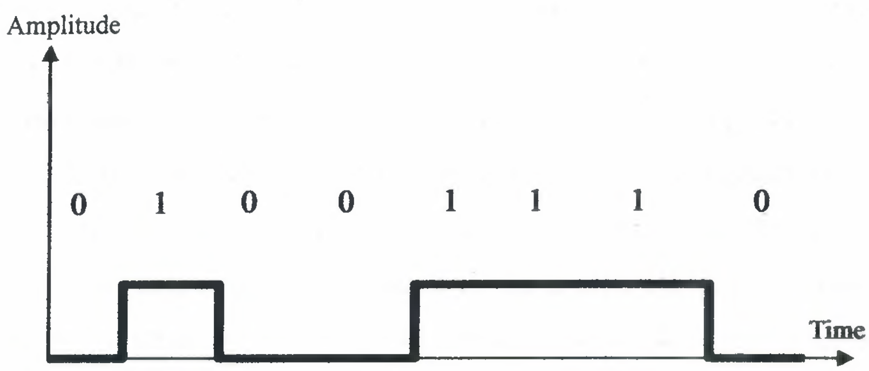


Figure 2.4 Unipolar Encoding

However, unipolar encoding has at least two problems that make it unusable: DC component and synchronization.

DC Component

The average amplitude of a unipolar encoded signal is nonzero. This creates what is called a direct current (DC) component (a component with zero frequency). When a signal contains a DC component, it cannot travel through media that cannot handle DC components, such as microwaves or transformers.

Synchronization

When a signal is unvarying, the receiver cannot determine the beginning and ending of each bit. Therefore, a synchronization problem in unipolar encoding can occur whenever the data stream includes a long uninterrupted series of 1s or 0s. Digital encoding schemes use changes in voltage level to indicate changes in bit type. A signal change also indicates that one bit has ended and a new bit has begun. In unipolar encoding. however, a series of one kind of bit, say seven 1s, occurs with no voltage changes, just an unbroken positive voltage that lasts seven times as long as a single 1 bit. Whenever there is no signal change to indicate the start of the next bit in a sequence, the receiver has to rely on a timer. Given an expected bit rate of 1000 bps, if the receiver detects a positive voltage lasting 0.005 seconds, it reads one 1 per 0.001 seconds, or five 1s.

Unfortunately, propagation delays can distort the timing of the signal so that, for example, five is can be stretched to 0006 seconds causing an extra 1 bit to be read by the receiver. That one extra bit in the data stream causes everything after it to be decoded erroneously. In addition, the receiver's clock can go out of synchronization, causing the receiver to read the bit stream erroneously. A solution developed to control the synchronization of unipolar transmission is to use a separate, parallel line that carries a clock pulse and allows the receiving device to resynchronize its timer to that of the signal. But doubling the number of lines used for transmission increases the cost and so proves uneconomical.

2.2.2 Polar

Polar encoding uses two voltage levels: one positive and one negative. By using both levels, in most polar encoding methods the average voltage level on the line is reduced and the DC component problem of unipolar encoding is alleviated In Manchester and Differential Manchester encoding, each bit consists of both positive and negative voltages, so the DC component is totally eliminated.

Polar encoding uses two levels (positive and negative) of amplitude.

Of the many existing variations of polar encoding, we will examine only the three most popular: non-return to zero (NRZ), return to zero (RZ), and biphase. NRZ encoding includes two methods: non-return to zero, level (NRZ-L), and non-return to zero-invert (NRZ-I). Biphase also refers to two methods. The first, Manchester, is the method used by Ethernet LANs. The second, Differential Manchester, is the method used by token ring LANs (see Figure 2.5).

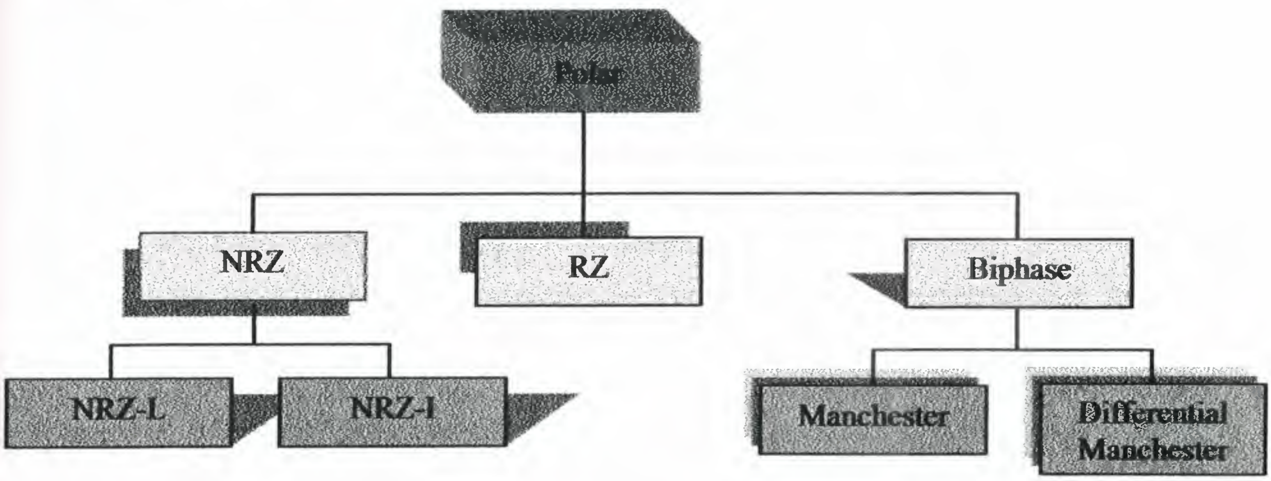


Figure 2.5 Types of Polar Encoding

Non-Return to Zero (NRZ)

In NRZ encoding, the level of the signal is always either positive or negative. Unlike in unipolar encoding, where a 0 bit is represented by an idle line, in NRZ systems if the line is idle it means no transmission is occurring at all. The two most popular methods of NRZ transmission are discussed below.

NRZ-L In NRZ-L encoding, the level of the signal depends on the type of bit it represents. A positive voltage means the bit is a 1, and a negative voltage means the bit is a 0; thus, the level of the signal is dependent upon the state of the bit. In NRZ-L the level of the signal is dependent upon the state of the bit.

NRZ-I In NRZ-I, an inversion of the voltage level represents a 1 bit. It is the transition between a positive and negative voltage, not the voltages themselves, that represents a 1 bit. A 0 bit is represented by no change. An advantage of NRZ-I over NRZ-L is that because the signal changes every time a 1 bit is encountered, it provides some synchronization.

A series of seven 1s will cause seven inversions. Each of those inversions allows the receiver to resynchronize its timer to the actual arrival of the transmission. Statistically, strings of 1s occur more frequently in transmissions than do strings of 0s. Synchronizing strings of 1s therefore goes a long way toward keeping the entire message synchronized. A string of 0s can still cause problems, but because 0s are not as likely, they are less of a threat to decoding. In NRZ-I the signal is inverted if a 1 is encountered.

Figure 2.6 shows the NRZ-L and NRZ-I representations of the same series of bits. In the NRZ-L sequence, positive and negative voltages have specific meanings: positive for 1 and negative for 0. In the NRZ-I sequence, the voltages per se are meaningless. Instead, the receiver looks for changes from one level to another as its basis for recognition of 1s.

Return to Zero (RZ)

As you can see, anytime the original data contain strings of consecutive 1s or 0s, the receiver can lose its place. As we mentioned in our discussion of unipolar encoding, one way to assure synchronization is to send a separate timing signal on a separate channel. However, this solution is both expensive and prone to errors of its own. A better solution is to somehow include synchronization in the encoded signal, something like the solution provided by NRZ-I, but one capable of handling strings of 0s as well as 1s.

To assure synchronization, there must be a signal change for each bit. The receiver can use these chances to build up, update. and synchronize its clock. As we saw above.

NRZ-I accomplishes this for sequences of 1s. But to chance with every bit, we need more than just two values. One solution is return to zero (RZ) encoding, which uses three values: positive, negative, and zero. In RZ, the signal changes not between bits but during each bit. Like NRZ-L, a positive voltage means 1 and a negative voltage means 0. But, unlike NRZ-L. halfway through each bit interval, the signal returns to zero. A 1 bit is actually represented by positive-to-zero, and a 0 bit by negative-to-zero, rather than by positive and negative alone. Figure 2.7 illustrates the concept.

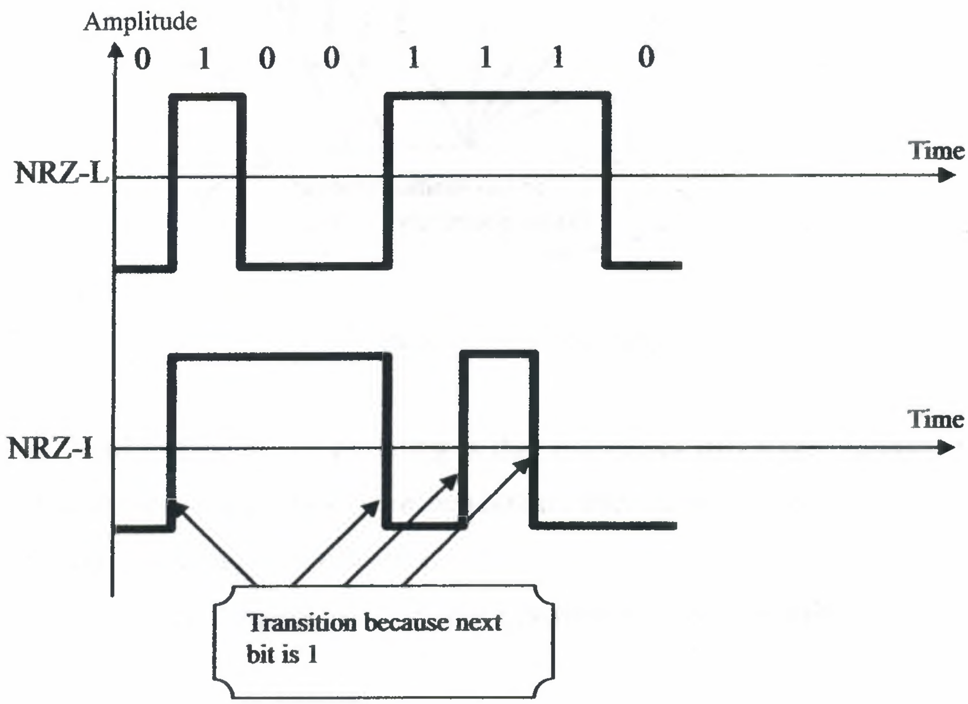


Figure 2.6 NRZ-L and NRZ-I Encoding

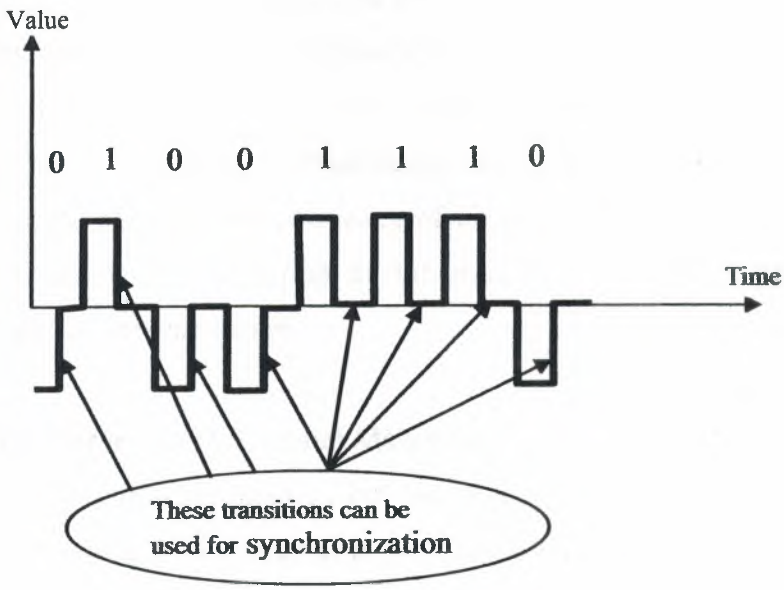


Figure 2.7 RZ Encoding

The main disadvantage of RZ encoding is that it requires two signal changes to encode one bit and therefore occupies more bandwidth. But of the three alternatives we have examined so far, it is the most effective.

A good encoded digital signal must contain a provision for synchronization.

Biphase

Probably the best existing solution to the problem of synchronization is biphase encoding. In this method, the signal changes at the middle of the bit interval but does not return to zero. Instead, it continues to the opposite pole. As in RZ, these midinterval transitions allow for synchronization.

As mentioned earlier, there are two types of biphase encoding in use on networks today: Manchester and Differential-Manchester.

Biphase encoding is implemented in two different ways: Manchester and Differential Manchester.

Manchester : Manchester encoding uses the inversion at the middle-of each bit interval for both synchronization and bit representation. A negative-to-positive transition represents binary 1 and a positive-to-negative transition represents binary 0. By using a single transition for a dual purpose, Manchester encoding achieves the same level of synchronization as RZ but with only two levels of amplitude.
In Manchester encoding the transition at the middle of the bit is used for both synchronization and bit representation.

Differential Manchester : In Differential Manchester, the inversion at the middle of the bit interval is used for synchronization, but the presence or absence of an additional transition at the beginning of the interval is used to identify the bit. A transition means binary 0 and no transition means binary 1. Differential Manchester requires two signal changes to represent binary 0 but only one to represent binary 1.
In Differential Manchester the transition at the middle of the bit is used only for synchronization. The bit representation is shown by the inversion or noninversion at the beginning of the bit. Figure 2.8 shows the Manchester and Differential Manchester signals for the same bit pattern.

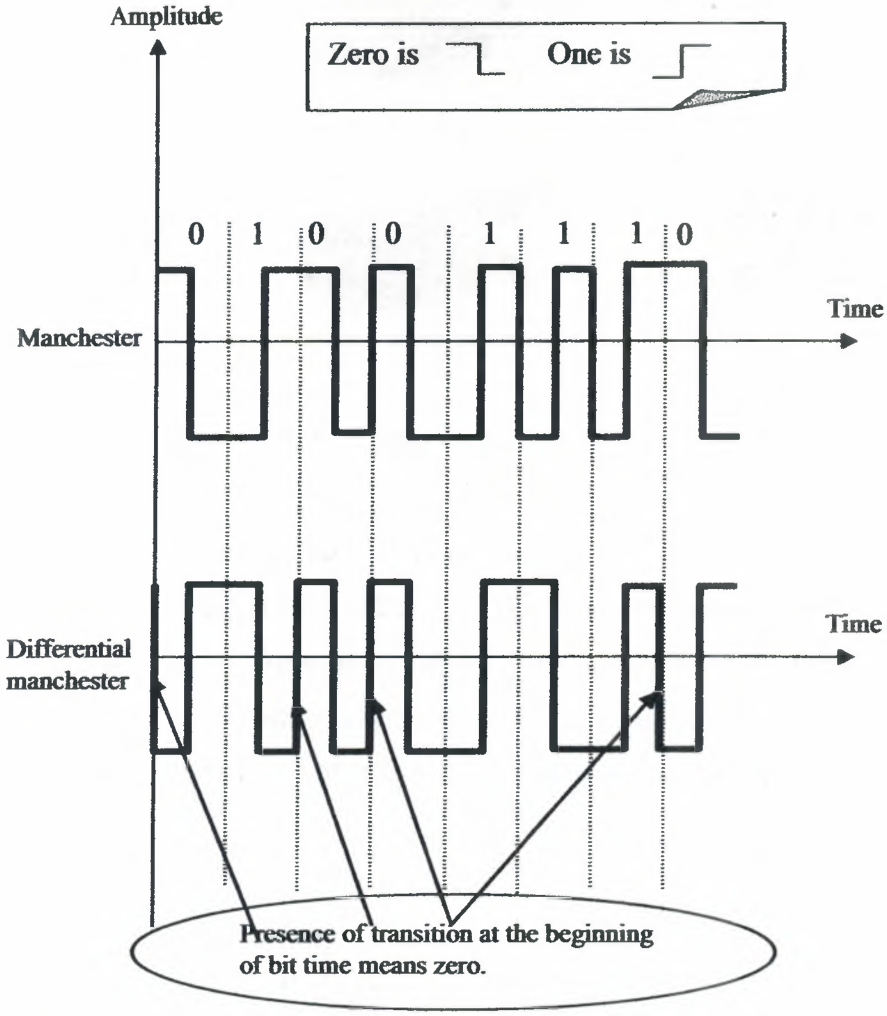


Figure 2.8 Manchester and Differential Manchester Encoding

2.2.3 Bipolar

Bipolar encoding, like RZ, uses three voltage levels: positive, negative, and zero. Unlike RZ, however, the zero level in bipolar encoding is used to represent binary 0. Positive and negative voltages represent alternating 1s. If the first t bit is represented by the positive amplitude, the second will be represented by the negative amplitude, the third by the positive amplitude, and so on. This alternation occurs even when the 1 bits are not consecutive.

Three types of bipolar encoding are in popular use by the data communications industry: AMI, B8ZS, and HDB3 (see Figure 2.9).

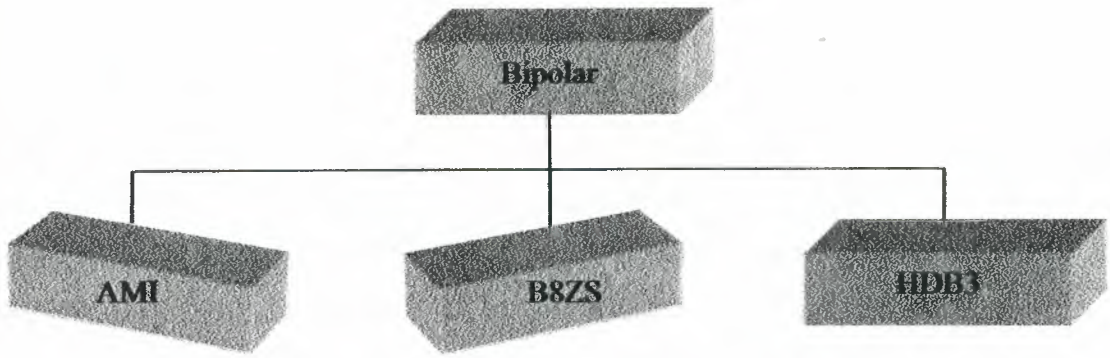


Figure 2.9 Types of bipolar encoding

Bipolar Alternate Mark Inversion (AMI)

Bipolar AMI is the simplest type of bipolar encoding. In the name alternate mark inversion, the word mark comes from telegraphy and means 1. So AMI means alternate 1 inversion. A neutral, zero voltage represents binary 0. Binary Is are represented by alternating positive and negative voltages. Figure 2.10 gives an example.

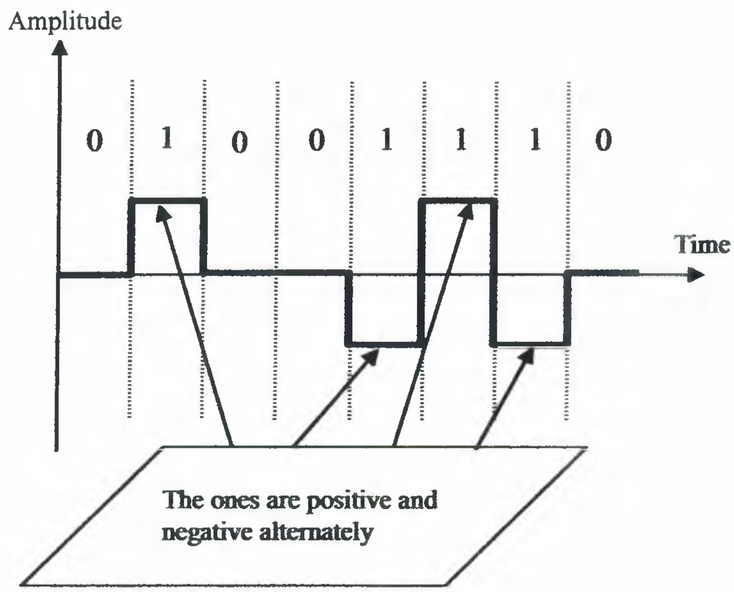


Figure 2.10 Bipolar AMI Encoding

By inverting on each occurrence of a 1, bipolar AMI accomplishes two things:
First, the DC component is zero, and second, a long-sequence of 1s stays synchronized. There is no mechanism to ensure the synchronization of a long string of 0s.
Two variations of bipolar AMI have been developed to solve the problem of synchronizing sequential 0s. The first, used in North America, is called bipolar 8-zero substitution (B8ZS). The second, used in Europe and Japan, is called high-density bipolar 3 (HDB3). Both are adaptations of bipolar AMI that modify the original pattern only in the case of multiple consecutive 0s.

Bipolar 8-Zero Substitution (B8ZS)

B8ZS is the convention adopted in North America to provide synchronization of long strings of 0s. In most situations, B8ZS functions identically to bipolar AMI. Bipolar AMI changes poles with every 1 it encounters. These changes provide the synchronization needed by the receiver. But the signal does not change during a string of 0s, so synchronization is often lost.
The difference between B8ZS and bipolar AMI occurs whenever eight or more consecutive 0s are encountered in the data stream. The solution provided by B8ZS is to force artificial signal changes, called violations, within the 0 string. Anytime eight 0s occur in succession, B8ZS introduces changes in the pattern based on the polarity of the previous 1 (the 1 occurring just before the 0s). See Figure 2.11

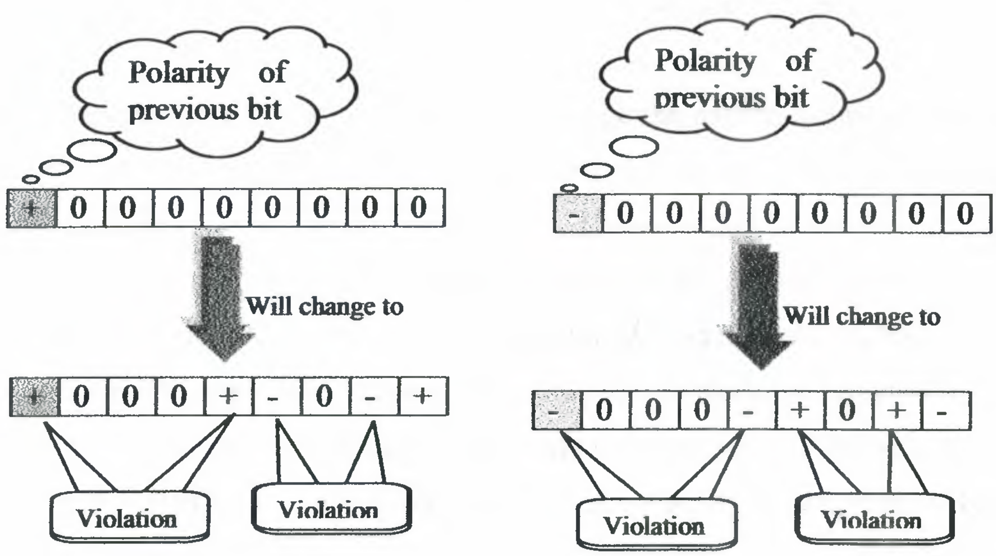


Figure 2.11 B8ZS Encoding

If the previous 1 bit was positive, the eight 0s will be encoded as zero, zero, zero. positive, negative, zero, negative, positive. Remember that the receiver is looking for alternating polarities to identify 1s. When it finds two consecutive positive charges surrounding three 0s, it recognizes the pattern as a deliberately introduced violation and not an error. It then looks for the second pair of the expected violations. When it finds them, the receiver translates all eight bits to 0s and reverts back to normal bipolar AMI mode.

If the polarity of the previous 1 is negative, the pattern of violations is the same but with inverted polarities. Both positive and negative patterns are shown in Figure 2.11.
In B8ZS if eight 0s come one after another, we change the pattern in one of two ways based on the polarity of the previous 1.

High-Density Bipolar 3 (HDB3)
The problem of synchronizing strings of consecutive 0s is solved differently in Europe and Japan than in the United States. This convention, called HDB3, introduces changes into the bipolar AMI pattern every time four consecutive 0s are encountered instead of waiting for the eight expected by B8ZS in North America. Although the name is HDB3. the pattern changes whenever there are four 0s in succession (see Figure 2.12).

In HDB3 if four 0s come one after another, we change the pattern in one of four ways based on the polarity of the previous 1 and the number of 1s since the last substitution.

As in B8ZS, the pattern of violations in HDB3 is based on the polarity of the previous 1 bit. But unlike B8ZS, HDB3 also looks at the number of 1s that have occurred in the bit stream since the last substitution. Whenever the number of 1s since the last substitution is odd, B8ZS puts a violation in the place of the fourth consecutive 0. If the polarity of the previous bit was positive, the violation is positive. If the polarity of the previous bit was negative, the violation is negative.

Whenever the number of 1s since the last substitution is even, B8ZS puts violations in the places of both the first and the fourth consecutive 0s. If the polarity of the previous bit was positive, both violations are negative. If the polarity of the previous bit was negative, both violations are positive. All four patterns are shown in Figure 2.15.

If the number of 1s since the last substitution is odd

If the number of 1s since the last substitution is even

Figure 2.12 HDB3 Encoding

2.3 Analog-To-Digital Encoding

Analog-to-digital encoding is the representation of analog information by a digital signal. To record a singer's voice onto a compact disc, for example, you use digital me to replicate analog information. To do so you need to reduce the potentially infinite number of values in an analog message so that they can be represented as a digital stream with a minimum loss of information. Several methods for analog-to-digital encoding will be discussed later in this chapter. Figure 2.13 shows the analog-to-digital encoder, called codec (coder-decoder).

In analog-to-digital encoding, we are representing the information contained continuous wave form as a series of digital pulses (1s or 0s).

So far, the encoding systems we have been examining have focused on the format of the transporting signal. Analog-to-digital encoding can make use of any of the digital signals discussed in Section 2.2. The structure of the transporting signal is not the problem.

Instead, the problem is how to translate information from an infinite number of values to a discrete number of values without sacrificing sense or quality.

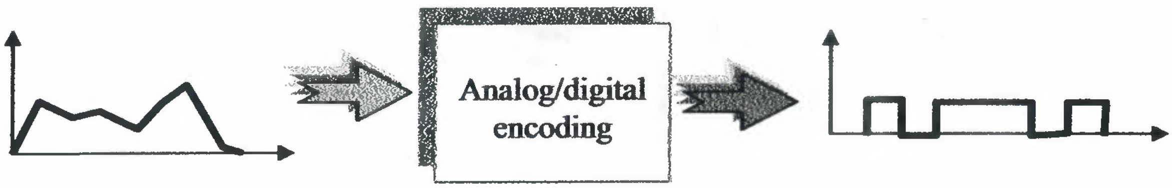


Figure 2.13 Analog-to-Digital Encoding

2.3.1 Pulse Amplitude Modulation (PAM)

The first step in analog-to-digital encoding is called pulse amplitude modulation (PAM). This technique takes analog information, samples it, and generates a series of pulses based on the results of the sampling. The term sampling means measuring the amplitude of the signal at equal intervals.

The method of sampling used in PAM is more useful to other areas of engineering than it is to data communication. However, PAM is the foundation of an important analog-to-digital encoding method called pulse code modulation (PCM).

In PAM, the original signal is sampled at equal intervals as shown in Figure 2.14. PAM uses a technique called sample and hold. At a given moment the signal level is read, then held briefly. The sampled value occurs only instantaneously in the actual wave form, but is generalized over a still short but measurable period in the PAM result.

The reason PAM is not useful to data communications is that, although it translates the original wave form to a series of pulses, these pulses are still of any amplitude (still an analog signal, not digital). To make them digital, we must modify them by using pulse code modulation (PCM).

Pulse amplitude modulation (PAM) has some applications, but it is not used by itself in data communication. However, it is the first step in another very popular encoding method called pulse code modulation (PCM).

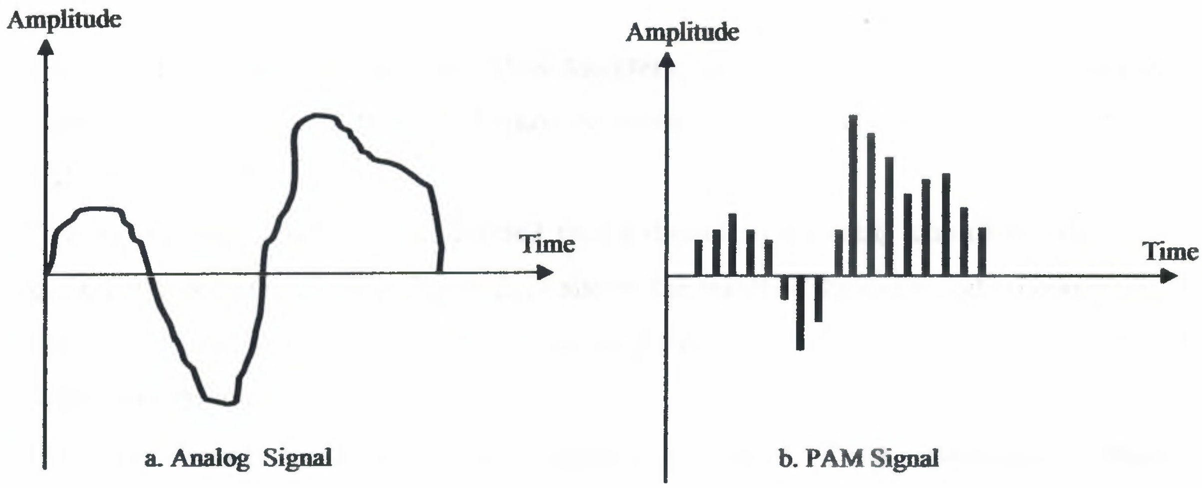


Figure 2.14 PAM

2.3.2 Pulse Code Modulation (PCM)

PCM modifies the pulses created by PAM to create a completely digital signal. To do so, PCM first quantizes the PAM pulses. Quantization is a method of assigning integral values in a specific range to sampled instances. The result of quantization is presented in Figure 2.15.

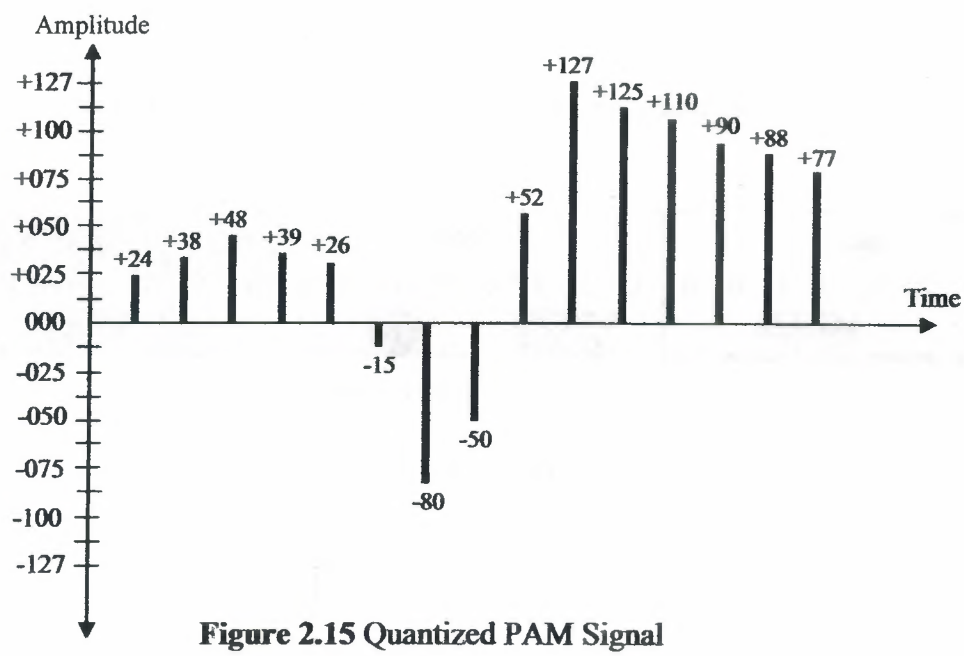


Figure 2.15 Quantized PAM Signal

Figure 2.16 shows a simple method of assigning sign and magnitude values quantized samples. Each value is translated into its seven-bit binary equivalent. The eighth bit indicates the sign.

The binary digits are then transformed into a digital signal using one of the digital to-digital encoding techniques. Figure 2.17 shows the result of the pulse code modulation of the original signal encoded finally into a unipolar signal. Only the first three sampled values are shown.

PCM is actually made up of four separate processes: PAM, quantization. binary encoding, and digital-to-digital encoding. Figure 2.18 shows the entire process graphic form. PCM is the sampling method used to digitize voice in T-line transmission in the North American telecommunication system.

+024	00011000	-015	10001111	+125	01111101
+038	00100110	-080	11010000	+110	01101110
+048	00110000	-050	10110010	+090	01011010
+039	00100111	+052	00110110	+088	01011000
+026	00011010	+127	01111111	+077	01001101

Sign bit
+ is 0 - is 1

Figure 2.16 Quantizing Using Sign and Magnitude

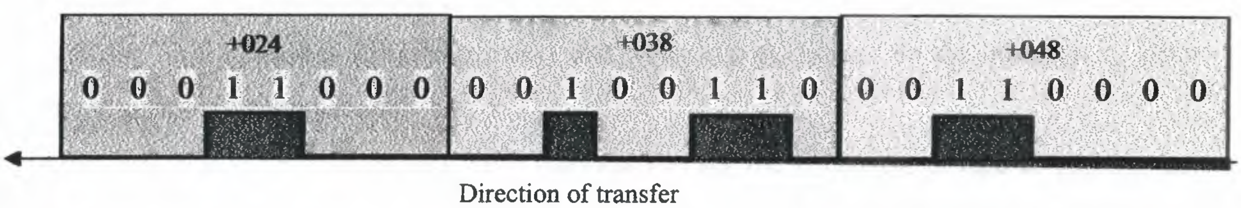


Figure 2.17 PCM

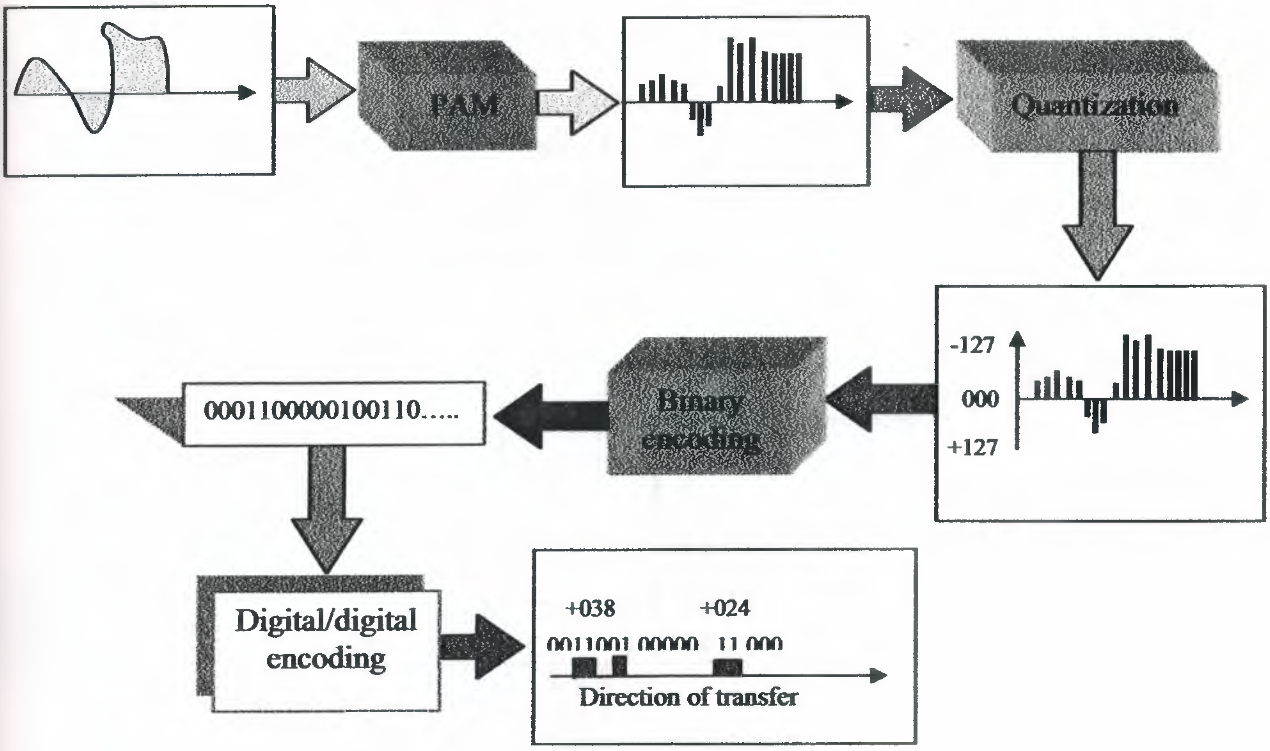


Figure 2.18 From Analog Signal PCM Digital Code

2.3.3 Sampling Rate

As you can tell from the preceding figures, the accuracy of any digital reproduction of an analog signal depends on the number of samples taken. Using PAM and PCM, we can reproduce the wave form exactly by taking infinite samples, or we can reproduce the barest generalization of its direction of change by taking three samples. Obviously, we prefer to find a number somewhere between these two extremes. So the question is, How many samples are sufficient?

Actually, it requires remarkably little information for the receiving device to reconstruct an analog signal. According to the Nyquist theorem, to ensure the accurate reproduction of an original analog signal using PAM, the sampling rate must be at least twice the highest frequency of the original signal. So if we want to sample telephone voice information with maximum frequency 3300 Hz, we need a sampling rate of 6600 samples per second.

In actual practice, 8000 samples are taken to compensate for imperfection in later processing. According to the Nyquist theorem, the sampling rate must be at least two times the highest frequency.

A sampling rate of twice a frequency of x Hz means that the signal must be sampled every 1/2x seconds. Using the voice-over-phone-lines example above, that means one sample every 1/8000 second. Figure 2.19 illustrates the concept.

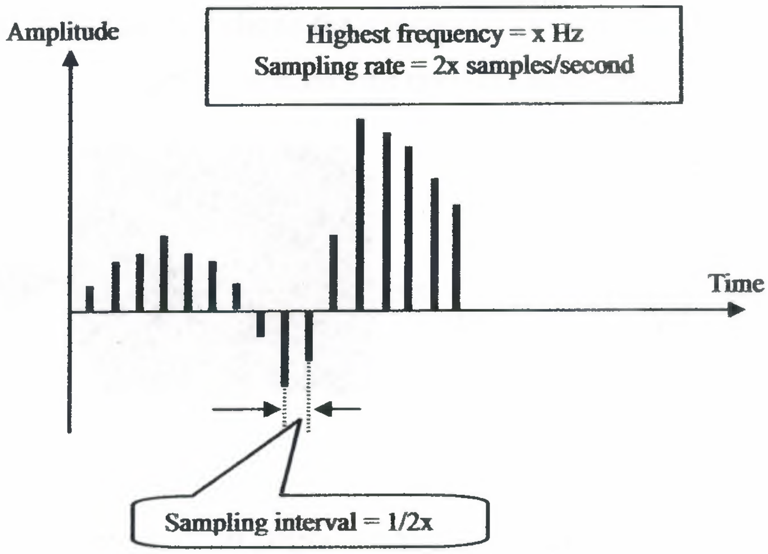


Figure 2.19 Nyquist Theorem

Example :

What sampling rate is needed for a signal with a bandwidth of 10,000 Hz (1000 to 11,000)Hz ? If the quantization is eight bits per sample, what is the bit rate?

Solution:

The sampling rate must be twice the highest frequency in the signal:

Sampling rate = 2(11,000) = 22,000 samples/sec.

Each sample is quantized to eight bits:

Data rate = (22,000 samples/sec.)(8 bits/sample) = 176 Kbps.

2.4 Digital-To-Analog Encoding

Digital-to-analog encoding is the representation of digital information by an analog signal. When you transmit data from one computer to another across a public access phone line, for example, the data start out as digital, but because telephone wires carry analog signals, the data must be converted. The digital data must be encoded on an analog signal that has been manipulated to look like two distinct values that correspond to binary 1 and binary 0. Figure 2.20 shows the relationship between the digital information, the digital-to-analog encoding hardware, and the resultant analog signal.

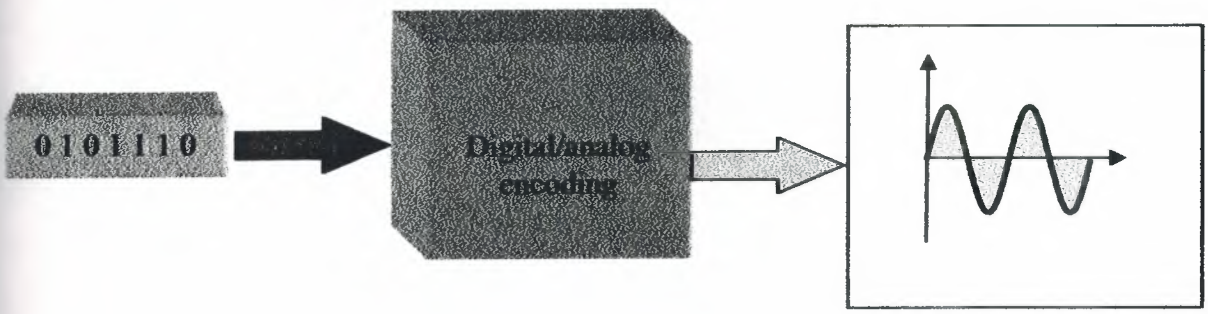


Figure 2.20 Digital-to-Analog Encoding

Of the many mechanisms for digital-to-analog encoding, we will discuss only those most useful for data communications.

As discussed previously, a sine wave is defined by three characteristics: amplitude, frequency, and phase. When we vary any one of these characteristics, we create a second version of that wave. If we then say that the original wave represents binary 1. the variation can represent binary 0, or vice versa. So, by changing one aspect of a simple electrical signal back and forth, we can use it to represent digital data. Any of the three characteristics listed above can be altered in this way, giving us at least three mechanisms for encoding digital data into an analog signal: amplitude shift keying (ASK), frequency shift keying (FSK), and phase shift keying (PSK). In addition, there is a fourth (and better) mechanism that combines changes in both amplitude and phase called quadrature amplitude modulation (QAM). QAM is the most efficient of these options and is the mechanism used in all modern modems (see Figure 2.21).

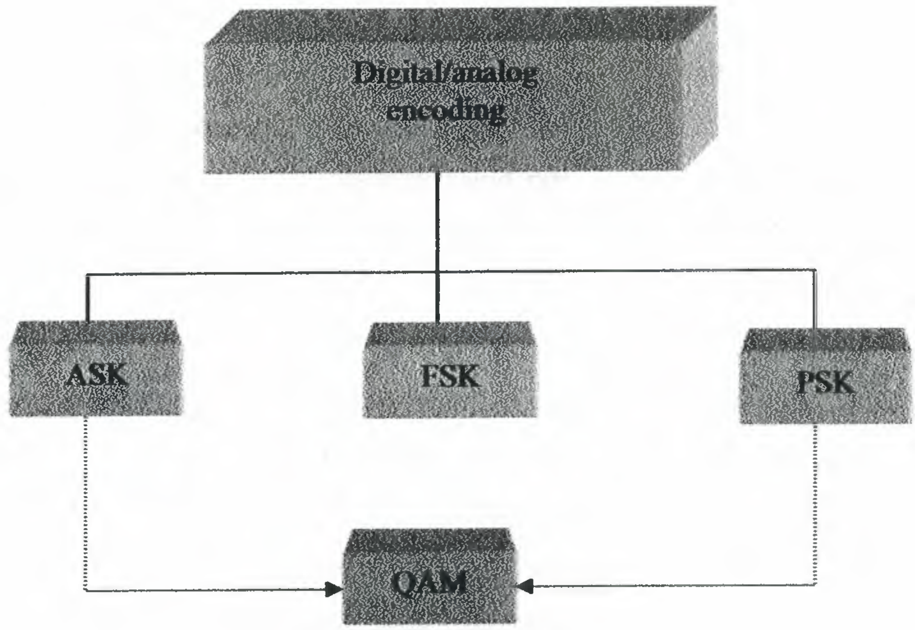


Figure 2.21 Types of Digital-to-Analog Encoding

2.4.1 Aspects of Digital-to-Analog Encoding

Before we discuss specific methods of digital-to-analog encoding, two basic issues must be defined: bit/baud rate and carrier signal.

Bit Rate and Baud Rate

Two terms used frequently in data communication are bit rate and baud rare. Bit rate is the number of bits transmitted during one second. Baud rate refers to the number of signal Units per second that are required to represent those bits. In discussions of computer efficiency the bit rate is the more important -we want to know how long it takes to process each piece of information. In data transmission, however, we are more concerned with how efficiently we can move that data from place to place, whether in pieces or blocks. The fewer signal units required, the more efficient the system and the less bandwidth required to transmit more bits; so we are more concerned with baud rate. The baud rate determines the bandwidth required to send the signal.

Bit rate equals the baud rate times the number of bits represented by each signal unit. The baud rate equals the bit rate divided by the number of bits represented by each signal shift. Bit rate is always greater than or equal to the baud rate.
Bit rate is the number of bits per second. Baud rate is the number of signal units per second. Baud rate is less than or equal to the bit rate.

Carrier Signal

In analog transmission the sending device produces a high-frequency signal that acts as a basis for the information signal. This base signal is called the carrier signal or carrier frequency. The receiving device is tuned to the frequency of the carrier signal that it expects from the sender. Digital information is then encoded onto the carrier signal by modifying one or more of its characteristics (amplitude, frequency, phase). This kind of modification is called modulation (or shift keying) and the information signal is called a modulating signal.

2.4.2 Amplitude Shift Keying (ASK)

In amplitude shift keying (ASK), the strength of the signal is varied to represent binary 1 or 0. Both frequency and phase remain constant while the amplitude changes. Which voltage represents 1 and which represents 0 is left to the system designers. A bit duration is the period of time that defines one bit. The peak amplitude of the signal during each bit duration is constant and its value depends on the bit (0 or 1). The speed of transmission using ASK is limited by the physical characteristics of the transmission medium. Figure 2.22 gives a conceptual view of ASK.

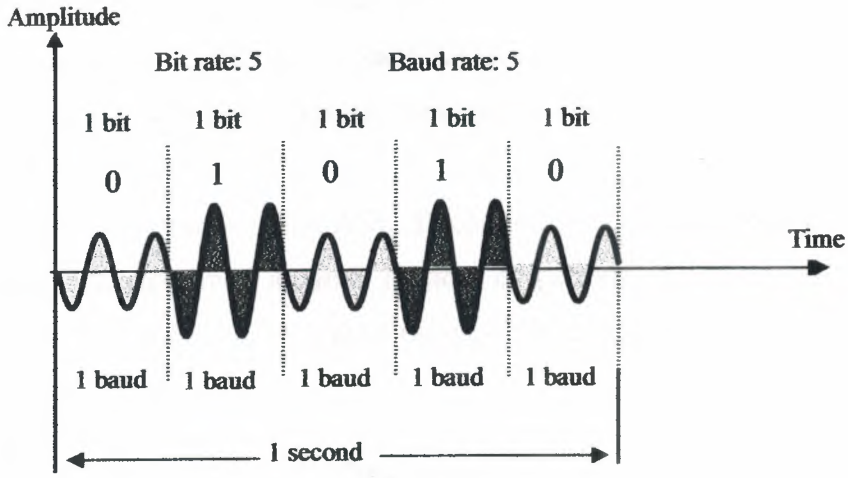


Figure 2.22 ASK Encoding

Unfortunately, ASK transmission is highly susceptible to noise interference. The term *noise* refers to unintentional voltages introduced onto a line by various phenomena such as heat or electromagnetic induction created by other sources. These unintentional voltages combine with the signal to change the amplitude.

Some types of noise, for example thermal noise, are constant enough not to interfere with the intelligibility of the signal. Impulse noise, however, is a sudden surge of energy that can wipe out an entire section of a transmission by inserting high-amplitude spikes where low amplitude was intended. In that case, a section of the signal that was intended to be received as one or more 0s will read as 1s. You can see how surges in voltage would be especially problematic for ASK, which relies solely on amplitude for recognition. Noise usually affects the amplitude; therefore, ASK is the encoding method most affected by noise.

A popular ASK technique is called on-off-keying (OOK). In OOK one of the bit values is represented by no voltage. The advantage is a reduction in the amount of energy required to transmit information.

2.4.4 Frequency Shift Keying (FSK)

In frequency shift keying (FSK), the frequency of the signal is varied to represent binary 1 or 0. The frequency of the signal during each bit duration is constant and its value depends on the bit (0 or 1): both peak amplitude and phase remain constant. Figure 2.23 gives the conceptual view of FSK.

FSK avoids most of the noise problems of ASK. Because the receiving device is looking for specific frequency changes over a given number of periods, it can ignore voltage spikes. The limiting factors of FSK are the physical capabilities of the carrier.

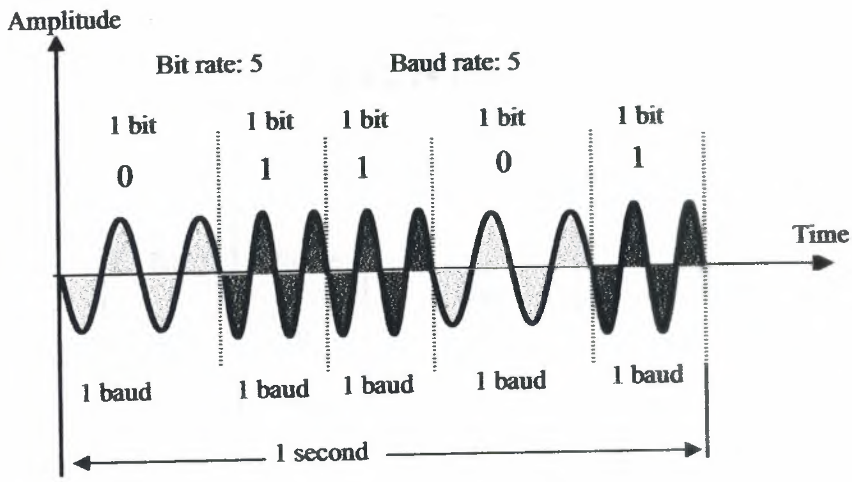


Figure 2.23 FSK Encoding

2.4.4 Phase Shift Keying (PSK)

In phase shift keying (PSK), the phase is varied to represent binary 1 or 0. Both peak amplitude and frequency remain constant as the phase changes. For example, if we start with a phase of 0 degrees to represent binary 0, then we can change the phase to 180 degrees to send binary 1. The phase of the signal during each bit duration is constant and its value depends on the bit (0 or 1). Figure 2.24 gives a conceptual view of PSK.

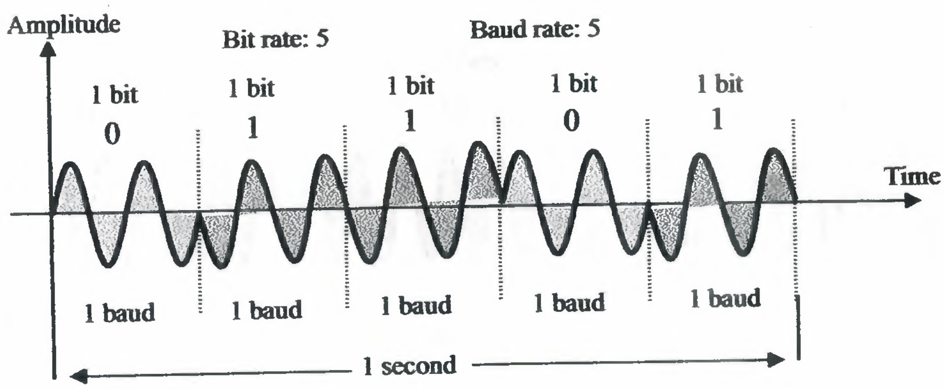
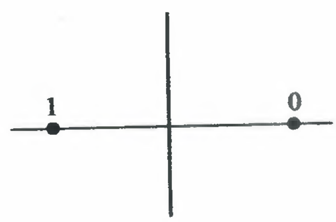


Figure 2.24 PSK

Bit	Phase
0	0
1	180



Bits

Constellation diagram

Figure 2.25 PSK Constellation

The above method is often called 2-PSK, or binary P5K, because two different phases (0 and 180 degrees) are used in the encoding. Figure 2.25 makes this point clearer by showing the relationship of phase to bit value. A second diagram, called a constellation or phase-state diagram, shows the same relationship by illustrating only the phases. PSK is not susceptible to the noise degradation that affects ASK, nor to the bandwidth limitations of FSK. This means that smaller variations in the signal can be detected reliably by the receiver. Therefore, instead of utilizing only two variations of a signal, each representing one bit, we can use four variations and let each phase shift represent two bits (see Figure 2.26).

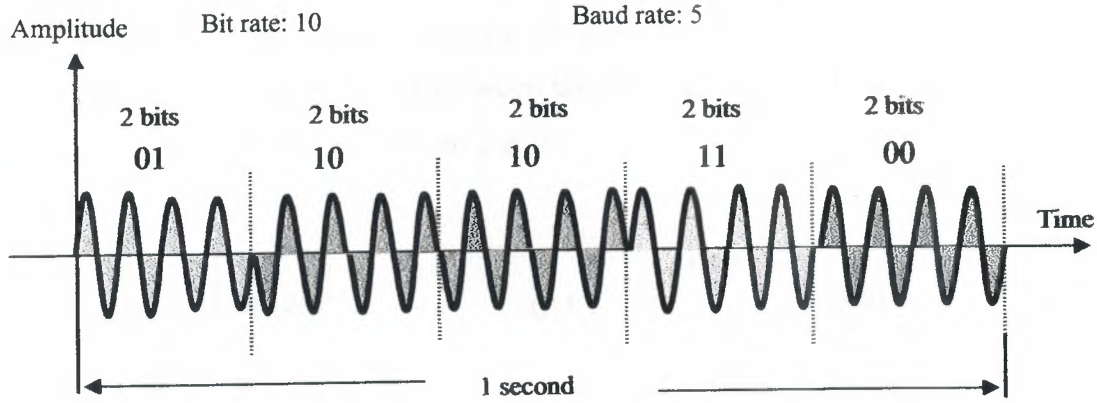
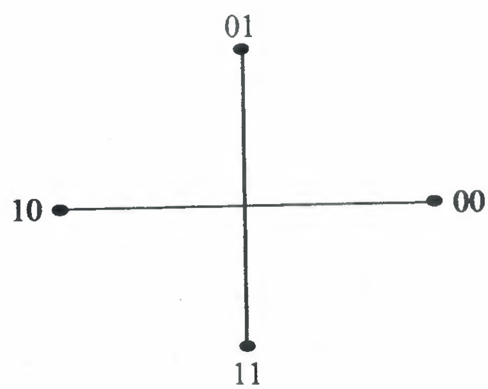


Figure 2.26 4-PSK

The constellation diagram for the signal in Figure 2.26 is given in Figure 2.27. A phase of 0 degrees now represents 00; 90 degrees represents 01; 180 degrees represents 10; and 270 degrees represents 11. This technique is called 4-PSK or Q-PSK. The pair of bits represented by each phase is. Called a dibit. We can transmit data two times as fast using 4-PSK as we can using 2-PSK.

Dibit	Phase
00	0
01	90
10	180
11	270

Dibit
(2 bits)



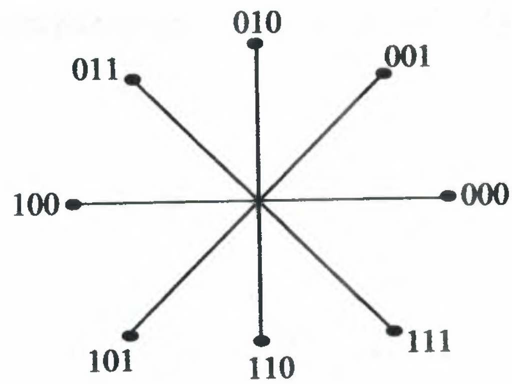
Constellation Diagram

Figure 2.27 4-PSK Characteristics

We can extend this idea to 8-PSK. Instead of 90 degrees, we now vary the signal by shifts of 45 degrees. With 8 different phases, each shift can represent three bits (one tribit) at a time. (As you can see, the relationship of number of bits per shift to number of phases is a power of two. When we have four possible phases, we can send two bits at a time—2^2 equals 4. When we have eight possible phases, we can send three bits at time—2^3 equals 8). Figure 2.28 shows the relationships between the phase shifts and the tribits each one represents. 8-PSK is three times faster than 2-PSK

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

Tribit
(3 bits)



Constellation Diagram

Figure 2.28 8-PSK Characteristics

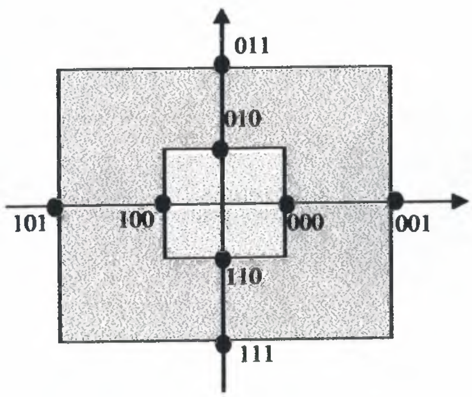
2.4.5 Quadrature Amplitude Modulation (QAM)

PSK is limited by the ability of the equipment to distinguish small differences in phase. This factor limits its potential bit rate.

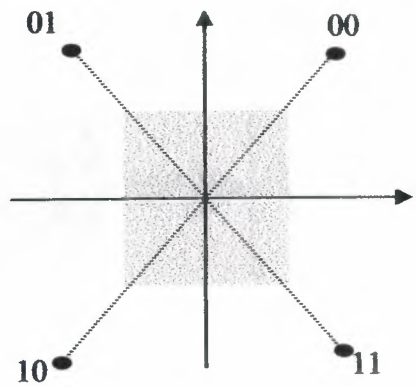
So far, we have been altering only one of the three characteristics of a sine wave at a time to achieve our encoding, but what if we alter two? Bandwidth limitations make combinations of FSK with other changes practically useless. But why not combine ASK and PSK? Then we could have x variations in phase and y variations in amplitude, giving us x times y possible variations and the corresponding number of bits per variation. Quadrature amplitude modulation (QAM) does just that. The term quadrature is derived from the restrictions required for minimum performance and is related to trigonometry.

Quadrature amplitude modulation (QAM) means combining ASK and PSK in such a way that we have maximum contrast between each bit, dibit, tribit, quadbit, and so on.

Possible variations of QAM are numerous. Theoretically any measurable number of changes in amplitude can be combined with any measurable number of changes in phase. Figure 2.29 shows two possible configurations, 4-QAM and 8-QAM. In both cases, the number of amplitude shifts is fewer than the number of phase shifts. Because amplitude changes are susceptible to noise and require greater shift differences than do phase changes, the number of phase shifts used by a QAM system is always larger than the number of amplitude shifts. The time-domain plot corresponding to the 8-QAM signal in Figure 2.29 is shown in Figure 2.30.



8-QAM
2 amplitudes, 4 phases



4 -QAM
1 amplitude, 4 phases

Figure 2.29 4-QAM and 8-QAM Constellation

Other geometric relationships besides concentric circles are also possible. The first example, three amplitudes and 12 phases, handles noise best because of a greater ratio of phase shift to amplitude. It is the ITU-T recommendation. The second example, four amplitudes an eight phases, is the OSI recommendation. If you examine the graph carefully, you will notice that although it is based on concentric circles, not every intersection of phase and amplitude is utilized. In fact, 4 times 8 should allow for 32 possible variations. But by using only half of those possibilities, the measurable differences between shifts are increased and greater signal readability is ensured. In addition, several QAM designs link specific amplitudes with specific phases. This means that even with the noise problems associated with amplitude shifting, the meaning of a shift can be recovered from phase information. In general, therefore, a second advantage of QAM encoding over ASK encoding is its lower susceptibility to noise.

2.5 Analog-To-Analog Encoding

Analog-to-analog encoding is the representation of analog information by an analog signal. Radio, that familiar utility, is an example of an analog-to-analog communication. Figure 2.30 shows the relationship between the analog information, the analog-to-analog conversion hardware, and the resultant analog signal.

Analog-to-analog encoding can be accomplished in three ways: amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). See Figure 2.31.

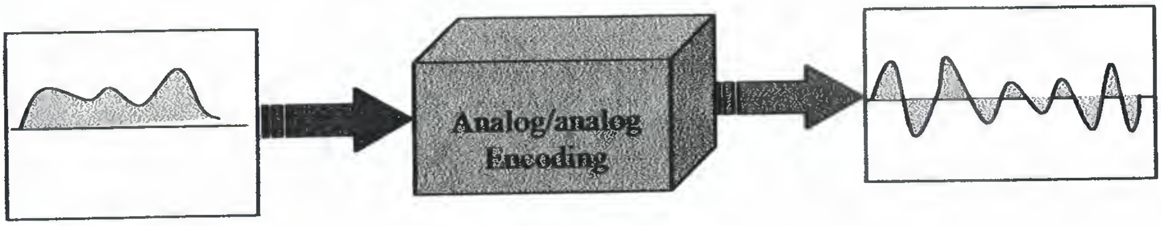


Figure 2.30 Analog-to-Analog Encoding

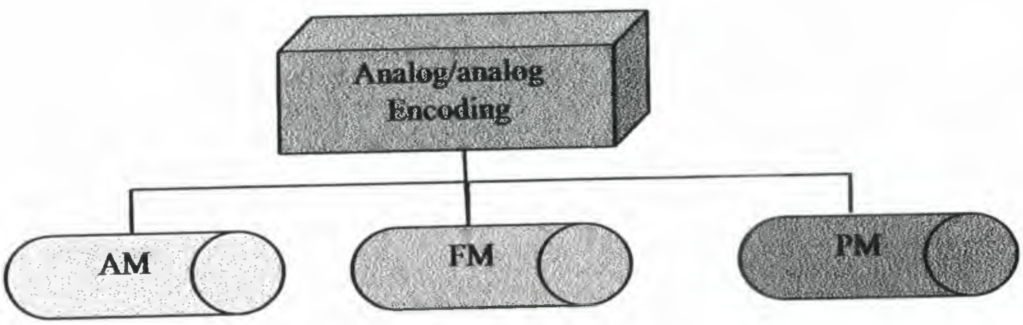
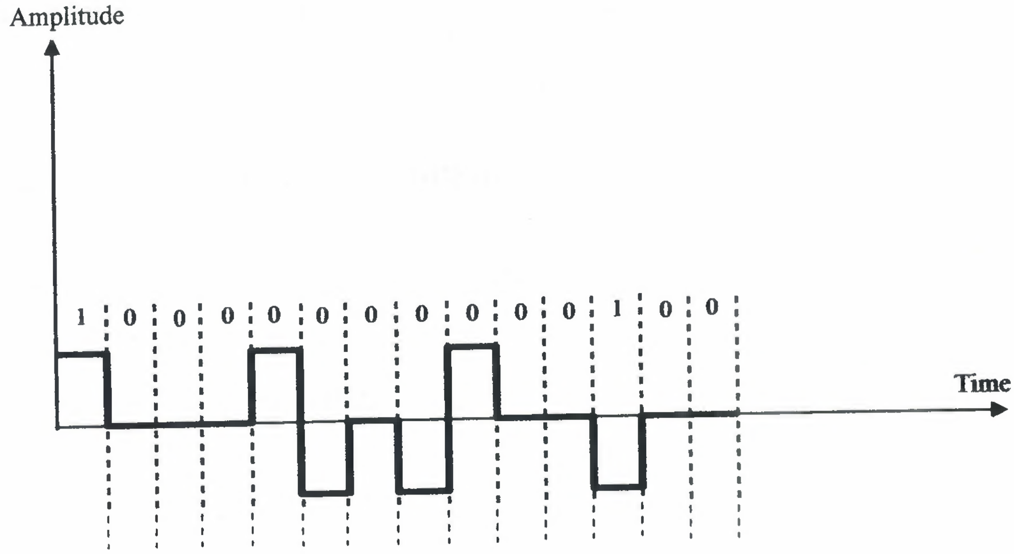


Figure 2.31 Types of Analog-to-Analog Encoding

Examples:

Using B8ZS, encode the bit stream 10000000000100. Assume that the polarity of the first 1 is positive.

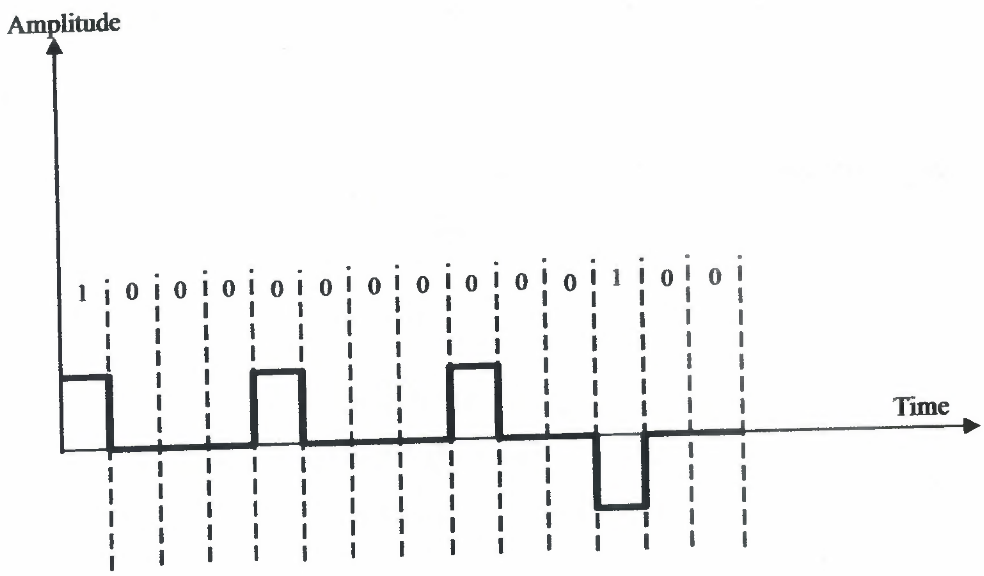
Solution:



Example:

Using HDB3, encode the bit stream 10000000000100. Assume that the number of 1s so far is odd and the first 1 is positive.

Solution:



Example:

Given a bandwidth of 10,000 Hz (1000 to 11,000), draw the full-duplex ASK diagram of the system. Find the carriers and the bandwidths in each direction. Assume there is no gap between the bands in two directions.

Solution:

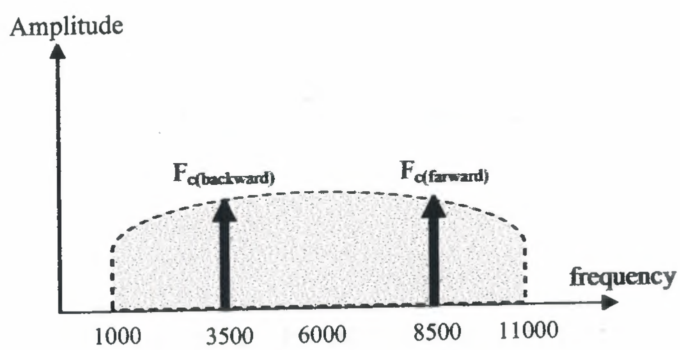
For full-duplex ASK the bandwidth for each direction is

BW = 10,000/2 = 5000 Hz

The carrier frequencies can be chosen at the middle of each band

$F_{c(forward)} = 1000 + 5000/2 = 3500$ Hz

$F_{c(backward)} = 1100 - 5000/2 = 8500$ Hz



Example:

Find the bandwidth for an FSK signal transmitting at 2000 bps. Transmission is in half-duplex mode and the carriers must be separated by 3000 Hz.

Solution:

For FSK, if f_{c1} and f_{c0} are the carrier frequencies, then

$$BW = \text{Baud rate} + (f_{c1} - f_{c0})$$

However, the baud rate here is the same as the bit rate. Therefore,

$$BW = \text{Bit rate} + (f_{c1} - f_{c0})$$
$$= 2000 + 3000 = 5000 \text{ Hz}$$

Example:

Find the maximum bit rates for an FSK signal if the bandwidth of the medium is 12,000 Hz and the distance between the two carriers must be at least 2000 Hz. Transmission is in full-duplex mode.

Solution:

Because the transmission is full-duplex, only 6000 Hz is allocated for each direction. For FSK, if f_{c1} and f_{c0} are the carrier frequencies.

$BW = \text{Baud rate} + (f_{c1} - f_{c0})$

$\text{Baud rate} = BW - ((f_{c1} - f_{c0}) = 6000 - 2000 = 4000$

But because the baud rate is the same as the bit rate, the bit rate is 4000 bps.

Example:

Find the bandwidth for a 4-PSK signal transmitting at 2000 bps. Transmission is in half-duplex mode.

Solution:

For 4-PSK the baud rate is half of the bit rate. The baud rate is therefore 1000. A PSK signal requires a bandwidth equal to its baud rate. Therefore, the bandwidth is 1000 Hz.

Example:

Given a bandwidth of 5000 Hz for an 8-PSK signal, what are the baud rate and bit rate?

Solution:

For PSK the baud rate is the same as the bandwidth, which means the baud rate is 5000. But in 8-PSK the bit rate is three times the baud rate. So the bit rate is 15,000 bps.

Example:

A constellation diagram consists of eight equally spaced points on a circle. If the bit rate is 4800 bps, what is the baud rate?

Solution:

The constellation indicates 8-PSK encoding with the points 45 degrees apart. Since $2^3 = 8$, three bits are transmitted with each signal element. Therefore, the baud rate is

$$4800/3 = 1600 \text{ baud}$$

Example:

Compute the bit rate for a 1000 baud 16-QAM signal.

Solution:

A 16-QAM signal means that there are four bits per signal element since $2^4 = 64$. Thus,

$$(1000)(4) = 4000 \text{ bps}$$

3. TRANSMISSION MEDIA

3.1 Mathematical Models for Communication Channels

In the design of communication systems we find it convenient to construct mathematical models that reflect the most important characteristics of the transmission medium. Below, we provide a brief description of the channel models that are frequently used to characterize many of the physical channels that we encounter in practice.

The additive noise channel. The simplest mathematical model for a communication channel is the additive noise channel, illustrated in Figure 3.1.

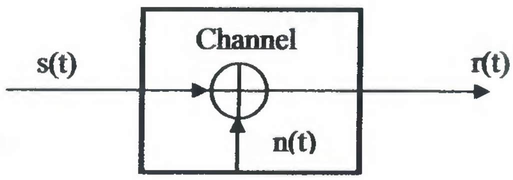


Figure 3.1 Mathematical Model for Communication Channel.

In this model, the transmitted signal s *(t)* is corrupted by an additive random noise process *n(t).* Physically, the additive noise process may arise from electronic components and amplifiers at the receiver of the communication system, or from interference encountered in transmission as in the case of radio signal transmission.

If the noise is introduced primarily by electronic components and amplifiers at the receiver, it may be characterized as thermal noise. This type of noise is characterized statistically as a *Gaussian noise process.* Hence, the resulting mathematical model for the channel is usually called the *additive Gaussian noise channel.* In this case the received signal is

$$r(t) = \alpha s(t) + n(t)$$

Where α represents the attenuation factor.

The linear filter channel. In some physical channels such as wire-line telephone channels, filters are used to ensure that the transmitted signals do not exceed specified bandwidth limitations and thus do not interfere with one another. Such channel (Figure 3.2) output can be characterized as

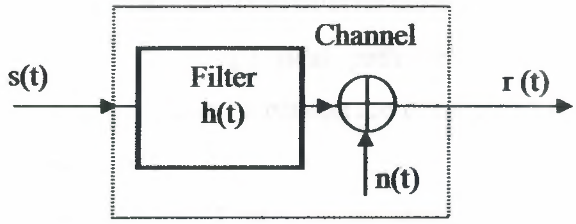


Figure 3.2 Output Channel Characterization

$$r(t) = s(t) * h(t) + n(t) = \int_{-\infty}^{\infty} h(\tau)s(t-\tau)d\tau + n(t) \tag{3.1}$$

Where h(t) is the impulse response of the linear filter and symbol * denotes convolution.

The linear time-variant filter channel. Physical channels such as underwater acoustic channels and ionosphere radio channels, which result in time-variant multi-path propagation of the transmitted signal, may be characterized mathematically as time-variant linear filters. Such system is characterized by a time-variant channel with impulse response h (τ; t) filters (Figure 3.3). For an input signal s (t), the channel output is

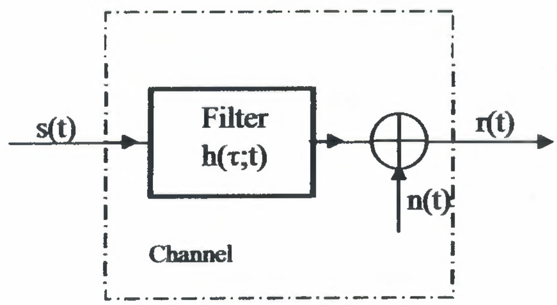


Figure 3.3 Time-Variant Channel with Impulse Response.

$$r(t) = s(t) * h(\tau; t) + n(t) \tag{3.2}$$

The three mathematical models described above adequately characterise a majority of physical channels encountered in practice.

3.2 Transmission Impairments

The transmission medium is the physical path between transmitter and receiver. The characteristics and quality of data transmission are determined both by the nature of the signal and the nature of the medium.

With any communication system, it must be recognized that the signal that is received will differ from the signal that is transmitted due to various transmission impairments. For analog signals, these impairments introduce various random modifications that degrade the signal quality. For digital signals, bit errors are introduced: a binary 1 is transformed into a binary 0 and vice versa.The most significant impairments are: Attenuation, Delay distortion and Noise. The various impairment effects that can degrade a signal during transmission are shown in Figure 3.4.

3.2.1 Attenuation

As a signal propagates along a transmission medium its amplitude decreases. This is known as signal attenuation. To compensate the attenuation, amplifiers are inserted at intervals along the cable to restore the received signal to its original level. Signal attenuation increases as a function of frequency. To overcome this problem, the amplifiers are designed to amplify different frequency by varying gains of amplifications. These devices are known as equaliser. For guided media (Twisted wires, Coaxial cables and Fiber optic cables) attenuation, is generally logarithmic and it is typically expressed as a constant number of decibels per unit distance

$$\mathbf{N,dB = 10\log\frac{P_2}{P_1}}, \quad \textbf{where N – number of decibels}$$

P_1, P_2 – input and output powers.

Example 3.1. A signal with power 10 mW is inserted into a transmission line. The power measured some distance is 5 mW. Find the loss.

$$\boldsymbol{Loss = 10\log_{10}\frac{5}{10} = 10(-0.3) = -3dB} \tag{3.3}$$

Taking into account that power is proportional to the square of voltage:

$P_1 = U_1^2/R$; $P_2 = U_2^2/R$ and

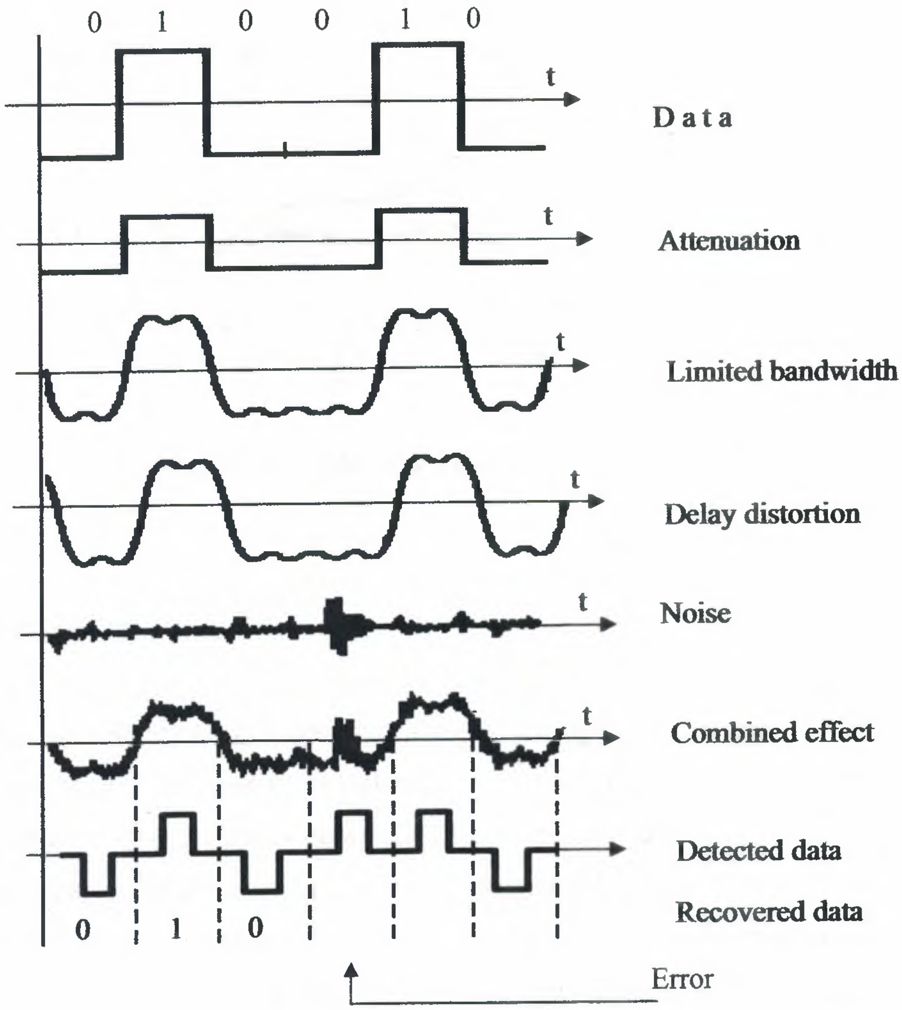


Figure 3.4 Various Impairment Effects

$$N, dB = 20 \log \frac{U_2}{U_1}$$

For unguided media attenuation is a more complex function of distance and the make-up of the atmosphere. An example is shown in Figure 3.5, which shows attenuation as a function of frequency for a typical wire line. In Figure 3.5, attenuation is measured relative to the attenuation at 1000 Hz. Positive values on the y-axis represent attenuation greater than that at 1000 Hz. For any other frequency f, the relative attenuation in decibels is $N_f = 10 \log_{10} P_f /$

P_{1000} .The solid line in Figure shows attenuation without equalization. The dashed line shows the effects of equalization.

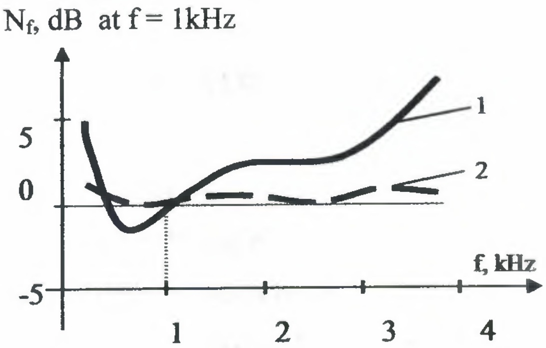


Figure 3.5. Attenuation without Equalization.

3.2.2 Delay Distortion

Delay distortion is a phenomenon peculiar to guided transmission media. The distortion is caused by the fact that the velocity of propagation of a signal through a guided medium varies with frequency. This effect is referred to as delay distortion, since the received signal is distorted due to variable delay in its components. Delay distortion is particularly critical for digital data. Consider that a sequence of bits is being transmitted, using either analog or digital signals. Because of delay distortion, some of the signal components of one bit position will spill over into other bit positions, causing inter-symbol interference, which is a major limitation to maximum bit rate over a transmission control. Equalizing techniques can also be used for delay distortion.

3.2.3 Noise

For any data transmission, the received signal will consist of the transmitted signal, modified by the various distortions imposed by the transmission system, plus additional unwanted signals that are inserted somewhere between transmission and reception. These undesired signals are referred to *Noise* and can be divided into four categories: Thermal noise, Inter-modulation noise, Cross-talk and Impulse noise.

The Thermal noise is due to thermal agitation of electrons in a conductor. It is present in all electronic devices and transmission media and is a function of temperature. Thermal noise is uniformly distributed across the frequency spectrum and hence is often referred to as *white* *noise.* Thermal noise cannot be eliminated and therefore places an upper bound on communications system performance. This noise is assumed to be independent of frequency. The thermal noise in watts present in a bandwidth of W-hertz can be expressed as

$$N = kTW$$

Or, in decibel-watts:

$N = 10 \log k + 10 \log T + 10 \log W$

$N = -228.6\ (dbW) + 10 \log T + 10 \log W$

Where No - noise power density, watts/hertz;

k - Boltzmann's constant $k = 1.3803 \times 10^{-23}$ $J/^{0}K$; T - temperature, degrees Kelvin

When signals at different frequencies share the same transmission medium, the result may be *inter-modulation noise.* The effect of inter-modulation noise is to produce signals at a frequency, which is the sum or difference of the two original frequencies or multiples of those frequencies. For example, the mixing of signals at frequencies f_1 and f_2 might produce energy at the frequency $f_1 + f_2$. This derived signal could interfere with an intended signal at the frequency $f_1 + f_2$.

Inter-modulation noise is produced when there is some non-linearity in the transmitter, receiver, or interviewing transmission system.

Cross-talk has been experienced by anyone who, while using the telephone, he/she is able to hear another conversation: it is an unwanted coupling between signal paths. It can occur by electrical coupling between nearby twisted pair or rarely coaxial cable lines carrying multiple signals. Among several types of cross-talk the most limiting impairment for data communication systems is near-end cross-talk (self-cross-talk or echo), since it is caused by the strong signal output by the transmitter output being coupled with much weaker signal at the input of the local receiver circuit. Adaptive noise canceller is used to overcome this type of impairment.

A Impulse noise, has short duration and have relatively high amplitude. It is generated from a variety of causes, including external electromagnetic disturbances, such as lightning, electrical impulses associated with the switching circuits used in the telephone exchange.

Impulse noise is generally only a minor annoyance for analog data. For example, voice transmission can be corrupted by short clicks and crackles with no loss of intelligibility. However, impulse noise is the primary source of error in digital data communication. For example, impulse noise of 0.01 s duration would not destroy any voice data, but would wash out about 50 bits of data is being transmitted at 4800 bps.

3.3 Channel Capacity

The rate at which data can be transmitted over a given communication channel, under given conditions, is referred to as the channel capacity.

There are four concepts here that we are trying to relate to one another.

- Data rate: This is the rate, in bits per second (bps), at which data can be transmitted.
- Bandwidth: This is the bandwidth of the transmitted signal as constrains by the transmitter and the nature of the transmission medium, expressed by Hertz.
- Noise: The average level of noise over the communications path.
- Error rate: The rate at which errors occur, where an error is the reception of a 1 when a 0 was transmitted or the reception of a 0 when a 1 was transmitted.

Communication facilities are expensive and, in general, the greater the bandwidth of a facility the greater the cost. Furthermore, all transmission channels of any practical interest are of limited bandwidth. The limitations arise from the physical properties of the transmission medium or from deliberate limitations at the transmitter on the bandwidth to prevent interference from other sources. Accordingly, we would like to make as efficient use as possible of a given bandwidth.

Let us consider the case of a channel that is noise-free. In this environment, the limitation on data rate is simply the bandwidth of the signal. A formulation of this limitation, due to Nyquist, states that if the rate of signal transmission is 2W, then a signal with frequencies no greater than W is sufficient to carry the data rate. The conserve is also true: Given a bandwidth of W, the highest signal rate that can be carried is 2W.

Example 3.2. A voice channel bandwidth is of $W = 3100$ Hz. Find the channel capacity. Solution: $C = 2\,W = 6200$ bps.

However, as we shall see in chapter 3, signals with more than two levels can be used; that is each signal element can represent more than one bit. For example; if M possible voltage levels are used, then each signal element can be represented by $n = \log_2 M$ numbers of bits. With multilevel signaling, the Nyquist formulation becomes

$$C = 2\,W \log_2 M$$

Thus, for $M = 8$, a value used with some modems, C becomes 18600 bps.

An important parameter associated with a channel is a signal-to-noise ratio (SNR) expressed as

$$SNR = 10\log_{10} (S/N) \text{ dB}$$

Where S/N – signal –to- noise powers ratio. Clearly a high S/N will mean a high - quality signal and a low number of required intermediate repeaters.

The signal - to noise ratio is important in the transmission of digital data because it sets the upper bound on the achievable data rate. The maximum channel capacity, in bits per second, obeys the equation attributed as the Shannon – Hartley law

$$C = W \log_2 (1+ S/N) \approx 3{,}32\ W\log_{10} (1+S/N),$$

3.4 Guided Media

The guided media includes: twisted pair, coaxial cable and fiber-optic cable (see Figure 3.6).

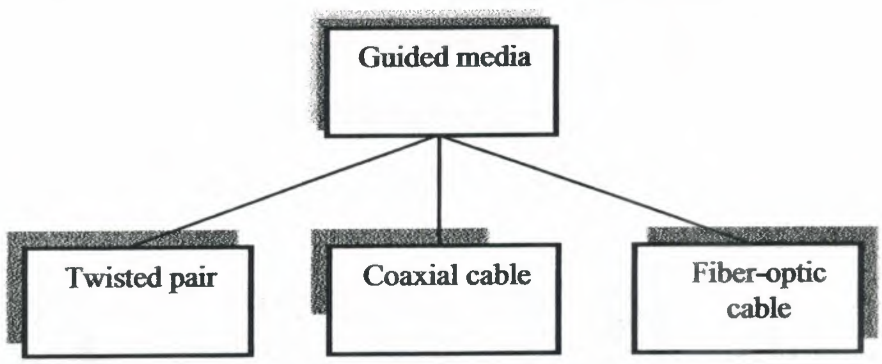


Figure 3.6 Categories of Guided Media

Table 3.1 contains the typical characteristics for guided media

Table 3.1 Typical characteristics for guided media

Medium Transmission	Total Data Rate	Bandwidth	Repeater Spacing
Twisted pair	1-100 Mbps	100Hz -5 MHz	2 - 10 km
Coaxial cable	1Mbps-1 Gbps	100 Hz – 500 MHz	1 - 10 km
Optical fiber	2 Gbps	2 GHz	10- 10 0 km

In the past two parallel flat wires were used for communications. Each wire is insulated from the other and both are open to free space. This type of line is used for connecting equipment that is up to 50 m apart using moderate rate (less than 20 kbps). The signal, typically a voltage or current level relative to some ground reference is applied to one wire while the ground reference is applied to the other. Although a two wire open line can be used to connect two computers directly, it is used mainly for connecting computers with modems. As shown in Figure 3.7 two simple wires more sensitive to noise interference.

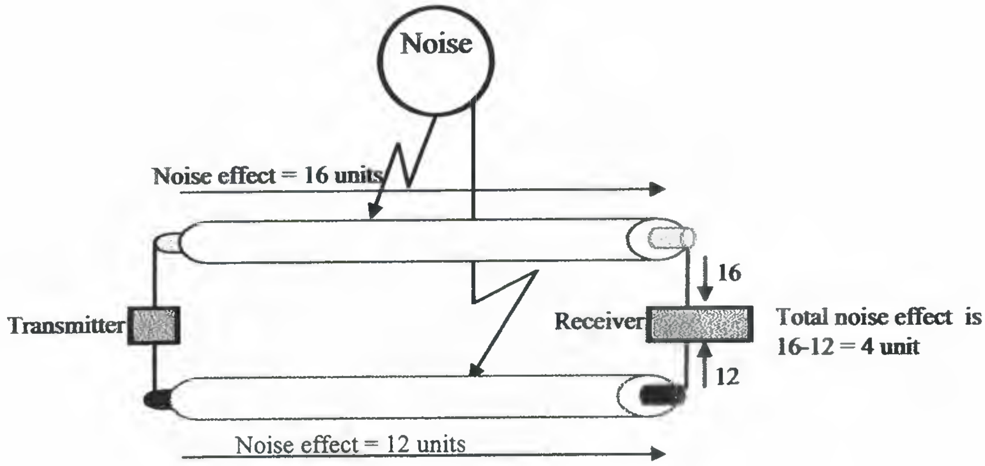


Figure 3.7 Effect of Noise in Parallel Lines.

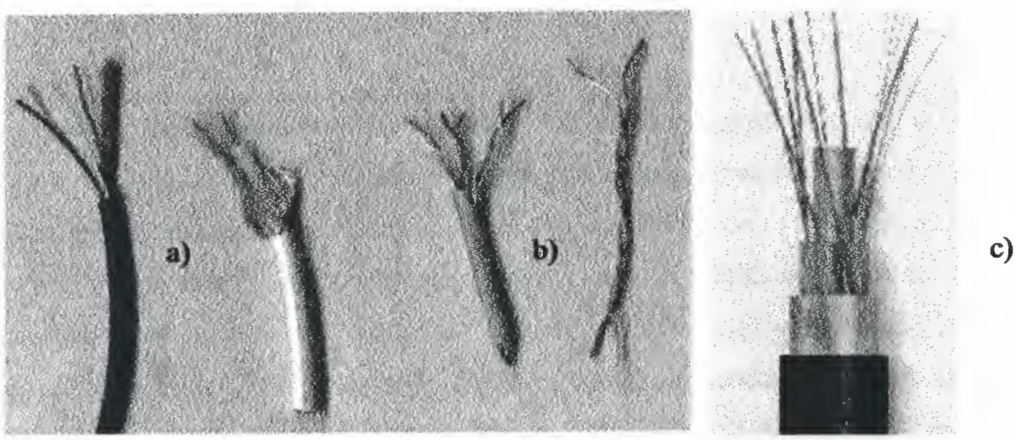


Figure 3.8 (a) Layouts of the Coaxial Cables, (b) Twisted Pairs and (c) Fiber-Optic Cables.

3.4.1 Twisted Pair

A twisted pair consists of two insulated copper wires. Over longer distances, cables may contain hundreds of pairs. The twisting of the individual pairs minimizes electromagnetic interference between the pairs (see Figure 3.9).

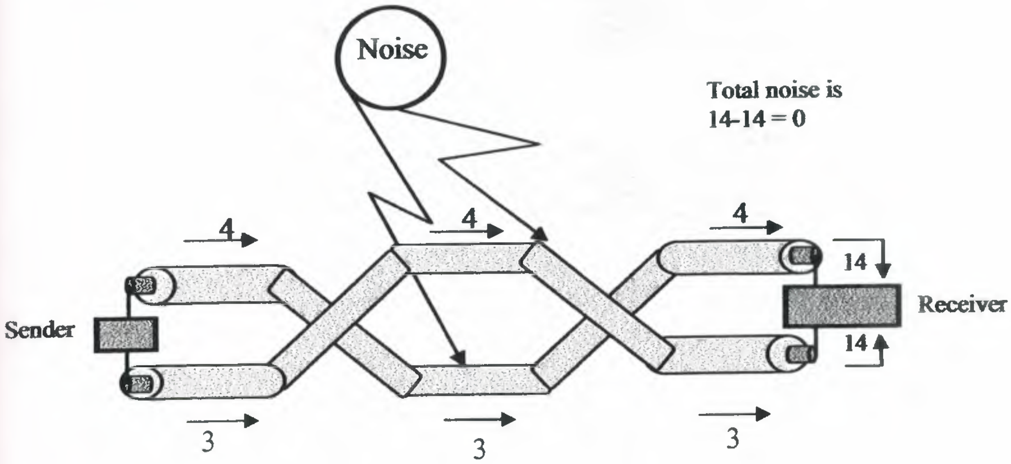


Figure 3.9 Effect of Noise on Twisted-Pair Lines

Wire pairs can be used to transmit both analog and digital signals. For analog signals, amplifiers are required about every 5 to 6 km. For digital signals, repeaters are used at every 2 or 3 km. It is the backbone of the telephone system as well as the low - cost microcomputer local network within a building. In the telephone system, individual telephone sets are connected to the local telephone exchange or "end office" by twisted - pair wire. These are referred to as "local loops". Within an office building, telephone service is often provided by means of a Private Branch Exchange (PBX). For modern digital PBX systems, data rate is about 64 kbps. Local loop connections typically require a modem, with a maximum data rate of 9600 bps. However, twisted pair is used for long - distance trucking applications and data rates of 100 Mbps or more may be achieved.

The twisted pair comes in two forms: shielded (STP) and unshielded (UTP). Figure 3.10 shows STP (a) and UTP (b, c). The metal casing prevents the penetration of electromagnetic noise and eliminates cross-talk. Materials and manufacturing requirements make STP more expensive than UTP but less susceptible to noise. UTP is cheap, flexible, and easy to use.

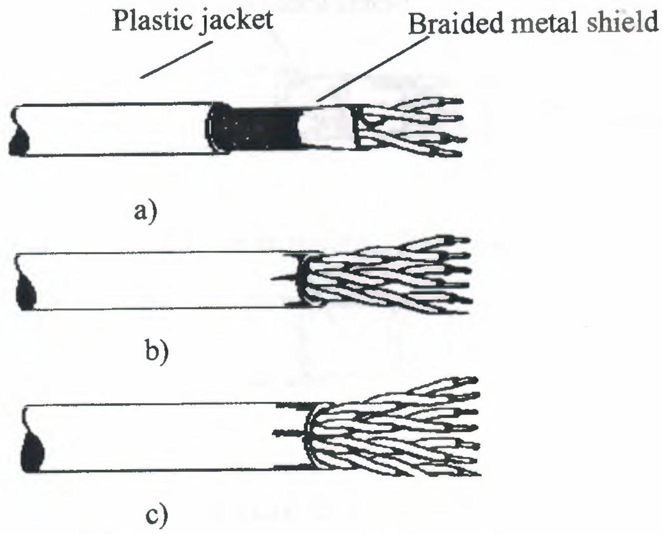


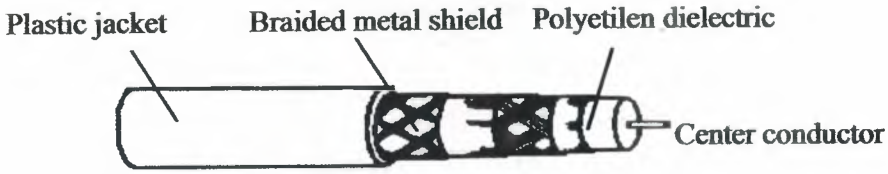
Figure 3.10 Shows STP (a) and UTP (b, c).

3.4.2 Coaxial Cable

The main limiting factor of a twisted pair line are its capacity and a phenomenon known as the skin effect. As the bit rate increases, the current flowing in the wires tends to flow only on the outer surface of the wire, thus using the less available cross-section. This increases the electrical resistance of the wires for higher frequency signals, leading to the attenuation In addition, at higher frequencies, more signal power is lost as a result of radiation effect.

Coaxial cables, like twisted pairs, consist of two conductors, but are constructed differently to permit it to operate over a wider range of frequencies. Coaxial cables have been perhaps the most versatile transmission medium and is enjoying increasing utilising in a wide variety of applications. The most important of these are long-distance telephone and television transmission, television distribution, and short-range connections between devices and local area networks. In Figure 3.11 are shown the constructions of the coaxial cables. Using frequency-division multiplexing a coaxial cable can carry over 10,000 voice channels simultaneously. Coaxial cables are used to transmit both analog and digital signals.

The principal constraints on performance are attention, thermal noise, and intermodulation noise.



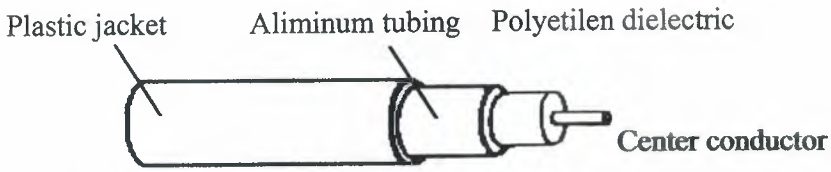


Figure 3.11 Coaxial Cable

3.5 Unguided Media

There are three basic modes of getting a radio wave from the transmitting to receiving antenna: ground wave, space wave, sky wave proportions (Figure 3.12)

The subdivision of the electromagnetic frequency range is given in the Table 3.2

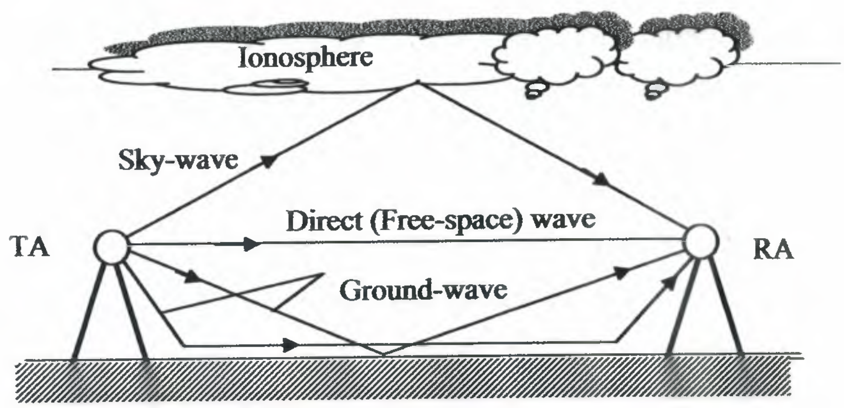


Figure 3.12 Sky Wave Proportion

Table 3.2 Frequency Range for Wireless Communication

Frequency Band	Name	Data rate	Principal applications
30 – 300 kHz	LF (Low Frequency)	0.1 – 100 bps	Navigation, Submarine
300 kHz - 3000 kHz	MF (Medium Frequency)	10 – 1000 bps	AM radio
3 – 30 MHz	HF (High Frequency)	10 – 3000 bps	Shortwave radio, CB radio
30 – 300 MHz	VHF (Very High Frequency)	To 100 kbps	VHF Television, FM radio
300 – 3000 MHz	UHF(Upper High Frequency)	To 10 Mbps	UHF Television Mobile communication Terrestrial Microwave
3 – 30 GHz	SHF (Super High Frequency)	To 100 Mbps	Satellite and Terrestrial microwaves, Radar

The frequency of the radio wave is of primary importance in considering the performance of each type of propagation.

Ground – Wave Propagation

A ground wave is a radio wave that travels along the earth's surface. It is sometimes referred to as a *surface wave.* Attenuation of ground waves is directly related to the surface impedance of the earth. This impedance is a function of conductivity and frequency. If the earth's surface is highly conductive, the absorption of wave energy, and thus its attenuation, will be reduced. Ground–wave propagation is much better over water (especially salt water) than say a very dry (poor conductivity) desert terrain. The ground losses increase rapidly with increasing frequency. For these reasons ground waves are not very effective at frequencies above 2 MHz. Ground- wave propagation is the only way to communicate into the ocean with submarines (about 100 miles distance). To minimise the attenuation of seawater, extremely low frequency (ELF) propagation is utilised. A typically used frequency is 100 Hz, the attenuation is about 0.3 dB/m.

Space – Wave (line-of-site propagation) Propagation

The two types of space waves are shown in Figure 3.12. They are the direct wave and ground reflected wave. Do not confuse these with the ground wave just discussed. The direct wave is by far the most widely used mode of antenna communications. The propagated wave is directed from transmitting to receiving antenna and does not travel along the ground. The earth's surface, therefore, does not attenuate it. The direct space wave has one severe limitation – it is basically limited to so called line–of –sight transmission distances. Thus, the antenna height and the curvature of the earth are the limiting factors. The actual radio horizon is about 1/3 times greater then the geometric line of sight due to diffraction effects and is empirically predicted by the following approximation:

$$d = \sqrt{2h_T} + \sqrt{2h_R} \qquad (3.4)$$

Where d - radio horizon (mi); h_T- transmitting antenna height (ft);

h_R - receiving antenna height (ft)

Ghosting in TV reception. Any tall or massive objects obstruct space waves. This results in diffraction (and subsequent shadow zones) and reflections. Reflections pose a specific problem since, for example, reception of a TV signal may be the combined result of a direct space wave and a reflected space waves. This condition results in ghosting, which manifests itself in the form of a double – image distortion. This is due to the two signals arriving at the receiver at two different times. A possible solution to the ghosting problem is to retune the receiving antenna orientation so that the reflected wave is too weak to be displayed.

Sky Wave Propagation

The sky wave has the ability to strike the ionosphere. It can be refracted from it to the ground, strike the ground, be reflected back toward the ionosphere, and so on. A frequency occurring problem is *signal multipath.* The multipath occurs when the transmitted signal arrives at the receiver via multipath paths at different delays. Signal multipath results intersymbol interference in a digital communication system The signal components arriving via different propogation paths may add destructively, resulting in a phenomenon called *signal fading.* Sky wave propogation ceases to exist at frequencies above 30 MHz. However it is possible to have atmospheric scatter propagation at the range of 30 MHz and troposphere scattering at 40 MHz to 300MHz.

Microwaves

Two general ranges of frequencies are of interest in discussion.

1. Microwave frequencies that cover a range of about 3 to 30 GHz. At these frequencies, highly directional beams are possible, and microwave is quite suitable for point-to-point transmission.
2. Radio waves that cover a range of about 30 MHz to 1 GHz. At these frequencies, omnidirectional transmission is possible, and microwave is quite suitable for broadcasting.

We will refer to signals in the range 30 MHz to 1 GHz as radio waves.

Omnidirectional transmission is used and signals at these frequencies are suitable for broadcast applications. The most common type of microwave antenna is the parabolic "dish". A typical size is about 10 ft in diameter. The antenna is fixed rigidly and focuses a narrow beam to achieve line-of-sight transmission to the receiving antenna. Microwave antennas are usually located at substantial heights above ground level in order to extend the range between antennas and to be able to transmit over intervening obstacles.

The primary use for terrestrial microwave systems is in long-haul telecommunications service, as an alternative to coaxial cable for transmitting television and voice. Like coaxial cable, microwave can support high data rates over long distances. The microwave facility requires far fewer amplifiers or repeaters than coaxial cable for the same distance, but requires line of sight transmission.

Another increasingly common use of microwave is for short point- to point links between buildings. This can be used for closed- circuit TV or as a data link between local networks. Finally, a potential use for terrestrial microwave is to provide digital data transmission in small regions (radius < 10 km). This concept has been termed as "local data distribution" and would provide an alternative to phone lines for digital networking.

The microwave transmission covers a substantial portion of the spectrum. Common frequencies used for transmission are in the range 2 to 40 GHz. The higher the frequency used the higher the potential bandwidth, and therefore, the higher the potential data rate.

As with any transmission system, a main source of loss for microwave is attenuation. For microwave (and radio frequency), the loss can be expressed as

$$L = 10\mathrm{Log}(\frac{4\pi d}{\lambda})^2 \ \mathrm{dB}; \quad \lambda[m] = \frac{3.10^8}{f[Hz]} \tag{3.5}$$

Where d is the distance and λ is the wavelength in the same units.

Thus loss varies as the square of the distance. This is in contrast to twisted pair and coaxial cable where the loss varies logarithmically with distance (linear in decibels). Thus repeaters or amplifiers may be placed farther apart for microwave systems – 10 to 100 km is typical. Attenuation is increased with rainfall. Another source of impairment for microwave is interference.

3.6 Overview of Fiber Optic Cable

The fiber-optic is defined as branch of optics that deals with the transmission of light through ultra pure glass, plastic or some other form of transparent media. One of first noted experiment that demonstrated the transmission of light through a dielectric medium has been created to John Tyndall. In 1854 John Tyndall demonstrated that light could be guided through stream of water based on the principle of total internal reflection.

In1880 Alexander Graham Bell invented the photo phone, a device that transmits voice signals over a beam of light.

Great interest in communication at optical frequencies was created in 1958 with the invention of the laser by Charles H. Townes.

In 1966 Charles K. Kao and George Hockham of Standard Telecommunications Laboratories of England performed several experiments to prove that, if glass could be made more transparent by reducing its impurities, light loss could be minimized. Their research led to a publication in which they predicted that optical fiber could be made pure enough to transmit light several kilometers. In the next two decades researchers worked intensively to reduce the attenuation to 0.16 dB/km.

In1988 the Synchronous Optical Network (SONET) was published by the American National Standards Institute (ANSI).

1995 Multimedia applications for business have become the major impetus for increased use of optical fiber within the LAN, MAN, and WAN environment.

3.6.1 Advantages and Disadvantages of the FOS

a) Advantages

The major advantages are:

Bandwidth One of the most significant advantages that fiber has over copper or other transmission media is a bandwidth. Bandwidth is directly related to the amount of information that can be transmitted per unit time. Today's advanced fiber optic systems are capable of transmitting several gigabits per second over hundreds of kilometers. Ten thousands of voice channels can now be multiplexed together and sent over a single fiber strand.

Less Lose. Currently, fiber is being manufactured to exhibit less than a few tenths of a decibel of loss per kilometer.

Less Weight and Volume. Fiber optic cables are substantially lighter in weight and occupy much less volume than copper cables with the same information capacity. For example, a 3-in. diameter telephone cable consisting of 900 twisted-pair wires can be replaced with a single fiber strand 0.005 inch in diameter (approximately the diameter of a hair strand) and retain the same information-carrying capacity. Even with a rugged protective jacket surrounding the fiber, it occupies enormously less space and weights considerably less.

Security. Since light does not radiate from a fiber optic cable, it is nearly impossible to secretly tap into it without detection. For this reason, several applications requiring communications security employ fiber-optic systems. Military information, for example, can be transmitted over fiber to prevent eavesdropping. In addition, metal detectors cannot detect fiber-optic cables unless they are manufactured with steel reinforcement for strength.

Flexibility. The surface of glass fiber is much more refined than ordinary glass. This, coupled with its small diameter, allows it to be flexible enough to wrap around a pencil. In terms of strength, a 0.005-in. strand of fiber is strong enough to cut one's finger before it breaks, if enough pressure is applied against it.

Economics. Presently, the cost of fiber is comparable to copper at approximately $0.20 to $0.50 per yard and is expected to drop as it becomes more widely used. Since transmission losses are considerably less than for coaxial cable, expensive repeaters can be spaced farther apart. Fewer repeaters mean a reduction in overall system cost and enhanced reliability.

Reliability. Once installed, a longer life span is expected with fiber over its metallic counterparts since it is more resistant to corrosion caused by environmental extremes such as temperature, corrosive gases, and liquids.

b) Disadvantages

In spite of the numerous advantages that fiber optic systems have over conventional methods of transmission, there are some disadvantages, particularly because of its newness. Many of these disadvantages are being overcome with new and competitive technology.

Interfacing costs. Electronic facilities must be converted to optics in order to interface to fiber. Often these costs are initially overlooked. Fiber-optic transmitter, receiver, couplers, and connectors, for example, must be employed as part of the communication system. Test and repair equipment is costly. If the fiber optic cable breaks, splicing can be a costly and tedious task

Strength. Fiber, by itself, has a tensile strength of approximately 1 lb, as compared the coaxial cable at 180 1b (RG59U) surrounding the fiber with stranded Kevlar and a protective PCV jacket can increase the pulling strength up to 500 1b. Installations requiring greater tensile strengths can be achieved with steel reinforcement.

Remote Powering of Devices. Occasionally it is necessary to provide electrical power to a remote device. Since this cannot be achieved through the fiber, metallic conductors are often included in the cable assembly. Several manufacturers now offer a complete line of cable types, including cables manufactured with both copper wire and fiber.

3.6.2 Theory of Light

In the seventeenth and eighteenth centuries, there were two schools of thought regarding the nature of light. Sir Isaac Newton and his followers believed that light consisted of rapidly moving particles (or corpuscles), whereas Dutch physicist Christian Huygens regarded light as being a series of waves.

The wave theory was strongly supported by an English doctor named Thomas Young. By 1905, quantum theory, introduced by Clark Maxwell, showed that when light is emitted or absorbed it is not only as a wave, but also as an electromagnetic particle called a photon. Photon is said to possess energy that is proportional to its frequency. This is known as Planck's law, which states:

$$E = h \times v$$

where E = photon's energy (J);

h = Planck's constants, 6.63×10^{-34} (J-s);

v = frequency of the photon (Hz).

Using the particle theory, Einstein and Planck were able to explain photoelectric effect: when visible light or electromagnetic radiation of a hire frequency shines on a metallic surface, electrons are emitted, which is turning an electric current.

Electromagnetic Spectrum

Fundamentally, light has been accepted as a form of electromagnetic radiation that can be categorized into a portion of the entire electromagnetic spectrum, as shown in Table 3.2. In addition, each frequency can be specified in terms of its equivalent wavelength. Frequency or wavelength are directly related to the speed of light.

$$C = f \times \lambda$$

Where c - speed of light in a vacuum or free space, 3×10^{8} (m/s);

f - frequency (Hz); λ-wavelength (m).

Example 3.2 Compute the wavelength for a frequency of 3.20 MHz.

Solution: $c = f \times \lambda$

Therefore, $F = c/\lambda = (3 \times 10^{8}) / (3.2 \times 10^{6}) = 41.7$ m

Table 3.2 Electromagnetic Spectrum

Range of wavelength, nm	Name of wavelength	
$10^{6} - 770$	*Infrared*	
770 - 662	Red	*Visible*
662 - 597	Orange	
597 - 577	Yellow	
577 - 492	Green	
492 - 455	Blue	
455 - 390	Violet	
390 - 10	*Ultraviolet*	

The portion of the electromagnetic spectrum regarded as light has been expanded in Table 3.1 to illustrate three basic categories of light:

1. Infrared: that portion of the electromagnetic spectrum having a wavelength ranging from 770 to 10^{6} nm. Fiber optic systems operate in this range.

2. Visible: that portion of the electromagnetic spectrum having a wavelength ranging from 390 to 770 nm. The human eye, responding to these wavelengths allows us to see the colours ranging from violet to red, respectively.

3. Ultraviolet: that portion of the electromagnetic spectrum ranging from 10 to 390 nm.

The light that we use for most fiber optic systems occupies a wavelength range from 800 to 1600 nm. This is slightly larger than visible red light and falls within the infrared portion of the spectrum.

Snell's Law: Total Interval Reflection

For light to propagate in any medium, the medium must be transparent to some degree. The degree of transparency determines how far light will propagate. Transparent materials can be in the form of a liquid, gas, or a solid. Some examples are glass, plastic, air, and water.

One of the most fundamental principles of light is that when it strikes the interface between two transparent mediums, such as air and water, a portion of the light energy is reflected back into the first medium and a portion is transmitted into the second medium. The path in which light travels from one point to another is commonly referred to as the ray. Figure 3.14 illustrates the classic example of a ray of light incident upon the surface of water. Notice that part of the light is *reflected* off the surface of water and part of it penetrates the water. The ray penetrating to water is said to be *refracted* or bent toward the normal. The amount of refracted light is determined by the medium's index of refraction, generally denoted by the letter n. Index of refraction is the ratio of the speed of light in a vacuum - c, to the speed of light in the given medium - v. This relationship is given by the equation: $n = c / v$. Since the speed of light is lower in mediums other than a vacuum, the index of refraction in such mediums is always greater than 1.

Example for air n = 1.003, for water n = 1.33, for fiber-optic n = 1.6.

In 1621, the Dutch mathematician Willebrard Snell established that rays of light could be traced as they propagate from one medium to another based on their indices of refraction. Snell's low is stated by the equation:

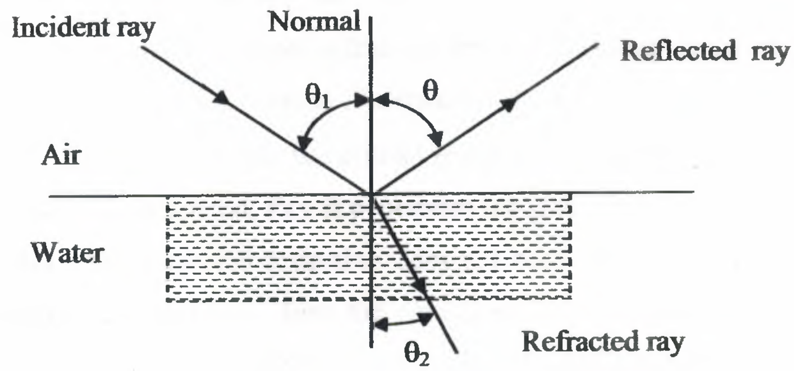


Figure 3.14 Ray Of Light Incident Upon The Surface Of Water.

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1}; \qquad n_1 \sin\theta_1 = n_2 \sin\theta_2$$

where n_1 - refractive index of material 1; θ_1 - angle of incidence; θ_2 - angle of refraction; n_2 - refractive index of material 2. When the angle of incidence, θ_1, becomes large enough to cause the sine of the refraction angle, θ_2, to exceed the value of 1, total internal reflection occurs. This angle is called the critical angle, θ_c. The critical angle, θ_c, can be derived from Snell's law as follows $\quad n_1 \sin\theta_1 = n_2 \sin\theta_2$

$$\sin\theta_1 = n_2 \sin\theta_2/n_1$$

When $\sin\theta_1 = \sin\theta_2$, then $\sin\theta_1 = n2 / n1$. Therefore, critical angle: $\theta_c = \sin^{-1}(n2 / n1)$

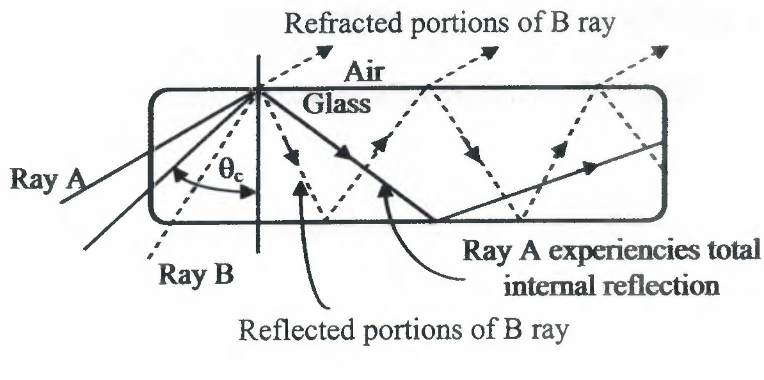


Figure 3.15 Ray A penetrates the glass-air interface at an angle exceeding the critical angle, θc.

For glass, n = 1.5; for air n = 1.0 and $\theta_c = \sin^{-1}(n2 / n1) = \sin^{-1}(1.0 / 1.5) = 41.8^0$.

By surrounding glass with material whose refraction index is less than that of the glass, total internal reflection can be achieved. This is illustrated in Figure 3.15. Ray A penetrates the glass-air interface at an angle exceeding the critical angle, θc, and therefore experiences total internal reflection. On the other hand, Ray B penetrates the glass air interface at an angle less than the critical angle. Total internal reflection does not occur. Instead, a portion of ray B escapes the glass and is refracted away from the normal as it enters the less dense medium of air. A portion is also reflected back into the glass. Ray B diminished in magnitude as it bounces back and forth between the glass-air interface. The foregoing principle is the basis for guiding light through optical fibers.

Two key elements that permit light guiding through optical fibers are its core and its cladding. The fiber's core is manufactured of ultra pure glass (silicon dioxide) or plastic. Surrounding the core is a material called cladding. A fiber cladding is also made of glass or plastic.

Its index of refraction, however, it is typically 1% less than that of its core. This permits total internal reflection of rays entering the fiber and striking the core-cladding interface above the critical angle of approximately 82-degree ($\sin^{-1}(1/1.01)$). The core of the fiber therefore guides the light and the cladding contains the light. The cladding material is much less transparent than the glass making up the core of the fiber. This causes light rays to be absorbed if they strike the core-cladding interface at an angle less than the critical angle.

In Total internal reflection occurs as it strikes the lower index cladding material.

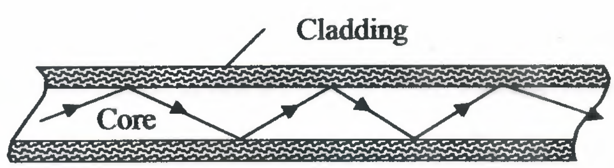


Figure 3.16 A Light Ray Is Transmitted Into The Core Of An Optical Fiber.

3.6.3 Block Diagram of the FOS

One of the main limitations of communication systems is their restricted information carrying capabilities. In more specific terms what this means is that the communications medium can only carry so many messages. And, as you have seen, this information-handling ability is directly proportional to the bandwidth of the communications channel. In telephone systems, the bandwidth is limited by the characteristics of the cable used to carry the signals. As the demand for telephones has increased, better cables and wiring systems have been developed. Further, multiplexing techniques have been used to transmit multiple telephone conversations over a single cable.

In radio communication systems, the information modulates a high frequency carrier. The modulation produces sidebands, and therefore, the signal occupies a narrow portion of the RF spectrum. However, the RF spectrum is finite. There is only so much space for radio signals. To increase the information capacity of a channel, the bandwidth of the channel must be increased. This reduces available spectrum space. Multiplexing techniques are used to send more signals in a given channel bandwidth, and methods have been developed to transmit more information in less bandwidth.

The information-carrying capacity of the radio signal can be increased tremendously if higher carrier frequencies are used. As the demand for increased communications capacity has gone up over the years, higher and higher RFs are being used. Today, microwaves are the preferred radio channels for this reason, but it is more complex and expensive to use these higher frequencies because of the special equipment required.

One way to expand communications capability further is to use light as the transmission medium. Instead of using an electrical signal traveling over a cable or electromagnetic waves traveling through space, the information is put on a light beam and transmitted through space or through a special cable. In the late nineteenth century, Alexander Graham Bell, the inventor of the telephone, demonstrated that information could be transmitted by light.

Light beam communication was made more practical with the invention of the laser. The laser is a special high-intensity, single frequency light source. It produces a very narrow beam of brilliant light of a specific wavelength (color). Because of its great intensity, the laser beam can penetrate atmospheric obstacles better than other types of light, thereby making light-beam communication more reliable over longer distances. The primary problem with such free-space light beam communication is that the transmitter and receiver must be perfectly aligned with one another.

Instead of using free space, some type of light carrying cable can also be used. For centuries it has been known that light is easily transmitted through various types of transparent media such as glass and water, but it wasn't until the early in 1900s that scientist were able to develop practical light carrying media. By the mid-1950s glass fibers were developed that permitted long light carrying cables to be constructed. Over the years, these glass fibers have been perfected. Further, low cost plastic fiber cable also developed. Developments in these cables permitted them to be made longer with less attenuation of the light.

Today the fiber optic cables have been highly refined. Cables many miles long can be constructed and interconnected for the purpose of transmitting information on a light beam over very long distances. Its great advantage is that light beams have an incredible information carrying capacity. Whereas hundreds of telephone conversations may be transmitted simultaneously at microwave frequencies, many thousands of signals can be carried on a light beam through a fiber optic cable. Using multiplexing techniques similar to those used in telephone and radio systems, fiber optic communications systems have an almost limitless capacity for information transfer.

The components of a typical fiber optic communications system are illustrated in Figure 3.17.

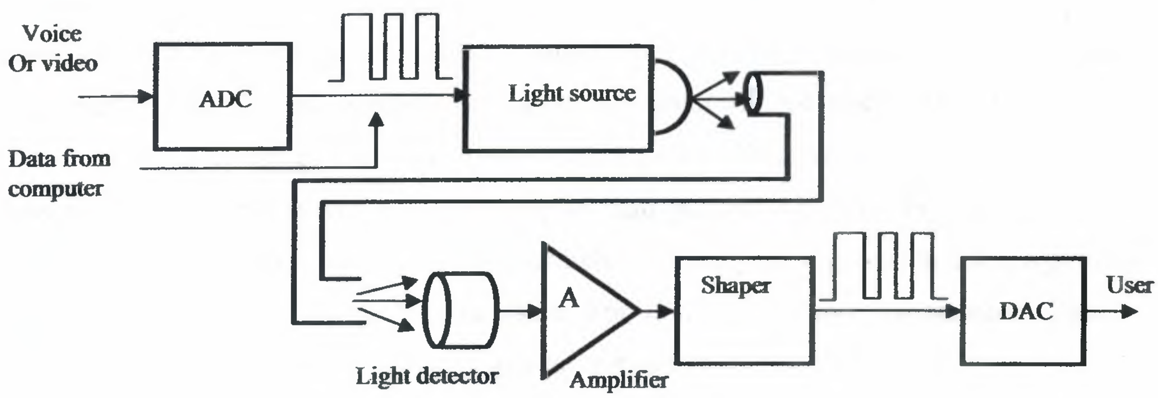


Figure 3.17 Typical Fiber Optic Communications System

The information signal to be transmitted may be voice, video, or computer data. The first step is to convert the information into a form compatible with the communications medium. This is usually done by converting continuous analog signals such as voice and video (TV) signals into a series of digital pulses. An Analog-to-Digital Converter (ADC) is used for this purpose. Computer data is already in digital form. These digital pulses are then used to flash a powerful light source off and on very rapidly. In simple low cost systems that transmit over short distances, the light source is usually a light-emitting diode (LED). This is a semiconductor device that puts out a low intensity red light beam. Other colors are also used. Infrared beams like those used in TV remote controls are also used in transmission. Another commonly used light source is the laser emitting diode. This is also a semiconductor device that generates an extremely intense single frequency light beam.

The light beam pulses are then fed into a fiber optic cable where they are transmitted over long distances. At the receiving end, a light sensitive device known as a photocell or light detector is used to detect the light pulses. This photocell or photo detector converts the light pulses into an electrical signal. The electrical pulses are amplified and reshaped back into digital form. They are fed to a decoder, such as a Digital-to-Analog Converter (DAC), where the original voice or video is recovered for user.

3.7 Fiber Optic Cables

Just as standard electric cables come in a variety of sizes, shapes, and types, fiber optic cables are available in different configurations. The simplest cable is just a single strand of fiber, whereas complex cables are made up of multiple fibers with different layers and other elements.

The portion of a fiber optic cable (core) that carries the light is made from either glass or plastic. Another name for glass is silica. Special techniques have been developed to create nearly perfect optical glass or plastic, which is transparent to light. Such materials can carry light over a long distance. Glass has superior optical characteristics over plastic. However, glass is far more expensive and more fragile than plastic. Although the plastic is less expensive and more flexible, its attenuation of light is greater. For a given intensity, light will travel a greater distance in glass than in plastic. For very long distance transmission, glass is certainly preferred. For shorter distances, plastic is much more practical.

All fibers consist of a number of substructures including (see Figure 3.18):

- A core, which carries most of the light, surrounded by
- A cladding, which bends the light and confines it to the core, surrounded by
- A substrate layer (in some fibers) of glass which does not carry light, but adds to the diameter and strength of the fiber, covered by
- A primary buffer coating, which provides the first layer of mechanical protection, covered by
- A secondary buffer coating, which protects the relatively fragile primary coating.

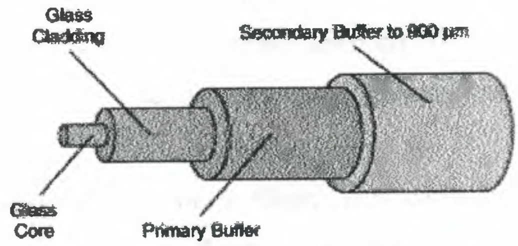


Figure 3.18 Fiber Optic Cable

The cladding is also made of glass or plastic but has a lower index of refraction. This ensures that the proper interface is achieved so that the light waves remain within the core. In addition to protecting the fiber core from nicks and scratches, the cladding adds strength. Some fiber optic cables have a glass core with a glass cladding. Others have a plastic core with a plastic cladding. Another common arrangement is a glass core with a plastic cladding. It is called plastic-clad silica (PCS) cable.

3.7.1 Basic Construction of the Fiber-Optic Cables

There are two basic ways of classifying fiber optic cables. The first way is an indication of how the index of refraction varies across the cross section of the cable. The second way of classification is by mode. Mode refers to the various paths that the light rays can take in passing through the fiber. Usually these two methods of classification are combined to define the types of cable. There are two basic ways of defining the index of refraction variation across a cable. These are step index and graded index. Step index refers to the fact that there is a sharply defined step in the index of refraction where the fiber core and the cladding interface. It means that the core has one constant index of refraction N1, while the cladding has another constant index of refraction N2.

The other type of cable has a graded index. In this type of cable, the index of refraction of the core is not constant. Instead, the index of refraction varies smoothly and continuously over the diameter of the core. As you get closer to the center of the core, the index of refraction gradually increases, reaching a peak at the center and then declining as the other outer edge of the core is reached. The index of refraction of the cladding is constant.

Mode refers to the number of paths for the light rays in the cable. There are two classifications: single mode and multimode. In single mode, light follows a single path through the core. In multimode, the light takes many paths through the core.

Each type of fiber optic cable is classified by one of these methods of rating the index or mode. In practice, there are three commonly used types of fiber optic cable: multimode step index, single mode step index and multimode graded index cables.

1. **The multimode step-index fiber.** This cable (see Figure 3.19(a)) is the most common and widely used type. It is also the easiest to make and, therefore, the least expensive. It is widely used for short to medium distances at relatively low pulse frequencies.

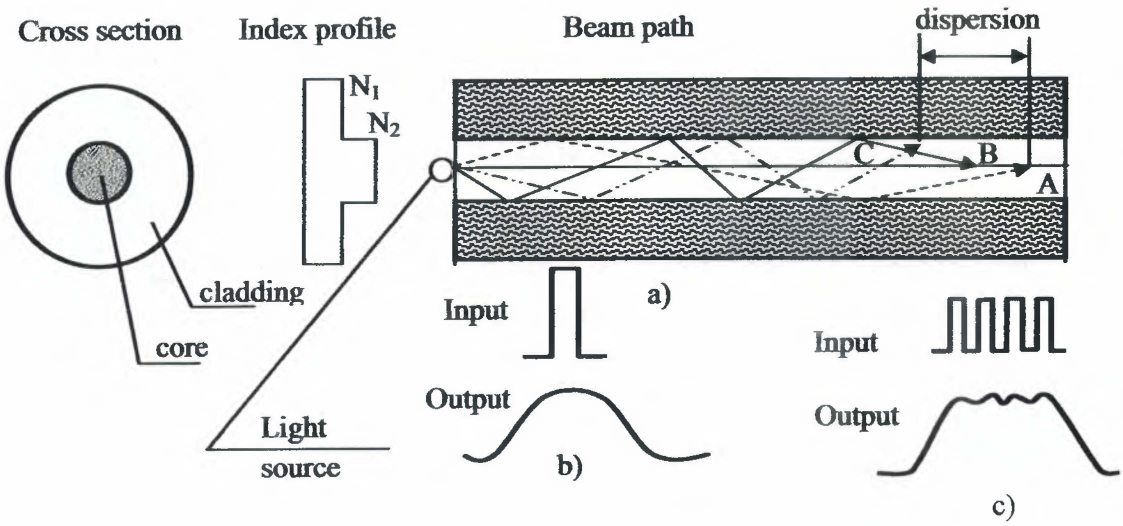


Figure 3.19 The Multimode Step-Index Fiber

The main advantage of a multimode step index fiber is the large size. Typical core diameters are in the 50-to-1000 micrometers (μm) range. Such large diameter cores are excellent at gathering light and transmitting it efficiently. This means that an inexpensive light source such as LED can be used to produce the light pulses. The light takes many hundreds of even thousands of paths through the core before exiting. Because of the different lengths of these paths, some of the light rays take longer to reach the end of the cable than others. The problem with this is that it stretches the light pulses (Figure 3.19 (b). In Figure 3.19 ray A reaches the end first, then B, and C. The result is a pulse at the other end of the cable that is lower in amplitude due to the attenuation of the light in the cable and increased in duration due to the different arrival times of the various light rays. The stretching of the pulse is referred to as modal dispersion. Because the pulse has been stretched, input pulses can not occur at a rate faster than the output pulse duration permits. Otherwise the pulses will essentially merge together as shown in Figure 3.19 (c). At the output, one long pulse will occur and will be indistinguishable from the three separate pulses originally transmitted. This means that incorrect information will be received. The only core for this problem is to reduce the pulse repetition rate. When this is done, proper operation occurs. But with pulses at a lower frequency, less information can be handled.

2. Single Mode Cable

In a single mode, or mono-mode, step-index fiber cable the core is so small that the total number modes or paths through the core are minimized and modal dispersion is essentially eliminated. The typical core sizes are 5 to 15 μm. The output pulse has essentially the same duration as the input pulse (see Figure 3.20).

The single mode step index fibers are by far the best since the pulse repetition rate can be high and the maximum amount of information can be carried. For very long distance transmission and maximum information content, single-mode step-index fiber cables should be used.

The main problem with this type of cable is that because of its extremely small size, it is difficult to make and is, therefore, very expensive. Handling, splicing, and making interconnections are also more difficult. Finally, for proper operation an expensive, super intense light source such as a laser must be used. For long distances, however, this is the type of cable preferred.

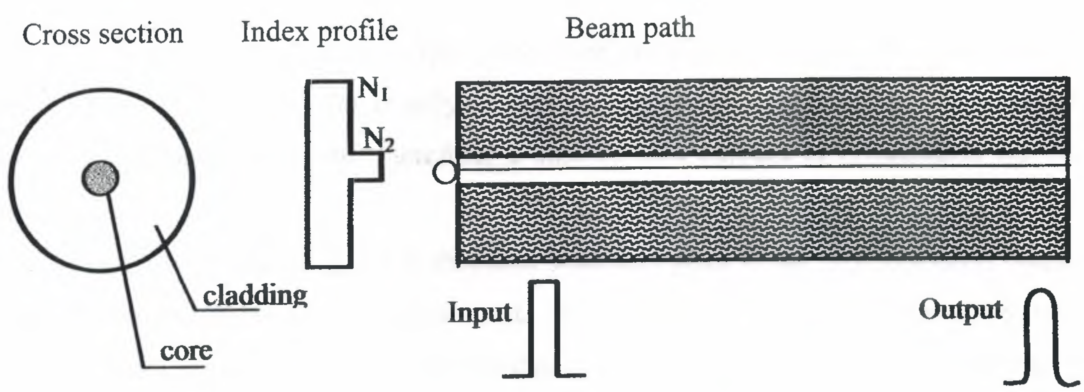


Figure 3.20 Single Mode Cable

3. Multimode Graded-Index Fiber Cables.

These cables have a several modes or paths of transmission through the cable, but they are much more orderly and predictable. Figure 3.8 shows the typical paths of the light beams. Because of the continuously varying index of refraction across the core, the light rays are bent smoothly and converge repeatedly at points along the cable.

The light rays near the edge of the core take a longer path but travel faster since the index of refraction is lower. All the modes or light paths tend to arrive at one point simultaneously. The result is that there is less modal dispersion.

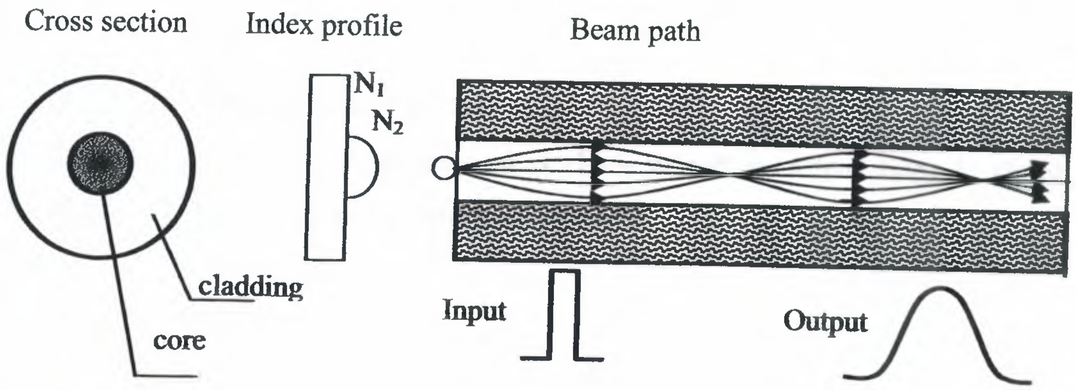


Figure 3.21 Multimode Graded-Index Fiber Cables

It is not eliminated entirely, but the output pulse is not nearly as stretched as in multimode step index cable. The output pulse is only slightly elongated. As a result, this cable can be used at very high pulse rates and, therefore, a considerable amount of information can be carried on it.

This type of cable is also much wider in diameter with core sizes in the 50 to100 (μm) range. Therefore, it is easier to splice and interconnect, and cheaper, less-intense light sources may be used. The most popular fiber-optic cables that are used in LAN: Multimode-step index cable -65.5/125; multimode-graded index cable - 50/125. The multimode-graded index cable - 100/140 or 200/300 are recommended for industrial control applications because its large size. In high data rate systems is used single mode fiber 9/125. Typical core and cladding diameters of these cables are shown in Figure 3.22.

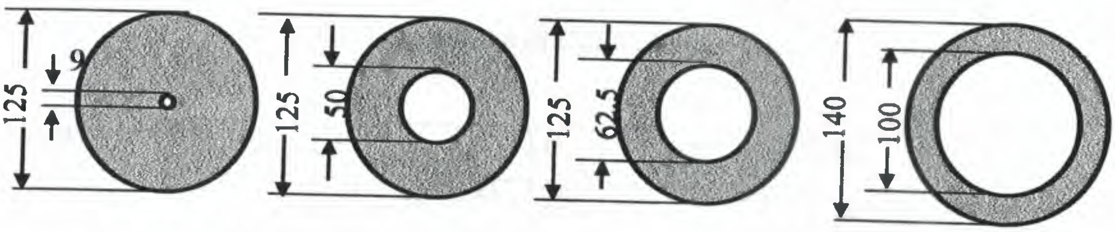


Figure 3.22 Typical core and cladding diameters of these cables

3.7.2 Specifications of the cables

The fiber as a transmission medium is characterized by Attenuation, A, db/km; Numeric aperture, NA and Dispersion, ns/km.

a) Attenuation

The main specification of a fiber optic cable is its attenuation.

Light power which does not reach the other end of the fiber has either left the fiber or been absorbed (converted to heat) in it. The amount of attenuation varies with the type of cable and its size. Glass has less attenuation than plastic. Wider cores have less attenuation than narrower cores. But more importantly, the attenuation is directly proportional to the length of the cable. It is obvious that the longer the distance the light has to travel the greater the loss due to absorption, scattering, and dispersion. Doubling the length of a cable doubles the attenuation, and so on.

The attenuation of a fiber optic cable is expressed in decibels per unit of length. The standard specification for fiber-optic cable is the attenuation expressed in terms of decibels per kilometers. The standard decibel formula used is

$$\text{Loss, dB} = 10 \log (P_0/P_I)$$

Where P_0 is the output power and P_I is the input power.

The table 3.2 shows the percentage of output power for various decibel loss. The attenuation ratings of fiber-optic cables vary over a considerable range.

The table 3.2 shows the percentage of output power for various decibel loss.

Table 3.3 The Percentage of Output Power Expressed by dB

Loss,(dB)	1	2	3	4	5	6	7	8	9	10	20	30
P_0 (%)	79	63	50	40	31	25	20	14	12	10	1	0.1

The finest single mode step-index cables have an attenuation of only 1 dB/km. However, a very large core plastic fiber cables can have an attenuation of several thousands decibels per kilometer.

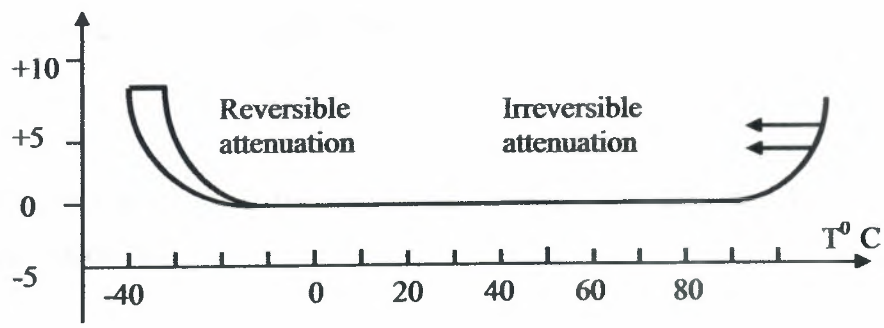


Figure 3.23 Temperature Dependence of the Attenuation of Fiber Optic Cable

The following contribute to the attenuation:

Reyleigh-scattering. A mechanism called Rayleigh scattering prevents any further improvement in attenuation loss. Rayleigh scattering is caused by micro irregularities in the random molecular structure of glass. These irregularities are formed as the fiber cools from a molten state. Normally, electrons in glass molecules interact with transmitted light by absorbing and reradiating light at the same wavelength. A portion of the light, however, strikes these micro irregularities and becomes scattered in all directions of the fiber, some of which is lost in the cladding. Consequently, the intensity of the beam is diminished.

Radiation Losses.

A phenomenon called micro bending can cause radiation losses in optical fibers in excess of its intrinsic losses. Micro bends are miniature bends and geometric imperfections along the axis of the fiber that occur during the manufacturing or installation of the fiber. Mechanical stress such as pressure, tension, and twist can cause micro bending. This geometric imperfection causes light to get coupled to various unguided electromagnetic modes that radiate and escape the fiber.

Absorption. The following contribute to the absorption: Intrinsic impurities, irregularities in core diameter, IR-absorption (infrared), OH- absorption (hydroxy, humidity) and molecular agitation.

b) Numerical Aperture

Numerical Aperture tells how much of the light can be pass into the fiber. An important characteristic of a fiber is its *numerical aperture* (NA). NA characterizes a fiber's light-gathering capability. Mathematically, it is defined as the sine of half the angle of a fiber's light *acceptance cone.* For multimode step index fiber

$$NA = \sqrt{N^2{}_1 - N^2{}_2}$$

Typical values for NA are 0.25 to 0.4 for multimode step-index fiber and 0.2 to 0.3 for multimode graded-index fiber.

c) Dispersion:

Dispersions classified: material dispersion, Ψ_{mat} (ns/km) and modal dispersion, Ψ_{mod} (ns/km).

Material dispersion.

A light pulse is composed of light of different wavelengths depending on the spectral width of the light source. The refractive index depends weakly on the wavelength. This causes the material dispersion.

Modal dispersion. As shown above the modal dispersion due to the different arrival times of the various light rays.

The Table 3.4 shows the characteristics of the dispersions.

Table 3.4 Characteristics of the Dispersions

Fiber type	Dispersion (ns / km)	
	Modal	Material
Step index	$\psi mod = t * (\Delta / 2)$	$\psi mat = 0.1\ \Delta\lambda$
Graded index	$\psi mod = t * (\Delta^2 / 2)$	
Single mode	$\psi mod = 0$	

Note: t- traveling time per km t=N/C, for N=1.5, t= 5μs/km;

$\Delta\lambda = \Delta N/N$, in practice Δ=0.01

$\Delta\lambda$ - Bandwidth of the light;

The total dispersion equals $\psi_{tot} = (\psi_{mat} + \psi_{mod})^{1/2}$

3.8 Optical Transmitters

In an optical communications system the transmitter consist of a modulator and the circuitry that generates the carrier. In this case, the carrier is a light beam that is modulated by digital pulses that turn it on and off. The basic transmitter is nothing more than a light source.

Several devices are emitters of light, both natural and artificial. Few of these devices, however, are suitable for fiber-optic transmitters. What we are interested in a light source that meets the following requirements:

- The light source must be able to turn on and off several tens of millions and even billions of time per second.
- The light source must be able to emit a wavelength that is transparent to the fiber.
- The light source must be efficient in terms of coupling light energy into the fiber.
- The power emitted must be sufficient enough to transmit through the optical fibers.
- Temperature variations should not affect the performance of the light source.
- The cost of manufacturing the light source must be relatively inexpensive.

Two commonly used devices that satisfy the above requirements are: monochromatic incoherent source-the light emitting diode (LED) and monochromatic coherent source-the injection laser diode (ILD).

3.8.1 Light Emitting Diode

The major difference between the LED and the ILD is the manner in which light is emitted from each source. The LED is an incoherent light source that emits light in a disorderly way. A laser is a light source that emits coherent monochromatic light. Monochromatic light has a pure single frequency. Coherent refers to the fact that all the light waves emitted are in phase with one another. Coherent light waves are focused into a narrow beam, which, as a result, is extremely intense. The effect is somewhat similar to that of using highly directional antenna to focus radio waves into a narrow beam, which also increases the intensity of the signal. Figure 3.24 illustrates the differences in radiation patterns. Both devices are extremely rugged, reliable, and small in size. In terms of spectral purity, the LED's half power spectral width is approximately 50 nm, whereas the ILD's spectral width is only a few nanometers. This is shown in Figure 3.24.

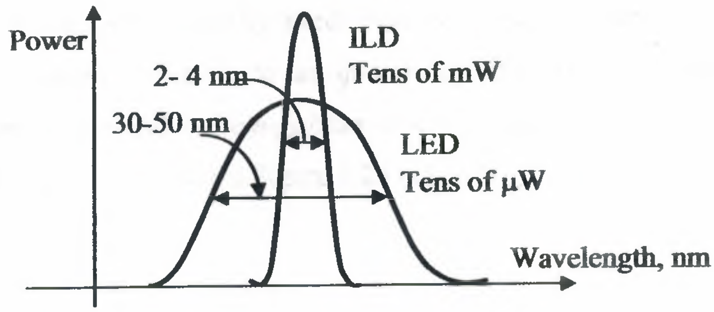


Figure 3.24 Differences In Radiation Patterns

Ideally, a single spectral line is desirable. As the spectral width of the emitter increases, attenuation and pulse dispersion increase. The spectral purity for the ILD and its ability to couple much more power into a fiber make it better suited for long-distances telecommunications links. In addition, injection laser can be turned on and off at much higher rates than an LED. The drawback, however, is its cost, which may approach several hundreds of dollars as compared to a few dollars for LED's in large quantities.

Table 3.4 lists the differences in operating characteristics between the LED and the ILD.

Table 3.4 Typical source characteristics for LED and ILD

	Output Power, µW	Peak wavelength, nm	Spectral width, nm	Rise time, ns
LED	250	820	35	12
	700	820	35	6
	1500	820	35	6
Laser	4000	820	4	1
	6000	1300	2	1

Various semiconductor materials are used to achieve this. Pure *gallium arsenide* (GaAs) emits light at a wavelength of about 900 nm. By adding a mixture of 10% aluminum (Al) to 90% GaAs, *gallium-aluminium-arsenide* (GaAlAs) is formed, which emits light at a wavelength of 820 nm. Recall that this is one of the optimum wavelengths for fiber optic transmission. By tailoring the amount of aluminium mixed with GaAs, wavelengths ranging from 800 to 900 nm can be obtained.

To take advantage of the reduced attenuation losses at longer wavelengths, it is necessary to include even more exotic materials. For wavelengths in the range 1000 to 1550 nm, a combination of four elements is typically used: *indium, gallium, arsenic* and *phosphorus.* These devices are commonly referred to as *quaternary* devices. Combining these four elements produces the compound *indium-gallium-arsenide-phosphide* (InGaAsP). Transfer characteristic of LED and ILD are shown Figure 3.25 (a) and (b).

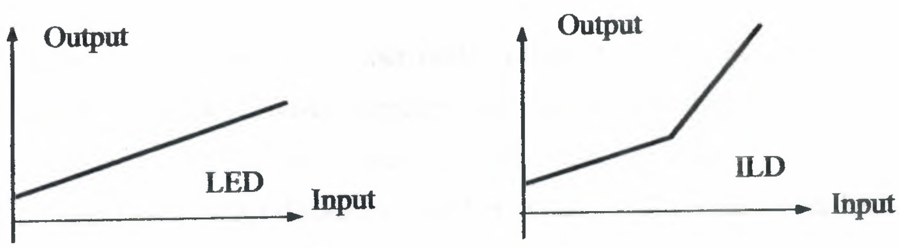


Figure 3.25 Transfer Characteristic of LED and ILD

3.8.2 Injection Laser Diode

The term *laser* is an acronym for *light amplification by stimulated emissions of radiation.* There are many types of lasers on the market. They are constructed of gases, liquids, and solids. Laser diodes are also called injection laser diodes (ILD), because when current is injected across the PN junction, light is emitted relatively large and sophisticated device that outputs a highly intense beam of visible light. Although this is in part true, the laser industry is currently devoting a great deal of effort toward the manufacture of miniature semiconductor laser diodes. Figure 3.24 illustrates the spectrum ILD. ILDs are ideally suited for use within the fiber-optic industry due to their small size, reliability, and ruggedness. Step response of ILD is shown Figure 3.26.

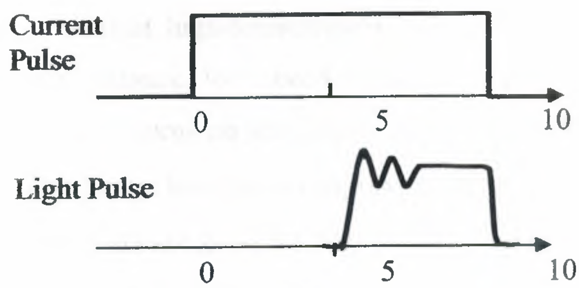


Figure 3.26 Step Response of ILD

The most widely used light source in fiber optic systems is ILD. Like the LED, it is a PN junction diode usually made of GaAs. Injection laser diodes are capable of developing light power up to several watts. They are far more powerful than LEDs and, therefore, are capable of transmitting over much longer distances. Another advantage ILDs have over LEDs is theirs speed. High-speed laser diodes are capable of gigabit per second digital data rates.

3.9 Optical Transmitter Circuits

The light transmitter consists of the LED and its associated driving circuitry. An optical transmitter circuit using the LED is shown in Figure 3.27. The binary pulses are applied to a logic gate, which, in turn, operates a transistor switch T that turns the LED off and on. A positive pulse at the NAND gate input causes the NAND output to go to zero.

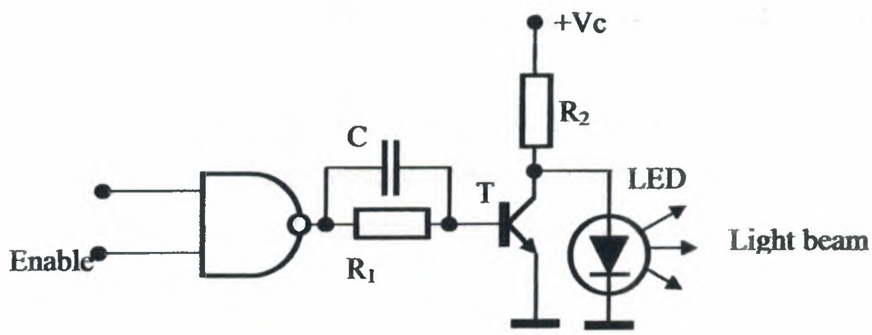


Figure 3.27 An Optical Transmitter Circuit using the LED

This turns off T, so the LED is forward-biased through R_1 and turns on. With zero input, the NAND output is 1, so T turns on and shunts current away from the LED. Very high current pulses are used to ensure a brilliant high-transmission rates are limited. Most LED like transmitters are used for short-distance, low speed digital fiber-optic systems. With zero input, the NAND output is 1, so T turns on and shunts current away from the LED. Most LEDs are capable of generating power levels up to approximately several hundred μW. With such low intensity, LED transmitters are good for only short distances. Further the speed of the LED is limited. Turn-of and turn-on times are no faster than several nanoseconds.

A typical injection laser transmitter circuit is shown in Figure 3.28.

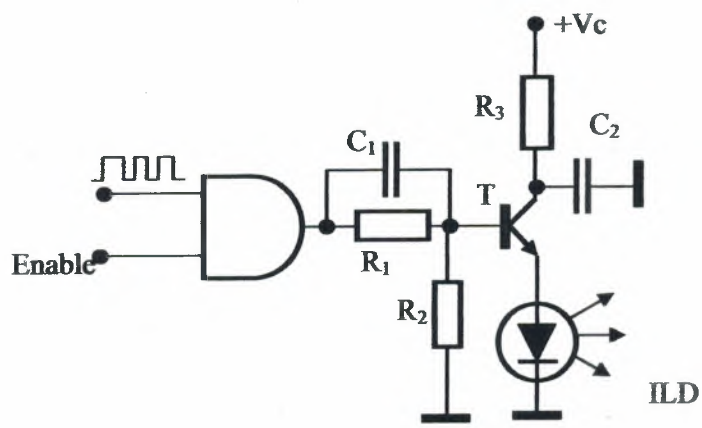


Figure 3.28 A Typical Injection Laser Transmitter Circuit

When the input is zero, the AND gate output is zero, so T is off and so is the laser. Capacitor C_2 charges through R_3 to the high voltage. When a binary 1 input occurs, T conducts connecting C_2 to the ILD. Then C_2 discharges a very high current pulse into the laser, turning it on briefly and creating the light pulse.

4. ERROR DETECTION AND CORRECTION

4.1 Overview

Networks must be able to transfer data from one device to another with complete accuracy. A system that cannot guarantee that the data received by one device are identical to the data transmitted by another device is essentially useless. Yet anytime data are transmitted from source to destination, they can become corrupted in passage. In fact, it is more likely that some part of a message will be altered in transit than that the entire contents will arrive intact. Many factros, including line noise, can alter or wipe out one or more bits of a given data unit. Reliable systems must have a mechanism for detecting and correcting such errors.

Data can be corrupted during transmission. For reliable communication, errors must be detected and corrected.

Error detection and correction are implemented either at data link layer or the transport layer of the OSI model.

4.2 Types of Errors

Whenever an electromagnetic signal flows from one point to another, it is subject to unpredictable interference from heat, magnetism, and other forms of electricity. This interference can change the shape or timing of the signal. If the signal is carrying encoded binary data, such changes can alter the meaning of the data, changing 0 to 1 or 1 to 0. bits can be changed singly or in clumps. For example, a 0.01 second burst of impulse noise on a transmission with a data rate of 1200 bps might change 12 bits of information. Other circumstances can alter just one bit of a data unit, or the first and third bits but not the second. These errors, though seemingly less significant, can make the data just as unreadable as wiping out 12 bits. So it is important to understand all three types of errors and how to detect them. These types are usually referred to as single-bit, multiple-bit, and burst errors. Of the three, a single-bit error is the most likely to occur and a burst error the least likely (see figure 4.1).

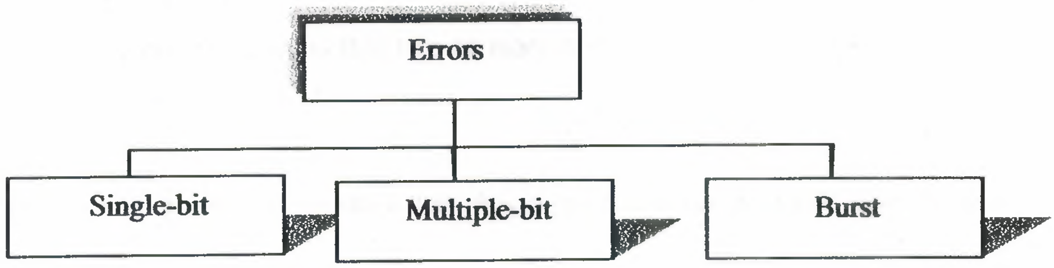


Figure 4.1 Types of Errors

Single-Bit Error

The term single-bit error means that only one bit of a given data unit (such as a byte, character, data unit, or packet) is changed from 1 to 0 or from 0 to 1.

A single-bit error is when only one bit in the data unit has changed.

Figure 4.2 shows the effect of a single-bit error on a data unit. To understand the impact of the change, imagine that each group of eight bits is an ASCII character with a 0 bit appended to the end. In the example, 00000010 was sent, meaning start of text, but 00001010 was received, meaning line feed.

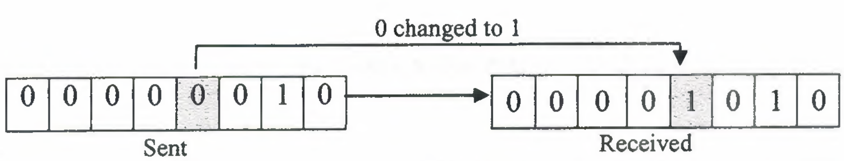


Figure 4.2 Single-Bit Error

Multiple-Bit Error

The term multiple-bit error means that two or more nonconsecutive bits in a data unit have changed from 1 to 0 or from 0 to 1.

غير متعاقب
غير مرتبطة ببعضها

متعاقب

A multiple-bit error is when two or more nonconsecutive bits in the data unit have changed.

Figure 4.3 shows the effect of a multiple-bit error on a byte of data. In this example, if we read the bit pattern as an ASCII character, 01000010 was sent, but 00001010 meaning line feed, was received.

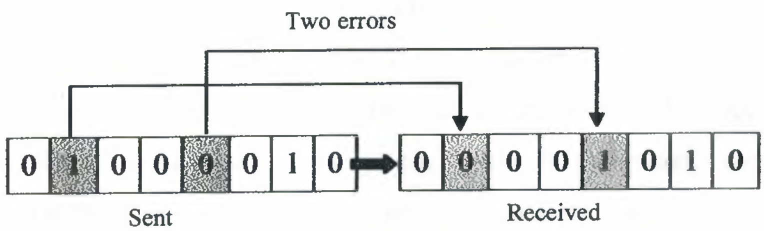


Figure 4.3 Multiple-Bit Error

Burst Error

The term burst error means that two or more consecutive bits in the data unit have changed from 1 to 0 or from 0 to 1.

Burst error means that two or more consecutive bits in the data unit have changed.

Figure 4.4 shows the effect of a burst error on the data unit. In this case, 0100010001000011 was sent, but 0101101101000011 was received.

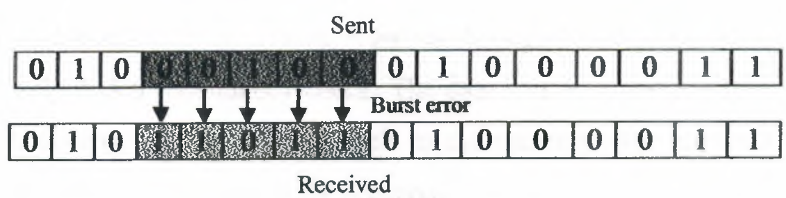


Figure 4.4 Burst Error

4.3 Detection

We will have no way of knowing we have received an error until we have decoded the transmission and failed to make sense of it. For a machine to check for errors this way would be slow, costly, and of questionable value. We don't need a system where computers decode whatever comes in, then sit around trying to decide if the sender really meant to use the word *glbrshnif* in the middle of any arry of weather statistics. What we need is a mechanism that is simple and completely objective.

Redundancy

One mechanism that would satisfy these requirements would be to send every data unit twice. The receiving device would then be able to do a bit for bit comparison between the two versions of the data. Any discrepancy would indicate an error, and an appropriate correction mechanism could be set in place. This system would be completely accurate (the odds of errors being introduced onto exactly the same bits in both sets of data are infinitesimally small), but it would also be insupportably slow. Not only would the transmission time double, but the time it takes to compare every unit bit by bit must be added.

The concept of including extra information in the transmission solely for the purposes of comparison is a good one. But instead of repeating the entire data stream, a shorter group of bits may be appended to the end of each unit. This technique is called redundancy because the extra bits are redundant to the information; they are discarded as soon as the accuracy of the transmission has been determined.

Error detection uses the concept of redundancy, which means adding extra bits for detecting errors at the destination.

Figure 4.5 shows the process of using redundant bits to check the accuracy of a data unit. Once the data stream has been generated, it passes through a device that analyzes it and adds on an appropriately coded redundancy check. The data unit, now enlarged by several bits (in this illustration, seven), travels over the link to the receiver. The receiver puts the entire stream through a checking function. If the received bit stream passes the checking criteria, the data portion of the data unit is accepted and the redundant bits are discarded.

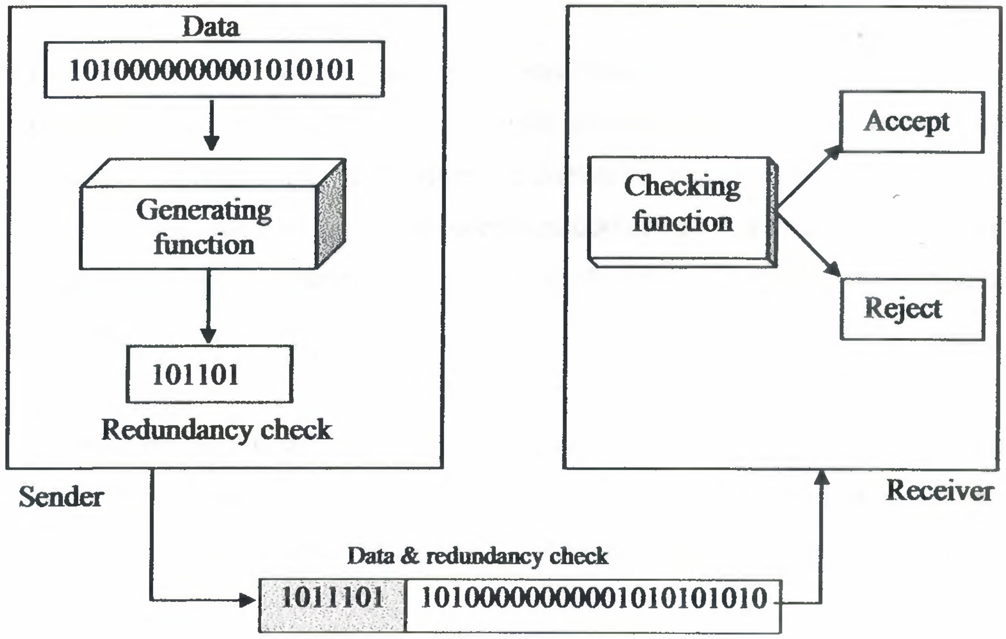


Figure 4.5 Redundancy

Four types of redundancy checks are used in data communications: vertical redundancy check (VRC) (also called parity check), longitudinal redundancy check (LRC), cyclic redundancy check (CRC), and checksum.

4.4 Repetition Method

When you try to talk someone across a noisy room, you may need to repeat yourself to be understood. A brute-force approach to binary communication over a noisy channel likewise employs repetition, so each message bit is represented by a codeword consisting of n identical bits. Any transmission error is a received codeword alters the repetition pattern by changing a 1 to a 0 vice versa.

If transmission errors occur randomly and independently with probability $P = \alpha$, then the binomial frequency distribution gives the probability of i errors in an n-bit codeword as;

$$P(i,n) = \binom{n}{i} \alpha^i (1-\alpha)^{n-1} \approx \binom{n}{i} \alpha^i \qquad \alpha << 1 \tag{4.1}$$

$$\text{where} \qquad \binom{n}{i} = n! = \frac{n(n-1)\cdots(n-i+1)}{i!}$$

Consider, for instance, a triple - repetition code with codeword 000 and 111. All the other received words, such as 001 or 101, clearly indicate the presence of errors. Depending on the decoding scheme, this code can detect or correct erroneous words. For error detection without correction, we say that any word other than 000 or 111 is a detected error. Single and double errors in a word are thereby detected, but triple errors result in an undetected word error with probability

$$P_{we} = P(3,3) = \alpha^3 \tag{4.2}$$

For error correction, we use majority - rule decoding based on the assumption that at least two of the three bits are correct. Thus, 001 and 101 are decoded as 000 and 111, respectively. This rule corrects words with single errors, but double or triple errors result in a decoding with probability.

$$P_{we} = P\,(2,3) + P(3,3) = 3\alpha^2\text{-}2\alpha^3 \tag{4.3}$$

Since Pe = α would be the error probability without coding, we see that either decoding scheme for the triple-repetition code greatly improves reliability if, say, $\alpha \leq 0.01$. However implementation is gained at the cost of reducing the message bit rate by a factor of 1/3.

4.5 Parity Check Codes

Parity is the simplest and oldest method of error detection. Although it is not very effective in data transmission, it is still widely used due to its simplicity. A single bit called the parity bit is added to a group of bits representing a letter, number, or symbol, ASCII characters on a keyboard for example, are typically encoded into seven bits with an eight bit acting as parity. The parity bit is computed by the transmitting device based on the number of "1"'s set in the character. Parity can be either odd or even. If odd parity is selected, the parity bit is set to a 1 or 0 to make the total number of 1-bits in the character, including the parity bit itself, equal to an odd value. The receiving device performs the same computation on the received number of 1-bits for each character and checks the computed parity against what was received. If they do not match, an error has been detected. Table 4.1 lists examples of even and odd parity.

The selection of even or odd parity is generally arbitrary. In most cases it is a matter of custom or preference. The transmitting and receiving stations, how ever, must be set to same mode.

Table 4.1 Even and Odd Parity

Data character	Odd parity bit	Data character	Even parity bit
1101000	0	1011101	1
0010111	1	1110111	0
1010110	1	0011010	1
1010001	0	1010111	1

Parity generating circuits can easily be implemented with a combination of Exclusive -OR gates (see Figure 4.6). Odd parity generating is obtained by adding an inverter (I) at the output.
The received data word and parity bit are applied at the circuit's input.

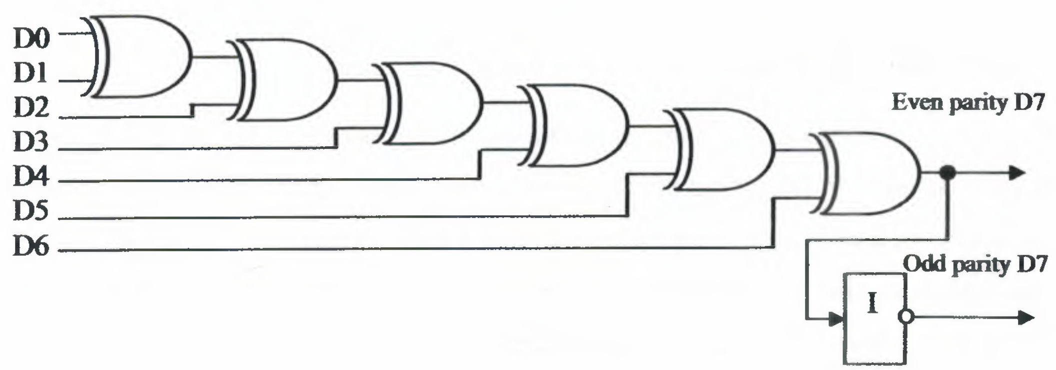


Figure 4.6 Parity Check

The Disadvantage with Parity

A major shortcoming with parity is that it is only applicable for detecting when one bit or an odd number of bits have been changed in a character. Parity checking does not detect when even numbers of bits have changed. For example, suppose that bit D2 in Example 4.1 were to change during the course of a transmission for an odd parity system. Example 1 shows how the bit error is detected.

Example 4.1

	Parity (odd)	D7	D6	D5	D4	D3	D2	D1	D0
Transmitted:	0	0	1	1	0	1	1	0	1
Received:	0	0	1	1	0	1	0	0	1

The received parity bit, a zero, is in conflict with the computed number of 1-bits that was received; in this case, four, an even number of 1-bits. The parity bit should have equal to a value making the total number of 1-bits odd. An error has been properly detected. If, in the other hand, bit D2 and bit D1 were both altered during the transmission, the computed parity bit would still be in agreement with the received parity bit. This error would go undetected, as shown in Example 4.2.

A little thought will reveal that an even number of errors in a character, for odd or even parity, will go undetected.

Example 4.2

	Parity (odd)	D7	D6	D5	D4	D3	D2	D1	D0
Transmitted:	0	0	1	1	0	1	1	0	1
Received:	0	0	1	1	0	1	0	1	1

Errors; two bits or an even number of bits go undetected

Parity, being a single-bit error-detection scheme, presents another problem in accommodating today's high-speed transmission rates. Many errors are the results of impulse noise (produced by lighting and switching transients) and rapid fading in radio-transmission systems, which tend to be bursty in nature. Noise impulses may last several milliseconds, consequently destroying several bits. The higher transmission rate, the greater the effect. Figure 4.7 depicts a 2-ms noise burst imposed on a 4800-bps signal. The bit time associated with this signal is 208 *ns* (1/4800). As many as 10 bits are effected. At least two characters are destroyed here, with the possibility of both character errors going undetected.

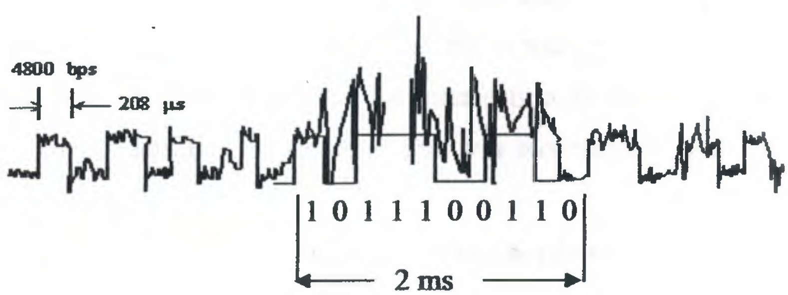


Figure 4.7 a 2-ms Noise Burst Imposed on a 4800-bps Signal

4.6 Vertical and Longitudinal Redundancy Check

Thus far, the discussion of parity has been on a per-character basis. This is often referred to as a Vertical Redundancy Check (VRC) method. Parity can also be computed an inserted at the end of a message block. In this case, the parity bit is computed based on an accumulation of the value of each character's LSB through MSB, including the VRC bit, as shown in Example 4.3. This method is referred to as a Longitudinal Redundancy Check (LRC) method. The resulting word is called the Block Check Character (BBC). Additional parity bits in LRC used to procedure the BCC provide extra error detection capabilities. Single-bit error can now be detected and corrected.

Example 4.3. Using ASC II and odd parity, encode the message "Have a nice day" given in the Table 4.2. Show how the combination LRC and VRC can detect an error if bit D0 in the letter "y" was received 1 instead of 0.

Table 4.2 VRC and LRC for Message

	H	a	v	e	sp	a	sp	n	i	c	e	sp	d	a	y	!	BCC (LRC)
D0	0	1	0	1	0	1	0	0	1	1	1	0	0	1	0*	1	0
D1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0
D2	0	0	1	1	0	0	0	1	0	0	1	0	1	0	0	0	0
D3	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1
D4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
D5	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
D6	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	0	1
VRC	1	0	0	1	0	0	0	0	1	1	1	0	0	0	0	1	1

The VRS and LRS in the column of the character "y" would be 1 instead 0. By itself, detected VRC or LRC does not specify which bit in the column has been received in error. A cross check, however, will reveal that the intersection of the detected parity error for the VRC and LRC check identifies the exact bit that was received in error. By inverting this bit, the error can be corrected.

Unfortunately, either the LRC or VRC does not detect an even number of bit errors.

4.7 Checksum Methods

There are three checksum methods:

1. Single - precision checksum
2. Double - precision checksum
3. Honeywell checksum

4.7.1 Single - Precision Checksum

The most fundamental checksum computation is the single-precision checksum. Here, the checksum is derived simply by performing a binary addition of each n-bit data word in the message block. Any carry or overflow during the addition process is ignored, thus the resultant checksum is derived, transmitted as the BCC, and used to verify the integrity of the received data. For simplicity, a four-byte (B1, B2, B3, B4) data block is used. Note that the um of the data exceeds $2^{n}-1$ and three fore a carry occurs out of the MSB. This carry is ignored and only the eight-bit (n-bit) checksum is sent.

Example 4.4
Drive checksum for the 4-byte data block shown in Figure 4.8

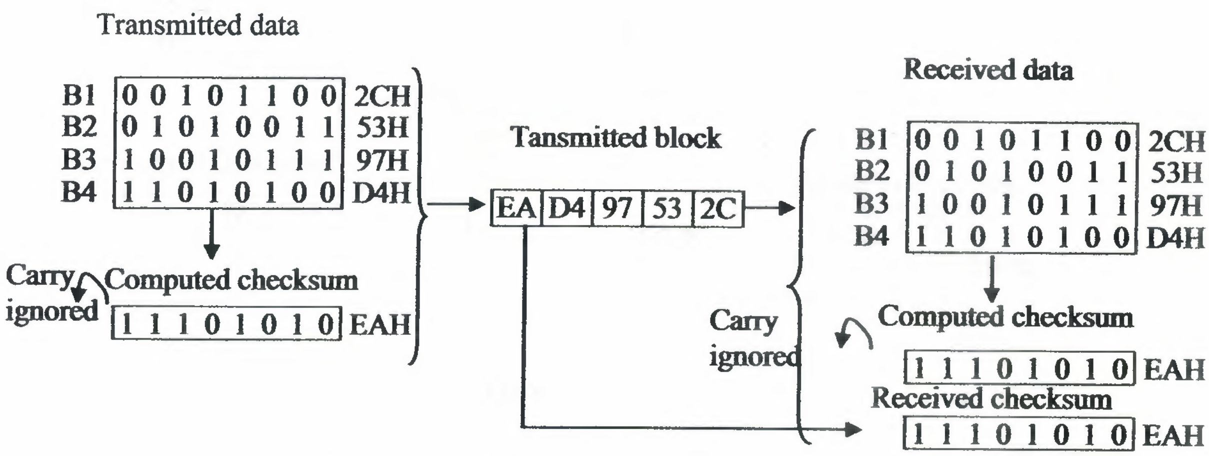


Figure 4.8 Single - Precision Checksum

An inherent problem with the single-precision checksum is if the MSB of the n-bit data word becomes logically stuck at 1 (SA 1), the checksum becomes SA1 as well. A little thought will reveal that the regenerated checksum on the received data will equal the original checksum and the SA1 fault will go undetected.

4.7.2 Double - Precision Checksum

As its name implies, the double-precision checksum extends the computed checksum to 2n bits in length, where n is the size of the data word in the message block. Note that the carry out of the MSB position of the low-order checksum byte is not ignored. For example, the 8-bit data words used in the single-precision checksum example above would have a 16-bit checksum. Message block with 16-bit data words would have a 32-bit checksum, and so forth. Summation of data words in the message block can now extend up to modulo 2^{2n}, thereby decreasing the probability of an erroneous checksum. In addition, the SA1 (stuck at 1) error discussed earlier would be detected as a checksum error at the receiver.

Example 4.5

Drive checksum for the 4-byte data block shown in Figure 4.9.

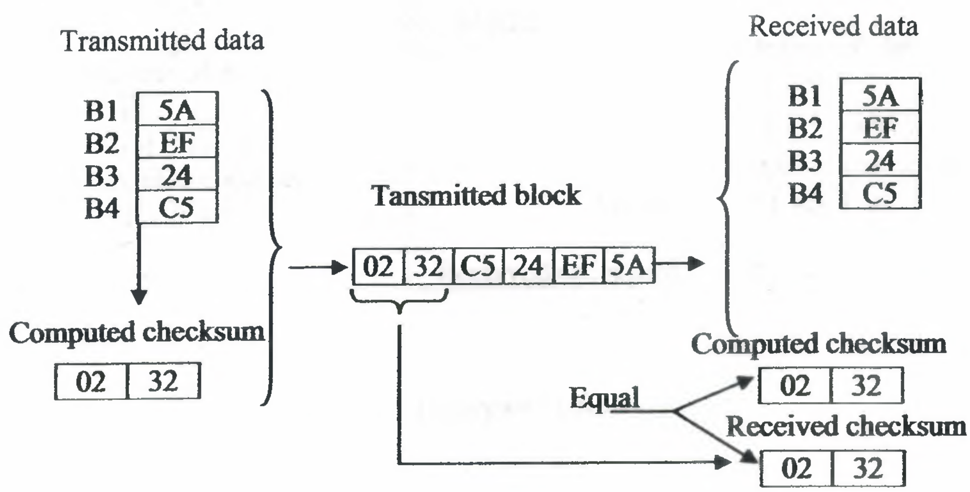


Figure 4.9 Double - Precision Checksum

4.7.3 Honeywell Checksum

The Honeywell checksum is an alternative form of the double-precision checksum. Its length is also 2n bits, where n is again the size of the data word in the message block. The difference is that the Honeywell checksum is based on interleaving consecutive data words to form double – length words. The double – length words are then summed together to form a double precision checksum. The advantage of the Honeywell checksum is that stuck at 1(SA 1) at 0 (SA 0) bit errors occurring in the same bit positions of all words can be detected during error detecting process. This is true since the interleaving process places the error in the upper and lower words of the checksum. At least two bit positions the checksum are affected.

Example 4.6

Drive checksum for the 4-byte data block shown in Figure 4.10

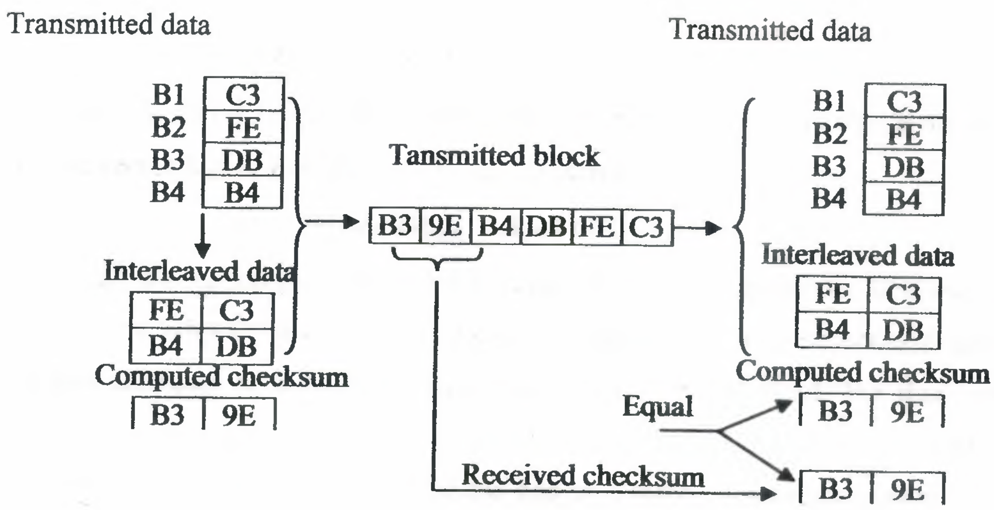


Figure 4.10 Honeywell Checksum

4.8 Block Code. Hamming Distance

The systematic block code with length of n, consist of k message bits m_1, m_2, ... m_k and q = n- k check bits c_1, c_2, ... , c_q. The vector notation of codeword takes the form

$$X = | \; M \;\; C \; |$$

Where M = (m_1, m_2 , ... m_k ; C= (c_1 , c_2 , ... , c_q)

The rate of block code is determined as $R_c = k/n$.

Since there are 2^k different k-bit message blocks and 2^n possible-bit vectors, the fundamental strategy of block coding is to choose the 2^k code vectors such that the minimum distance between codewords is as large as possible.

Let an arbitrary code vector be represented by

$$X = (x_1 \ x_2 \ \dots \ x_n)$$

Where the elements x_1 , x_2 , … , are , of course , binary digits . A code is linear if it includes the all-zero vector and if the sum of any two code vectors produces another vector in the code. The sum of two vectors, say X and Z, is defined as

$$X + Z \cong (x_1 \oplus z_1 \ \ x_2 \oplus z_2 \ \dots \ x_n \oplus z_n)$$

As a consequence of linearity, we can determine a code's minimum distance by the fallowing argument. Let the number of nonzero elements of a vector X be symbolised by the w (x), called the vector weight. The Hamming distance between any two-code vectors X and Z is then

$$D (X, Z) = w (X + Z)$$

Since $x_1 \oplus z1 = 1$ if $x_1 \neq z_1$, etc. the distance between X and Z therefore equals the weight of another code vector X + Z. But if Z = (0 0 … 0) then X +Z = X ; hence ,

$$D_{min} = [\ w\ (X)]_{min} \quad X \neq (0 \ \ 0 \ \dots \ 0)$$

In other words, the minimum distance of a linear block code equals the smallest nonzero vector weight. An arbitrary n-bit codeword can be visualised in an n-dimensional space as a vector whose elements or coordinates equal the bits in the codeword. We thus write the codeword 101 in row notation as X = (1 0 1). Figure 4.11 portrays all possible 3-bit codeword as dots corresponding to the vector tips in a three–dimensional space. The solid dots in a part (a) represent the triple-repetition code, while those in part (b) represent a parity-check code.

Notice that the triple-repetition code vectors have greater separation than the parity-check code vectors. This separation, measured in terms of the Hamming distance, has direct bearing on the error-control power of a code.

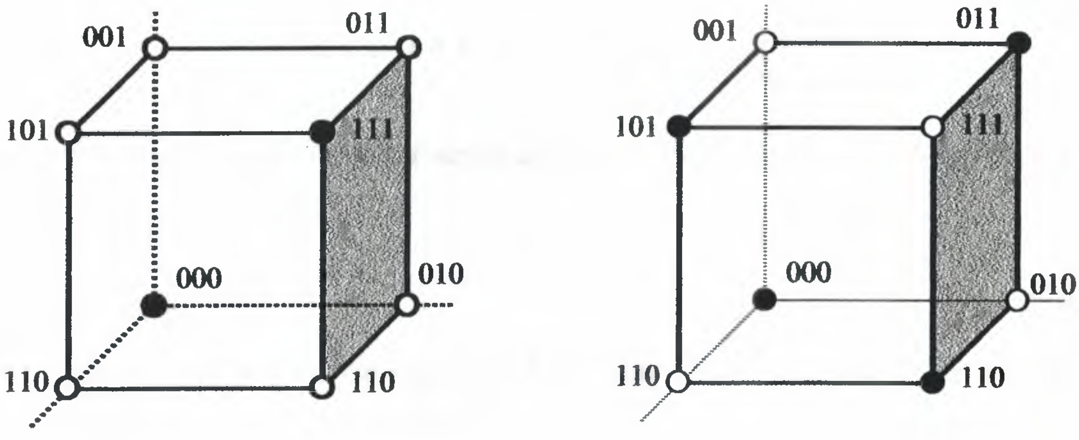


Figure 4.11 Portrays all Possible 3-bit Codeword as Dots Corresponding to the Vector Tips in a three–Dimensional Space.

The Hamming distance d (X, Y) between two vectors X and Y is defined to equal the number of different elements. For instance, if X = (1 0 1) and Y = (1 1 0) then d (X , Y) = 2 because the second and third elements are different .

The minimum distance d_{min} of a particular code is the smallest Hamming distance between valid code vectors. Consequently, error detection is always possible when the number of transmission errors in a codeword is less than d_{min} so the erroneous word is not a valid vector. Conversely, when the number of errors equals or exceeds dmin, the erroneous word may correspond to another valid vector and the errors cannot be detected. The following distance requirements for various degrees of error control capability:

Detect up to l errors per word	$dmin \geq l + 1$
Correct up to t errors per word	$dmin \geq 2t + 1$
Correct up to t errors and detect $l > t$ errors per words	$dmin \geq t + l + 1$

By way of example, we see from Figure 4.11 that the triple-repetition code has dmin = 3. Hence, this code could be used to detect $l \leq 3 - 1 = 2$ errors per word or to correct $t \leq (3 - 1) / 2 = 1$ error per word in agreement with our previous observation. A more powerful code with dmin = 7 could correct triple errors or it could correct double errors and detect quadruple errors. The power of a code obviously depends on the number of bits added to each codeword for error-control purposes. In particular, suppose that the codeword consist of k < n message bits and n – k parity checking bits. The minimum distance of an (n, k) block code is upper-bounded by

$$d_{min} \le n - k + 1$$

Regrettably, the upper bound of dmin is satisfied only by repetition codes, which have $k = 1$ and very inefficient code rate $R_c = 1 / n$.

4.9 FEC Systems

The forward error correction system diagrammed in Figure 7.8. Message bits come from an information source at rate R_b. The encoder takes blocks of *k* message bits and constructs an (n, k) block code with code rate $R_c = k / n < 1$.

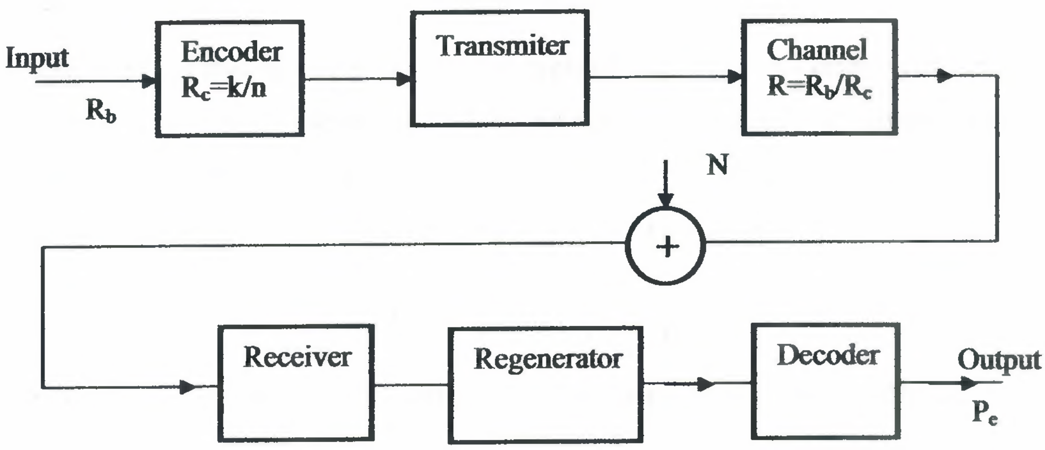


Figure 4.12 FEC System

The bit rate on the channel therefore must be greater than R_b, namely

$$R = (n / k) R_b = R_b / R_c$$

The code has $d_{min} = 2t + 1 \le n - k + 1$, and the decoder operates strictly in an error-correction mode. We'll investigate the performance of this FEC system when additive white noise N causes random errors with probability $\alpha << 1$. The value of α depends on the SNR at the receiver. If E_b represents the average energy per message bit, then the average energy per code bit is R_c. E_b and the ratio of bit energy to noise density is

$$\gamma_c = R_c E_b / N = R_c \gamma_b \qquad (4.4)$$

Where $\gamma_b = E_b / N$. Our performance criterion will be the probability of output message-bit errors, denoted by P_{be} to distinguish it from the word error probability P_{we}.

The code always corrects up to t errors per word and some patterns of more than t errors may also be correctable, depending upon the specific code vectors. Thus, the probability of a decoding word error is upper-bounded by

$$P_{WE} \leq \sum_{i=t+1}^{n} P(i,n) \tag{4.5}$$

For a rough but reasonable performance estimate, we'll take the approximation

$$P_{we} \approx P(t+1,n) \approx \begin{bmatrix} n \\ t+1 \end{bmatrix} \alpha^{t+1} \tag{4.6}$$

Which means that an uncorrected word typically has t + 1 bit errors. On the average, there will be (k / n)(t + 1) message-bit errors per uncorrected word; the remaining errors being in check bits. When Nk bits are transmitted in N >> 1 word, the expected total number of erroneous message bits at the output is (k / n) (t + 1) NP_{we}. Hence,

$$P_{be} = \frac{t+1}{n} P_{we} \approx \begin{bmatrix} n-1 \\ t \end{bmatrix} \alpha^{t+1} \tag{4.7}$$

In which we have used P_{be} to combine *(t +1) / n* with the binomial coefficient.

4.10 Hamming Code

In FEC, a return path is not used for requesting the retransmissions of a message block in error, hence the name forward error correction. Several codes have been developed to suit applications requiring FEC. Those most commonly recognised have been based on the research of mathematician Richard W. Hamming. These codes are referred to as Hamming codes. Hamming codes employ the use of redundant bits that are inserted into the message stream for error correction. The position of these bits are established and known by the transmitter. If the receiver detects an error in the message block, the Hamming bits are used to identify the position of the error. This position, known as the syndrome, is the underlying principle of the Hamming code.

If *k* is a number of bits in the message stream m_0, m_1, ...m_k and q is the number of Hamming bits, then the total numbers of the bits *n* in the transmitted bit stream is defined as

$$2^q \geq n + 1 \;; n = k + q$$

Example 4.7 For a message of 10 data bits M=1101001110. Using the Hamming code compute the flowing:

a) Find q and n;

b) Check bits C_0, C_1, C_2, C_4 that are inserted into a message stream;

c) Total transmitted bit stream;

d) Syndrome if bit position 7 is corrupted.

Solution:

$$\text{a) } 2^4 \geq (10+4)+1;$$

$$q = 4;\ \ n = 14.$$

The Table 4.3 shows the binary representation of each bit position forms an alternating bit pattern in the vertical direction. Each column proceeding from the LSB to the MSB alternates at one-half the rate of the previous column. The LSB alternates with every position. The next bit alternates every two-bit position, and so forth.

Table 4.3 Bit Positions

Bit position in message	**Binary representation**	**Check bit**	**Position set**
1	0001	C0	1, 3, 5, 7, 9, 11, 13
2	0010	C1	2, 3, 6, 7, 10, 11, 14
3	0011		
4	0100	C2	1, 5, 6, 7, 12, 13, 14
5	0101		
6	0110		
7	0111		
8	1000	C3	8, 9, 10, 11, 12, 13, 14
9	1001		
10	1010		
11	1011		
12	1100		
13	1101		
14	1110		
15	1111		

The original message bit stream is

m_9	m_8	m_7	m_6	m_5	m_4	m_3	m_2	m_1	m_0
1	1	0	1	0	0	1	1	1	0

Thus, the bits positions are

m_9	m_8	m_7	m_6	m_5	m_4	C_3	m_3	m_2	m_1	C_2	m_0	C_o	C_1

To determine C_0 , C_1 , C_2 , C_3 we use the following.

$$C_o = m_8 \oplus m_6 \oplus m_4 \oplus m_3 \oplus m_1 \oplus m_0 = 1 \oplus 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = \mathbf{0}$$

$$C_1 = m_9 \oplus m_6 \oplus m_5 \oplus m_3 \oplus m_2 \oplus m_0 = 1 \oplus 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 = \mathbf{0}$$

$$C_2 = m_9 \oplus m_8 \oplus m_7 \oplus m_3 \oplus m_2 \oplus m_1 = 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 1 = \mathbf{1}$$

$$C_3 = m_9 \oplus m_8 \oplus m_7 \oplus m_6 \oplus m_5 \oplus m_4 = 1 \oplus 1 \oplus 0 \oplus 1 \oplus 0 \oplus 0 = \mathbf{1}$$

Transmitted bit stream

m_9	m_8	m_7	m_6	m_5	m_4	C_3	m_3	m_2	m_1	C_2	m_0	C_1	C_0
1	1	0	1	0	0	1	1	1	1	1	0	0	0

Received bit stream

Error

m_9	m_8	m_7	m_6	m_5	m_4	C_3	m_3	m_2	m_1	C_2	m_0	C_1	C_0
1	1	0	1	0	0	1	**0**	1	1	1	0	0	0

The receiver computes the following syndrome:

$$C^*_{0,} = m_8 \oplus m_6 \oplus m_4 \oplus m_3 \oplus m_1 \oplus m_0 = 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 0 = \mathbf{1}$$

$$C^*_1 = m_9 \oplus m_6 \oplus m_5 \oplus m_3 \oplus m_2 \oplus m_0 = 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 0 = \mathbf{1}$$

$$C^*_2 = m_9 \oplus m_8 \oplus m_7 \oplus m_3 \oplus m_2 \oplus m_1 = 1 \oplus 1 \oplus 0 \oplus 0 \oplus 1 \oplus 1 = \mathbf{0}$$

$$C^*_3 = m_9 \oplus m_8 \oplus m_7 \oplus m_6 \oplus m_5 \oplus m_4 = 1 \oplus 1 \oplus 0 \oplus 1 \oplus 0 \oplus 0 = \mathbf{1}$$

The resulting syndrome is 1 0 1 1. Exclusive-ORing received check bits and computed syndrome, we detect bit position that corrupted.

$$\begin{array}{lll} \quad C_3 \;\; C_2 \;\; C_1 \;\; C_0 & \longrightarrow \quad 1\;1\;0\;0 & \\ \oplus\; C^*_3 \;\; C^*_2 \;\; C^*_1 \;\; C^*_0 & \qquad \oplus\; 1\;0\;1\;1 & \\ & \qquad\;\; 0\;1\;1\;1 & 7^{th}\ \text{Bit.} \end{array}$$

Thus, corrupted bit position (7 th bit position-m_3) is detected and it can be corrected by written *1* instead of *0*. The check bits C_3, C_2, C, C_0 are removed from positions 1, 2, 4, and 8, thereby resulting in the original message.

4.11 Algorithm Encoding

Generating of the transmitted Cyclic Code is performed by the procedure described below:

1. Data bits are represented by message polynomial M (x). Example for the message M = 10011

   $M(x) = X^4 + X + 1$

2. Message polynomial M (x) is multiplied by a factor X^q.
3. The product M (x) X^{n-k} is then divided by the generator polynomial G (x). The quotient is discarded and the remainder B (x), the BCC, is transmitted at the end of the message bits.

discarded

$$\mathbf{M(x).X^{n-k} / G(x) = \cancel{C(x)} \oplus B(x) / G(x)}$$

where

C (x)	- quotient
$\oplus$	- Exclusive OR operation
G (x)	- generator polynomial
X^q	- multiplication factor
B (x)	- BCC

4. The transmitted block code can be represented by the polynomial $T(x) = X^q . M(x) \oplus B(x)$

Example 4.11 Given a message M = 101001101 and Generator polynomial $G(x) = X^5 + X^2 + X + 1$, compute the total transmitted block T(x).

Solution:

M = 101001101 (k = 9)

Message polynomial, $M(x) = X^8 + X^6 + X^3 + X^2 + 1$

Generator polynomial $G(x) = X^5 + X^2 + X + 1$

The number of bits in the BCC, q = 5, (the highest degree of the generator polynomial). The bits of total transmitted block T(x) n = k + q = 14.

1.Multiply the message polynomial, M (x) , by X^{n-k}

$$X^{q}[M(x)] = X^{5}(X^{8}+X^{6}+X^{3}+X^{2}+1) = X^{13}+X^{11}+X^{8}+X^{7}+X^{5}$$
$$= 10100110100000$$

2.Divide X^{q} [M (x)] by the generator polynomial and discard the quotient. The remainder is the BCC, B (x).

$$\begin{array}{cr|l}
 & X^{13}+X^{11}+X^{8}+X^{7}+X^{5} & X^{5}+X^{2}+X+1 \\ \hline
\oplus & X^{13}+X^{10}+X^{9}+X^{8} & X^{8}+X^{6}+X^{5}+X^{4}+X^{3}+X^{2}+X \\ \cline{1-2}
 & X^{11}+X^{10}+X^{9}+X^{7}+X^{5} & \\
\oplus & X^{11}+X^{8}+X^{7}+X^{6} & \\ \cline{1-2}
 & X^{10}+X^{9}+X^{8}+X^{6}+X^{5} & \\
\oplus & X^{10}+X^{7}+X^{6}+X^{5} & \\ \cline{1-2}
 & X^{9}+X^{8}+X^{7} & \\
\oplus & X^{9}+X^{6}+X^{5}+X^{4} & \\ \cline{1-2}
 & X^{8}+X^{7}+X^{6}+X^{5}+X^{4} & \\
\oplus & X^{8}+X^{5}+X^{4}+X^{3} & \\ \cline{1-2}
 & X^{7}+X^{6}+X^{3} & \\
\oplus & X^{7}+X^{4}+X^{3}+X^{2} & \\ \cline{1-2}
 & X^{6}+X^{4}+X^{2} & \\
\oplus & X^{6}+X^{3}+X^{2}+X & \\ \cline{1-2}
 & X^{4}+X^{3}+X & \rightarrow\ 11010 = B(x)
\end{array}$$

3. The total transmitted message block, T (x) = B (x) $\oplus$ X^{n-k} [M (x)];

$$T(x) = X^{n-k}[M(x)] + B(x) = X^{13}+X^{11}+X^{8}+X^{7}+X^{5}+X^{4}+X^{3}+X$$

$$\begin{array}{crl}
 & 10100110100000 & \\
\oplus & \underline{11010} & \text{BCC, B (x)} \\
 & 10100110111010 & \text{transmitted block, T (x)}
\end{array}$$

The G (x) for different cyclic codes are derived from the Table 4.5.

Table 4.5 Generator Polynomials

Type	n	k	Rc	dmin	G(x)
Hamming Codes	7	4	0.57	3	X^3+X^2+1
	15	11	0.73	3	X^4+X^2+1
	31	26	0.84	3	X^5+X^2+1
Golay code	23	12	0.52	7	$X^{11}+X^9+X^7+X^6+X^5+X+1$
Base, Chaudhuri Hocquenghem (BCH) code	15	7	0.46	5	$X^8+X^7+X^6+X^4+1$
	31	21	0.68	5	$X^{10}+X^9+X^8+X^6+X^5+X^3+1$
	63	45	0.71	7	$X^{17}+X^{16}+X^{15}+X^{14}+X^8+X^7+X^6+$ $+X^3+X^2+X+1$

Example 4.12

For simplicity for k=4 bits message M=1101, we use $G(x) = X^3+X^2+1$

The total number of bits in the transmitted message block, n=7, $q = n-k = 3$.

Generator polynomial; $G(x) = X^3+X^2+1$

Message polynomial; $M(x) = X^2+X+1$

1) Multiply the message polynomial,

$M(x)\,X^3 = (X^2+X+1)\,X^3 = X^5+X^4+X^3$

2) Divide $M(x)\,X^3$ by the generator polynomial and discard the quotient we have

$X^5+X^4+X^3$	X^3+X^2+1
$\oplus\ X^5+X^4+X^2$	X^2+1
X^3+X^2	
$\oplus\ X^3+X^2+1$	

$$\overline{\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad 1} \quad \rightarrow \quad B(x)$$

3) Transmitted message block is;

$T(X) = X^3M(X) \oplus B(x);$

$T(X) = X^5+X^4+X^3+1;$ $\quad\quad$ $T(X) = 111001$

4.11.1 Algorithm Decoding

At the receiving and the received block T*(x), is divided by the same generating polynomial G (x). If the remainder is zero, the received block T*(x) = T (x), and the block received without errors otherwise we get no zero remainder indicating detected errors.

Example 4.13

Realise the syndrome decoding for the problem given in Example 4.11.

$$\begin{array}{rl|l}
 & X^{13}+X^{11}+X^8+X^7+X^5+X^4+X^3+X & X^5+X^2+X+1 \\
\oplus & X^{13}+X^{10}+X^9+X^8 & X^8+X^6+X^5+X^4+X^3+X \\
 & X^{11}+X^{10}+X^9+X^7 & \\
\oplus & X^{11}+X^8+X^7+X^6 & \\
 & X^{10}+X^9+X^8+X^6+X^5 & \\
\oplus & X^{10}+X^7+X^6+X^5 & \\
 & X^9+X^8+X^7+X^4 & \\
\oplus & X^9+X^6+X^5+X^4 & \\
 & X^8+X^7+X^6+X^5 & \\
\oplus & X^8+X^5+X^4+X^3 & \\
 & X^7+X^6+X^4+X^2 & \\
\oplus & X^7+X^4+X^3+X^2 & \\
 & X^6+X^3+X^2+X & \\
\oplus & X^6+X^3+X^2+X & \\
 & 0 \rightarrow & B(x) = 0
\end{array}$$

Example 4.14 Consider the message given in Example 4.12. Suppose 7 th bit is distorted during the transmission and instead of T = 0111001 we receive T* = 1111001

$T^*(X) = X^6+X^5+X^4+X^3+1$

$$\begin{array}{rl|l}
 & X^6+X^5+X^4+X^3+1 & X^3+X^2+1 \\
\oplus & X^6+X^5+X^3 & X^3+X+1 \\
 & X^4+1 &
\end{array}$$

$$
\begin{array}{ll}
\oplus\ X^4+X^3+X & \\
\quad X^3+X+1 & \\
\oplus\ X^3+X^2+1 & \\
\quad X^2+X & B(x) = X^2+X \rightarrow 110
\end{array}
$$

Example 4.15 Consider the cyclic (7, 4) code generated by G (x) = $X^3 + 0 + X + 1$. We'll use long division to calculate the check bit polynomial C (x) when M = (1 1 0 0). We first write the message-bit polynomial

$$M(x) = X^3 + X^2 + 0 + 0$$

So $X^q M(x) = X^3 M(x) = X^6 + X^5 + 0 + 0 + 0 + 0 + 0$. Next, we divide $X^q M(x)$ into G (x), keeping in mind that subtraction is the same as addition in mod-2 arithmetic. Thus

$$
\begin{array}{l|l}
X^6+X^5 & X^3+X+1 \\
\oplus\ X^6+X^4+X^3 & X^3+X^2+X \\
\quad X^5+X^4+X^3 & \\
\oplus\ X^5+X^3+X^2 & \\
\quad X^4+X^2 & \\
\oplus\ X^4+X^2+X & \\
\quad X & B(x) = X \quad 10
\end{array}
$$

So the complete transmitted code polynomial is

$$T(x) = X^3 M(x) + B(x) = X^6 + X^5 + 0 + 0 + 0 + X + 0$$

Table 4.5 Shift Register Encoder for (7,4) Code

Input bits m	Register bits before shift			Register bits after shift		
	R_2	R_1	R_0	$R_2' = R_1$	$R'_1 = R_0 \oplus R'_0$	$R'_0 = R_2 \oplus m$
1	0	0	0	0	1	1
1	0	1	1	1	0	1
0	1	0	1	0	0	1
0	0	0	1	0	1	0

4.12 ARQ System

The ARQ strategy for error control is based on error detection and retransmissions. Consequently, ARQ system differs from FEC systems in three important respects.

1. An (n, k) block code designed for error detection generally requires fewer check bits and has a higher k/n ratio than a code designed for error correction.
2. ARQ system needs a return transmission path and additional hardware in order to implement repeat transmission of codeword with detects errors.
3. The forward transmission bit rate must make allowance for repeated word transmissions.

Each codeword constructed by the encoder is stored temporarily and transmitted to the destination where the decoder looks for errors. The decoder issues a positive acknowledgment (ACK) if no errors are detected, or a negative acknowledgment (NAK) if errors are detected. A negative acknowledgment causes the input controller to retransmit the appropriate word from those stored by the input buffer. A particular word may be transmitted just two or more times, depending on the occurrence of transmission errors. The function of the output controller and buffer is to assemble the output bit stream from the codewords that have been accepted by the decoder.

When an error is detected in a word, the receiver signals back to the transmitter and the word is transmitted again. There are three basic ARQ systems: *stop-and-wait* ARQ, *go-back N ARQ,* and *selective-repeat ARQ.*

a) The *stop-and-wait* ARQ system is the simplest to implement to implement and is represented in Figure 4.13(a). The transmitter sends a codeword to the receiver during the time T_t. The receiver receives and processes the received word and if the receiver detects no error, it than sends back to the transmitter a *positive-acknowledgment* (ACK) signal. Upon receipt of the ACK signal, the transmitter sends the next word. The elapsed time between the end of transmission of one word and the start of transmission of the next word is T_1. Clearly the limitation of such a system is that it must stand by idly without transmission while waiting for an ACK or NAK. Nonetheless this system is effectively used in many data systems including IBM's Binary Synchronous Communication (BISYNC) protocol.

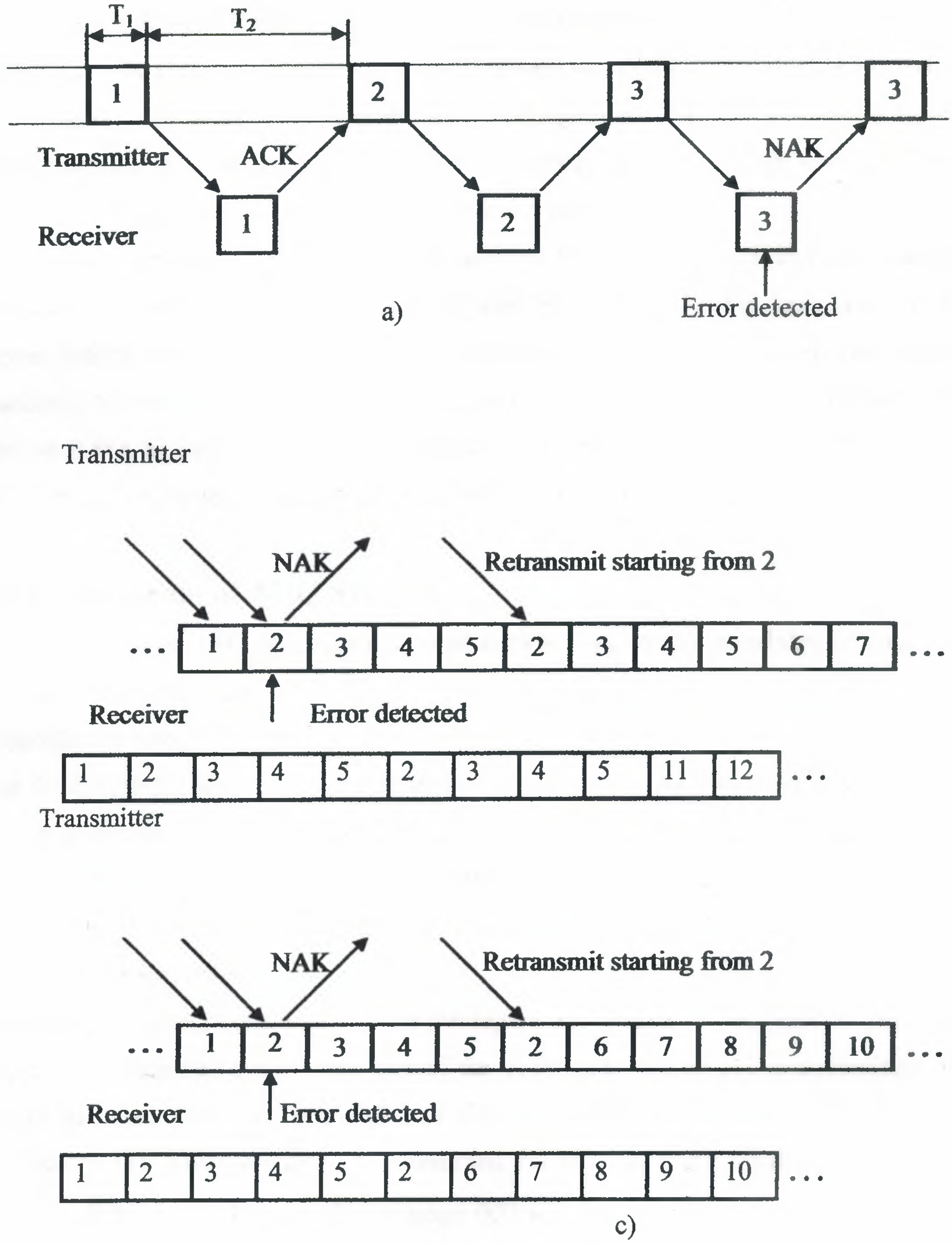


Figure 4.13 *Stop-and-Wait* ARQ System

b) The *go-back* N ARQ scheme is represented in Figure 7.13 (b). The transmitter sends messages, one after another, without delay and does not wait for an ACK signal. When, however, the receiver detects an error in a message, say message i, a NAK signal is returned to the transmitter. In response to the NAK the transmitter returns to codeword i and starts all over again at that word. Figure 7.13 (b) we have assumed that the propagation delay and the processing at the receiver occupies an interval of such length that when an error is detected in word i the number N or words that are sent over again is N=4. The go-back-N system is readily implemented, and, as we shall see, it is a significant improvement over the stop-and-wait system. The go-back-N scheme is used in many data systems.

c) The selective-repeat ARQ system is represented in Figure 4.13 (c). Here the transmitter sends messages in succession, again without waiting for an ACK after each message. If the receiver detects that there is an error in codeword i, the transmitter is notified. The transmitter retransmits codeword i and thereafter returns immediately to its sequential transmission. The selective ARQ, as might well be anticipated, has the highest transmission efficiency of the three systems but, on the other hand, it is the most costly to implement.

4.13 Performance of ARQ Systems

The performance of ARQ system is measured in two ways, by the probability of error and by the transmission efficiency.

Probability of Error

In an ARQ system, block codes are used for error detection. As discussed earlier, whenever an error is detected a NAK is returned to the transmitter and the message is repeated. Thus, the only time that a received message is in error is when a received message has a sufficient number of errors and looks like a different codeword. In such a case an ACK is returned and therefore an error is made.

To illustrate this point considers that the message 0 and 1 are encoded prior to transmission using a (3, 1) repeated code. Thus, one of the codewords 000 or 111 is transmitted. If the receiver 001, 010, 100, 110, 101 or 011, a NAK is returned to the transmitter. If, however, either 000 or 111 is received an ACK is returned. If a 000 was transmitted and a 111 received an error will be made. Thus, if the message 000 was transmitted there are $2^3 = 8$ possible received words and an error is made only when the message 111 is received. If all eight messages were equally likely to be received the probability of error would be $P_e = 1/8$. However, in order for a received word to be in error, all three bits must be in error. Since such an event is less likely than any of the other seven possible received words, the probability of error is less than 1/8 and we say that P_e is upper bounded by 1/8, that is $P_e \leq 1/8$.

Let us now generalize our discussion to an arbitrary (n, k) block code. There are now 2^n possible received words and of these there are 2^k codewords. Thus, if a codeword is transmitted, an error will occur if one of the other $2^k - 1$ codewords is received. To upper bound P_e we again assume that all 2^n possible received words are equally likely. Then P_e is

$$P_e \leq \frac{2^k - 1}{2^n} \approx 2^{-(n-k)}$$

If a BCH (1023,973) code is used for error detection, the probability of error Pe is bounded by

$$P_e \leq 2^{-50} \cong 10^{-15}$$

Throughput

The throughput efficiency is defined as the ratio of the average number of information bits accepted at the receiver per unit of time to the number of information bits that would be accepted per unit of time if ARQ were not used. While all the ARQ systems yield the same error rate, the throughput efficiencies are different.

Throughput of the Stop-and-Wait ARQ

Let P_A is the probability that the receiver accepts the message on any particular transmission. Then the probability that only a single transmission is all that is needed for acceptance is PA. The probability that two transmission will be required is $(1-P_A)P_A$ that is, the product of the probability $(1-P_A)P_A$ that is , the product of the probability $(1 - P_A)$ that the first transmission was rejected and P_A the probability that it was accepted on the second try . The average number of transmission required for acceptance of a single word is the sum of the products of the number of transmission j and the probability of requiring j transmissions, $P_A(1-P_A)^{j-1}$.

$N_{sw} = 1 \cdot P_A + 2 \cdot P_A(1-P_A) + 3P_A(1-P_A)^2 + \ldots = 1/P_A$

The average time required to transmit one word is:

$$T_{sw} = \frac{T_1 + T_2}{P_A}$$

If ARQ is not used and no coding bits were added to the k information bits, the time needed to transmit the k bits would be

$$T_k = kT_{sw}/n$$

The throughput efficiency of the stop-and-wait ARQ system is

$$\eta_{sw} = \frac{T_k}{T_{sw}} = \frac{k}{n}\frac{P_A}{1+\frac{T_1}{T_2}}$$

Throughput of Go-Back-N ARQ

In this system, when transmitter is informed that an error has been detected in particular word, retransmissions are required of that word and of the N-1 words that followed. Hence the retransmissions involves N words. Thus, if a word is received in error, N words are retransmitted. Thus, the total number of words transmitted is N + 1. If the same word is again in error, the N words are repeated once again etc. Following the analysis used in the stop-and-wait ARQ we find that average number of word transmissions required for the acceptance of a single word is

$$N_{GBN} = 1.P_A + (N+1)P_A(1-P_A) + (2N+1)P_A(1-P_A)^2 + \ldots = 1+N(1-P_A)/P_A$$

Correspondingly the average time to transmit one word is

$$T_{GBN} = T_2\left(1+\frac{N(1-P_A)}{P_A}\right)$$

and

$$\eta_{GBN} = \frac{T_K}{T_{GBN}} = \frac{k}{n}\frac{1}{1+\frac{N(1-P_A)}{P_A}} \qquad (4.6)$$

Selective Repeat ARQ

The mean time for transmission of a word T_{SR} in this selective-repeat case is calculated exactly as in the stop-and-wait case except that T_2 is set to zero. Hence we find

$$T_{SR} = T_2 / P_A$$

and

$$\eta_{SR} = T_K / T_{SR} = \frac{k}{n}P_A$$

As an example, to compare throughput efficiencies, let us assume that $T_1 = 10$ ns, that a BCH (1023,973) code is used, that $P_A = 0.99$ and that $T_2 = 40$ ns. Let us further that N = 4 so that the retransmissions time in the go-back-N system is the same as the idle in the stop-and-wait system We then find:

$$\eta_{sw} = \frac{k}{n}\frac{P_A}{1+\frac{T_1}{T_2}} = \frac{973}{1023}\cdot\frac{0.99}{1+4} = 0.188$$

$$\eta_{GBN} = \frac{k}{n}\frac{1}{1+\frac{N(1-P_A)}{P_A}} = \frac{973}{1023}\cdot\frac{1}{1+\frac{4(0.01)}{0.99}} = 0.915$$

$$\eta_{SR} = \frac{k}{n}P_A = \frac{973}{1023}(0.99) = 0.942$$

Thus there is a significant improvement obtained by using the go-back-N algorithm rather than the stop-and-wait algorithm. However, the improvement made in using the selective repeat algorithm is often not deemed worth the additional complexity.

4.14 Data Compression

Why Do We Need Data Compression?

As an example, imagine the process encoding a television picture. The black and white television picture contains 313.633 pixels (485 rows and (3/4) 485 columns). When we want to send the brightness each pixel we quantize brightness into 128 levels or 7 bits. This means that we need $(2.2)10^6$ bits for each picture. Data compression is a method of using fewer bits to represent same information. Application data compression algorithm reduces the amount of information and allows increasing overall system performance (reducing the time of transmission and processing, memory capacity to store information)

4.14.1 Source-Coding Theorem

The source-coding theorem is one of the three fundamentals of information theory introduced by Shannon. The source-coding theorem establishes a fundamental limit on the rate at which the output of an information source can be compressed without causing a large error probability. In the statement of the source-coding theorem an entropy of the source plays a major role. From communication point of view the entropy denoted by H of a message is the theoretical minimum average number of bits per codeword

$$H = \sum_{1}^{n} P_i \log_2 \frac{1}{P_i} = -\sum_{1}^{n} P_i \log_2 P_i \qquad (4.7)$$

where P_I is the probability of character (i) occurring in the codeword.

The entropy can be thought of as the average information content per source message.

Theorem: A source with entropy (or entropy rate) H can be encoded with arbitrarily small error probability at any rate R (bit/source output) as long as $R < H$, the error probability will be bounded away from zero, independent of the complexity of the encoder and the decoder employed.

4.14.2 Huffman Encoding

Huffman coding provides an organised technic for finding the best possible variable–length code for a given set of messages. The Huffman coding take advantages that not all symbols in the transmitted frame occur with same frequency. As an example low-case letters **a, e, s** are the most frequently occurring characters in English-language text, but the capital letters **X, Q, Z** are rarely occurring. In fact a letter **E** occurs 100 times more often than letter **Q**, or the word **"the"** occurs 10 times more often than the word **"be"**.

The idea of Huffman coding is to encode the more-frequently-occurring fixed-length sequences to shorter binary sequences and the less-frequently-occurring ones to longer binary sequences. Thus, in Huffman coding fixed-length blocks of the source output are mapped to variable-length binary blocks. This is called fixed-to variable-length coding. In variable-length coding synchronization is a problem. The following example clarifies points.

Example 7.18

Let us assume that the possible outputs of an information source are (α_1, α_2, α_3, α_4, α_5), and consider the following four codes for this source.

Table 4.6 Source Output

Letter	Probability	Code 1	Code 2	Code 3	Code 4
α_1	$P_1 = \frac{1}{2}$	1	1	0	00
α_2	$P_2 = \frac{1}{4}$	01	10	10	01

α_3	$p_3 = \frac{1}{8}$	001	100	110	10
α_4	$p_4 = \frac{1}{16}$	0001	1000	1110	11
α_5	$p_5 = \frac{1}{16}$	00001	10000	1111	110

In the first code each codeword ends with a 1. Therefore, as soon as the decoder observes a 1, it knows that the codeword has ended and a new codeword will start. This means that the code is a self-synchronising code. In the second code each codeword starts with a 1. Therefore, upon observing a 1, the decoder knows that a new codeword has started and, hence, the previous bit was the last bit of the previous codeword. This code is again self-synchronising but not as desirable as the first code because with this code we have to wait to receive the first bit of the next codeword to recognise that a new codeword has started, whereas in code 1 we recognise the last bit without having to receive the first bit of the next codeword. Both codes 1 and 2 therefore are uniquely decodable. However, only code 1 is instantaneous. Codes 1 and 3 have the nice property that no codeword is the prefix of another codeword it is said that they satisfy the prefix condition. It can be proved that a necessary and sufficient condition for a code to be uniquely decodable and instantaneous is that it satisfies the prefix condition. This means that both codes 1 and 3 are uniquely decodable and instantaneous. However, code 3 has the advantage of having a smaller average number bit per codeword E(L). In fact, for code E(L) is

$$\mathbf{E(L)} = \sum_{i=1}^{1} \mathbf{m_i P_i} \quad \mathbf{E(L)} = \sum_{i=1}^{n} \mathbf{m_i p_i}$$

where m_I is the number of bits representing i character

$$E(L) = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{16}\right) + 5\left(\frac{1}{16}\right) = \frac{31}{16}$$

and for code 3

$$E(L) = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{16}\right) + 4\left(\frac{1}{16}\right) = \frac{30}{16}$$

Code 4 has a major disadvantage because it is not uniquely decodable. For example, the sequence 110110 can be decoded n two ways, as $\alpha_5\alpha_5$ or $\alpha_1\alpha_2\alpha_3$. Codes that are not uniquely decodable are not desirable and should be avoided in practice. From the discussion above it is seen that the most desirable of the above four codes is code 3, which is uniquely decodable, is instantaneous, and has the least average codeword length. This code is an example of Huffman code, to be discussed- to be discussed shortly.

As already mentioned the idea in Huffman coding is to choose codeword lengths such that more-probable sequences have longer codewords. If we can map each source output of probability P_i to a codeword of length approximately log l/P_i and at the same time ensure unique decodability, we can achieve an average codeword length of approximately H(X). Huffman codes are uniquely decodable instantaneous codes with minimum average codeword length. In this sense they are optimal. By optimal we mean that, among all codes that satisfies, the prefix condition (and therefore are uniquely decodable and instantaneous). Huffman codes have the minimum-average codeword length. We present below the algorithm for the design of the Huffman code. From the algorithm, it is obvious that the resulting code satisfies the prefix condition.

4.14.3 Algorithm Design of Huffman Code

The Huffman code is a source code whose average word length approaches the fundamental limit set by the entropy of discrete memory less source, namely entropy *H*. The Huffman encoding algorithm proceeds as follows:

1. The source symbols are listed in order of decreasing probability. The two source symbols of lowest probability are assigned a 0 and a 1. This part of the step is referred to as a splitting stage.
2. These two sources, symbols are regarded as being combined into a new source symbol with probability equal to the sum of the two original probabilities, The list of source symbols, and therefore source statistics, is thereby reduced in size by one. The probability of the new symbol is placed in the list in accordance with its value.
3. The procedure is repeated until we are left with a final list of source statistics (symbols) of only or only two for which, a 0 and a 1 are assigned.

The code for each (original) source symbol is found by working backward and tracing the sequence or 0s and Is assigned to that symbol as well as its successors.

Example 7.19 The source output is given by table 7.7.

a) Design a Huffman code for the source;

b) Find the entropy of source;

c) Find the average number bits per codeword

Table 4.7 Source Output

Symbol	α_1	α_2	α_3	α_4	α_5
Probability	0.4	0.2	0.2	0.1	0.1
Source code	00	10	11	010	011

Solution:

a) Using the steps 2 and 3 of the algorithm we construct the Huffman code given by

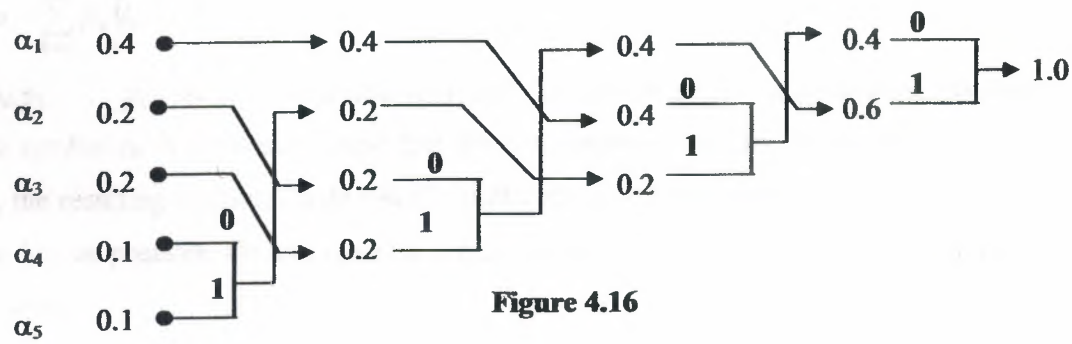


Figure 4.16

the 3rd row of the table.

b) The theoretical minimum average number bits per character

Entropy, $H = \sum_{1-1}^{8} p_1 \log_2 p_1$ bits per codewords

$$H(y) = 0.4\log_2\left(\frac{1}{0.4}\right) + 0.2\log_2\left(\frac{1}{0.2}\right) + 0.2\log_2\left(\frac{1}{0.2}\right) + 0.1\log_2\left(\frac{1}{0.1}\right) + 0.1\log_2\left(\frac{1}{0.1}\right)$$

$$= 0.52877 - 0.46439 - 0.46439 - 0.33219 - 0.33219$$

$$= 2.12193$$

c) The obtained average number of bits per character

Therefore:

$$E(L) = 0.4(2) - 0.2(2) - 0.2(2) - 0.1(3) - 0.1(3) = 2.2$$

For the example we may make two observations: The average code-word length E(L)exceeds the entropy $H(y)$ only 3.67 percent.

It is noteworthy that the Huffman encoding process is not unique. First, at each splitting stage in the construction of a Huffman code, there is arbitrariness in the way a 0 and a 1 are assigned to the last two source symbols. So, the resulting differences are trivial. Second, ambiguity arises when the probability of a combined symbol (obtained by adding the last two probabilities) , is found to equal another probability in the list. We may proceed by placing the probability of the new symbol as high as possible, as in Example 7.14. Alternatively, we may place it as low as possible. This time, noticeable differences arise in that the code words in the resulting source codes can have different lengths. Nevertheless, the average code-word length remains the same.

As a measure of the variability in code-word lengths of a source code, we define the variance of the average code-word length $\overline{L}$ over the ensemble of source symbols as

$$\sigma = \sum_{k=0}^{K-1} \rho_k \left(l_k - \overline{L}\right)^2$$

where $p_0, p_1, \ldots, p_{K-1}$ are the source statistics, and l_k is the length or the code-word assigned to source symbol s_k. It is usually round that when a combined symbol is moved as high as possible, the resulting Huffman code has a significantly smaller variance σ^2 than when it is moved as low as possible. On this basis, it is reasonable to choose the former Huffman code over the latter.

CONCLUSION

Huffman coding provides an organized technique for finding the best possible variable–length code for a given set of messages. The Huffman coding take advantages that not all symbols in the transmitted frame occur with same frequency.

From communication point of view the entropy denoted by H of a message is the theoretical minimum average number of bits per codeword.

The throughput efficiency is defined as the ratio of the average number of information bits accepted at the receiver per unit of time to the number of information bits that would be accepted per unit of time if ARQ were not used. While all the ARQ systems yield the same error rate, the throughput efficiencies are different.

The ARQ strategy for error control is based on error detection and retransmissions. Consequently, ARQ system differs from FEC systems in three important respects.

Whenever an electromagnetic signal flows from one point to another, it is subject to unpredictable interference from heat, magnetism, and other forms of electricity. This interference can change the shape or timing of the signal. If the signal is carrying encoded binary data, such changes can alter the meaning of the data, changing 0 to 1 or 1 to 0. bits can be changed singly or in clumps.

REFERENCES

[1] Alan. V. Oppenheim & Ronald. W. Schafer. (1998). Discrete-Time Signal Processing. International Edition. Prentice Hall

[2] Coutautiue H. Houpis, Gary B. Lamout. (1992). Digital Control Systems. International Edition. McGraw-Hill

[3] Garne E. Brayn. (1999). Telecommunications Primer, Data, Voice and Video Communications. International Edition. Prentice Hall

[4] Gibson J. D. (1993). Principles Of Digital and Analog Communication. 2'nd Edition. Macmillan

[5] Haykin S.S. (1986). Adaptive Filter Theory. Prentice Hall

[6] I. A. Glover & P. M Gran. (1998). Digital Communications. Prentice Hall

[7] I.E.E.E. Communication Magazine Vol.20 No:3 .(1982). The Discrete Fourier Transform Applied to a Time Domain Signal Processing.

[8] John.G.Pproakis and Masoud Salehi. (1994). Communication Systems Engineering. Prentice Hall

[9] J. S. Lim & A. V. Oppenheim. (1988). Efficient Fourier Transform and Convolution Algorithims Advanced Topics in Signal Processing. Prentice Hall

[10] L. W Couch. (1990). Digital and Analog Communication Systems. 3'rd Edition. Macmillan

[11] Mamedov F. S. (1981). Adaptive Sampling. Patent No:792582. 1981

[12] Mamedov F. S. (1982). Reconstruction of Analog Signals from its Samples. Patent No: 877572. 1982

[13] R. J. Marks. (1991). Introduction to Shannon Sampling and Interpolation Theory. Spinger-Valey

[14] R. I. Damper. (1995). Introduction to Discrete-Time Signals and Systems. Charpman & Hall

[15] Simon Haykin. (1994). Communication Systems. 3'rd Edition. WILEY

[16] Steven C. Chatra and Reymond P. Canale. (1988). Numerical Methods for Engineers. International Edition. McGraw-Hill

[17] Widrow B. (1986). Adaptive Filters Fundamentals. Tech.Report No:6764-6 (From E.M.Univercity Bookstore)

[18] Ziemer R. E & Tranter W. H. (1990). Principles of Communication Systems Modulation and Noise. Macmillan