DATA COMMUNICATIONS & COMPUTER NETWORK

LECTURE NOTES

Department of Information Technology

Jharsuguda Engineering School, Jharsuguda

Chapter-1

Introduction to Data Communications:

In Data Communications, *data* generally are defined as information that is stored in digital form. *Data communications* is the process of transferring digital information between two or more points. *Information* is defined as the knowledge or intelligence. Data communications can be summarized as the transmission, reception, and processing of digital information. For data communications to occur, the communicating devices must be part of a communication system made up of a combination of hardware (physical equipment) and software (programs). The effectiveness of a data communications system depends on four fundamental characteristics: delivery, accuracy, timeliness, and jitter.

1. **Delivery.** The system must deliver data to the correct destination. Data must be received by the intended device or user and only by that device or user.
2. **Accuracy.** The system must deliver the data accurately. Data that have been altered in transmission and left uncorrected are unusable.
3. **Timeliness**. The system must deliver data in a timely manner. Data delivered late are useless. In the case of video and audio, timely delivery means delivering data as they are produced, in the same order that they are produced, and without significant delay. This kind of delivery is called real-time transmission.
4. **Jitter**. Jitter refers to the variation in the packet arrival time. It is the uneven delay in the delivery of audio or video packets. For example, let us assume that video packets are sent every 30 ms. If some of the packets arrive with 30-ms delay and others with 40-ms delay, an uneven quality in the video is the result.

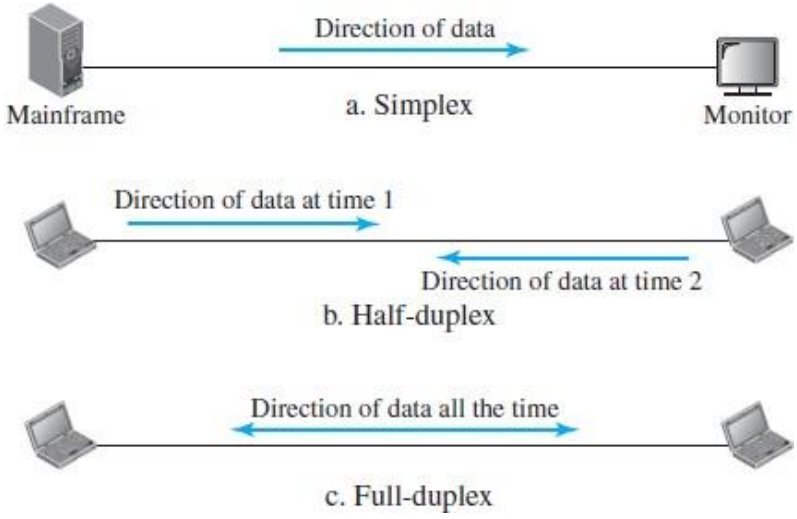
COMPONENTS:

A data communications system has five components:

1. **Message:** The message is the information (data) to be communicated. Popular forms of information include text, numbers, pictures, audio, and video.
2. **Sender:** The sender is the device that sends the data message. It can be a computer, workstation, telephone handset, video camera, and so on.
3. **Receiver:** The receiver is the device that receives the message. It can be a computer, workstation, telephone handset, television, and so on.
4. **Transmission medium:** The transmission medium is the physical path by which a message travels from sender to receiver. Some examples of transmission media include twisted- pair wire, coaxial cable, fiber-optic cable, and radio waves.
5. **Protocol**: A protocol is a set of rules that govern data communications. It represents an agreement between the communicating devices.

Data Flow:

Communication between two devices can be simplex, half-duplex, or full-duplex



- **Simplex:**

In simplex mode, the communication is unidirectional, as on a one-way street. Only one of the two devices on a link can transmit; the other can only receive. Keyboards and traditional monitors are examples of simplex devices. The keyboard can only introduce input; the monitor can only accept output. The simplex modem can use the entire capacity of the channel to send data in one direction.

- **Half-Duplex:**

In half-duplex mode, each station can both transmit and receive, but not at the same time. When one device is sending, the other can only receive, and vice versa.The half-duplex mode is like a one-lane road with traffic allowed in both directions. In a half-duplex transmission, the entire capacity of a channel is taken over by whichever of the two devices is transmitting at the time. Walkie-talkies and CB (citizens band) radios are both half-duplex systems.
The half-duplex mode is used in cases where there is no need for communication in both directions at the same time; the entire capacity of the channel can be utilized for each direction.

- **Full-Duplex:**

In full-duplex mode (also called duplex), both stations can transmit and receive simultaneously. The full-duplex mode is like a two-way street with traffic flowing in both directions at the same time. In full-duplex mode, signals going in one direction share the capacity of the link with signals going in the other direction. This sharing can occur in two ways: Either the link must contain two physically separate transmission paths, one for sending and the other for receiving; or the capacity of the channel is divided between signals traveling in both directions. One common example of full-duplex communication is the telephone network. When two people are communicating by a telephone line, both can talk and listen at the same time. The full-duplex mode is used when communication in both directions is required all the time. The capacity of the channel, however, must be divided between the two directions.

NETWORKS:

INTRODUCTION:

A network is the interconnection of a set of devices capable of communication. In this definition, a device can be a host (or an end system as it is sometimes called) such as a large computer, desktop, laptop, workstation, cellular phone, or security system. A device in this definition can also be a connecting device such as a router, which connects the network to other networks, a switch, which connects devices together, a modem (modulator-demodulator), which changes the form of data, and so on. These devices in a network are connected using wired or wireless transmission media such as cable or air. When we connect two computers at home using a plug-and-play router, we have created a network, although very small.

Network Components, Functions, and Features:

The major components of a network are end stations, applications and a network that will support traffic between the end stations. Computer networks all share common devices, functions, and features, including servers, clients, transmission media, shared data, shared printers and other peripherals, hardware and software resources, network interface card (NIC), local operating system (LOS) and the network operating system (NOS).

Servers: Servers are computers that hold shared files, programs and the network operating system. Servers provide access to network resources to all the users of the network and different kinds of servers are present. Examples include file servers, print servers, mail servers, communication servers etc.

Network interface card: Every computer in the network has a special expansion card called network interface card (NIS), which prepares and sends data, receives data, and controls data flow between the computer and the network. While transmitting, NIC passes frames of data on to the physical layer and on the receiver side, the NIC processes bits received from the physical layer and processes the message based on its contents**.**

Local operating system: A local operating system allows personal computers to access files, print to a local printer, and have and use one or more disk and CD drives that are located on the computer. Examples are MS-DOS, PC-DOS, UNIX, Macintosh, OS/2, Windows 95, 98, XP and Linux.

Network operating system: the NOS is a program that runs on computers and servers that allows the computers to communicate over a network. The NOS provides services to clients such as log-in features, password authentication, printer access, network administration functions and data file sharing.

NETWORK MODELS:

Computer networks can be represented with two basic network models: peer-to-peer client/server and dedicated client/server. The client/server method specifies the way in which two computers can communicate with software over a network.

Peer-to-peer client/server network: Here, all the computers share their resources, such as hard drives, printers and so on with all the other computers on the network. Individual resources like disk drives, CD-ROM drives, and even printers are transformed into shared, collective resources that are accessible from every PC. Unlike client-server networks, where network information is stored on a centralized file server PC and made available to tens, hundreds, or thousands client PCs, the information stored across peer-to-peer networks is uniquely decentralized. Because peer-to-peer PCs have their own hard disk drives that are accessible by all computers, each PC acts as both a client (information requestor) and a server (information provider). The peer-to-peer network is an appropriate choice when there are fewer than 10 users on the network, security is not an issue and all the users are located in the same general area.

The advantages of peer-to-peer over client-server NOSs include

- No need for a network administrator.
- Network is fast/inexpensive to setup & maintain.
- Each PC can make backup copies of its data to other PCs for security.
- Easiest type of network to build, peer-to-peer is perfect for both home and office use.

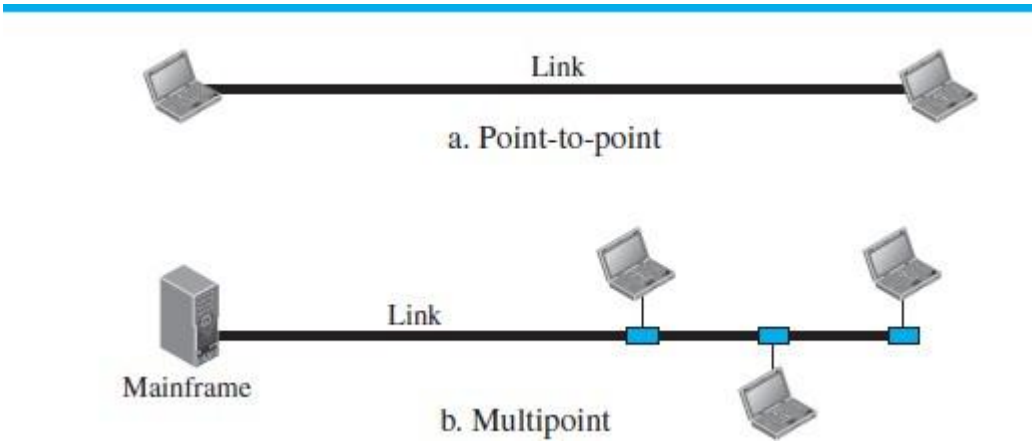
Dedicated client/server network: Here, one computer is designated as server and the rest of the computers are clients. Dedicated Server Architecture can improve the efficiency of client server systems by using one server for each application that exists within An organization. The designated servers store all the networks shared files and applications programs and function only as servers and are not used as a client or workstation. Client computers can access the servers and have shared files transferred to them over the transmission medium. In some client/server networks, client computers submit jobs to one of the servers and once they process the jobs, the results are sent back to the client computer.

NOTE: In general, **the dedicated client/server** model is preferable to **the peer-to-peer client/server** model for general purpose data networks.

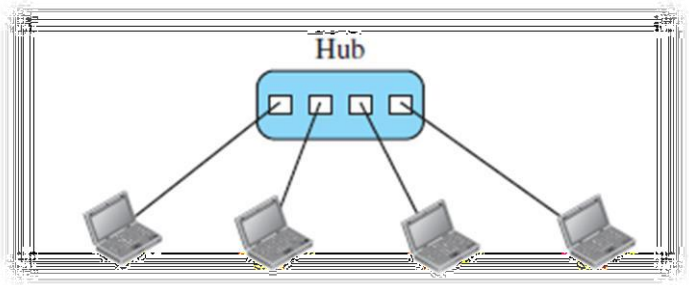
PHYSICAL STRUCTURE (Physical Topology):

In computer networking, Physical structure or topology refers to the layout of connected devices, i.e. how the computers, cables, and other components within a data communications network are interconnected, both physically and logically. The physical topology describes how the network is actually laid out, and the logical topology describes how the data actually flow through the network.

Two most basic topologies are point-to- point and multipoint. A point to-point topology usually connects two mainframe computers for high-speed digital information. A multipoint topology connects three or more stations through a single transmission medium and some examples are star, bus, ring, mesh and hybrid.

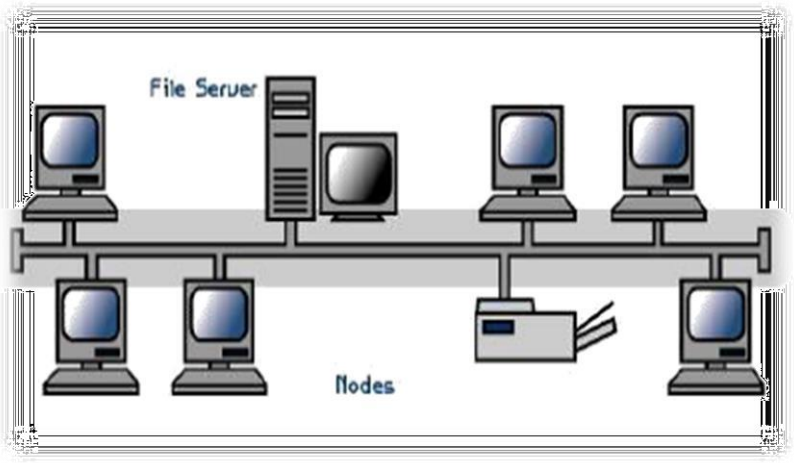


Star topology: A star topology is designed with each node (file server, workstations, and peripherals) connected directly to a central network hub, switch, or concentrator. Data on a star network passes through the hub, switch, or concentrator before continuing to its destination. The hub, switch, or concentrator manages and controls all functions of the network. It also acts as a repeater for the data flow.



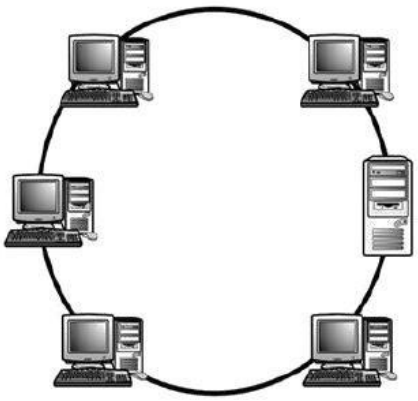
ADVANTAGES	DISADAVANTAGES
• Easily expanded with out any disruption of the network. • Cable failure affects only a single user. • Easy to trouble shoot and isolate a problem	• Requires more cable. • A central connecting device allows for a single point of failure. • More difficult to implement

Bus topology: Bus networks use a common backbone to connect all devices. A single cable, (the backbone) functions as a shared communication medium that devices attach or tap into with an interface connector. A device wanting to communicate with another device on the network sends a broadcast message onto the wire that all other devices see, but only the intended recipient actually accepts and processes the message. The bus topology is the simplest and most common method of interconnecting computers. The two ends of the transmission line never touch to form a complete loop. A bus topology is also known as multidrop or linear bus or a horizontal bus.



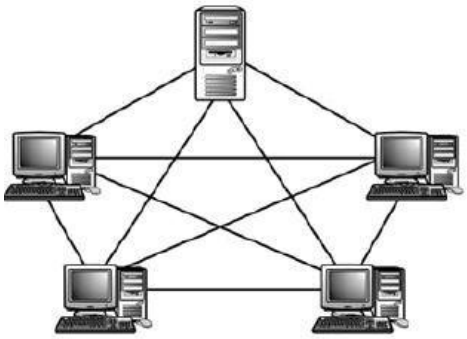
ADVANTAGES	DISADVANTAGES
• Cheap and easy to implement. • Required less cable. • Does not use any specialized network equipment.	• Network disruptions occur when hosts are added or removed • A fault in the cable prevents all system to go offline. • Difficult to trouble shoot

Ring topology: In a ring network (sometimes called a loop), every device has exactly two neighbours for communication purposes. All messages travel through a ring in the same direction (either "clockwise" or "counter clockwise"). All the stations are interconnected in tandem (series) to form a closed loop or circle. Transmissions are unidirectional and must propagate through all the stations in the loop. Each computer acts like a repeater and the ring topology is similar to bus or star topologies.



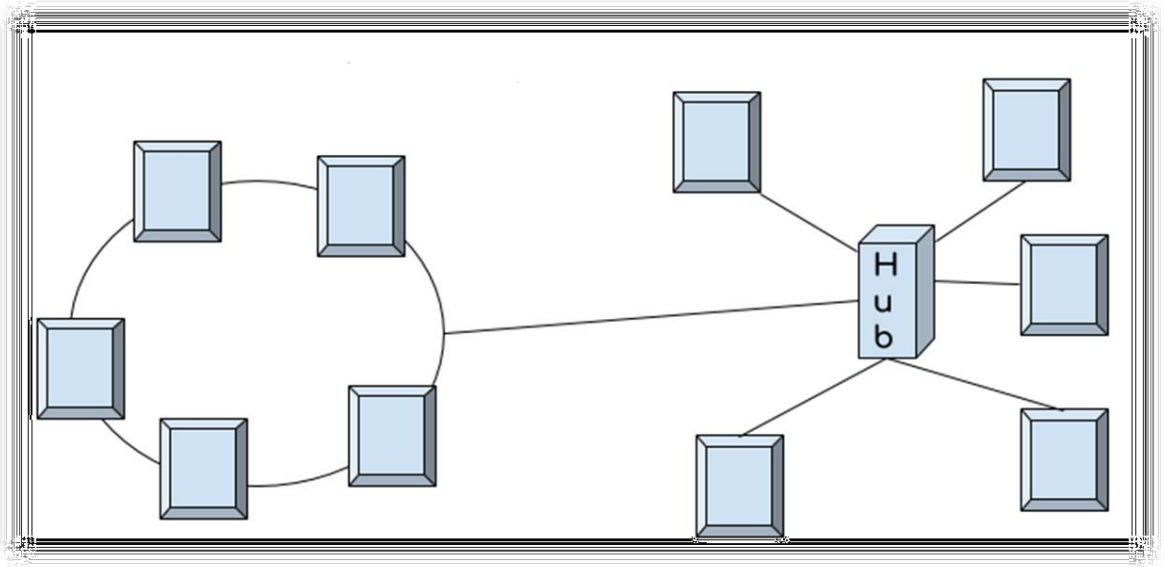
Advantages	Disadvantages
Cable faults are easily located, making troubleshooting easier	Expansion to the network can cause network disruption
Ring networks are moderately easy to install	A single break in the cable can disrupt the entire network.

Mesh topology: The mesh topology incorporates a unique network design in which each computer on the network connects to every other, creating a point-to-point connection between every device on the network. Unlike each of the previous topologies, messages sent on a mesh network can take any of several possible paths from source to destination.
Mesh network in which every device connects to every other is called a full mesh. A disadvantage is that, a mesh network with n nodes must have n(n-1)/2 links and each node must have n-1 I/O ports (links).



Advantages	Disadvantages
Provides redundant paths between devices	Requires more cable than the other LAN topologies
The network can be expanded without disruption to current uses	Complicated implementation

Hybrid topology: This topology (sometimes called mixed topology) is simply combining two or more of the traditional topologies to form a larger, more complex topology. Main aim is being able to share the advantages of different topologies.



Advantages	Disadvantages
• We can choose the topology based on the requirement for example, scalability is our concern then we can use star topology instead of bus technology. • Scalable as we can further connect other computer networks with the existing networks with different topologies.	• Fault detection is difficult. • Installation is difficult.Design is complex so maintenance is high thus expensive.

STANDARDS:

An Internet standard is a thoroughly tested specification that is useful to and adhered to by those who work with the Internet. It is a formalized regulation that must be followed.

There is a strict procedure by which a specification attains Internet standard status. A specification begins as an Internet draft. An Internet draft is a working document (a work in progress) with no official status and a six-month lifetime. Upon recommendation from the Internet authorities, a draft may be published as a Request for Comment (RFC). Each RFC is edited, assigned a number, and made available to all interested parties. RFCs go through maturity levels and are categorized according to their requirement level.

Standards Organizations for Data Communications

1. **International Standard Organization (ISO):**
   ISO is the international organization for standardization on a wide range of subjects. It is comprised mainly of members from the standards committee of various governments throughout the world. It is even responsible for developing models which provides high level of system compatibility, quality enhancement, improved productivity and reduced costs. The ISO is also responsible for endorsing and coordinating the work of the other standards organizations.

2. **International Telecommunications Union-Telecommunication Sector (ITU-T)**

   ITU-T is one of the four permanent parts of the International Telecommunications Union based in Geneva, Switzerland. It has developed three sets of specifications: the V series for modem interfacing and data transmission over telephone lines, the X series for data transmission over public digital networks, email and directory services; the I and Q series for Integrated Services Digital Network (ISDN) and its extension Broadband ISDN. ITU-T membership consists of government authorities and representatives from many countries and it is the present standards organization for the United Nations.

3. **Institute of Electrical and Electronics Engineers (IEEE**)
   IEEE is an international professional organization founded in United States and is compromised of electronics, computer and communications engineers. It is currently the world's largest professional society with over 200,000 members. It develops communication and information processing standards with the underlying goal of advancing theory, creativity, and product quality in any field related to electrical and electronics Engineering.

4. **American National Standards Institute (ANSI)**
   ANSI is the official standards agency for the United States and is the U.S voting representative for the ISO. ANSI is a completely private, non-profit organization comprised of equipment manufacturers and users of data processing equipment and services. ANSI membership is comprised of people form professional societies, industry associations, governmental and regulatory bodies, and consumer goods.
5. **Electronics Industry Association (EIA)**
   EIA is a non-profit U.S. trade association that establishes and recommends industrial standards.EIA activities include standards development, increasing public awareness, and lobbying and it is responsible for developing the RS (recommended standard) series of standards for data and communications.
6. **Telecommunications Industry Association (TIA)**
   TIA is the leading trade association in the communications and information technology industry. It facilitates business development opportunities through market development, trade promotion, trade shows, and standards development. It represents manufacturers of communications and information technology products and also facilitates the convergence of new communications networks.

7. **Internet Architecture Board (IAB)**
   IAB earlier known as Internet Activities Board is a committee created by ARPA (Advanced Research Projects Agency) so as to analyze the activities of ARPANET whose purpose is to accelerate the advancement of technologies useful for U.S military.

8. **Internet Engineering Task Force (IETF)**

   The IETF is a large international community of network designers, operators, vendors and researchers concerned with the evolution of the Internet architecture and smooth operation of the Internet.

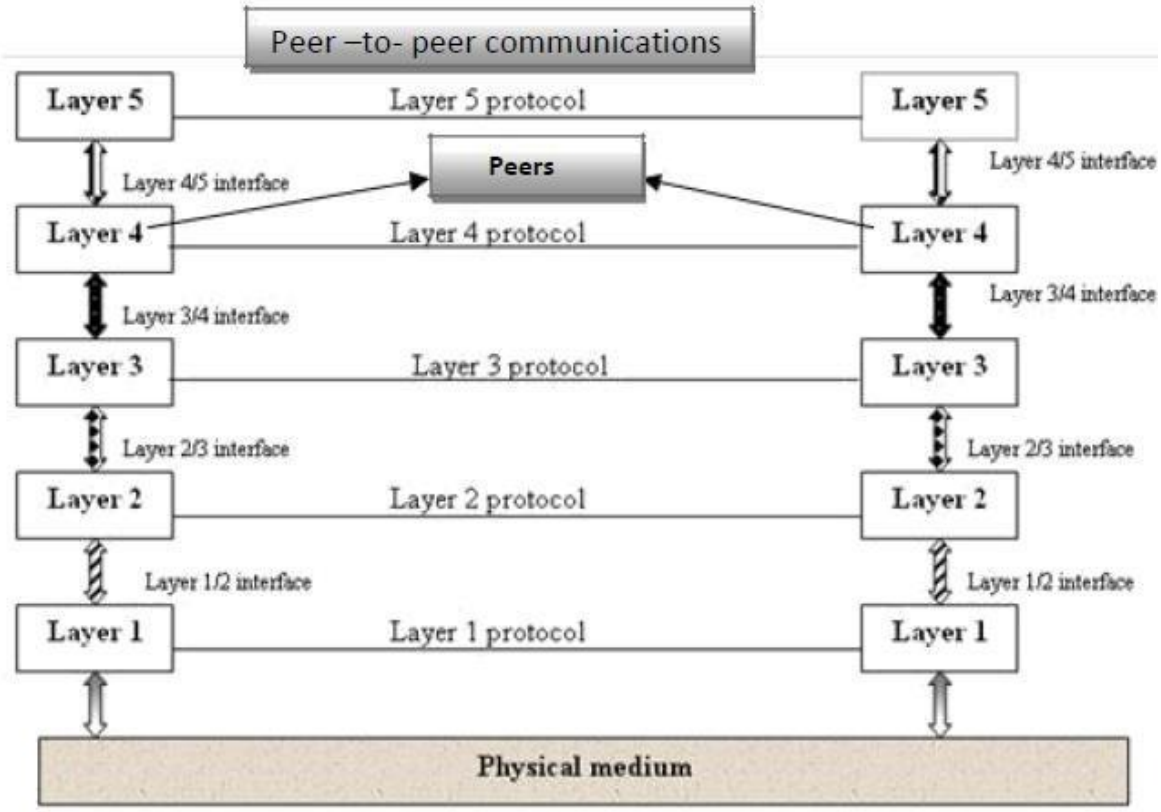
9. **Internet Research Task Force (IRTF)**
   The IRTF promotes research of importance to the evolution of the future Internet by creating focused, long-term and small research groups working on topics related to Internet protocols, applications, architecture and technology.

LAYERED ARCHITECTURE:

To reduce the design complexity, most of the networks are organized as a series oflayers or levels, each one build upon one below it. The basic idea of a layered architecture isto divide the design into small pieces. Each layer adds to the services provided by the lower layers in such a manner that the highest layer is provided a full set of services to manage communications and run the applications. The benefits of the layered models are modularity and clear interfaces, i.e. open architecture and comparability between the different providers' components. A basic principle is to ensure independence of layers by defining services provided by each layer to the next higher layer without defining how the services are to be performed. This permits changes in a layer without affecting other layers.

The basic elements of a layered model are **services, protocols** and **interfaces. A service** is a set of actions that a layer offers to another (higher) layer. **Protocol** is a set of rules that a layer uses to exchange information with a peer entity. These rules concern both the contents and the order of the messages used. Between the layers service interfaces are defined. The messages from one layer to another are sent through those interfaces.

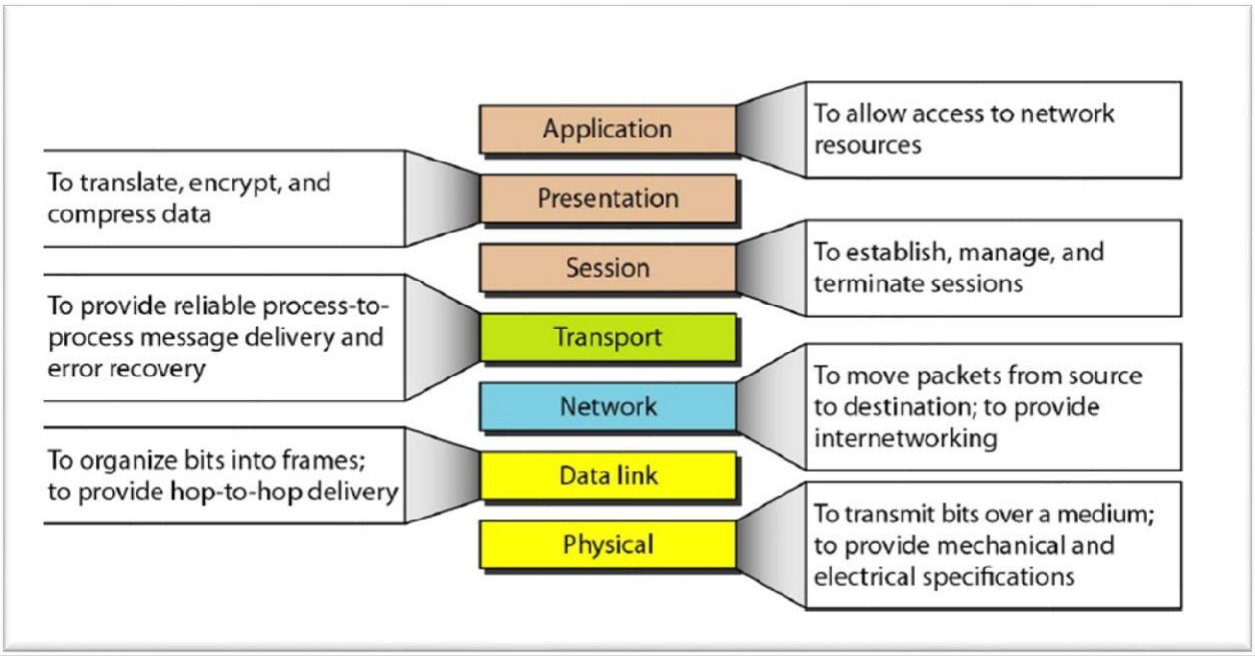
In a n-layer architecture, layer n on one machine carries on conversation with the layer n on other machine. The rules and conventions used in this conversation are collectively known as the layer-n protocol. Basically, a protocol is an agreement between the communicating parties on how communication is to proceed. Five-layer architecture is shown below; the entities comprising the corresponding layers on different machines are called peers. In other words, it is the peers that communicate using protocols. In reality, no data is transferred from layer n on one machine to layer n of another machine. Instead, each layer passes data and control information to the layer immediately below it, until the lowest layer is reached. Below layer-1 is the physical layer through which actual communication occurs.



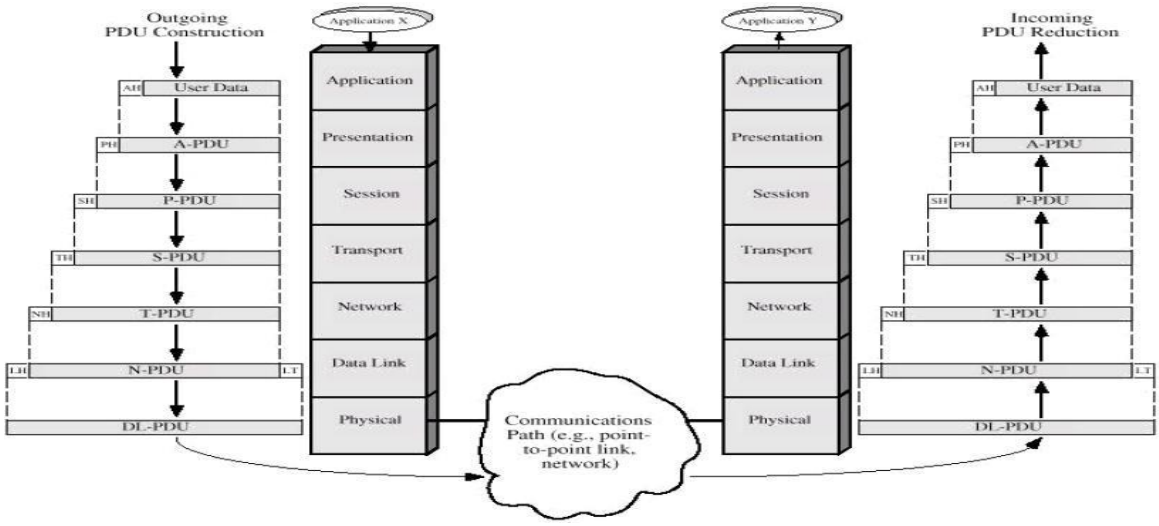
With layered architectures, communications between two corresponding layers requires a unit of data called a **protocol data unit (PDU**). A PDU can be a header added at the beginning of a message or a trailer appended to the end of a message. Data flows downward through the layers in the source system and upwards at the destination address. As data passes from one layer into another, headers and trailers are added and removed from the PDU. This process of adding or removing PDU information is called **encapsulation/decapsulation.** Between each pair of adjacent layers there is an **interface**. The interface defines which primitives operations and services the lower layer offers to the upper layer adjacent to it. A set of layers and protocols is known as **network architecture**. A list of protocols used by a certain system, one protocol per layer, is called **protocol stack**.

Open Systems Interconnection (OSI)

International standard organization (ISO) established a committee in 1977 to develop architecture for computer communication and the OSI model is the result of this effort. In 1984, the Open Systems Interconnection (OSI) reference model was approved as an international standard for communications architecture. The term "open" denotes the ability to connect any two systems which conform to the reference model and associated standards. The OSI model describes how information or data makes its way from application programmes (such as spreadsheets) through a network medium (such as wire) to another application programme located on another network. The OSI reference model divides the problem of moving information between computers over a network medium into SEVEN smaller and more manageable problems. The seven layers are:

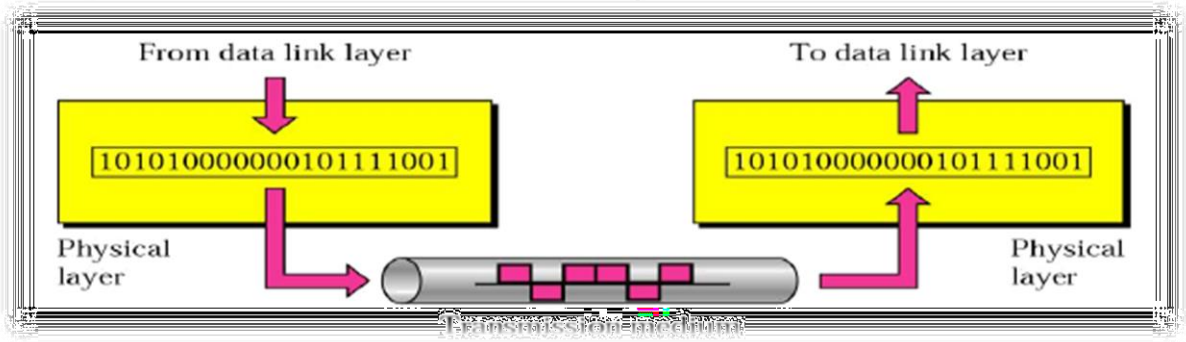


The lower 4 layers (transport, network, data link and physical —Layers 4, 3, 2, and 1) are concerned with the flow of data from end to end through the network. The upper four layers of the OSI model (application, presentation and session—Layers 7, 6 and 5) are orientated more toward services to the applications. Data is Encapsulated with the necessary protocol information as it moves down the layers before network transit.



Physical Layer :

The physical layer is the lowest layer of the OSI hierarchy and coordinates the functions required to transmit a bit stream over a physical medium. It also defines the procedures and functions that physical devices and interfaces have to perform for transmission occur. The physical layer specifies the type of transmission medium and the transmission mode (simplex, half duplex or full duplex) and the physical, electrical, functional and procedural standards for accessing data communication networks.

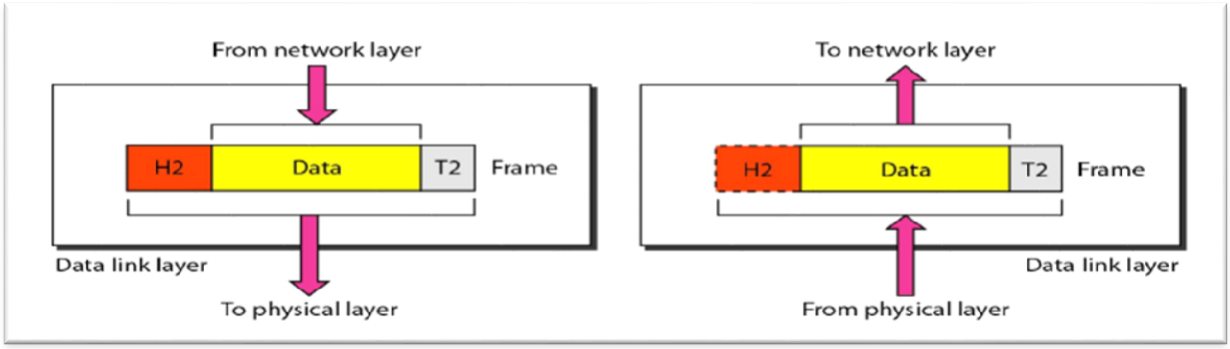


Transmission media defined by the physical layer include metallic cable, optical fiber cable or wireless radio-wave propagation. The physical layer also includes the carrier system used to propagate the data signals between points in the network. The carrier systems are simply communication systems that carry data through a system using either metallic or optical fiber cables or wireless arrangements such as microwave, satellites and cellular radio systems.

Data-link Layer:

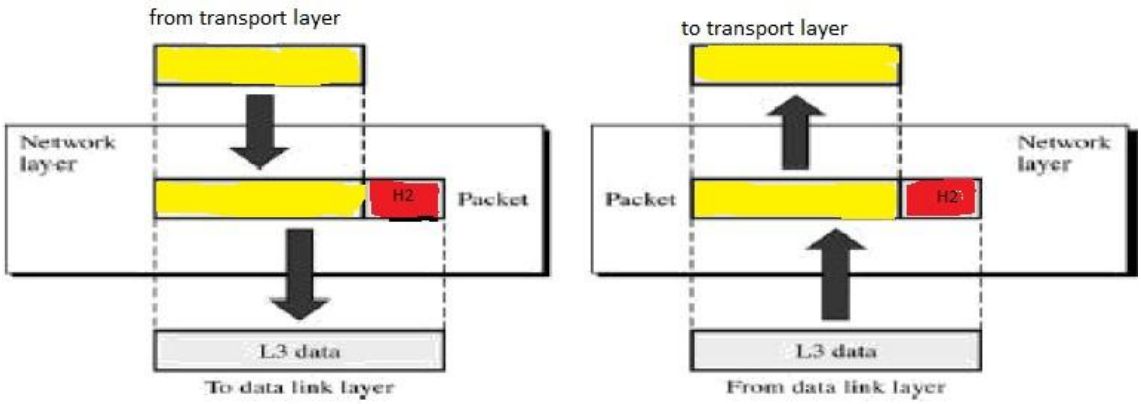
The data link layer transforms the physical layer, a raw transmission facility, to a reliable link and is responsible for node-to-node delivery. It makes the physical layer appear error free to the upper layer (network layer).

The data link layer packages data from the physical layer into groups called blocks, frames or packets. If frames are to be distributed to different systems on the network, the data link layer adds a header to the frame to define the physical address of the sender (source address) and/or receiver (destination address) of the frame. The data-link layer provides flow-control, access-control, and error-control.



Network Layer :

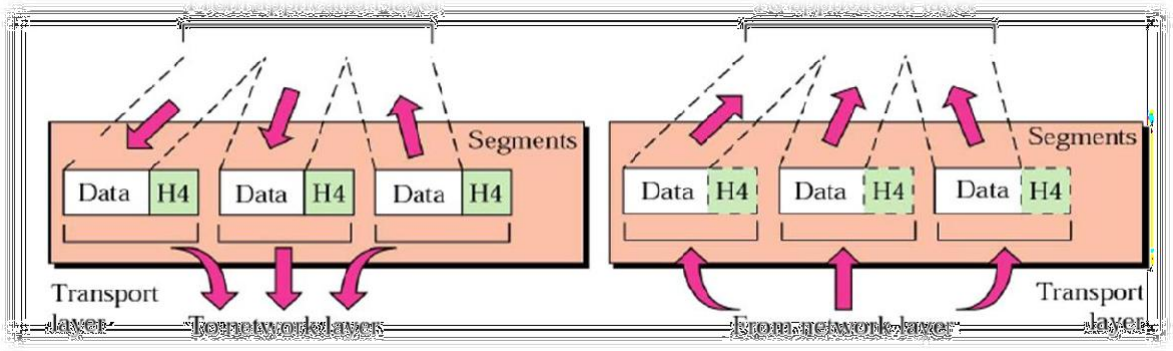
The network layer provides details that enable data to be routed between devices in an environment using multiple networks, subnetworks or both. This is responsible for addressing messages and data so they are sent to the correct destination, and for translating logical addresses and names (like a machine name FLAME) into physical addresses. This layer is also responsible for finding a path through the network to the destination computer



The network layer provides the upper layers of the hierarchy with independence from the data transmission and switching technologies used to interconnect systems. Networking components that operate at the network layer include routers and their software.

Transport Layer:

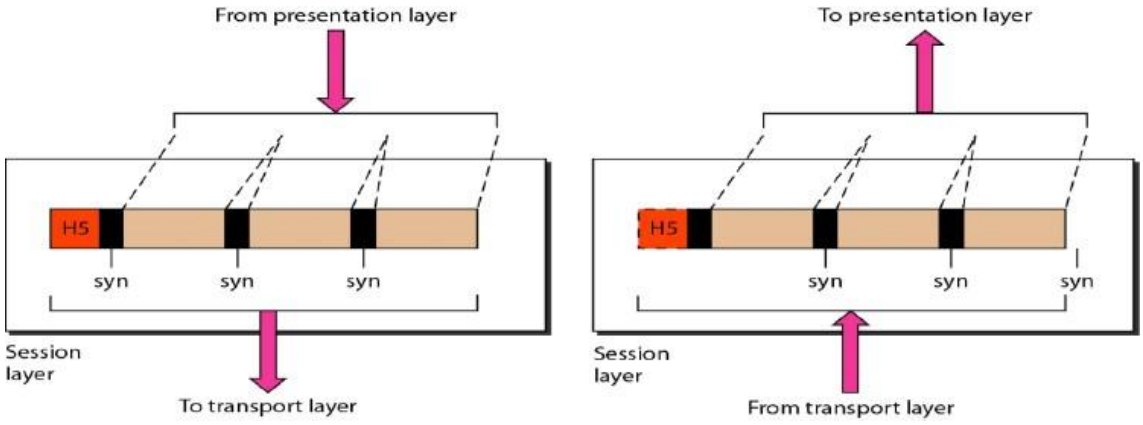
The transport layer controls and ensures the end-to-end integrity of the data message propagated through the network between two devices, providing the reliable, transparent transfer between two end points.



Transport layer responsibilites includes message routing, segmenting, error recovery and two types of basic services to an upper-layer protocol: connection oriented and connectionless. The transport layer is the highest layer in the OSI hierarchy in terms of communicatons and may provide data tracking, connection flow control, sequencing of data, error checking, and application addressing and identification.

Session Layer:

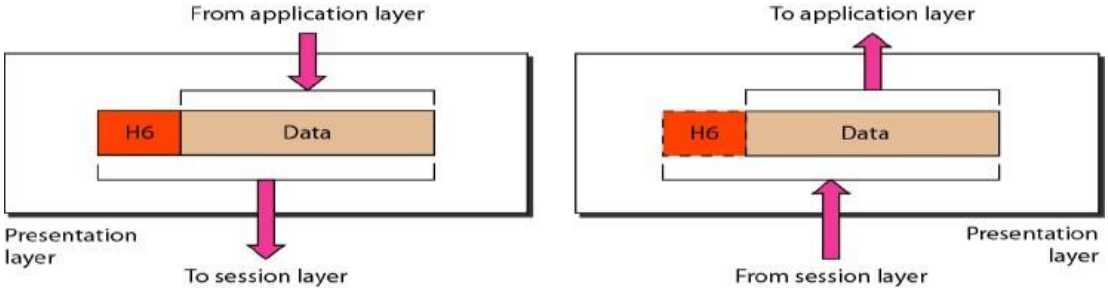
Session layer, some times called the dialog controller provides mechanism for controlling the dialogue between the two end systems. It defines how to start, control and end conversations (called sessions) between applications.



Session layer protocols provide the logical connection entities at the application layer. These applications include file transfer protocols and sending email. Session responsibilities include network log-on and log-off procedures and user authentication. Session layer characteristics include virtual connections between applications, entities, synchronization of data flow for recovery purposes, creation of dialogue units and activity units, connection parameter negotiation, and partitioning services into functional groups.

Presentation Layer:

The presentation layer provides independence to the application processes by addressing any code or syntax conversion necessary to present the data to the network in a common communications format. It specifies how end-user applications should format the data.



The presentation layer translated between different data formats and protocols. Presentation functions include data file formatting, encoding, encryption and decryption of data messages, dialogue procedures, data compression algorithms, synchronization, interruption, and termination.

Application Layer:

The application layer is the highest layer in the hierarchy and is analogous to the general manager of the network by providing access to the OSI environment. The applications layer provides distributed information services and controls the sequence of activities within and application and also the sequence of events between the computer application and the user of another application. The application layer communicates directly with the user's application program. User application processes require application layer service elements to access the networking environment. The service elements are of two types: CASEs (common application service elements) satisfying particular needs of application processes like association control, concurrence and recovery. The second type is SASE (specific application service elements) which include TCP/IP stack, FTP, SNMP, Telnet and SMTP.

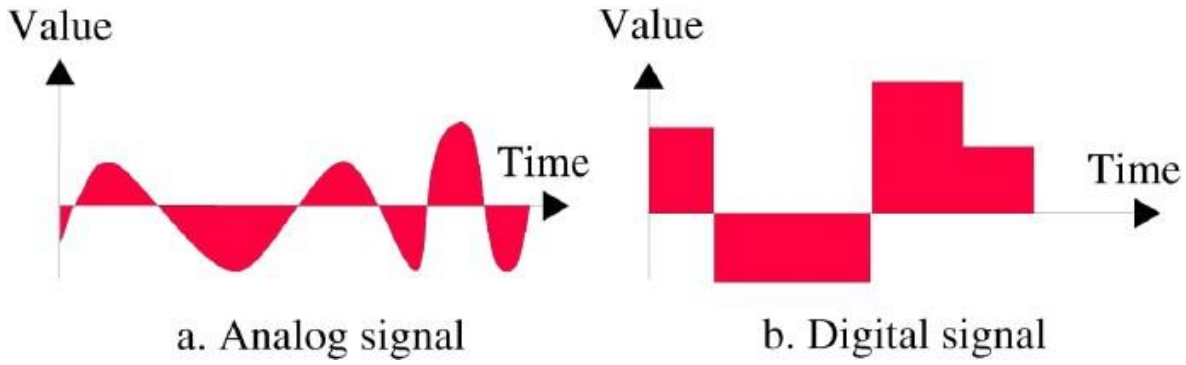
Chapter 2

ANALOG AND DIGITAL DATA:

Data can be analog or digital. The term analog data refers to information that is continuous; digital data refers to information that has discrete states. For example, an analog clock that has hour, minute, and second hands gives information in a continuous form; the movements of the hands are continuous. On the other hand, a digital clock that reports the hours and the minutes will change suddenly from 8:05 to 8:06. Analog data, such as the sounds made by a human voice, take on continuous values. When someone speaks, an analog wave is created in the air. This can be captured by a microphone and converted to an analog signal or sampled and converted to a digital signal. Digital data take on discrete values. For example, data are stored in computer memory in the form of 0s and 1s. They can be converted to a digital signal or modulated into an analog signal for transmission across a medium.

Analog and Digital Signals:

Like the data they represent, signals can be either analog or digital. An analog signal has infinitely many levels of intensity over a period of 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, can have only a limited number of defined values. Although each value can be any number, it is often as simple as 1 and 0. The simplest way to show signals is by plotting them on a pair of perpendicular axes. The vertical axis represents the value or strength of a signal. The horizontal axis represents time. The below Figure illustrates an analog signal and a digital signal. The curve representing the analog signal passes through an infinite number of points. The vertical lines of the digital signal, however, demonstrate the sudden jump that the signal makes from value to value.



a. Analog signal b. Digital signal

Periodic and Nonperiodic Signals:

A periodic signal completes a pattern within a measurable time frame, called a period, and repeats that pattern over subsequent identical periods. The completion of one full pattern is called a cycle. A nonperiodic signal changes without exhibiting a pattern or cycle that repeats over time. Both analog and digital signals can be periodic or nonperiodic. In data communications, we commonly use periodic analog signals and nonperiodic digital signals.

Signal Analysis:

Mathematical signal analysis is used to analyze and predict the performance of the circuit on the basis of the voltage distribution and frequency composition of the information signal.

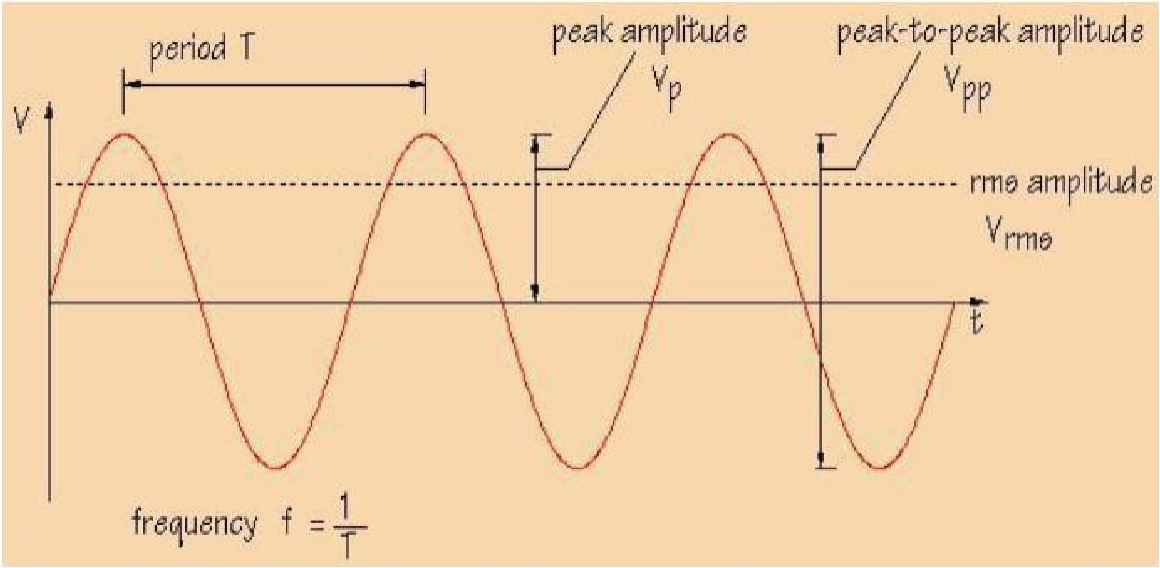
Amplitude, Frequency and Phase:

A **cycle** is one complete variation in the signal, and the **period** is the time the waveform takes to complete on **cycle**. One cycle constitutes 360 degrees (or 2π radians).

Sine waves can be described in terms of three parameters: amplitude, frequency and phase.

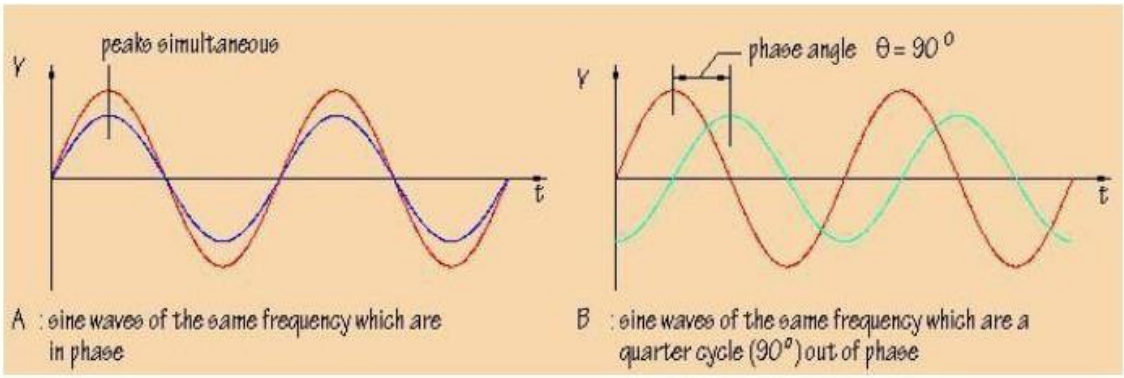
Amplitude (A):

It is analogous to magnitude or displacement. The amplitude of a signal is the magnitude of the signal at any point on the waveform. The amplitude of electrical signal is generally measured in voltage. The maximum voltage of a signal in respect to its average value is called its peak amplitude or peak voltage.



Frequency (f):

The time of one cycle of a waveform is its period, which is measured in seconds. Frequency is the number of cycles completed per second. The measurement unit for frequency is the hertz, Hz. 1 Hz = 1 cycle per second. The frequency of the signal can be calculated From T=1/f.



A phase shift of 360 degrees corresponds to a shift of one complete cycle. If two sine waves have the same frequency and occur at the same time, they are said to be in phase, or they are said to out of phase. The difference in phase can be measured in degrees, and is called the phase angle,Θ.

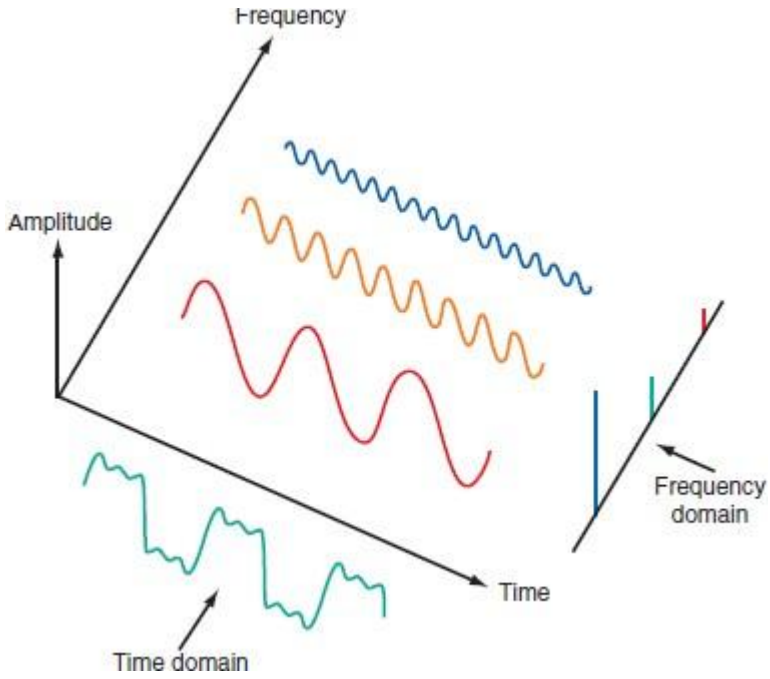
Time domain:

Time domain is a term used to describe the analysis of mathematical functions, or physical signals, with respect to time. In the time domain, the signal or function's value is known for all real numbers, for the case of continuous time, or at various separate instants in the case of discrete time. An oscilloscope is a time-domain tool commonly used to visualize real-world signals in the time domain. A time domain graph shows how a signal changes over time.

Frequency Domain:

Frequency domain is a term used to describe the analysis of mathematical functions or signals with respect to frequency, rather than time. A spectrum analyzer is a frequency domain instrument which displays amplitude-versus frequency plot (called a frequency

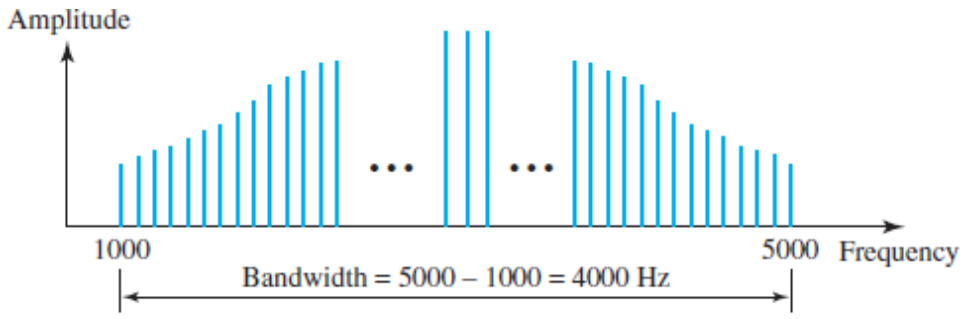
spectrum).The horizontal axis represents frequency and the vertical axis amplitude showing a vertical deflection for each frequency present in the waveform, which is proportional to the amplitude of the frequency it represents.



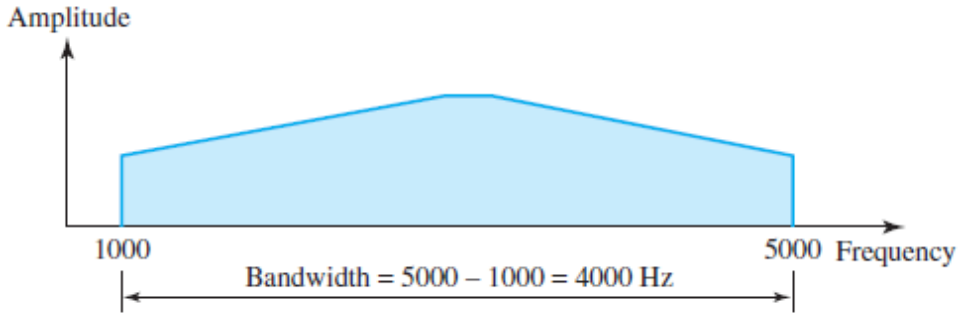
Frequency Spectrum and Bandwidth:

The frequency spectrum of a waveform consists of all the frequencies contained in the waveform and their respective amplitudes plotted in the frequency domain.

Bandwidth of an information signal is simply the difference between the highest and lowest frequencies contained in the information and the bandwidth of a communication channel is the difference between the highest and lowest frequencies that the channel will allow to pass through it.



a. Bandwidth of a periodic signal



b. Bandwidth of a nonperiodic signal

Electrical Noise and Signal-To-Noise Ratio

Noise is any disturbance or distortion that comes in the process of communication. Electrical noise is defined as any undesirable electrical energy that falls within the passband of the signal. A noise signal consists of a mixture of frequencies with random amplitudes. Noise can originate in various ways. The most prevalent and most interfering to data communication signals are man-made noise, thermal noise, correlated noise, and impulse noise.

Man-made noise: It is the kind of noise produced by mankind. The main sources are spark producing mechanisms like commutators in electric motors, automobile ignition systems, ac power-generating and switching equipment, and fluorescent lights. It is impulsive in nature and contains a wide range of frequencies propagated in the free space like the radio waves. Man-made noise is most intense in more densely populated areas and sometimes is referred to as industrial noise.

Thermal noise: This is the noise generated by thermal agitation of electrons in a conductor. It is also referred to as white noise because of its uniform distribution across the entire electromagnetic frequency spectrum. Noise power density is the thermal noise power present in a 1-Hz bandwidth and is given by No = KT.

Correlated noise: this noise is correlated to the signal and cannot be present in a circuit unless there is a signal. Correlated noise is produced by nonlinear amplification and includes harmonic distortion and inter modulation distortion. Harmonic distortion occurs when unwanted harmonics of a signal are produced through nonlinear amplification and is also called amplitude distortion. Inter modulation distortion is the generation of unwanted sum and difference frequencies produced when two or more signals are amplified in a nonlinear device.

Impulse noise: This noise is characterized by high-amplitude peaks of short duration in the total noise spectrum. It consists of sudden bursts of irregularly shaped pulses that generally last between a few microseconds and several milliseconds, depending on their amplitude and origin. In case of voice communications, impulse noise is very annoying as it generates a sharp popping or crackling sound where as it is devastating in data circuits.

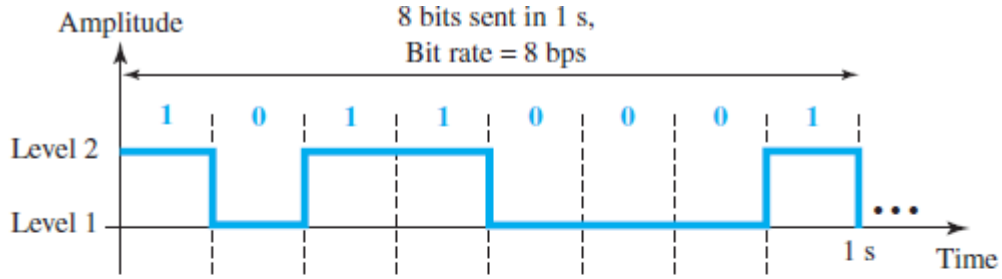
Signal-to-noise power ratio: Signal-to-noise ratio (often abbreviated SNR or S/N) is defined as the ratio of signal power to the noise power corrupting the signal. A ratio higher than 1:1 indicates more signal than noise. Signal-to-noise ratio is defined as the power ratio between signal (meaningful information) and the background noise (unwanted signal)

$$SNR = \frac{P_{SIGNAL}}{P_{NOISE}}$$

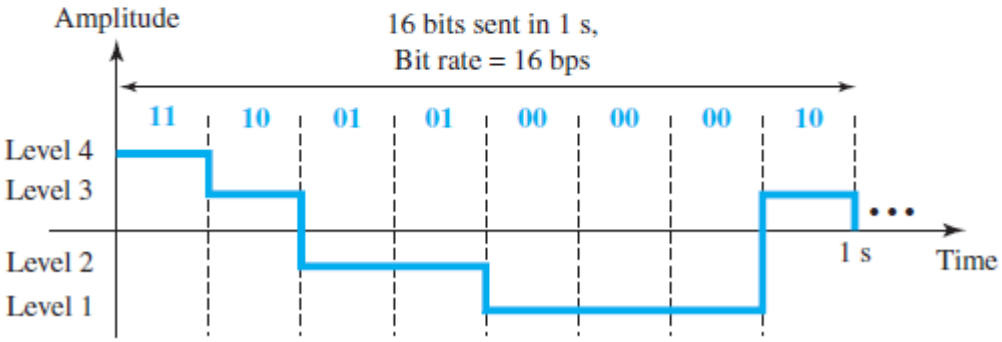
where P is average power in watts. The ratio often expressed in decibels as S/N (dBm) = 10 log(PS/PN).

DIGITAL SIGNALS:

In addition to being represented by an analog signal, information can also be represented by a digital signal. For example, a 1 can be encoded as a positive voltage and a 0 as zero voltage. A digital signal can have more than two levels. In this case, we can send more than 1 bit for each level.



a. A digital signal with two levels



b. A digital signal with four levels

The above Figure shows two signals, one with two levels and the other with four. We send 1 bit per level in part a of the figure and 2 bits per level in part b of the figure. In general, if a signal has L levels, each level needs log2 L bits. For this reason, we can send log24 = 2 bits in part b.

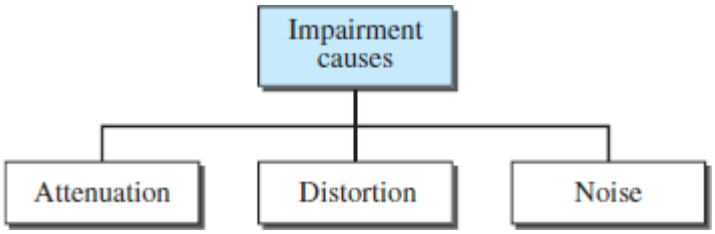
Analog Versus Digital Transmission

Analog Transmission: An analog wave form (or signal) is characterized by being continuously variable along amplitude and frequency. In the case of telephony, for instance, when you speak into a handset, there are changes in the air pressure around your mouth. Those changes in air pressure fall onto the handset, where they are amplified and then converted into current, or voltage fluctuations. Those fluctuations in current are an analog of the actual voice pattern hence the use of the term analog to describe these signals.

Digital transmission: it is quite different from analog transmission. For one thing, the signal is much simpler. Rather than being a continuously variable wave form, it is a series of discrete pulses, representing one bits and zero bits . Each computer uses a coding scheme that defines what combinations of ones and zeros constitute all the characters in a character set.

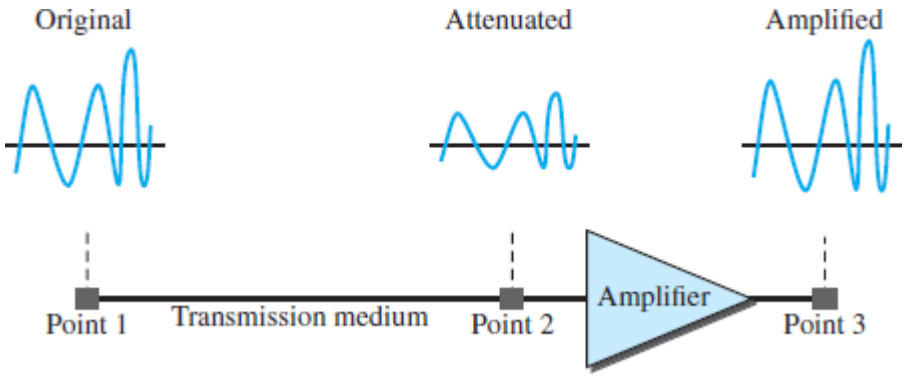
Transmission Impairment

Signals travel through transmission media, which are not perfect. The imperfection causes signal impairment. This means that the signal at the beginning of the medium is not the same as the signal at the end of the medium. What is sent is not what is received causes of impairment are attenuation, distortion, and noise



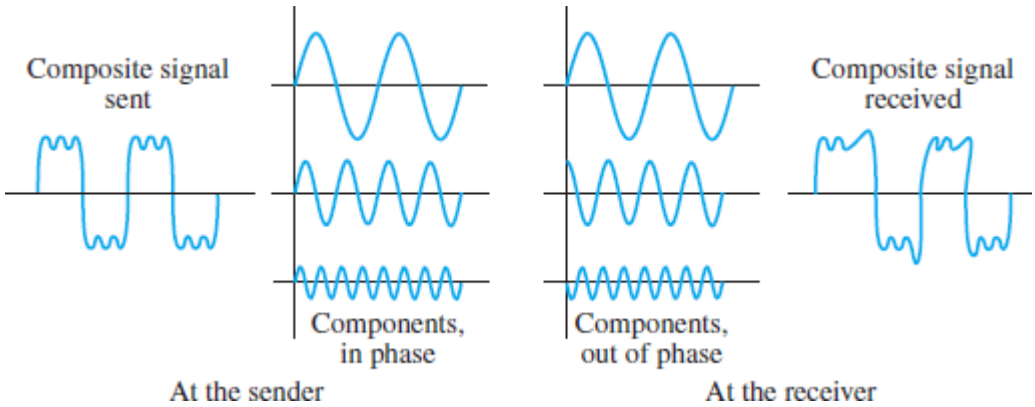
Attenuation:

Attenuation means a loss of energy. When a signal, simple or composite, travels through a medium, it loses some of its energy in overcoming the resistance of the medium. That is why a wire carrying electric signals gets warm, if not hot, after a while. Some of the electrical energy in the signal is converted to heat. To compensate for this loss, amplifiers are used to amplify the signal. The below Figure shows the effect of attenuation and amplification.



Distortion:

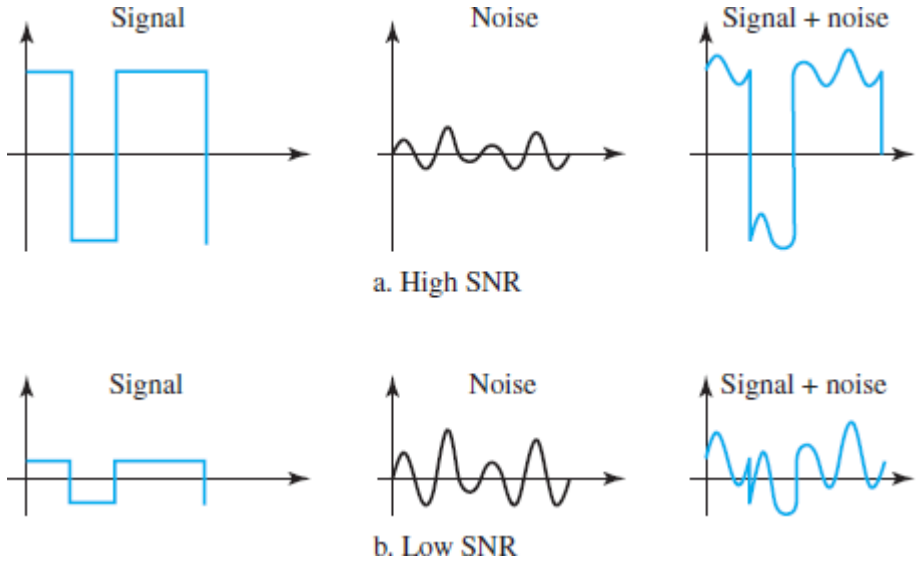
Distortion means that the signal changes its form or shape. Distortion can occur in a composite signal made of different frequencies. Each signal component has its own propagation speed (see the next section) through a medium and, therefore, its own delay in arriving at the final destination. Differences in delay may create a difference in phase if the delay is not exactly the same as the period duration. In other words, signal components at the receiver have phases different from what they had at the sender. The shape of the composite signal is therefore not the same. Figure below shows the effect of distortion on a composite signal.



Noise:

Noise is another cause of impairment. Several types of noise, such as thermal noise, induced noise, crosstalk, and impulse noise, may corrupt the signal. Thermal noise is the random motion of electrons in a wire, which creates an extra signal not originally sent by the transmitter. Induced noise comes from sources such as motors and appliances. These devices

act as a sending antenna, and the transmission medium acts as the receiving antenna. Crosstalk is the effect of one wire on the other. One wire acts as a sending antenna and the other as the receiving antenna. Impulse noise is a spike (a signal with high energy in a very short time) that comes from power lines, lightning, and so on. The below Figure shows the effect of noise on a signal



DATA RATE LIMITS:

A very important consideration in data communications is how fast we can send data, in bits per second, over a channel. Data rate depends on three factors:

1. The bandwidth available
2. The level of the signals we use
3. The quality of the channel (the level of noise).

Noiseless Channel: Nyquist Bit Rate

For a noiseless channel, the Nyquist bit rate formula defines the theoretical maximum bit rate is defined as

$$\textbf{BITRATE} = 2 \times \textbf{bandwidth} \times \log_2 \textbf{L}$$

In this formula, bandwidth is the bandwidth of the channel, L is the number of signal levels used to represent data, and BitRate is the bit rate in bits per second.

Noisy Channel: Shannon Capacity:

In reality, we cannot have a noiseless channel; the channel is always noisy. In 1944, Claude Shannon introduced a formula, called the Shannon capacity, to determine the theoretical highest data rate for a noisy channel:

$$\boldsymbol{CAPACITY} = \boldsymbol{bandwidth} \times \boldsymbol{log}(1 + \boldsymbol{SNR})$$ **Transmission**

Media:

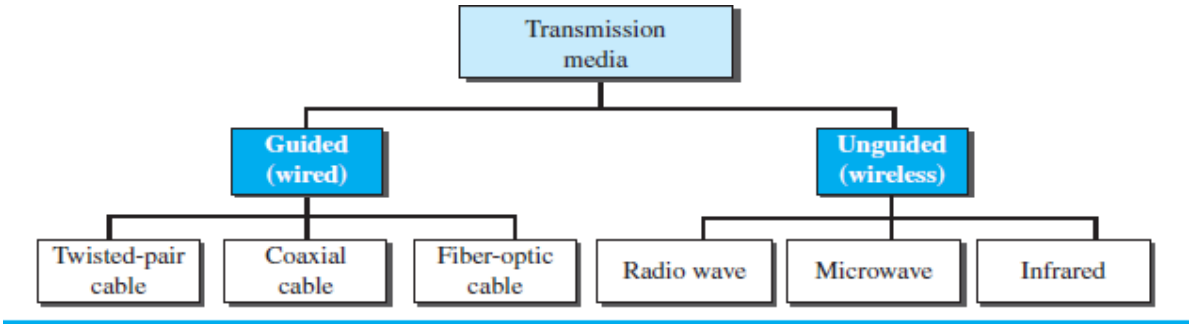
Transmission media are actually located below the physical layer and are directly controlled by the physical layer. A transmission medium can be broadly defined as anything that can carry information from a source to a destination

In data communications the definition of the information and the transmission medium is more specific. The transmission medium is usually free space, metallic cable, or fiber-optic cable. The information is usually a signal that is the result of a conversion of data from another form.

The use of long-distance communication using electric signals started with the invention of the telegraph by Morse in the 19th century. Communication by telegraph was slow and dependent on a metallic medium.

Extending the range of the human voice became possible when the telephone was invented in 1869. Telephone communication at that time also needed a metallic medium to carry the electric signals that were the result of a conversion from the human voice. The communication was, however, unreliable due to the poor quality of the wires. The lines were often noisy and the technology was unsophisticated. Wireless communication started in 1895 when Hertz was able to send high frequency signals. Later, Marconi devised a method to send telegraph-type messages over the Atlantic Ocean. Better metallic media have been invented (twisted-pair and coaxial cables, for example). The use of optical fibers has increased the data rate incredibly. Electromagnetic energy, a combination of electric and magnetic fields vibrating in relation to each other, includes power, radio waves, infrared light, visible light, ultraviolet light, and X, gamma, and cosmic rays. Each of these constitutes a portion of the electromagnetic spectrum. Not all portions of the spectrum are currently usable for telecommunications, however. The media to harness those that are usable are also limited

In telecommunications, transmission media can be divided into two broad categories: guided and unguided. Guided media include twisted-pair cable, coaxial cable, and fiber-optic cable. Unguided medium is free space.



Guided Media:

Guided media, which are those that provide a conduit from one device to another, include twisted-pair cable, coaxial cable, and fiber-optic cable. A signal traveling along any of these media is directed and contained by the physical limits of the medium. Twisted-pair and coaxial cable use metallic (copper) conductors that accept and transport signals in the form of electric current. Optical fiber is a cable that accepts and transports signals in the form of light.

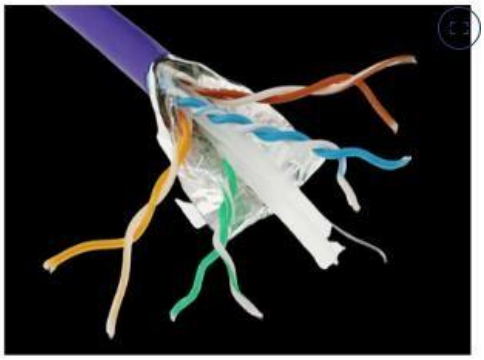
Twisted-Pair Cable:

A twisted pair consists of two conductors (normally copper), each with its own plastic insulation, twisted together. One of the wires is used to carry signals to the receiver, and the other is used only as a ground reference. The receiver uses the difference between the two
In addition to the signal sent by the sender on one of the wires, interference (noise) and crosstalk may affect both wires and create unwanted signals. If the two wires are parallel, the effect of these unwanted signals is not the same in both wires because they are at different

locations relative to the noise or crosstalk sources This results in a difference at the receiver. By twisting the pairs, a balance is maintained. There are two type of twisted pair cables

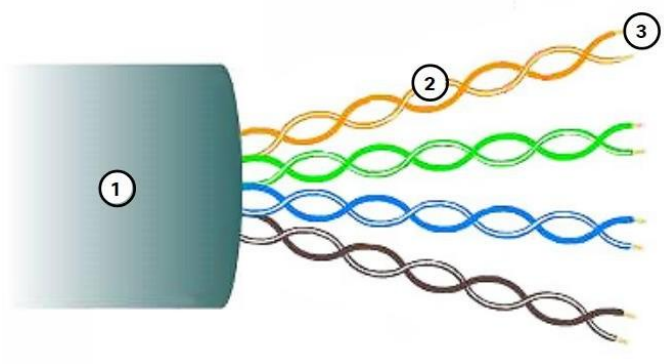
1. Unshielded twisted pair cable
2. Shielded twisted pair cable

Unshielded Twisted-Pair (UTP) Cable



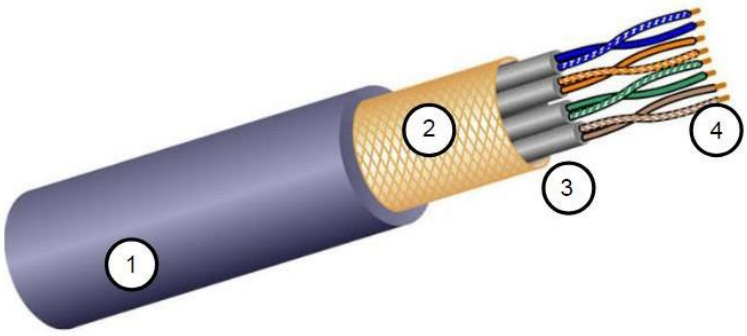
Shielded Twisted-Pair (STP) Cable

The most common twisted-pair cable used in communications is referred to as unshielded twisted-pair (UTP). IBM has also produced a version of twisted-pair cable for its use, called shielded twisted-pair (STP). STP cable has a metal foil or braided mesh covering that encases each pair of insulated conductors. Although metal casing improves the quality of cable by preventing the penetration of noise or crosstalk, it is bulkier and more expensive.



The numbers in the figure identify some of the characteristics of UTP cable

1. The outer jacket protects the copper wires from physical damage.
2. Twisted-pairs protect the signal from interference.
3. Color-coded plastic insulation electrically isolates wires from each other and identifies each pair.



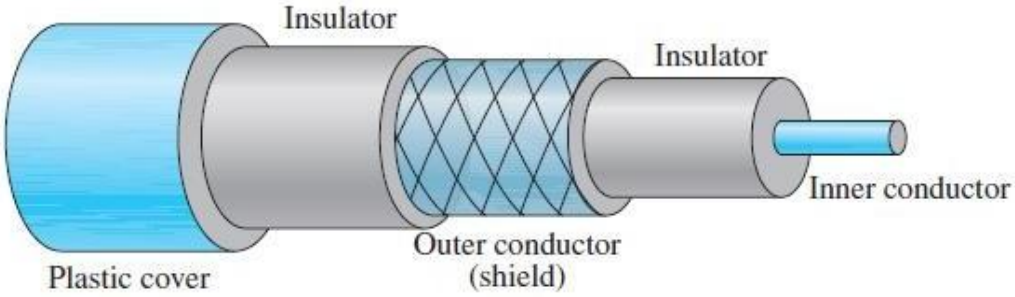
The numbers in the figure identify some key features of shielded twisted pair cable

1. Outer jacket
2. Braided or foil shield
3. Foil shields
4. Twisted pairs

UTP cable is usually terminated with an RJ-45 connector.

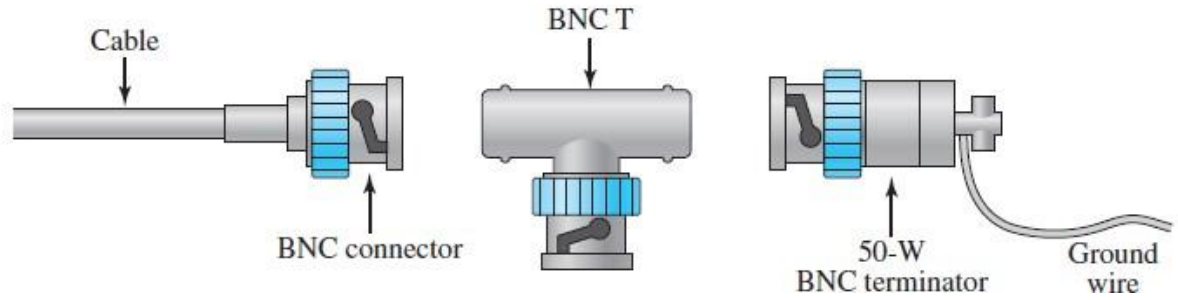
Coaxial Cable:

Coaxial cable (or coax) carries signals of higher frequency ranges than those in twisted pair cable, in part because the two media are constructed quite differently. Instead of having two wires, coax has a central core conductor of solid or stranded wire (usually copper) enclosed in an insulating sheath, which is, in turn, encased in an outer conductor of metal foil, braid, or a combination of the two. The outer metallic wrapping serves both as a shield against noise and as the second conductor, which completes the circuit.
This outer conductor is also enclosed in an insulating sheath, and the whole cable is protected by a plastic cover.



Coaxial Cable Connectors:

To connect coaxial cable to devices, we need coaxial connectors. The most common type of connector used today is the Bayonet Neill-Concelman (BNC) connector.

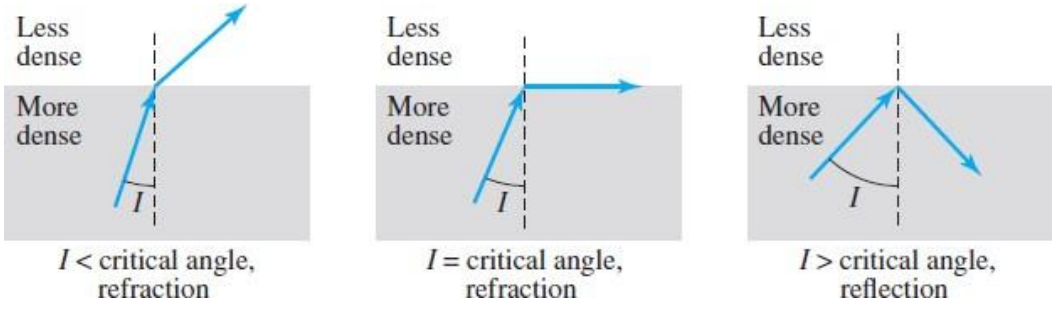


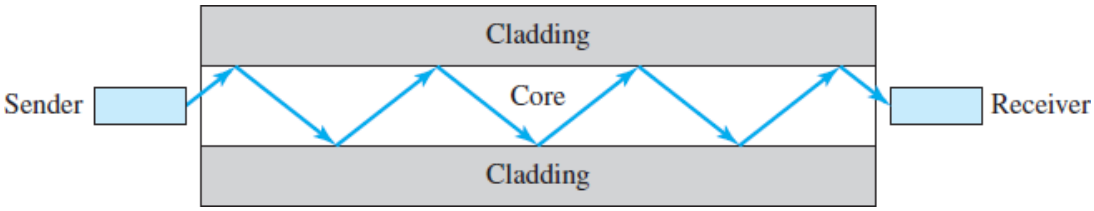
The BNC connector is used to connect the end of the cable to a device, such as a TV set. The BNC T connector is used in Ethernet networks to branch out to a connection to a computer or other device. The BNC terminator is used at the end of the cable to prevent the reflection of the signal.

Fiber-Optic Cable:

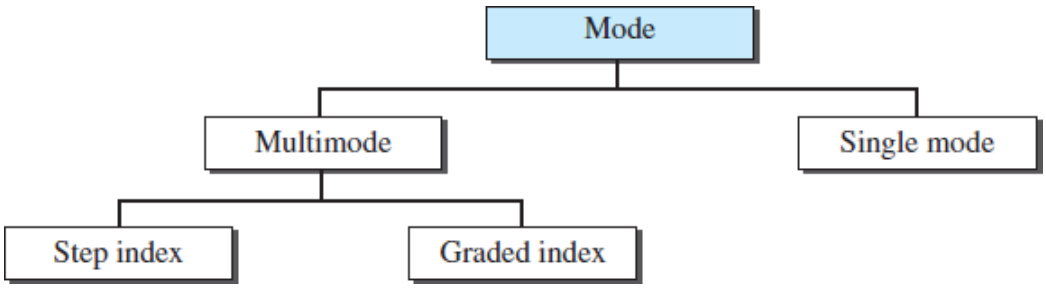
A fiber-optic cable is made of glass or plastic and transmits signals in the form of light. Light travels in a straight line as long as it is moving through a single uniform substance. If a ray of light traveling through one substance suddenly enters another substance (of a different density), the ray changes direction. if the angle of incidence I (the angle the ray makes with the line perpendicular to the interface between the two substances) is less than the critical angle, the ray refracts and moves closer to the surface. If the angle of incidence is equal to the critical angle, the light bends along the interface. If the angle is greater than the critical angle, the ray reflects (makes a turn) and travels again in the denser substance. The critical angle is a property of the substance, and its value differs from one substance to another.

Optical fibers use reflection to guide light through a channel. A glass or plastic core is surrounded by a cladding of less dense glass or plastic. The difference in density of the two materials must be such that a beam of light moving through the core is reflected off the cladding instead of being refracted into it.





Propagation mode of optical fibre:



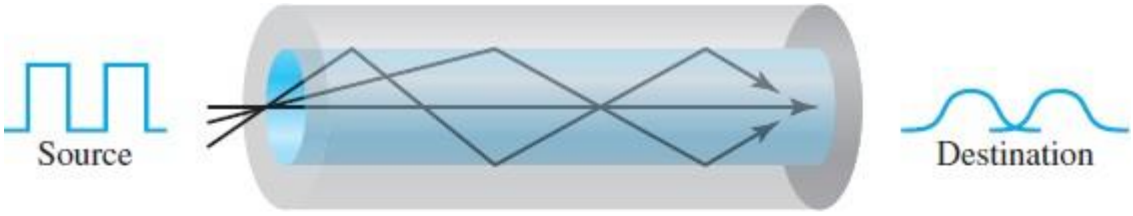
Multimode:

Multimode is so named because multiple beams from a light source move through the core in different paths In **multimode step-index fiber**, the density of the core remains constant from the center to the edges. A beam of light moves through this constant density in a straight line until it reaches the interface of the core and the cladding. At the interface, there is an abrupt change due to a lower density; this alters the angle of the beam's motion. The term step-index refers to the suddenness of this change, which contributes to the distortion of the signal as it passes through the fiber.

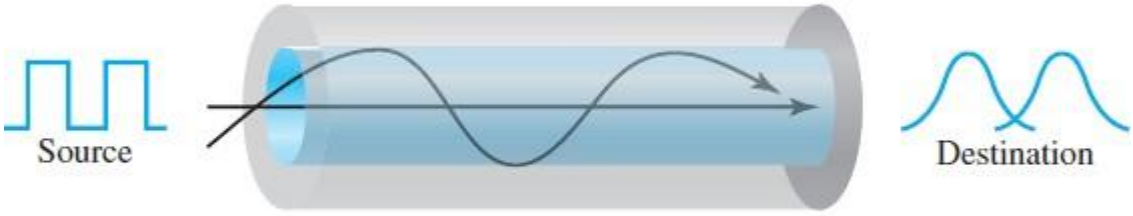
A second type of fiber, called **multimode graded-index fiber**, decreases this distortion of the signal through the cable. The word index here refers to the index of refraction. As we saw above, the index of refraction is related to density. A graded index fiber, therefore, is one with varying densities. Density is highest at the center of the core and decreases gradually to its lowest at the edge.

Single mode :

Single-mode uses step-index fiber and a highly focused source of light that limits beams to a small range of angles, all close to the horizontal. The single-mode fiber itself is manufactured with a much smaller diameter than that of multimode fiber, and with substantially lower density (index of refraction). The decrease in density results in a critical angle that is close enough to 90° to make the propagation of beams almost horizontal. In this case, propagation of different beams is almost identical, and delays are negligible. All the beams arrive at the destination "together" and can be recombined with little distortion to the signal.



a. Multimode, step index



b. Multimode, graded index

c. Single mode

Fiber optic connectors:

Straight tip connectors(ST):

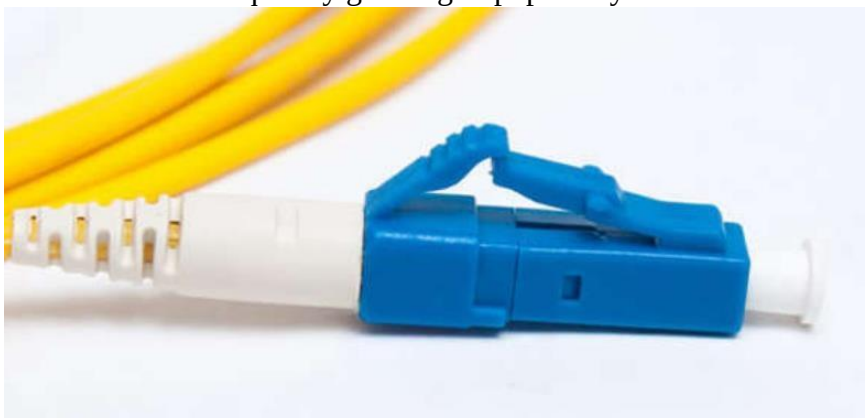
ST connectors were one of the first connector types used. The connector locks securely with a "Twist-on/twist-off" bayonet-style mechanism.

Subscriber channel (SC) connector:

SC connectors are sometimes referred to as square connector or standard connector. They are a widely-adopted LAN and WAN connector that uses a push-pull mechanism to ensure positive insertion. This connector type is used with multimode and single-mode fiber.

Lucent simplex connector (LC)

LC simplex connectors are a smaller version of the SC connector. These are sometimes called little or local connectors and are quickly growing in popularity due to their smaller size.



Duplex multimode LC connector

A duplex multimode LC connector is similar to a LC simplex connector, but uses a duplex connector.

Advantages and Disadvantages of Optical Fiber:

Advantages:

Fiber-optic cable has several advantages over metallic cable (twisted-pair or coaxial).

1. **Higher bandwidth.** Fiber-optic cable can support dramatically higher bandwidths (and hence data rates) than either twisted-pair or coaxial cable. Currently, data rates and bandwidth utilization over fiber-optic cable are limited not by the medium but by the signal generation and reception technology available.
2. **Less signal attenuation**. Fiber-optic transmission distance is significantly greater than that of other guided media. A signal can run for 50 km without requiring regeneration. We need repeaters every 5 km for coaxial or twisted-pair cable.
3. **Immunity to electromagnetic interference.** Electromagnetic noise cannot affect fiber-optic cables.
4. **Resistance to corrosive materials**. Glass is more resistant to corrosive materials than copper.
5. **Light weight.** Fiber-optic cables are much lighter than copper cables.
6. **Greater immunity to tapping.** Fiber-optic cables are more immune to tapping than copper cables. Copper cables create antenna effects that can easily be tapped.

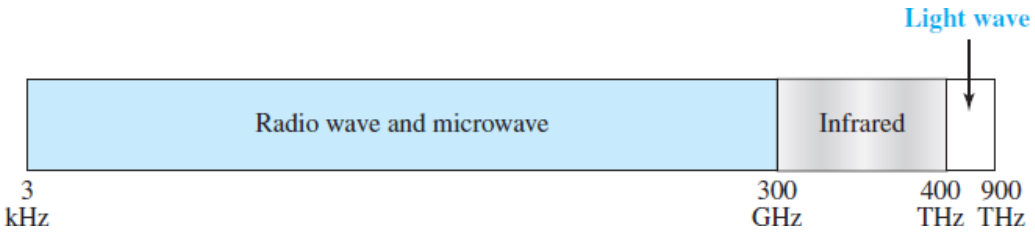
Disadvantages

There are some disadvantages in the use of optical fiber.

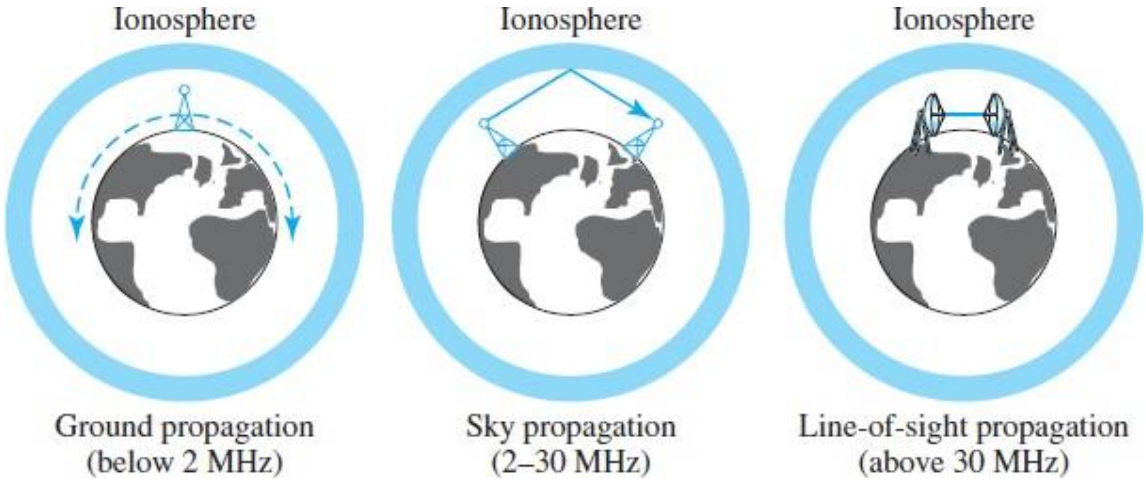
1. **Installation and maintenance**. Fiber-optic cable is a relatively new technology. Its installation and maintenance require expertise that is not yet available everywhere.
2. **Unidirectional light propagation**. Propagation of light is unidirectional. If we need bidirectional communication, two fibers are needed.
3. **Cost**. The cable and the interfaces are relatively more expensive than those of other guided media. If the demand for bandwidth is not high, often the use of optical fiber cannot be justified.

Unguided Media(wireless):

Unguided medium transport electromagnetic waves without using a physical conductor. This type of communication is often referred to as wireless communication. Signals are normally broadcast through free space and thus are available to anyone who has a device capable of receiving them. The electromagnetic spectrum, ranging from 3 kHz to 900 THz, used for wireless communication.



Unguided signals can travel from the source to the destination in several ways: ground propagation, sky propagation, and line-of-sight propagation In ground propagation, radio waves travel through the lowest portion of the atmosphere, hugging the earth. These low-frequency signals emanate in all directions from the transmitting antenna and follow the curvature of the planet. Distance depends on the amount of power in the signal: The greater the power, the greater the distance. In sky propagation, higher-frequency radio waves radiate upward into the ionosphere (the layer of atmosphere where particles exist as ions) where they are reflected back to earth. This type of transmission allows for greater distances with lower output power. In line-of-sight propagation, very high-frequency signals are transmitted in straight lines directly from antenna to antenna. Antennas must be directional, facing each other, and either tall enough or close enough together not to be affected by the curvature of the earth. Line-ofsight propagation is tricky because radio transmissions cannot be completely focused.



BAND	RANGE	PROPAGATION	APPLICATION
very low frequency (VLF)	3–30 kHz	Ground	Long-range radio navigation
low frequency (LF)	30–300 kHz	Ground	Radio beacons and navigational locators
middle frequency (MF)	300 kHz–3 MHz	SKY	AM radio
high frequency (HF)	3–30 MHz	SKY	Citizens band (CB), ship/aircraft
very high frequency (VHF)	30–300 MHz	SKY & Line of sight	VHF TV, FM radio
ultrahigh frequency (UHF)	300 MHz–3 GHz	Line-of-sight	UHF TV, cellular phones, paging, satellite
superhigh frequency (SF)	3–30 GHz	Line-of-sight	Satellite
extremely high frequency (EHF)	30–300 GHz	Line-of-sight	Satellite

Radio Waves:

Electromagnetic waves ranging in frequencies between 3 kHz and 1 GHz are normally called radio waves; waves ranging in frequencies between 1 and 300 GHz are called microwaves. Radio waves, for the most part, are omni-directional. When an antenna transmits radio waves, they are propagated in all directions. This means that the sending and receiving antennas do not have to be aligned. A sending antenna sends waves that can be received by any receiving antenna. The omni-directional property has a disadvantage, too. The radio waves transmitted by one antenna are susceptible to interference by another antenna that may send signals using the same frequency or band. Radio waves, particularly those waves that propagate in the sky mode, can travel long distances. This makes radio waves a good candidate for long-distance broadcasting such as AM radio. Radio waves, particularly those of low and medium frequencies, can penetrate walls. This characteristic can be both an advantage and a disadvantage. It is an advantage because, for example, an AM radio can receive signals inside a building. It is a disadvantage because we cannot isolate a communication to just inside or outside a building. The radio wave band is relatively narrow, just under 1 GHz, compared to the microwave band. When this band is divided into sub bands, the sub bands are also narrow, leading to a low data rate for digital communications.

NOTE:Radio waves are used for multicast communications, such as radio and television, and paging systems.

Microwaves:

Electromagnetic waves having frequencies between 1 and 300 GHz are called microwaves. Microwaves are unidirectional. When an antenna transmits microwaves, they can be narrowly focused. This means that the sending and receiving antennas need to be aligned. The unidirectional property has an obvious advantage. A pair of antennas can be aligned without interfering with another pair of aligned antennas. The following describes some characteristics of microwave propagation:

1. Microwave propagation is line-of-sight. Since the towers with the mounted antennas need to be in direct sight of each other, towers that are far apart need to be very tall.
The curvature of the earth as well as other blocking obstacles do not allow two short towers to communicate by using microwaves. Repeaters are often needed for long distance communication.
2. Very high-frequency microwaves cannot penetrate walls. This characteristic can be a disadvantage if receivers are inside buildings.
3. The microwave band is relatively wide, almost 299 GHz. Therefore wider sub bands can be assigned, and a high data rate is possible.
4. Use of certain portions of the band requires permission from authorities.

Microwaves need unidirectional antennas that send out signals in one direction.

NOTE: Microwaves are used for unicast communication such as cellular telephones, satellite networks, and wireless LANs.

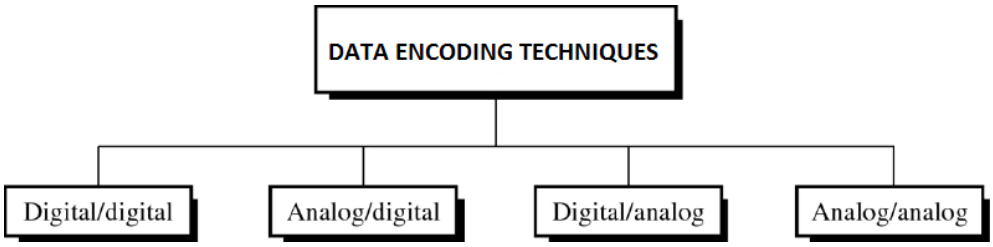
Infrared :

Infrared waves, with frequencies from 300 GHz to 400 THz (wavelengths from 1 mm to 770 nm), can be used for short-range communication. Infrared waves, having high frequencies, cannot penetrate walls. This advantageous characteristic prevents interference between one system and another; a short-range communication system in one room cannot be affected by another system in the next room.

Chapter-3

DATA ENCODING:

Data encoding is of 4 types

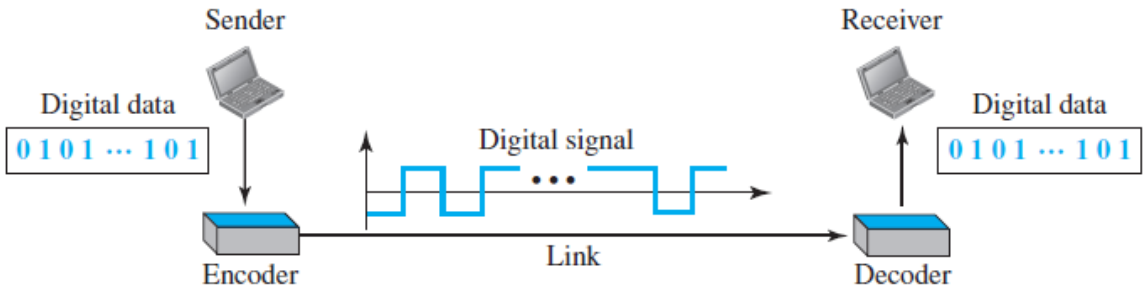


DIGITAL-TO-DIGITAL CONVERSION:

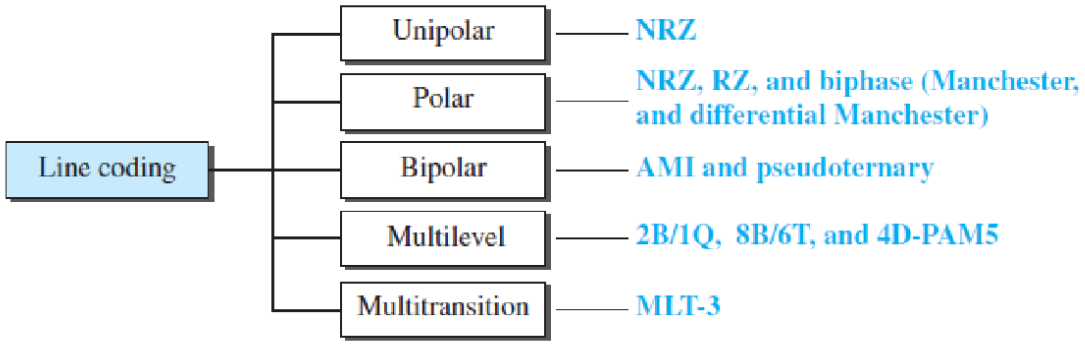
The conversion involves three techniques: line coding, block coding, and scrambling. Line coding is always needed; block coding and scrambling may or may not be needed.

LINE CODING:

Line coding is the process of converting digital data to digital signals. We assume that data, in the form of text, numbers, graphical images, audio, or video, are stored in computer memory as sequences of bits. Line coding converts a sequence of bits to a digital signal. At the sender, digital data are encoded into a digital signal; at the receiver, the digital data are recreated by decoding the digital signal.



Line Coding Schemes:

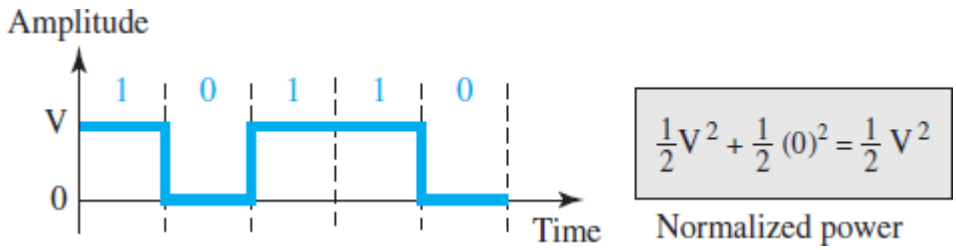


Unipolar Scheme:

In a unipolar scheme, all the signal levels are on one side of the time axis, either above or below.

NRZ (Non-Return-to-Zero):

Traditionally, a unipolar scheme was designed as a non-return-to-zero (NRZ) scheme in which the positive voltage defines bit 1 and the zero voltage defines bit 0. It is called NRZ because the signal does not return to zero at the middle of the bit. Figure shows a unipolar NRZ scheme.



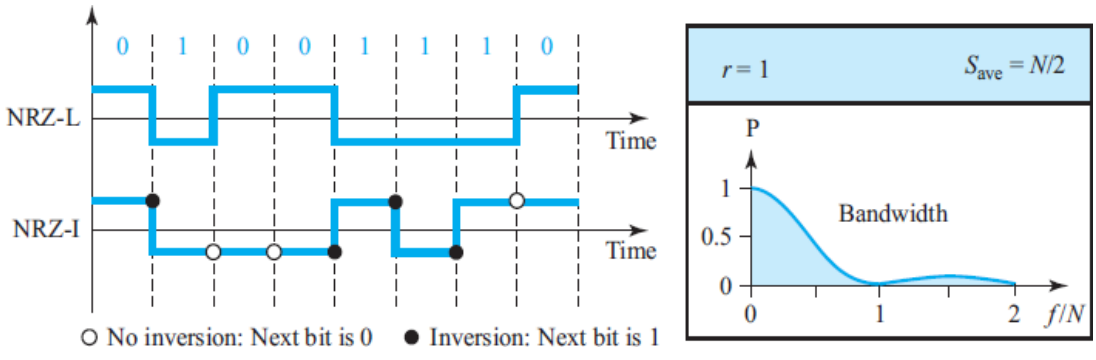
Polar Schemes:

In polar schemes, the voltages are on both sides of the time axis. For example, the voltage level for 0 can be positive and the voltage level for 1 can be negative.

Polar Non-Return-to-Zero (NRZ):

In polar NRZ encoding, we use two levels of voltage amplitude. We can have two versions of polar NRZ: NRZ-L and NRZ-I, as shown in the figure . The figure also shows the value of r,
the average baud rate, and the bandwidth. In the first variation, NRZ-L (NRZ-Level), the level of the voltage determines the value of the bit. In the second variation, NRZ-I (NRZ-Invert), the change or lack of change in the level of the voltage determines the value of the bit. If there is no change, the bit is 0; if there is a change, the bit is 1.

Polar NRZ-L AND Polar NRZ-I

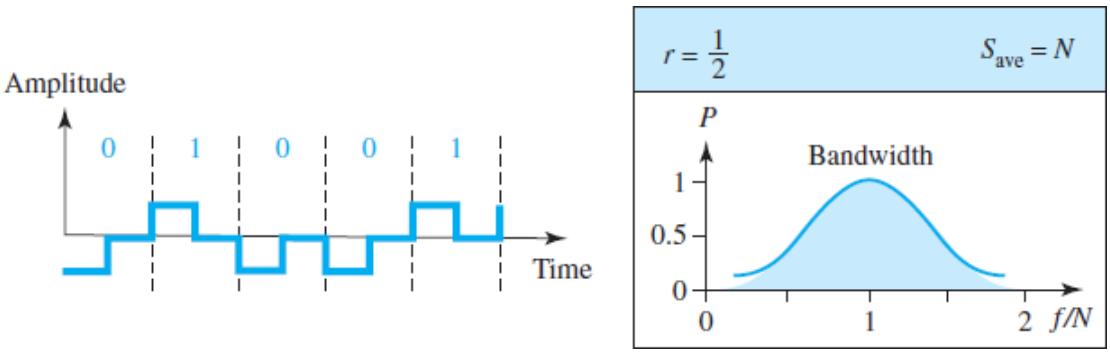


Return-to-Zero (RZ):

The main problem with NRZ encoding occurs when the sender and receiver clocks are not synchronized. The receiver does not know when one bit has ended and the next bit is starting. One solution is the return-to-zero (RZ) scheme, which uses three values: positive, negative, and zero. In RZ, the signal changes not between bits but during the bit. In Figure can we see that the signal goes to 0 in the middle of each bit. It remains

there until the beginning of the next bit. The main disadvantage of RZ encoding is that it requires two signal changes to encode a bit and therefore occupies greater bandwidth.

Polar RZ Scheme

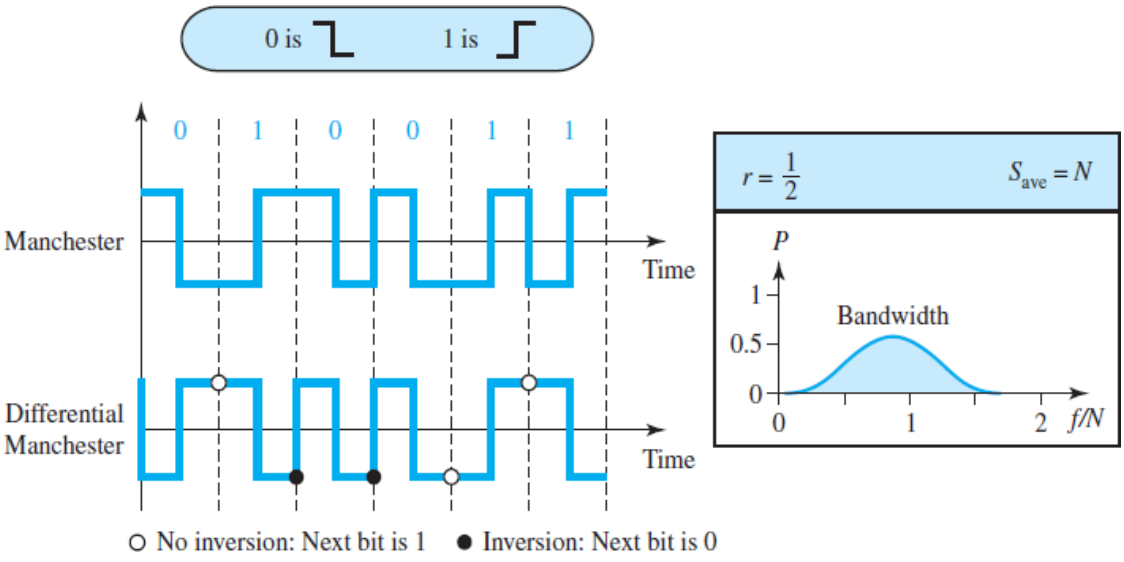


Biphase: Manchester and Differential Manchester:

The idea of RZ (transition at the middle of the bit) and the idea of NRZ-L are combined into the Manchester scheme. In Manchester encoding, the duration of the bit is divided into two halves. The voltage remains at one level during the first half and moves to the other level in the second half. The transition at the middle of the bit provides synchronization.

Differential Manchester, on the other hand, combines the ideas of RZ and NRZ-I. There is always a transition at the middle of the bit, but the bit values are determined at the beginning of the bit. If the next bit is 0, there is a transition; if the next bit is 1, there is none. Figure shows both Manchester and differential Manchester encoding.

Polar biphase: Manchester and differential Manchester schemes



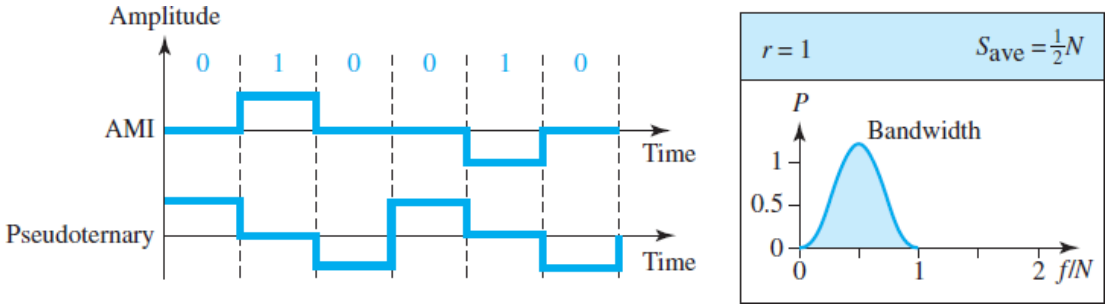
NOTE: In Manchester and differential Manchester encoding, the transition at the middle of the bit is used for synchronization. The minimum bandwidth of Manchester and differential Manchester is 2 times that of NRZ.

Bipolar Schemes:

In bipolar encoding (sometimes called multilevel binary), there are three voltage levels: positive, negative, and zero. The voltage level for one data element is at zero, while the voltage level for the other element alternates between positive and negative

Figure shows two variations of bipolar encoding: AMI and pseudo-ternary. A common bipolar encoding scheme is called bipolar alternate mark inversion (AMI). In the term 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 1s are represented by alternating positive and negative voltages. A variation of AMI encoding is called pseudo- ternary in which the 1 bit is encoded as a zero voltage and the 0 bit is encoded as alternating positive and negative voltages.

Bipolar schemes: AMI and pseudoternary



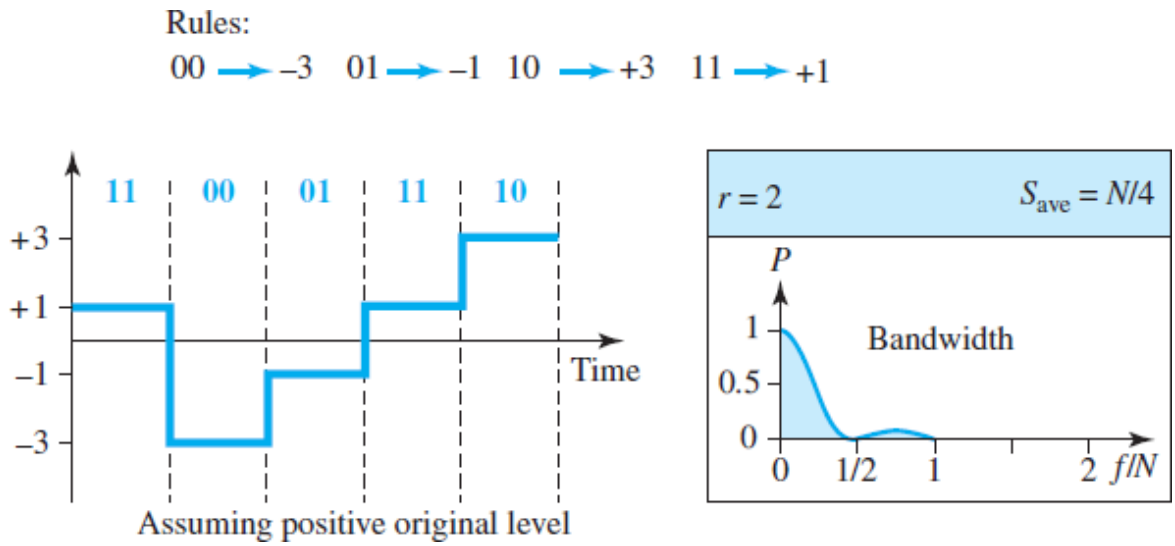
Multilevel Schemes:

The desire to increase the data rate or decrease the required bandwidth has resulted in the creation of many schemes. The goal is to increase the number of bits per baud by encoding a pattern of m data elements into a pattern of n signal elements. We only have two types of data elements (0s and 1s), which means that a group of m data elements can produce a combination of 2m data patterns. We can have different types of signal elements by allowing different signal levels. If we have L different levels, then we can produce Ln combinations of signal patterns. If 2m = Ln, then each data pattern is encoded into one signal pattern. If 2m < Ln, data patterns occupy only a subset of signal patterns. The subset can be carefully designed to prevent baseline wandering, to provide synchronization, and to detect errors that occurred during data transmission. Data encoding is not possible if 2m > Ln because some of the data patterns cannot be encoded.

2B1Q:

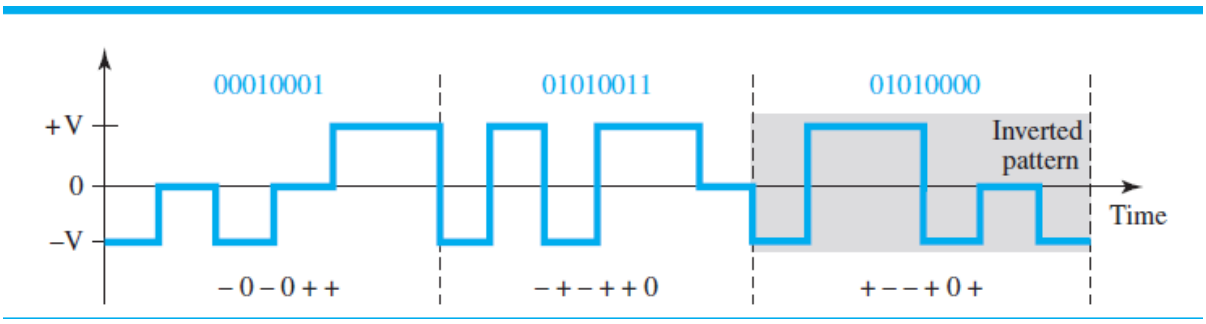
two binary, one quaternary (2B1Q), uses data patterns of size 2 and encodes the 2-bit patterns as one signal element belonging to a four-level signal. In this type of encoding m = 2, n = 1, and L = 4 (quaternary). Figure shows an example of a 2B1Q signal.

Multilevel: 2B1Q scheme



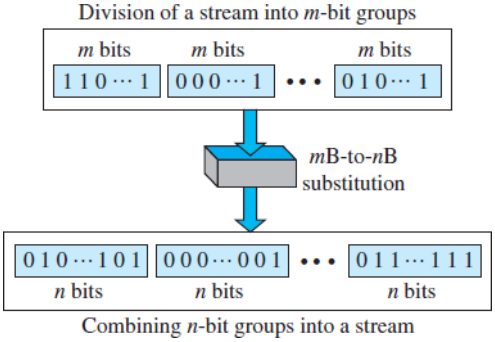
8B6T:

A very interesting scheme is eight binary, six ternary (8B6T). This code is used with 100BASE-4T cable, as we will see in Chapter 13. The idea is to encode a pattern of
8 bits as a pattern of six signal elements, where the signal has three levels (ternary). In this type of scheme, we can have 28 = 256 different data patterns and 36 = 729 different signal patterns.



Block Coding:

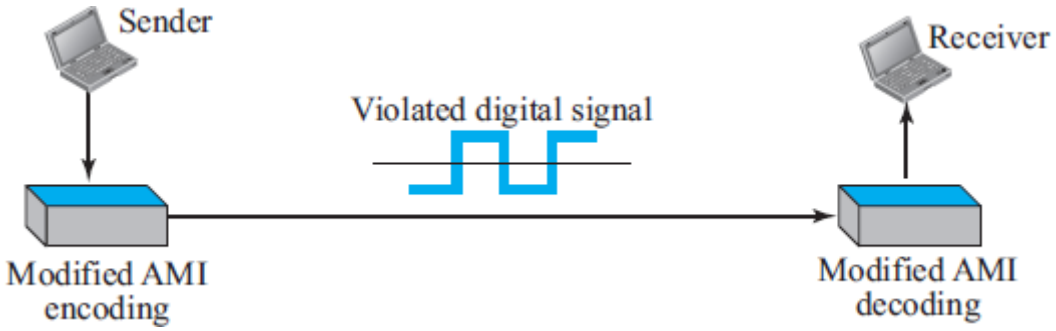
We need redundancy to ensure synchronization and to provide some kind of inherent error detecting. Block coding can give us this redundancy and improve the performance of line coding. In general, block coding changes a block of m bits into a block of n bits, where n is larger than m. Block coding is referred to as an mB/nB encoding technique.



Scrambling:

Biphase schemes that are suitable for dedicated links between stations in a LAN are not suitable for long-distance communication because of their wide bandwidth requirement.

The combination of block coding and NRZ line coding is not suitable for long-distance encoding either, because of the DC component. Bipolar AMI encoding, on the other hand, has a narrow bandwidth and does not create a DC component. However, a long sequence of 0s upsets the synchronization. If we can find a way to avoid a long sequence of 0s in the original stream, we can use bipolar AMI for long distances. We are looking for a technique that does not increase the number of bits and does provide synchronization. We are looking for a solution that substitutes long zero-level pulses with a combination of other levels to provide synchronization. One solution is called scrambling. We modify part of the AMI rule to include scrambling, as shown in Figure. Note that scrambling, as opposed to block coding, is done at the same time as encoding. The system needs to insert the required pulses based on the defined scrambling rules. Two common scrambling techniques are B8ZS and HDB3.

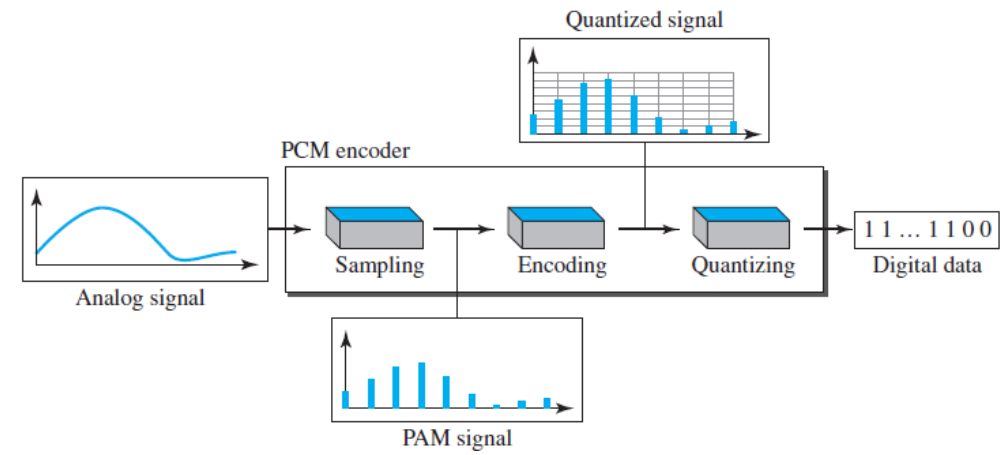


ANALOG-TO-DIGITAL CONVERSION

Sometimes we have an analog signal such as one created by a microphone or camera. But in today's world the analog signal is preferably converted to digital data for fast transmission and error free purpose.

Pulse Code Modulation (PCM):

The most common technique to change an analog signal to digital data (digitization) is called pulse code modulation (PCM). A PCM encoder has three processes, as shown in Figure.



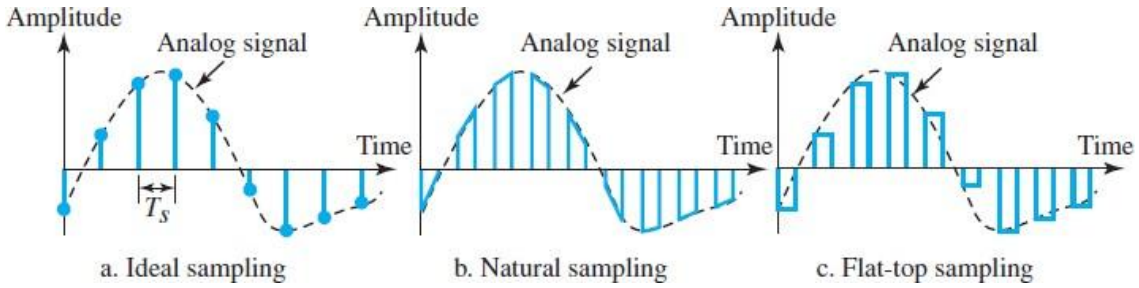
1. The analog signal is sampled.
2. The sampled signal is quantized.
3. The quantized values are encoded as streams of bits.

Sampling:

The first step in PCM is sampling. The analog signal is sampled every Ts s, where Ts is the sample interval or period. The inverse of the sampling interval is called the sampling rate or sampling frequency and denoted by fs, where fs = 1/Ts. There are three sampling methods—ideal, natural, and flat-top.

In ideal sampling, pulses from the analog signal are sampled. This is an ideal sampling method and cannot be easily implemented. In natural sampling, a high-speed switch is turned on for only the small period of time when the sampling occurs. The result is a sequence of samples that retains the shape of the analog signal. The most common sampling method, called sample and hold, however, creates flat-top samples by using a circuit.

The sampling process is sometimes referred to as pulse amplitude modulation (PAM).However, that the result is still an analog signal with nonintegral values.



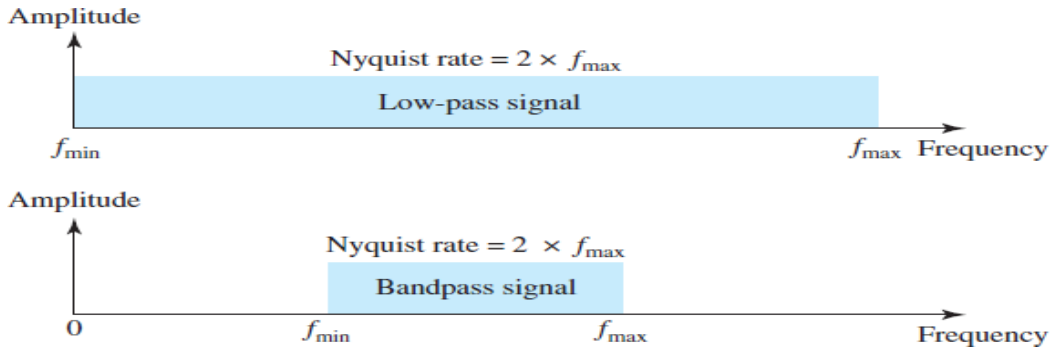
Sampling Rate(T_s)

One important consideration is the sampling rate or frequency.

. This question was elegantly answered by Nyquist. According to the Nyquist theorem, to reproduce the original analog signal, one necessary condition is that the sampling rate be at least twice the highest frequency in the original signal.

According to the Nyquist theorem, the sampling rate must beat least 2 times the highest frequency contained in the signal.

We need to elaborate on the theorem at this point. First, we can sample a signal only if the signal is band-limited. In other words, a signal with an infinite bandwidth cannot be sampled. Second, the sampling rate must be at least 2 times the highest frequency, not the bandwidth. If the analog signal is low-pass, the bandwidth and the highest frequency are the same value. If the analog signal is bandpass, the bandwidth value is lower than the value of the maximum frequency



Quantization:

The result of sampling is a series of pulses with amplitude values between the maximum

and minimum amplitudes of the signal. The set of amplitudes can be infinite with nonintegral values between the two limits. These values cannot be used in the encoding process. The following are the steps in quantization:

1. We assume that the original analog signal has instantaneous amplitudes between V_{MIN} and V_{MAX}.
2. We divide the range into L zones, each of height Δ (delta).

$$\Delta = \frac{V_{MAX} - V_{MIN}}{L}$$

3. We assign quantized values of 0 to L − 1 to the midpoint of each zone.
4. We approximate the value of the sample amplitude to the quantized values.

As a simple example, assume that we have a sampled signal and the sample amplitudes are between −20 and +20 V. We decide to have eight levels (L = 8). This means that Δ = 5 V

Quantization Levels:

In the previous example, we showed eight quantization levels. The choice of L, the number of levels, depends on the range of the amplitudes of the analog signal and how accurately we need to recover the signal. If the amplitude of a signal fluctuates between two values only, we need only two levels; if the signal, like voice, has many amplitude values, we need more quantization levels. In audio digitizing, L is normally chosen to be 256; in video it is normally thousands. Choosing lower values of L increases the quantization error if there is a lot of fluctuation in the signal.

Quantization Error:

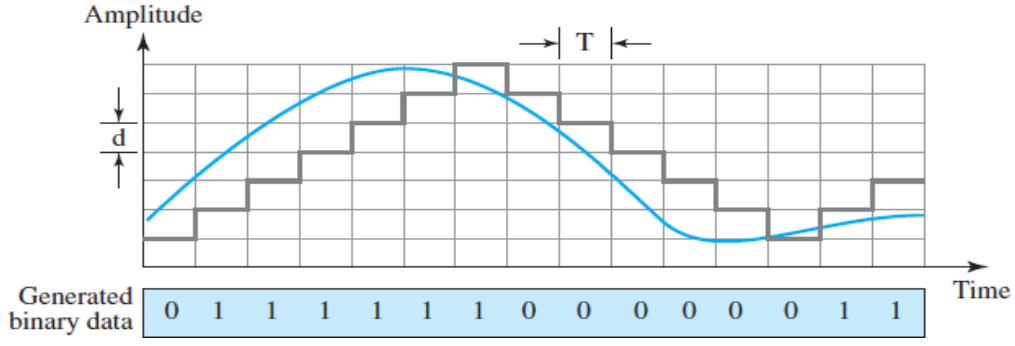
One important issue is the error created in the quantization process. Quantization is an approximation process. Theinput values to the quantizer are the real values; the output values are the approximated values. The output values are chosen to be the middle value in the zone. If the input value is also at the middle of the zone, there is no quantization error; otherwise, there is an error. In the previous example, the normalized amplitude of the third sample is 3.24, but the normalized quantized value is 3.50. This means that there is an error of +0.26. The value of the error for any sample is less than Δ/2. other words, we have $-\Delta/2 \leq \text{error} \leq \Delta/2$. The quantization error changes the signal-to-noise ratio of the signal, which in turn reduces the upper limit capacity according to Shannon.

It can be proven that the contribution of the quantization error to the SNRdB of the signal depends on the number of quantization levels L, or the bits per sample n_b, as shown in the following formula:

$$\textbf{SNRdB} = 6.02 n_b + 1.76 \text{ dB}$$

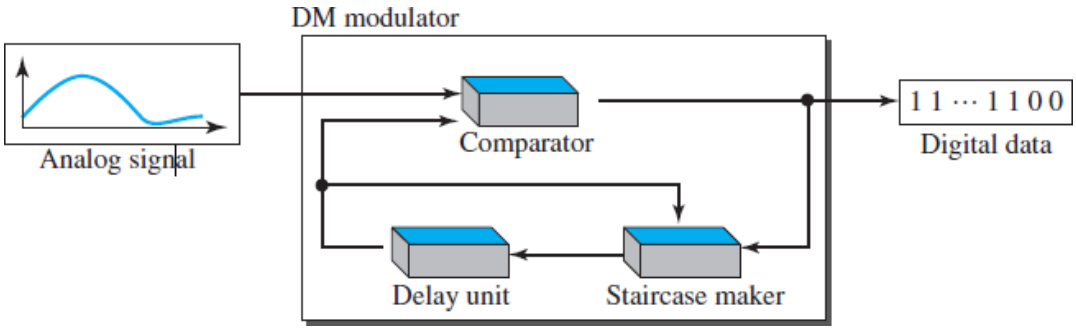
Delta Modulation (DM):

PCM is a very complex technique. Other techniques have been developed to reduce the complexity of PCM. The simplest is delta modulation. PCM finds the value of the signal amplitude for each sample; DM finds the change from the previous sample. Figure shows the process. Note that there are no code words here; bits are sent one after another



Modulator

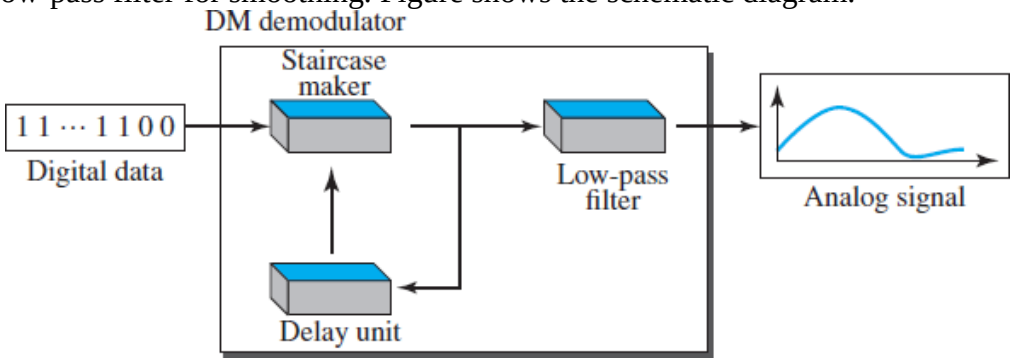
The modulator is used at the sender site to create a stream of bits from an analog signal. The process records the small positive or negative changes, called delta δ. If the delta is positive, the process records a 1; if it is negative, the process records a 0. However, the process needs a base against which the analog signal is compared. The modulator builds a second signal that resembles a staircase. Finding the change is then reduced to comparing the input signal with the gradually made staircase signal. Figure shows a diagram of the process.



The modulator, at each sampling interval, compares the value of the analog signal with the last value of the staircase signal. If the amplitude of the analog signal is larger, the next bit in the digital data is 1; otherwise, it is 0. The output of the comparator, however, also makes the staircase itself. If the next bit is 1, the staircase maker moves the last point of the staircase signal δ up; if the next bit is 0, it moves it δ down. Note that we need a delay unit to hold the staircase function for a period between two comparisons.

Demodulator:

The demodulator takes the digital data and, using the staircase maker and the delay unit, creates the analog signal. The created analog signal, however, needs to pass through a low-pass filter for smoothing. Figure shows the schematic diagram.



Adaptive DM

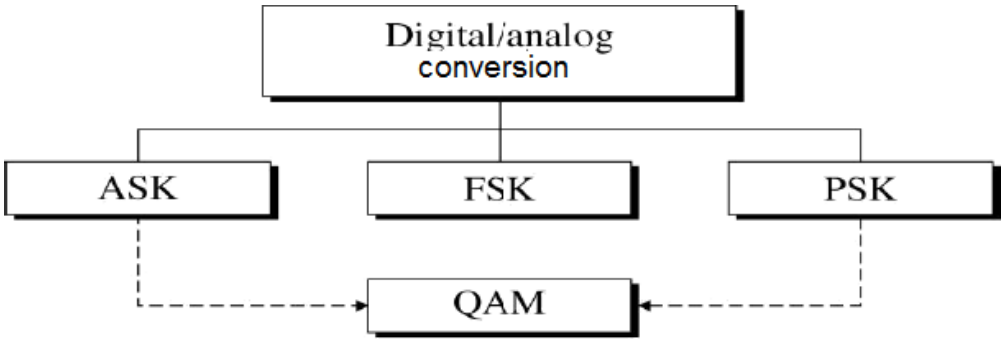
A better performance can be achieved if the value of δ is not fixed. In adaptive delta modulation, the value of δ changes according to the amplitude of the analog signal.

Quantization Error:

It is obvious that DM is not perfect. Quantization error is always introduced in the process. The quantization error of DM, however, is much less than that for PCM.

DIGITAL-TO-ANALOG CONVERSION:

Digital-to-analog conversion is the process of changing one of the characteristics of an analog signal based on the information in digital data.

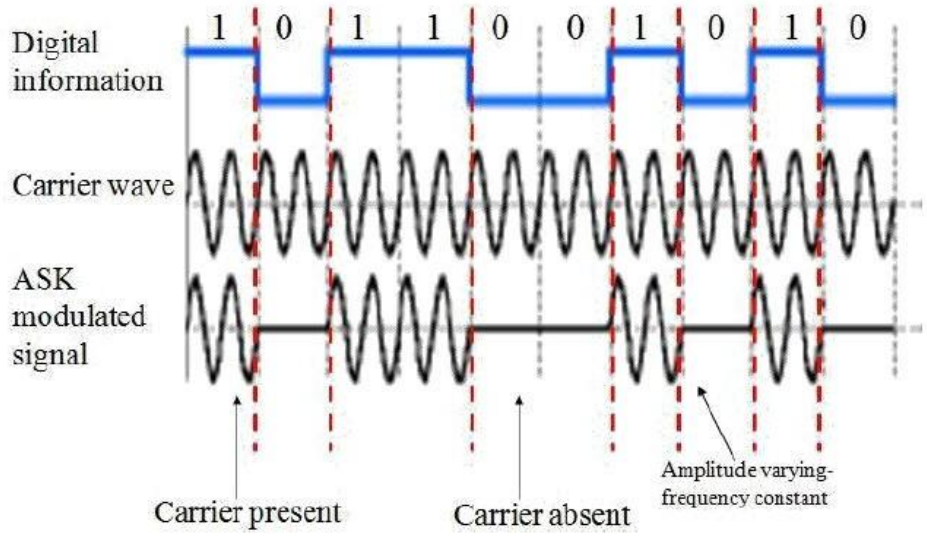


Amplitude-Shift Keying:

It is the simplest digital modulation technique where a binary information signal directly modulates the amplitude of an analog carrier. Only two output amplitudes are possible and ASK is sometimes called as digital amplitude modulation (DAM). Amplitude shift keying is given in mathematical terms as:

$$vask(t) = [\ 1 + vm(t)\ ][\ A/2 \cos(\omega ct)]$$

Where vask(t) is amplitude-shift keying wave, vm(t) is digital modulation (modulating) signal in volts, A/2 is unmodulated carrier amplitude in volts and ωc is analog carrier radian frequency in radians per second.



In the above equation, for the modulating signal vm(t), logic 1 is represented by +1V and logic 0 is represented by -1V. So the modulated wave vask(t) is either Acos(ωct) or 0 i.e., the carrier is either on or off. ASK is sometimes referred as on-off keying (OOK). The rate of change of the ASK waveform (baud) is the same as the rate of change of the binary input making bit rate equal to baud. With ASK, the bit rate is also equal to the minimum Nyquist bandwidth.

Frequency Shift Keying:

FSK is another simple, low-performance type of digital modulation. It is similar to FM, except the modulating signal is a binary signal varying between two discrete voltage levels.

FSK is sometimes called as binary FSK (BFSK). FSK is generally expressed as

$$v_{fsk}(t) = V_c \cos\{ 2\pi[f_c + v_m(t)\Delta f]t \}$$

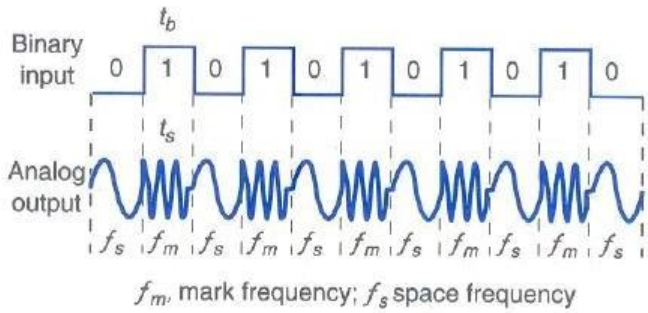
Where vfsk(t) is binary FSK waveform, Vc is peak analog carrier amplitude in volts, fc is analog carrier center frequency in hertz, f is peak change or shift in the analog carrier frequency and vm(t) is binary input(modulating) signal in volts. For logic 1, vm(t) = +1 and for logic 0, vm(t) = -1 reducing the equation to

$$v_{fsk}(t) = V_c \quad \cos\{ 2\pi[f_c \quad + \quad f]t \}$$

AND

$$v_{fsk}(t) = V_c \cos\{ 2\pi[f_c - \quad f]t \}$$

As the binary signal changes from a logic 0 to a logic 1 and vice versa, the output frequency shifts between two frequencies: a mark, or logic 1 frequency (fm) and a space or logic 0 frequency (fs). The mark and space frequencies are separated from the carrier frequency by the peak frequency deviation (f) and from each other by 2 f.



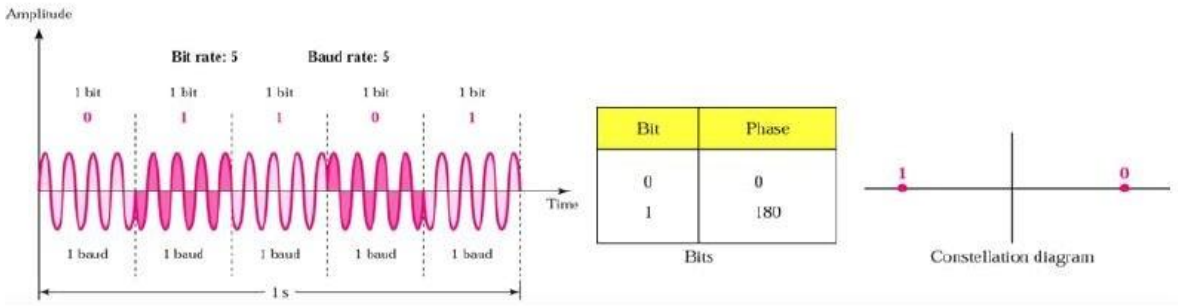
binary input	frequency output
0	space (f_s)
1	mark (f_m)

(FSK waveform and truth table)

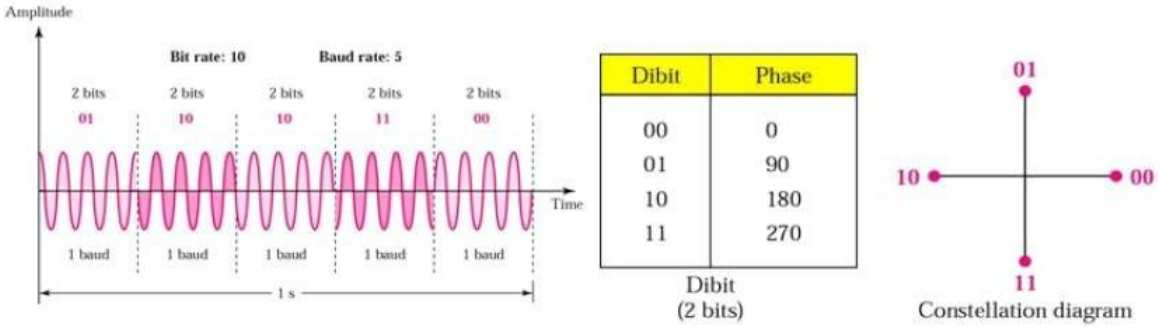
Phase-Shift Keying:

Phase-shift keying (PSK) is a digital modulation scheme that conveys data by changing, or modulating, the phase of a reference signal (the carrier wave). PSK uses a finite number of phases; each assigned a unique pattern of binary digits. Usually, each phase encodes an equal number of bits. PSK is not susceptible to the noise degradation that affects ASK or to the bandwidth limitations of FSK.

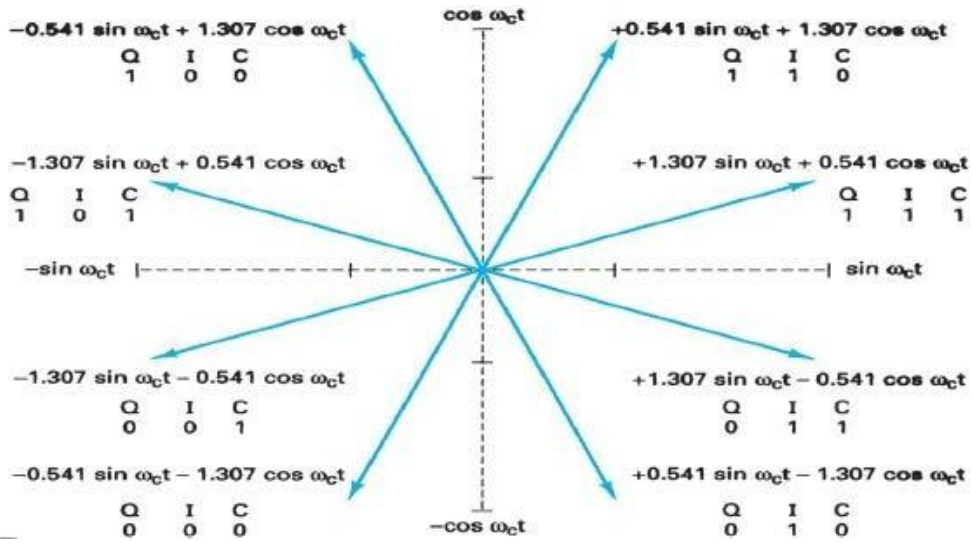
Binary phase-shift keying: The simplest PSK technique is called binary phase-shift keying (BPSK), where N = 1 and M = 2. Therefore, with BPSK two phases are possible for the carrier. It uses two opposite signal phases (0 and 180 degrees). The digital signal is broken up timewise into individual bits (binary digits). The state of each bit is determined according to the state of the preceding bit. If the phase of the wave does not change, then the signal state stays the same (0 or 1). If the phase of the wave changes by 180 degrees -- that is, if the phase reverses -- then the signal state changes (from 0 to 1 or from 1 to 0). Because there are two possible wave phases, BPSK is sometimes called **biphase modulation or phase-reversal keying (PRK).**



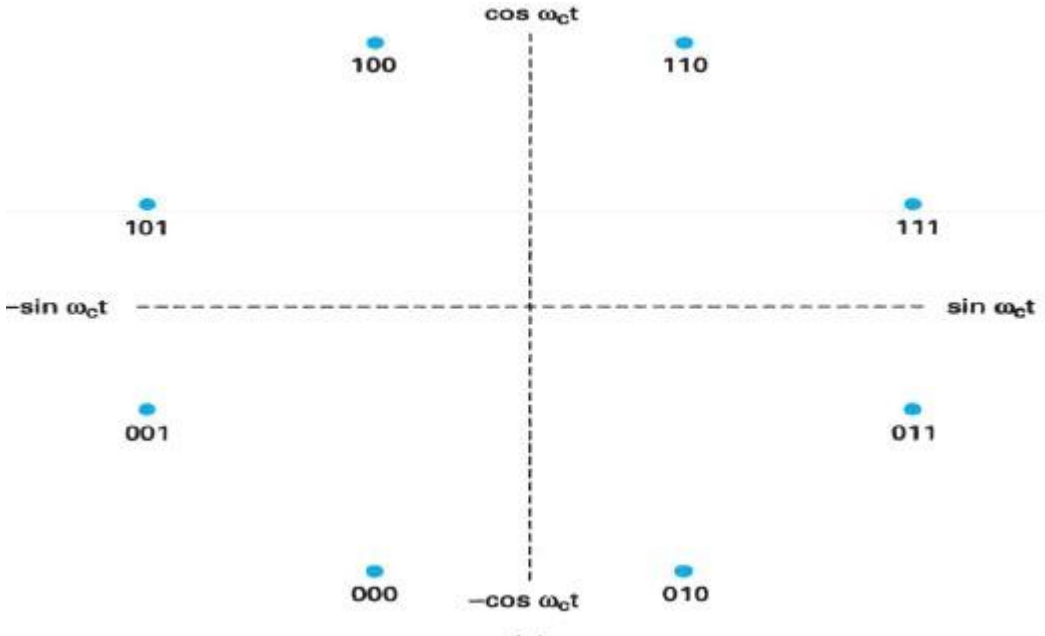
More sophisticated forms of PSK exist. In M-ary or multiple phase-shift keying (MPSK), there are more than two phases, usually four (0, +90, -90, and 180 degrees) or eight (0, +45, -45,
+90, -90, +135, -135, and 180 degrees). If there are four phases (m = 4), the MPSK mode is called **quadrature phase-shift keying** or **quaternary phase-shift keying (QPSK),** and each phase shift represents two signal elements. If there are eight phases (m = 8), the MPSK mode is known as **octal phase-shift keying (OPSK),** and each phase shift represents three signal elements. In MPSK, data can be transmitted at a faster rate, relative to the number of phase changes per unit time, than is the case in BPSK.
QPSK is an M-ary encoding scheme where N = 2 and M = 4, which has four output phases are possible for a single carrier frequency needing four different input conditions. With two bits, there are four possible conditions: 00, 01, 10, and 11. With QPSK, the binary input data are combined into groups of two bits called **dibits.**



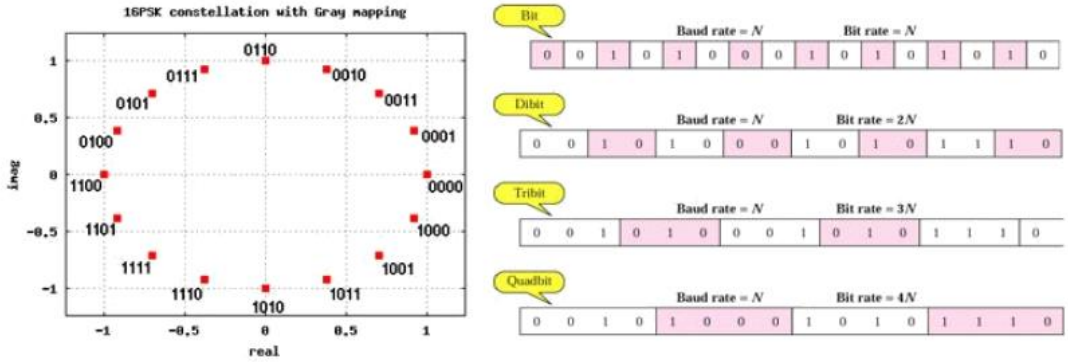
The above figure shows the output phase-versus-time relationship, truth table, and constellation diagram for QPSK. A phase of 0^0 now represents 00; 90^0represents 01; 180^0 represents 10; and 270^0 represents 11. Data can be transmitted twice as efficiently using 4-PSK than 2-PSK. With 8-PSK, three bits are encoded forming tribits and producing eight different output phases. With 8-PSK, N = 3, M = 8, and the minimum bandwidth and baud equal one third the bit rate (fb /3). 8-PSK is 3 times as efficient as 2-PSK.



Binary input			8-PSK output phase
Q	I	C	
0	0	0	−112.5°
0	0	1	−157.5°
0	1	0	−67.5°
0	1	1	−22.5°
1	0	0	+112.5°
1	0	1	+157.5°
1	1	0	+67.5°
1	1	1	+22.5°

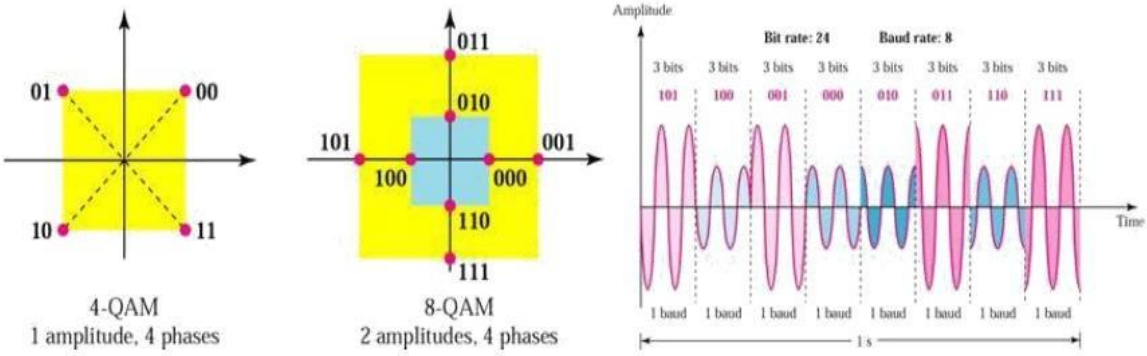


With 16-PSK, four bits called quadbits are combined, producing 16 different outputs phases. With 16-PSK, N = 4, M = 16, and the minimum bandwidth and baud equal one-fourth the bit rate (fb /4).



Modulation	Bit Rate	Encoding Scheme	Bandwidth Efficiency	Outputs Possible	Minimum Bandwidth	Baud
ASK	N	Single bit	1	2	$f^a{}_b$	f_b
FSK	N	Single bit	1	2	$>f_b$	f_b
BPSK	N	Single bit	1	2	f_b	f_b
QPSK	2N	Dibits	2	4	$f_b/2$	$f_b/2$
8-PSK	3N	Tribits	3	8	$f_b/3$	$f_b/3$
16-PSK	4N	Quadibits	4	16	$f_b/4$	$f_b/4$

Quadrature Amplitude Modulation (QAM):
PSK is limited by the ability of the equipment to distinguish small differences in phase. Bandwidth limitations make combinations of FSK with other changes practically useless. Quadrature amplitude modulation is a combination of ASK and PSK so that a maximum contrast between each signal unit (bit, dibit, tribit, and so on) is achieved. QAM is used extensively as a modulation scheme for digital telecommunication systems. The primary advantage of QAM over PSK is immunity to transmission impairments, especially phase impairments that are inherent in all communication systems.
In 4-QAM and 8-QAM, 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.



With 16-QAM, there are 12 phases and three amplitudes that are combined to produce 16

different output conditions. With QAM, there are always more phases possible than amplitude.

Bandwidth Efficiency:

Bandwidth efficiency is often used to compare the performance of one digital modulation technique to another. It is the ration of transmission bit rate to the minimum bandwidth required for a particular modulation scheme. Mathematically represented as:

Bη = transmission bit rate (bps) / minimum bandwidth (Hz)

Modulation	Encoding Scheme	Outputs Possible	Minimum Bandwidth	Baud	Bη
ASK	Single bit	2	f_b	f_b	1
FSK	Single bit	2	f_b	f_b	1
BPSK	Single bit	2	f_b	f_b	1
QPSK	Dibits	4	$f_b/2$	$f_b/2$	2
8-PSK	Tribits	8	$f_b/3$	$f_b/3$	3
8-QAM	Tribits	8	$f_b/3$	$f_b/3$	3
16-PSK	Quadbits	16	$f_b/4$	$f_b/4$	4
16-QAM	Quadbits	16	$f_b/4$	$f_b/4$	4
32-PSK	Five bits	32	$f_b/5$	$f_b/5$	5
64-QAM	Six bits	64	$f_b/6$	$f_b/6$	6

Note: f_b indicates a magnitude equal to the input bit rate.

ASK, FSK, PSK, and QAM Summary

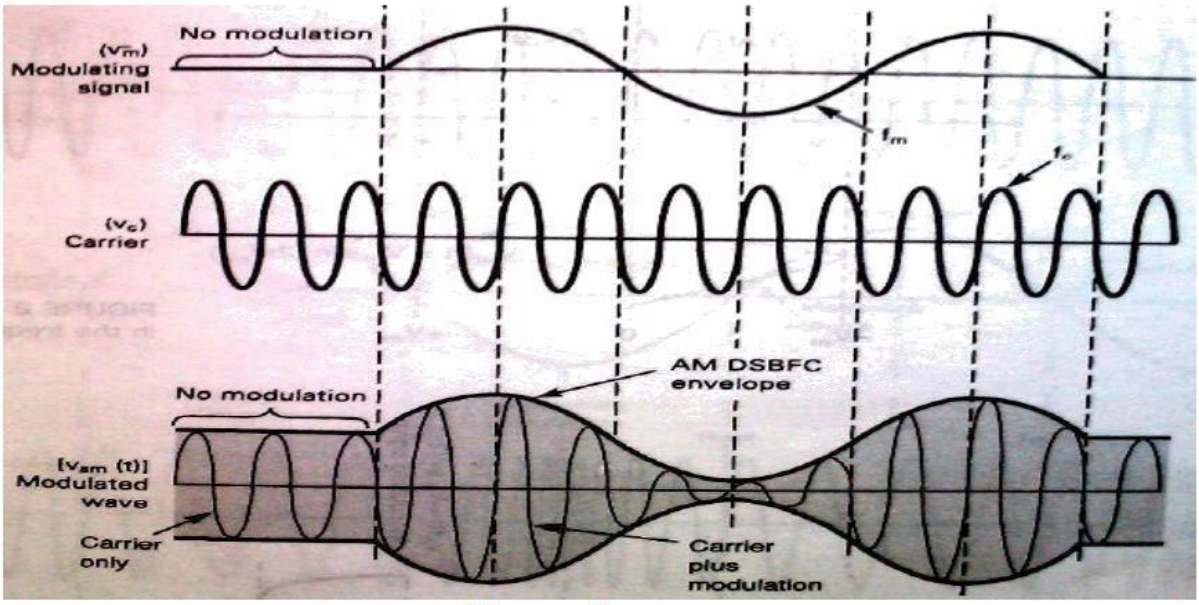
ANALOG TO ANALOG CONVERSION:

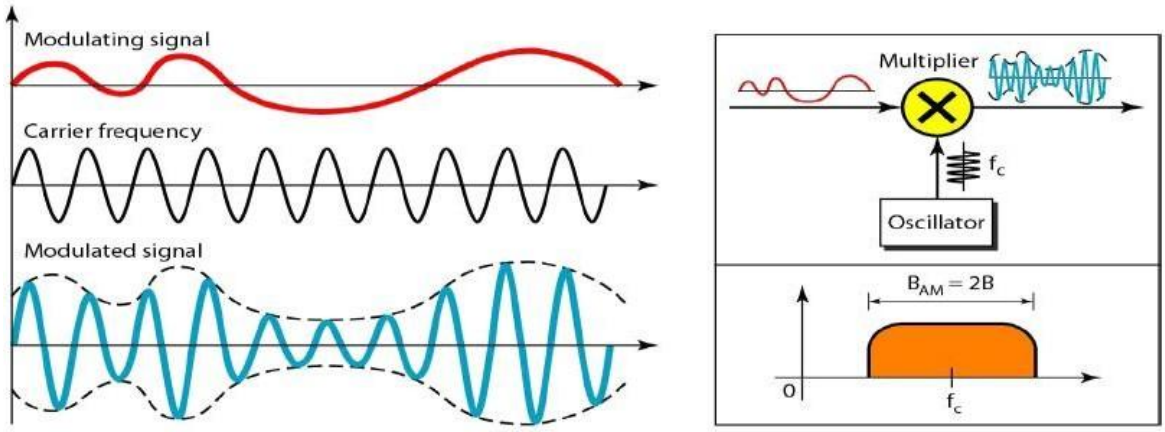
Analog-to-analog conversion, or analog modulation, is the representation of analog information by an analog signal. One may ask why we need to modulate an analog signal; it is already analog. Modulation is needed if the medium is bandpass in nature or if only a bandpass channel is available to us. An example is radio. The government assigns a narrow bandwidth to each radio station. The analog signal produced by each station is a low-pass signal, all in the same range. To be able to listen to different stations, the low-pass signals need to be shifted, each to a different range.

Analog-to-analog conversion can be accomplished in three ways: **amplitude modulation (AM)**, **frequency modulation (FM**), and **phase modulation (PM).**
FM and PM are usually categorized together as **Angle modulation.**

Amplitude Modulation:

Amplitude modulation is the process of changing the amplitude of a relatively high frequency carrier signal in proportion to the instantaneous value of the modulating signal (information). AM modulators are two-input devices, one of them is a single, relatively high frequency carrier signal of constant amplitude and the second is the relatively lowfrequency information signal. The following figure shows generation of AM waveform when a single-frequency modulating signal acts on a high frequency carrier signal.



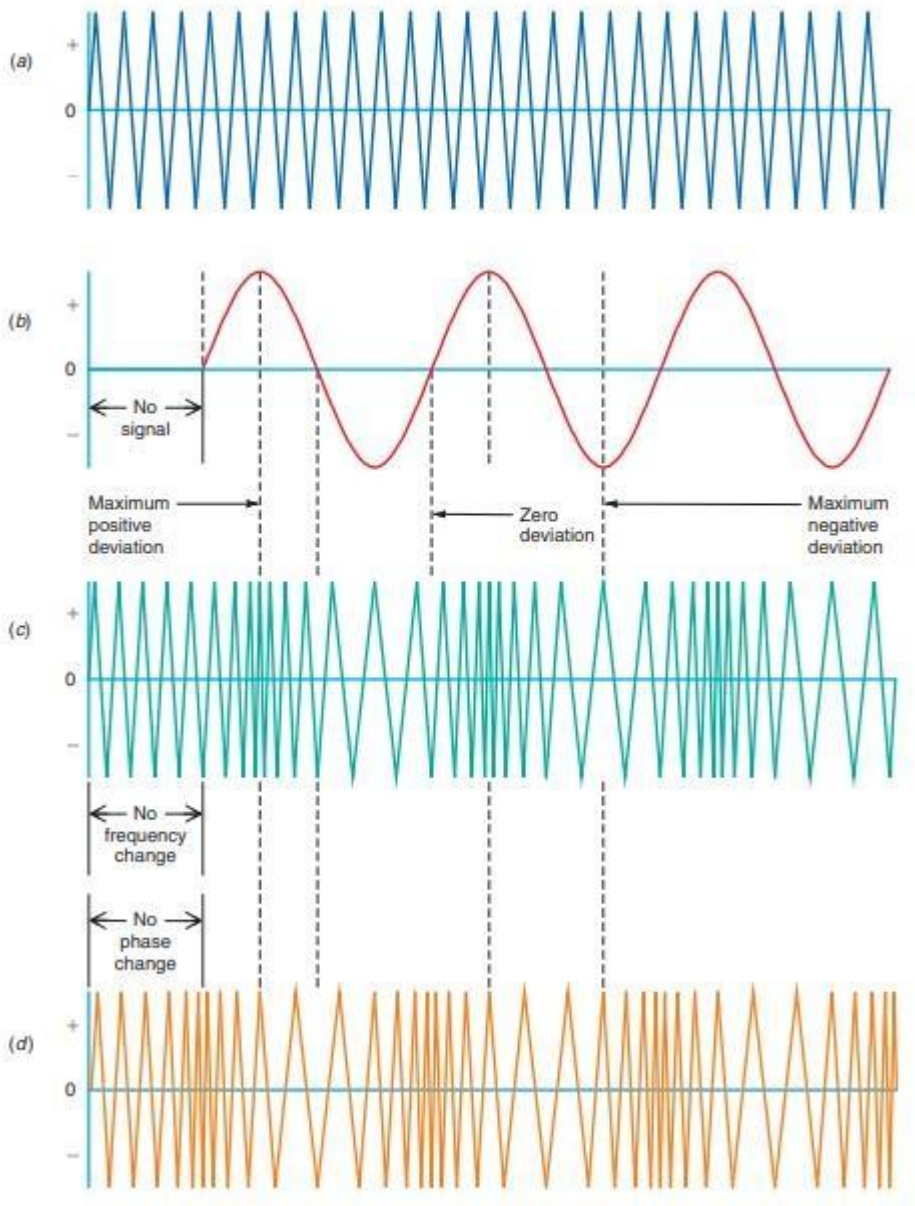


AM GENERATION

Advantages of AM are simple to implement, needs a circuit with very few components and inexpensive. The disadvantages include inefficient power usage and use of bandwidth and also prone to noise. The total bandwidth required for AM can be determined from the bandwidth of the audio signal: $B_{AM} = 2B$.

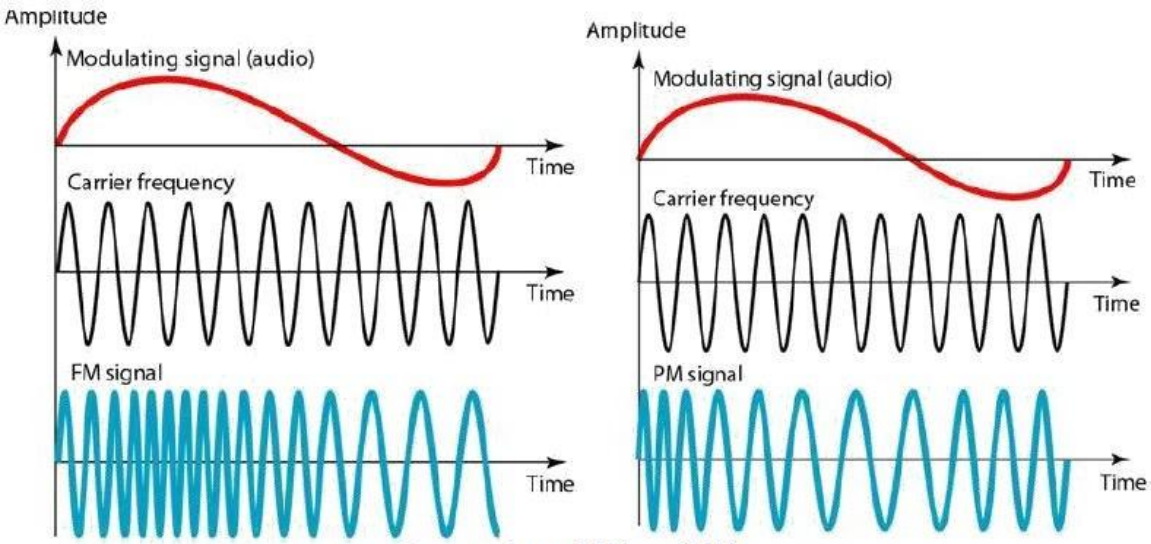
Angle Modulation:

Angle modulation results whenever the phase angle of a sinusoidal signal is varied with respect to time and includes both FM and PM. Whenever the frequency of a carrier signal is varied, the phase is also varied and vice versa. If the frequency of the carrier is varied directly in accordance with the information signal, FM results, whereas if the phase is varied directly, PM results.



(a) Carrier,(b) Modulating signal,(c) FM signal(d) PM signal

The above figure shows the FM and PM of a sinusoidal carrier by a single-frequency modulating signal. Both FM and PM waveforms are identical except for their time relationship (phase). With FM, the maximum frequency deviation occurs during the maximum positive and negative peaks of the modulating signal. With FM, the maximum frequency deviation occurs during the zero crossings in the modulating signal.



Comparison of FM and PM

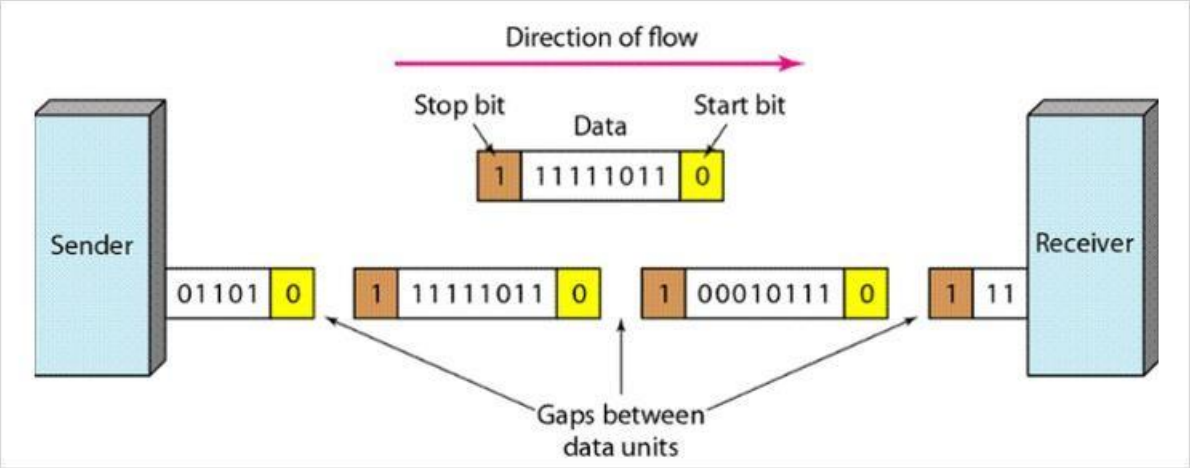
An important feature of FM and PM is that they can provide much better protection to the message against channel noise when compared to AM. Also because of their constant amplitude nature, they can withstand nonlinear distortion and amplitude fading.

Chapter-4

Asynchronous Transmission:

In asynchronous transmission process transmitted information is encoded with start and stop bits, specifying the beginning and end of each character. As long as some pattern is followed, the receiving device can retrieve the information without regard to which it is sent. Patterns are based on grouping the bit-streams into a byte usually eight bits is sent along with the link as a unit. The sender sends each group of data independently, relaying it to the link whenever ready, without regard to a timer.

1. The receiver cannot use timing to predict the arrival time of the next group so that synchronizing pulse is required. To notify the receiving system to the arrival of a new group, therefore, an extra bit is added to the beginning of each byte.
2. This 0s bit is referred to as the start bit. Telling the receiver that the byte is finished, one or more additional bits are appended to the end of the byte. This is bit is called a stop bit.
3. By this method, each byte is increased in size to at least 10 bits, of which 8 are information and two or more are signals to the receiver.
4. Also, the transmission of each byte may gap of varying duration. This gap between information can be represented either by an idle channel or by a stream of additional stop bits.
5. The bits of a byte that is 8 bits are transmitted simultaneously on separate wires. If two devices are close together computer or printer so the communication within the computer.



This figure shows the schematic illustration of asynchronous transmission. In this example, the start bits are 0s and the stop bits are 1s and the gap is represented by an idle line rather than by additional stop bits.

In this diagram, the ASCII character would be transmitted using 10 bits. In this transmission scheme "0100 0001" changes into "1 0100 0001 0". The extra bits, depending on the parity bit, at the start and end of the data transmission.

This starts and stops bits tells the receiver that character is coming and also the character has ended. This scheme of transmission is used when data are transmitted irregularly as opposed to in a solid stream.

Characteristics of asynchronous communication are as follows:

1. Extra bits are added to the start and end of the character stream.
2. Between two characters there may exist gaps or spaces.
3. The idle time is not constant between bytes as only one byte is sent at a time.

4. The reception of data is done at different clock frequencies.

Advantages:

1. Synchronization between devices is not necessary.
2. It is a low-cost scheme.

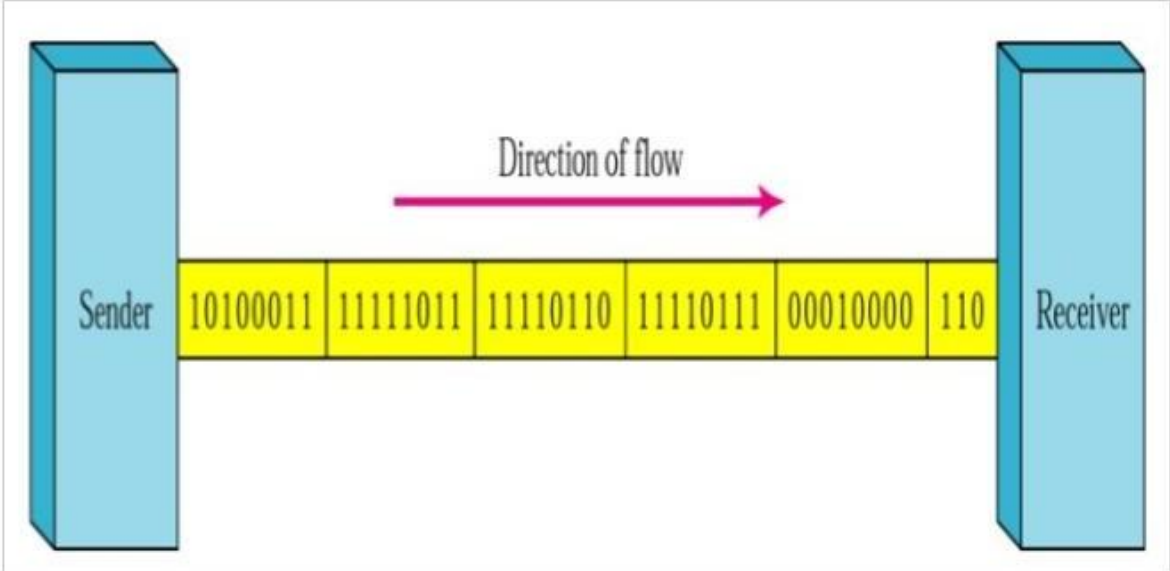
Disadvantages:

1. Low transmission due to the use of 'start' and 'stop' bits and gaps between data chunks.
2. Timing errors take place.

Examples of asynchronous transmission: emails, forums, letters, radios and televisions

Synchronous Transmission:

In Synchronous Transmission, data is sent in form of blocks or frames. This transmission is the full duplex type. Between sender and receiver the synchronization is compulsory. In Synchronous transmission, There is no gap present between data. It is more efficient and more reliable than asynchronous transmission to transfer the large amount of data. In this transmission, we send bits one after another without start/stop bits or gaps. Grouping of bits is the receiver's responsibility.



Synchronous Transmission

Characteristics of Synchronous Transmission:

1. Between transmitted characters, there are no spaces.
2. Timing is very important as the accuracy of the received information is completely dependent on the ability of the receiving device to keep an accurate count of the bits as they come in.

3. Special ‘syn’ characters are sent before the data being sent.
4. These syn characters are placed between chunks of data for timing functions.

Advantages:

1. Data speed is much higher because of no extra bits at the sending end and at the receiving end.
2. Timing errors are reduced due to syn.
3. More useful for high-speed applications.

Disadvantages:

1. Timing is responsible for the accuracy of data.
2. It is required that transmitter and receiver be properly synchronized.

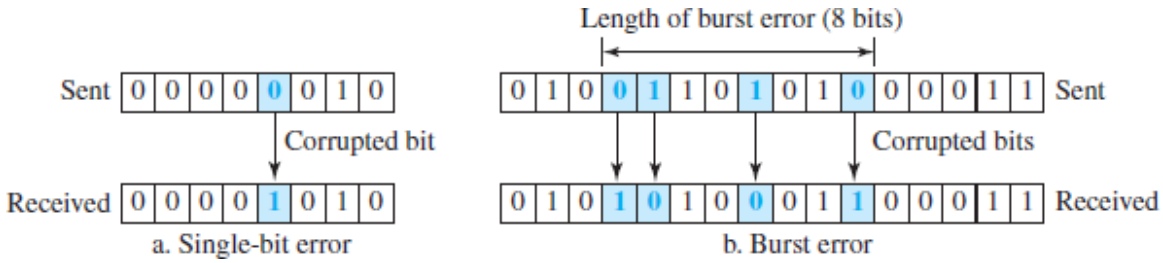
Examples of Synchronous Transmission: chat rooms, video conferencing, telephonic conversations, face-to-face interactions.

Synchronous Transmission	Asynchronous Transmission
1. In Synchronous transmission, Data is sent in form of blocks or frames.	1. In asynchronous transmission, Data is sent in form of byte or character.
2. Synchronous transmission is fast.	2. Asynchronous transmission is slow.
3. Synchronous transmission is costly.	3. Asynchronous transmission economical.
4. In Synchronous transmission, time interval of transmission is constant.	4. In asynchronous transmission, time interval of transmission is not constant, it is random.
5. Synchronous transmission needs precisely synchronized clocks for the information of new bytes.	5.Asynchronous transmission have no need of synchronized clocks as parity bit is used in this transmission for information of new bytes.

Error Detection and Correction:

Types of Errors:

Whenever bits flow from one point to another, they are subject to unpredictable changes because of interference. This interference can change the shape of the signal. The term single-bit error means that only 1 bit of a given data unit (such as a byte, character, or packet) is changed from 1 to 0 or from 0 to 1. The term burst error means that 2 or more bits in the data unit have changed from 1 to 0 or from 0 to 1. Figure shows the effect of a single-bit and a burst error on a data unit.



A burst error is more likely to occur than a single-bit error because the duration of the noise signal is normally longer than the duration of 1 bit, which means that when noise affects data, it affects a set of bits. The number of bits affected depends on the data rate and duration of noise.

The central concept in detecting or correcting errors is redundancy. To be able to detect or correct errors, we need to send some extra bits with our data. These redundant bits are added by the sender and removed by the receiver. Their presence allows the receiver to detect or correct corrupted bits.

The correction of errors is more difficult than the detection. In error detection, we are only looking to see if any error has occurred. The answer is a simple yes or no. We are not even interested in the number of corrupted bits. A single-bit error is the same for us as a burst error. In error correction, we need to know the exact number of bits that are corrupted and, more importantly, their location in the message. The number of errors and the size of the message are important factors. If we need to correct a single error in an 8-bit data unit, we need to consider eight possible error locations; if we need to correct two errors in a data unit of the same size, we need to consider 28 (permutation of 8 by 2) possibilities. You can imagine the receiver's difficulty in finding 10 errors in a data unit of 1000 bits.

Redundancy is achieved through various coding schemes. The sender adds redundant bits through a process that creates a relationship between the redundant bits and the actual data bits. The receiver checks the relationships between the two sets of bits to detect errors. The ratio of redundant bits to data bits and the robustness of the process are important factors in any coding scheme.

We can divide coding schemes into two broad categories: block coding and convolution Coding.

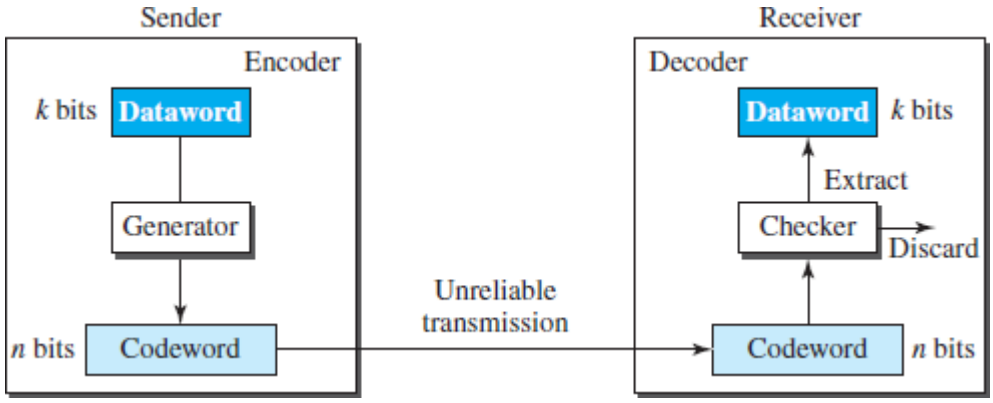
BLOCK CODING:

In block coding, we divide our message into blocks, each of k bits, called datawords. We add r redundant bits to each block to make the length n = k + r. The resulting n-bit blocks are called codewords. How the extra r bits are chosen or calculated is something we will discuss later. For the moment, it is important to know that we have a set of datawords, each of size k, and a set of codewords, each of size of n. With k bits, we can create a combination of 2^k datawords; with n bits, we can create a combination of 2^n codewords. Since n > k, the number of possible codewords is larger than the number of possible datawords. The block coding process is one-to-one; the same dataword is always encoded as the same codeword. This means that we have $2^n - 2^k$ codewords that are not used. We call these codewords invalid or illegal. The trick in error detection is the existence of these invalid codes, as we discuss next. If thereceiver receives an invalid codeword, this indicates that the data was corrupted during transmission.

If the following two conditions are met, the receiver can detect a change in the original codeword.

1. The receiver has (or can find) a list of valid codewords.
2. The original codeword has changed to an invalid one.

Each codeword sent to the receiver may change during transmission. If the received codeword is the same as one of the valid codewords, the word is accepted; the corresponding dataword is extracted for use. If the received codeword is not valid, it is discarded. However, if the codeword is corrupted during transmission but the received word still matches a valid codeword, the error remains undetected.



Dataword	*Codeword*	*Dataword*	*Codeword*
00	000	10	101
01	011	11	110

Assume the sender encodes the dataword 01 as 011 and sends it to the receiver. Consider the
following cases:

1. The receiver receives 011. It is a valid codeword. The receiver extracts the dataword 01 from it.
2. The codeword is corrupted during transmission, and 111 is received (the leftmost bit is corrupted).This is not a valid codeword and is discarded.
3. The codeword is corrupted during transmission, and 000 is received (the right two bits are corrupted). This is a valid codeword. The receiver incorrectly extracts the dataword 00. Two corrupted bits have made the error undetectable.

An error-detecting code can detect only the types of errors for which it is designed; other types of errors may remain undetected.

Hamming Distance:

One of the central concepts in coding for error control is the idea of the Hamming distance. The Hamming distance between two words (of the same size) is the number of differences between the corresponding bits. We show the Hamming distance between two words x and y as d(x, y). We may wonder why Hamming distance is important for error detection. The reason is that the Hamming distance between the received codeword and the sent codeword is the number of bits that are corrupted during transmission. For example,if the codeword 00000 is sent and 01101 is received, 3 bits are in error and the Hamming distance between the two is d(00000, 01101) = 3. In other words, if the Hamming distance between the sent and the received codeword is not zero, the codeword has been corrupted during transmission. The Hamming distance can easily be found if we apply the XOR operation ($\oplus$)

on the two words and count the number of 1s in the result Note that the Hamming distance is a value greater than or equal to zero.

Minimum Hamming Distance for Error Detection:
In a set of codewords, the minimum Hamming distance is the smallest Hamming distance between all possible pairs of codewords. Now let us find the minimum Hamming distance in a code if we want to be able to detect up to s errors. If s errors occur during transmission, the Hamming distance between the sent codeword and received codeword is s. If our system is to detect up to s errors, the minimum distance between the valid codes must be (s + 1), so that the received codeword does not match a valid codeword. In other words, if the minimum distance between all valid codewords is (s + 1), the received codeword cannot be erroneously mistaken for another codeword. The error will be detected. We need to clarify a point here: Although a code with dmin = s + 1 may be able to detect more than s errors in some special cases, only s or fewer errors are guaranteed to be detected.

CYCLIC CODES:

Cyclic codes are special linear block codes with one extra property. In a cyclic code, if a codeword is cyclically shifted (rotated), the result is another codeword. For example, if 1011000 is a codeword and we cyclically left-shift, then 0110001 is also a codeword. In this case, if we call the bits in the first word a0 to a6, and the bits in the second word b0 to b6, we can shift the bits by using the following:

$$b_1 = a_0 \qquad b_2 = a_1 \qquad b_3 = a_2 \qquad b_4 = a_3 \qquad b_5 = a_4 \qquad b_6 = a_5 \qquad b_0 = a_6$$

Cyclic Redundancy Check:
We can create cyclic codes to correct errors. However, the theoretical background required is beyond the scope of this book. In this section, we simply discuss a subset of cyclic codes called the cyclic redundancy check (CRC), which is used in networks such as LANs and WANs.

Dataword	*Codeword*	*Dataword*	*Codeword*
0000	0000000	1000	1000101
0001	0001011	1001	1001110
0010	0010110	1010	1010011
0011	0011101	1011	1011000
0100	0100111	1100	1100010
0101	0101100	1101	1101001
0110	0110001	1110	1110100
0111	0111010	1111	1111111

A CRC code with C(7, 4)

LINK CONFIGURATION:

A network is two or more devices connected through links. A link is a communications

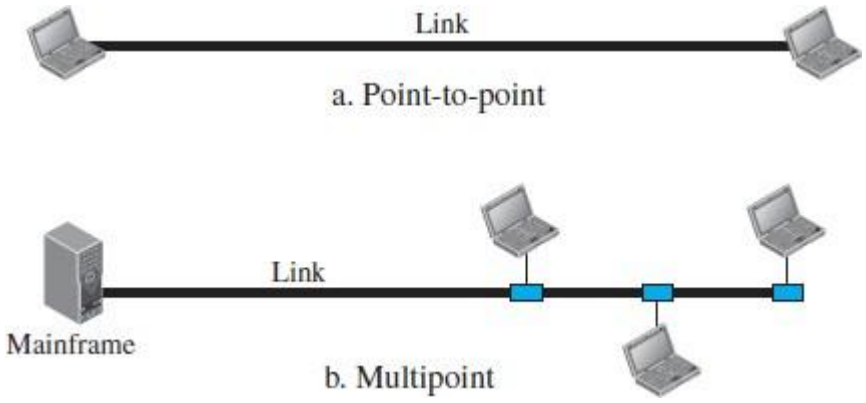
pathway that transfers data from one device to another. For visualization purposes, it is simplest to imagine any link as a line drawn between two points. For communication to occur, two devices must be connected in some way to the same link at the same time. There are two possible types of connections: point-to-point and multipoint.

Point-to-Point:

A point-to-point connection provides a dedicated link between two devices. The entire capacity of the link is reserved for transmission between those two devices. Most point-to-point connections use an actual length of wire or cable to connect the two ends, but other options, such as microwave or satellite links, are also possible When we change television channels by infrared remote control, we are establishing a point-to-point connection between the remote control and the television's control system.

Multipoint:

A multipoint (also called multidrop) connection is one in which more than two specific devices share a single link In a multipoint environment, the capacity of the channel is shared, either spatially or temporally. If several devices can use the link simultaneously, it is a spatially shared connection. If users must take turns, it is a timeshared connection



Flow Control

Whenever we have a producer and a consumer, we need to think about flow control. If the producer produces items that cannot be consumed, accumulation of items occurs. The sending data-link layer at the end of a link is a producer of frames; the receiving data-link layer at the other end of a link is a consumer. If the rate of produced frames is higher than the rate of consumed frames, frames at the receiving end need to be buffered while waiting to be consumed (processed). Definitely, we cannot have an unlimited buffer size at the receiving side. We have two choices.

1. The first choice is to let the receiving data-link layer drop the frames if its buffer is full.
2. The second choice is to let the receiving data-link layer send a feedback to the sending data-link layer to ask it to stop or slow down.

Different data-link-layer protocols use different strategies for flow control.

Error Control:

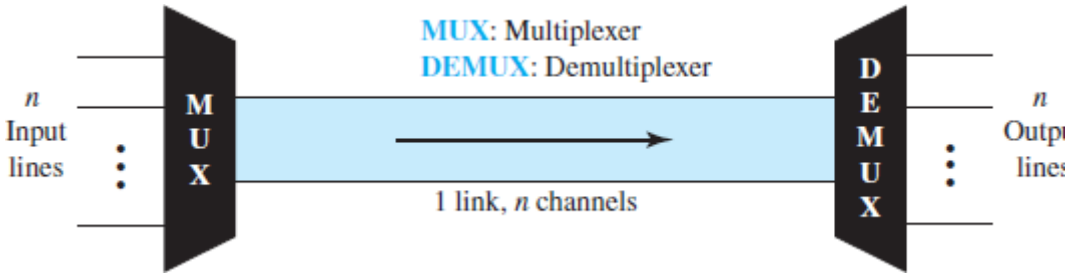
At the sending node, a frame in a data-link layer needs to be changed to bits, transformed

to electromagnetic signals, and transmitted through the transmission media. At the receiving node, electromagnetic signals are received, transformed to bits, and put
Together to create a frame. Since electromagnetic signals are susceptible to error, a frame is susceptible to error. The error needs first to be detected. After detection, it needs to be either corrected at the receiver node or discarded and retransmitted by the sending node.

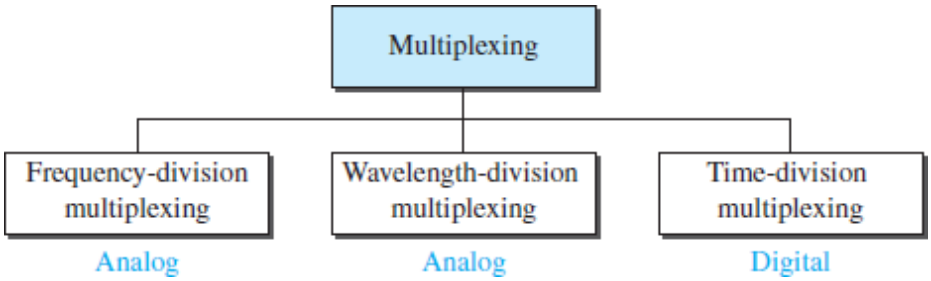
MULTIPLEXING:

Whenever the bandwidth of a medium linking two devices is greater than the bandwidth needs of the devices, the link can be shared. Multiplexing is the set of techniques that allow the simultaneous transmission of multiple signals across a single data link. As data and telecommunications use increases, so does traffic. We can accommodate this increase by continuing to add individual links each time a new channel is needed; or we can install higher-bandwidth links and use each to carry multiple signals. In today's technology includes high-bandwidth media such as optical fiber and terrestrial and satellite microwaves. Each has a bandwidth far in excess of that needed for the average transmission signal. If the bandwidth of a link is greater than the bandwidth needs of the devices connected to it, the bandwidth is wasted. An efficient system maximizes the utilization of all resources; bandwidth is one of the most precious resources we have in data communications.

In a multiplexed system, n lines share the bandwidth of one link. Figure shows the basic format of a multiplexed system. The lines on the left direct their transmission streams to a multiplexer (MUX), which combines them into a single stream (many-to-one). At the receiving end, that stream is fed into a demultiplexer (DEMUX), which separates the stream back into its component transmissions (one-to-many) and directs them to their corresponding lines. In the figure, the word link refers to the physical path. The word channel refers to the portion of a link that carries a transmission between a given pair of lines. One link can have many (n) channels.

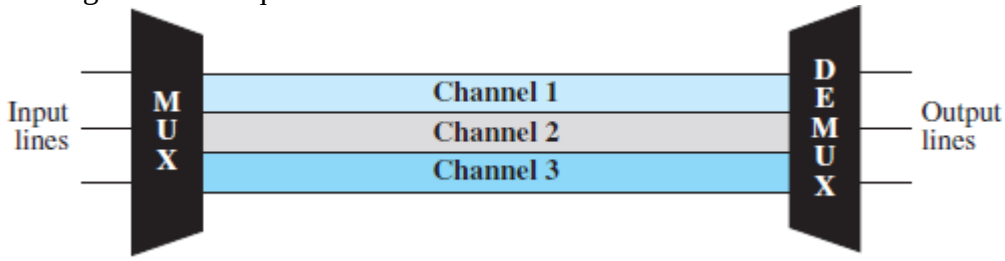


There are three basic multiplexing techniques: frequency-division multiplexing, wavelength-division multiplexing, and time-division multiplexing. The first two are techniques designed for analog signals, the third, for digital signals.



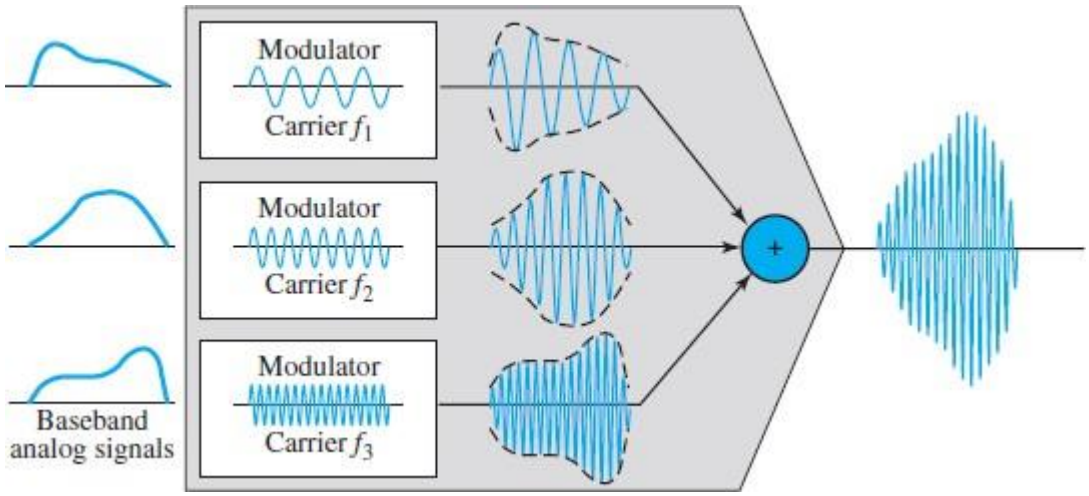
Frequency-Division Multiplexing:

Frequency-division multiplexing (FDM) is an analog technique that can be applied when the bandwidth of a link (in hertz) is greater than the combined bandwidths of the signals to be transmitted. In FDM, signals generated by each sending device modulate different carrier frequencies. These modulated signals are then combined into a single composite signal that can be transported by the link. Carrier frequencies are separated by sufficient bandwidth to accommodate the modulated signal. These bandwidth ranges are the channels through which the various signals travel. Channels can be separated by strips of unused bandwidth— guard bands—to prevent signals from overlapping. In addition, carrier frequencies must not interfere with the original data frequencies.



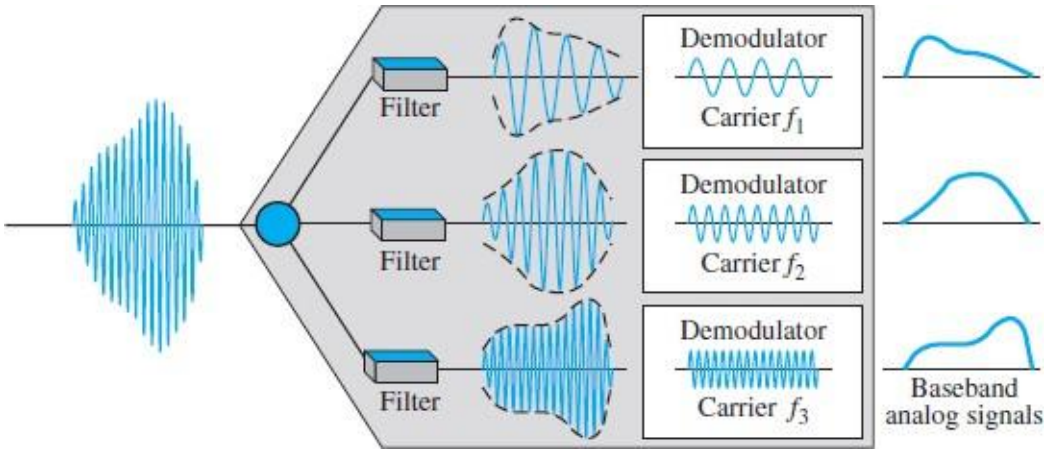
Multiplexing Process:

Each source generates a signal of a similar frequency range. Inside the multiplexer, these similar signals modulate different carrier frequencies (f1, f2, and f3). The resulting modulated signals are then combined into a single composite signal that is sent out over a media link that has enough bandwidth to accommodate it.



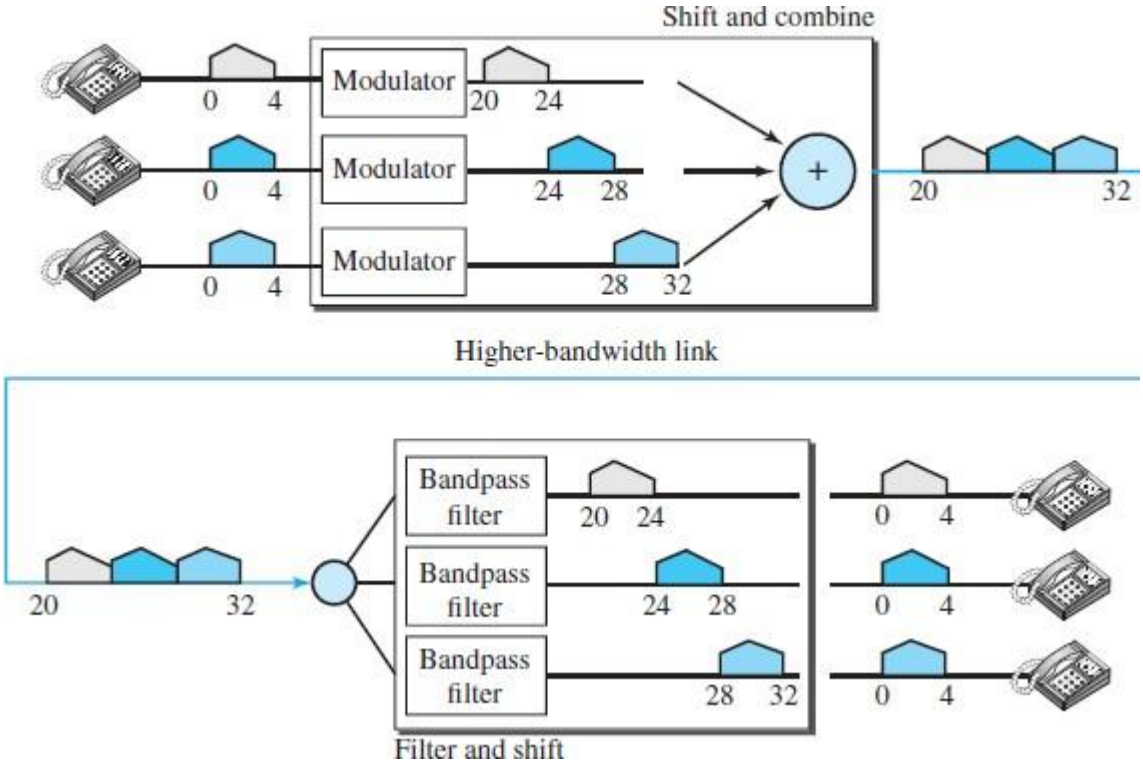
Demultiplexing Process:

The demultiplexer uses a series of filters to decompose the multiplexed signal into its constituent component signals. The individual signals are then passed to a demodulator that separates them from their carriers and passes them to the output lines.



Q.Assume that a voice channel occupies a bandwidth of 4 kHz. We need to combine three voice channels into a link with a bandwidth of 12 kHz, from 20 to 32 kHz. Show the configuration, using the frequency domain. Assume there are no guard bands.

Solution:

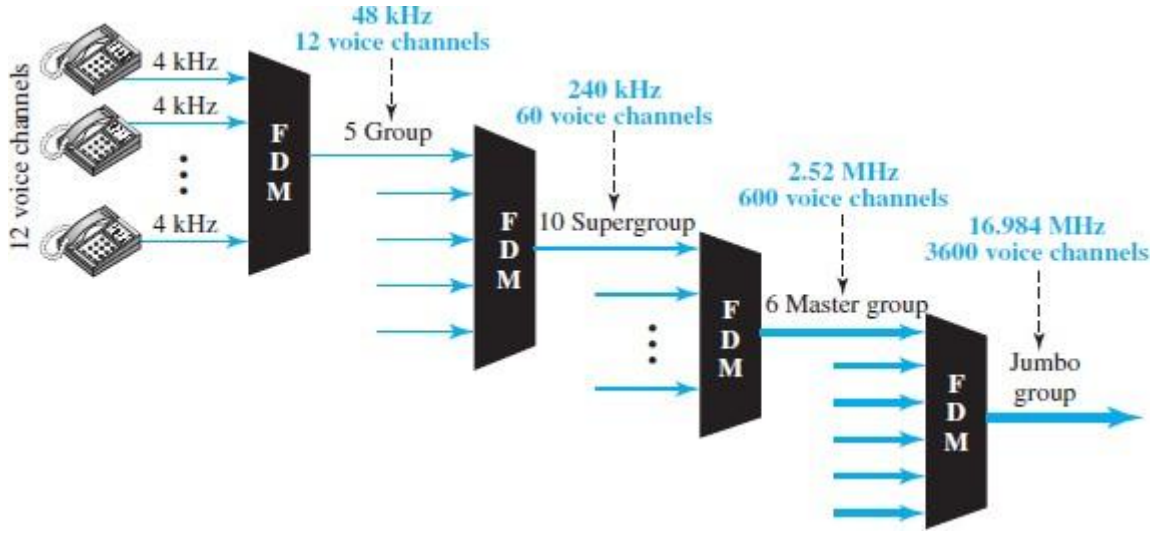


The three frequencies are shifted to 3 different frequencies by the process of modulation
We use the 20- to 24-kHz bandwidth for the first channel, the 24- to 28-kHz bandwidth for the second channel, and the 28- to 32-kHz bandwidth for the third one. At the receiver, each channel receives the entire signal, using a filter to separate out its own signal. The first channel uses a filter that passes frequencies between 20 and 24 kHz and filters out (discards) any other frequencies. The second channel uses a filter that passes frequencies between 24 and 28 kHz, and the third channel uses a filter that passes frequencies between 28 and 32 kHz. Each channel then shifts the frequency to start from zero.

The Analog Carrier System:
To maximize the efficiency of their infrastructure, telephone companies have traditionally multiplexed signals from lower-bandwidth lines onto higher-bandwidth lines. In this way, many switched or leased lines can be combined into fewer but bigger channels. For analog

lines, FDM is used. One of these hierarchical systems used by telephone companies is made up of groups, supergroups, master groups, and jumbo groups.

1. In this analog hierarchy, 12 voice channels are multiplexed onto a higher-bandwidth line to create a group.
2. A group has 48 kHz of bandwidth and supports 12 voice channels. At the next level, up to five groups can be multiplexed to create a composite signal called a supergroup.
3. A supergroup has a bandwidth of 240 kHz and supports up to 60 voice channels. Supergroups can be made up of either five groups or 60 independent voice channels.
4. At the next level, 10 supergroups are multiplexed to create a master group.
5. A master group must have 2.40 MHz of bandwidth, but the need for guard bands betweenthe supergroups increases the necessary bandwidth to 2.52 MHz. Master groups supportup to 600 voice channels.
6. Finally, six master groups can be combined into a jumbo group. A jumbo group must have 15.12 MHz (6 × 2.52 MHz) but is augmented to 16.984 MHz to allow for guard bands between the master groups.



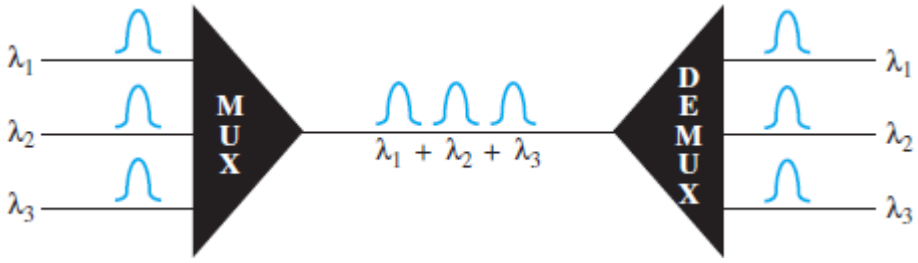
Other Applications of FDM:

1. A very common application of FDM is AM and FM radio broadcasting.
2. Radio uses theair as the transmission medium. A special band from 530 to 1700 kHz is assigned to AM radio. All radio stations need to share this band.
3. Each AM station needs 10 kHz of bandwidth. Each station uses a different carrier frequency, which means it is shifting its signal and multiplexing. The signal that goes to the air is a combination of signals. A receiver receives all these signals, but filters (by tuning) only the one which is desired. Without multiplexing, only one AM station could broadcast to the common link, the air. However, we need to know that there is no physical multiplexer or demultiplexer.
4. FM has a wider band of 88to 108 MHz because each station needs a bandwidth of 200 kHz.
5. Another common use of FDM is in television broadcasting. Each TV channel has its own bandwidth of 6 MHz.
6. The first generation of cellular telephones (See Chapter 16) also uses FDM. Each user is assigned two 30-kHz channels, one for sending voice and the other for receiving. The voice signal, which has a bandwidth of 3 kHz (from 300 to 3300 Hz), is modulated by using FM.

Wavelength-Division Multiplexing:

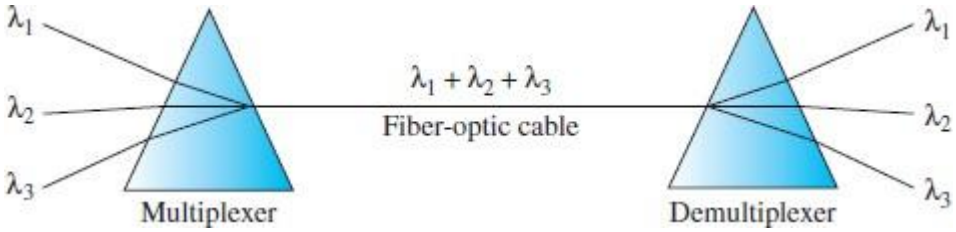
Wavelength-division multiplexing (WDM) is designed to use the high-data-rate capability of fiber-optic cable. The optical fiber data rate is higher than the data rate of

metallic transmission cable, but using a fiber-optic cable for a single line wastes the available bandwidth. Multiplexing allows us to combine several lines into one.
WDM is conceptually the same as FDM, except that the multiplexing and demultiplexing involve optical signals transmitted through fiber-optic channels. The idea is the same: We are combining different signals of different frequencies. The difference is that the frequencies are very high.



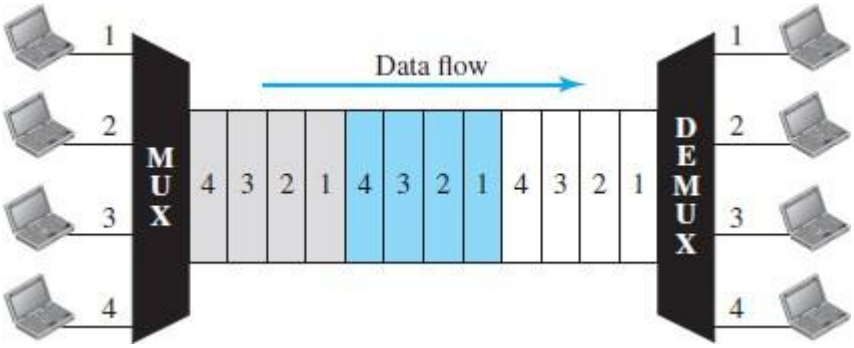
The above Figure gives a conceptual view of a WDM multiplexer and demultiplexer. Very narrow bands of light from different sources are combined to make a wider band of light. At the receiver, the signals are separated by the demultiplexer.

Although WDM technology is very complex, the basic idea is very simple. We want to combine multiple light sources into one single light at the multiplexer and do the reverse at the demultiplexer. The combining and splitting of light sources are easily handled by a prism. Recall from basic physics that a prism bends a beam of light based on the angle of incidence and the frequency. Using this technique, a multiplexer can be made to combine several input beams of light, each containing a narrow band of frequencies, into one output beam of a wider band of frequencies.



Time-Division Multiplexing

Time-division multiplexing (TDM) is a digital process that allows several connections to share the high bandwidth of a link. Instead of sharing a portion of the bandwidth as in FDM, time is shared. Each connection occupies a portion of time in the link. The below Figure gives a conceptual view of TDM. Note that the same link is used as in FDM; here, however, the link is shown sectioned by time rather than by frequency. In the figure, portions of signals 1, 2, 3, and 4 occupy the link sequentially.



We also need to remember that TDM is, in principle, a digital multiplexing technique. Digital data from different sources are combined into one timeshared link. However, this does not mean that the sources cannot produce analog data; analog data can be sampled, changed to digital data, and then multiplexed by using TDM.

Note:TDM is a digital multiplexing technique for combining several low-rate channels into one high-rate one.

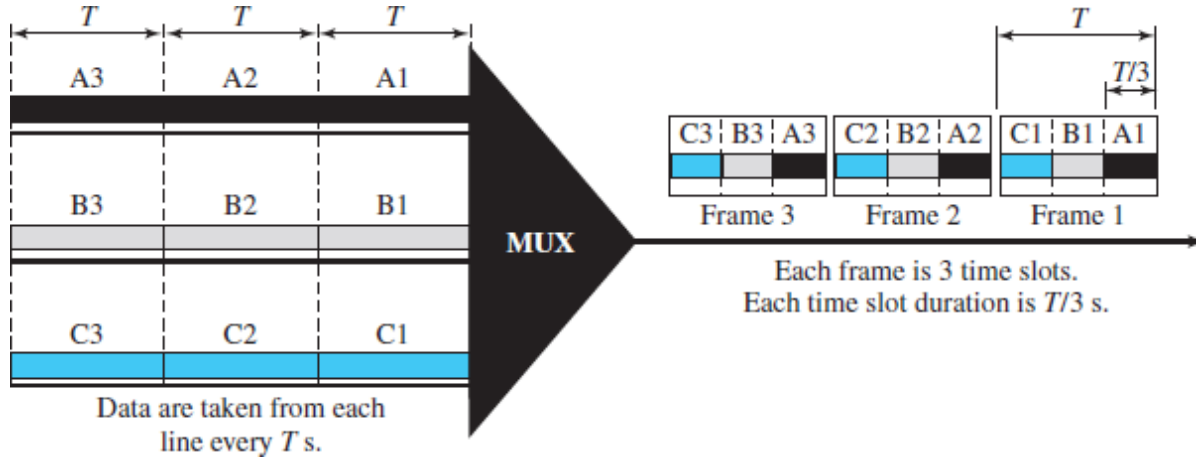
We can divide TDM into two different schemes: synchronous and statistical.

Synchronous TDM:

In synchronous TDM, each input connection has an allotment in the output even if it is not sending data.

Time Slots and Frames:

In synchronous TDM, the data flow of each input connection is divided into units, where each input occupies one input time slot. A unit can be 1 bit, one character, or one block of data. Each input unit becomes one output unit and occupies one output time slot. However, the duration of an output time slot is n times shorter than the duration of an input time slot. If an input time slot is T s, the output time slot is T/n s, where n is the number of connections. In other words, a unit in the output connection has a shorter duration; it travels faster . The below figure shows an example of synchronous TDM where n is 3.



In synchronous TDM, a round of data units from each input connection is collected into a frame (we will see the reason for this shortly). If we have n connections, a frame is divided into n time slots and one slot is allocated for each unit, one for each input line. If the duration of the input unit is T, the duration of each slot is T/n and the duration of each frame is T .

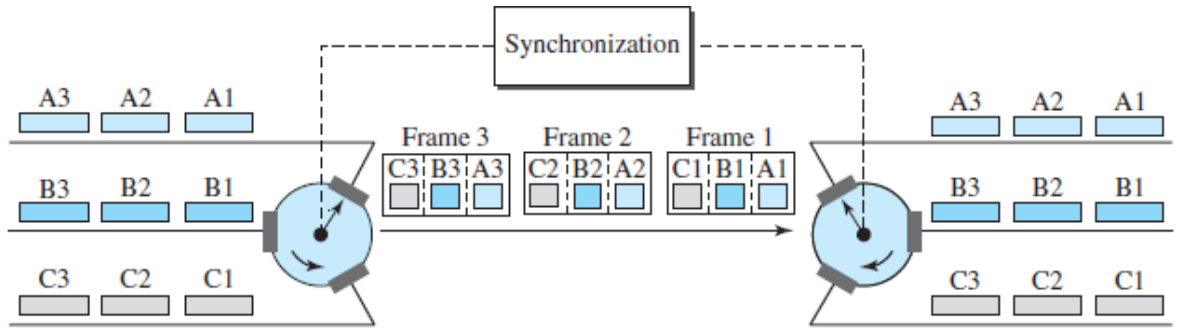
The data rate of the output link must be n times the data rate of a connection to guarantee the flow of data. In above Figure the data rate of the link is 3 times the data rate of a connection; likewise, the duration of a unit on a connection is 3 times that of the time slot (duration of a unit on the link). In the figure we represent the data prior to multiplexing as 3 times the size of the data after multiplexing. This is just to convey the idea that each unit is 3 times longer in duration before multiplexing than after.

NOTE: In synchronous TDM, the data rate of the link is n times faster, and the unit duration is n times shorter.

Time slots are grouped into frames. A frame consists of one complete cycle of time slots, with one slot dedicated to each sending device. In a system with n input lines, each frame has n slots, with each slot allocated to carrying data from a specific input line.

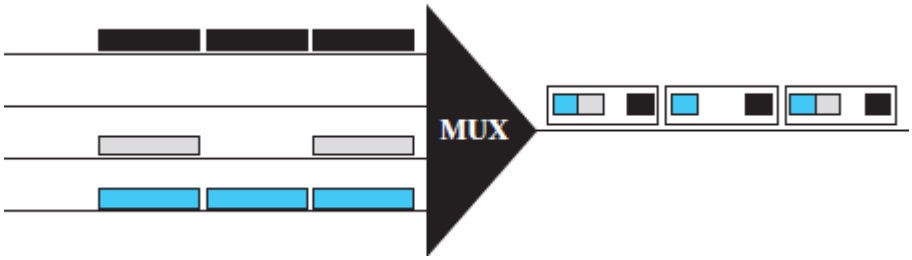
INTERLEAVING:

TDM can be visualized as two fast-rotating switches, one on the multiplexing side and the other on the demultiplexing side. The switches are synchronized and rotate at the same speed, but in opposite directions. On the multiplexing side, as the switch opens in front of a connection, that connection has the opportunity to send a unit onto the path. This process is called interleaving. On the demultiplexing side, as the switch opens in front of a connection, that connection has the opportunity to receive a unit from the path.



Empty Slots:

Synchronous TDM is not as efficient as it could be. If a source does not have data to send, the corresponding slot in the output frame is empty. The below Figure shows a case in which one of the input lines has no data to send and one slot in another input line has discontinuous data.

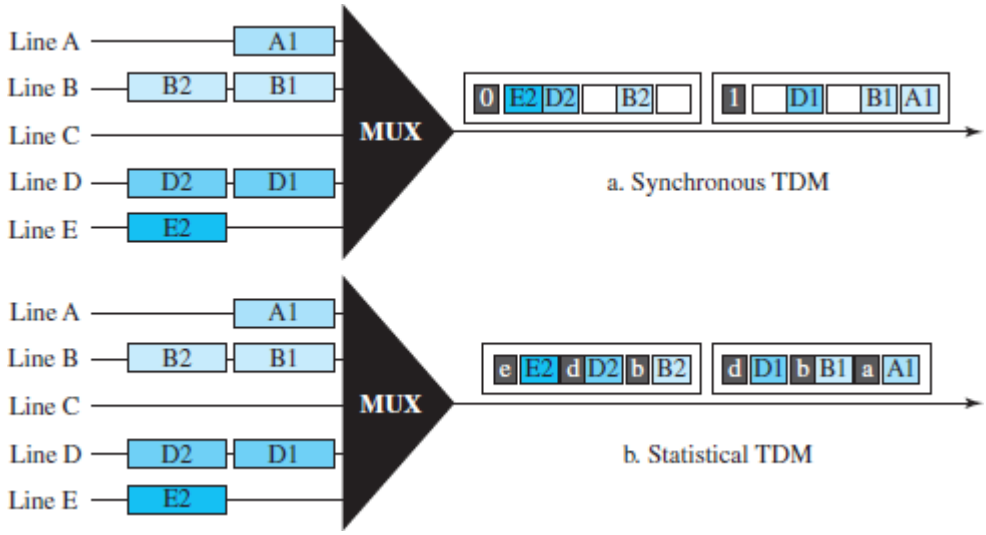


The first output frame has three slots filled, the second frame has two slots filled, and the third frame has three slots filled. No frame is full.

So to the remedy of this problem is using statistical TDM which in turn can improve the efficiency by removing the empty slots from the frame.

Statistical Time-Division Multiplexing:

in synchronous TDM, each input has a reserved slot in the output frame. This can be inefficient if some input lines have no data to send. In statistical time-division multiplexing, slots are dynamically allocated to improve bandwidth efficiency. Only when an input line has a slot's worth of data to send is it given a slot in the output frame. In statistical multiplexing, the number of slots in each frame is less than the number of input lines. The multiplexer checks each input line in round robin fashion; it allocates a slot for an input line if the line has data to send; otherwise, it skips the line and checks the next line.



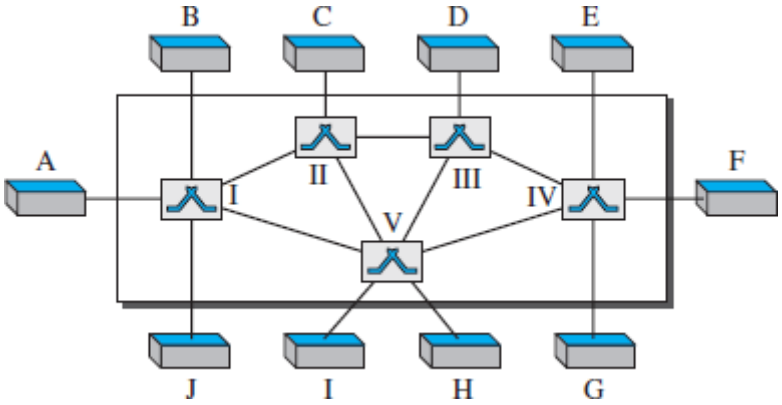
The above Figure shows a synchronous and a statistical TDM example. In the synchronous TDM some slots are empty because the corresponding line does not have data to send. In the statistical time-division multiplexing , however, no slot is left empty as long as there are data to be sent by any input line.

Chapter-5

Switching & Routing:

network is a set of connected devices. Whenever we have multiple devices, we have the problem of how to connect them to make one-to-one communication possible. One solution is to make a point-to-point connection between each pair of devices (a mesh topology) or between a central device and every other device (a star topology). These methods, however, are impractical and wasteful when applied to very large networks. The number and length of the links require too much infrastructure to be cost-efficient, and the majority of those links would be idle most of the time. Other topologies employing multipoint connections, such as a bus, are ruled out because the distances between devices and the total number of devices increase beyond the capacities of the media and equipment.

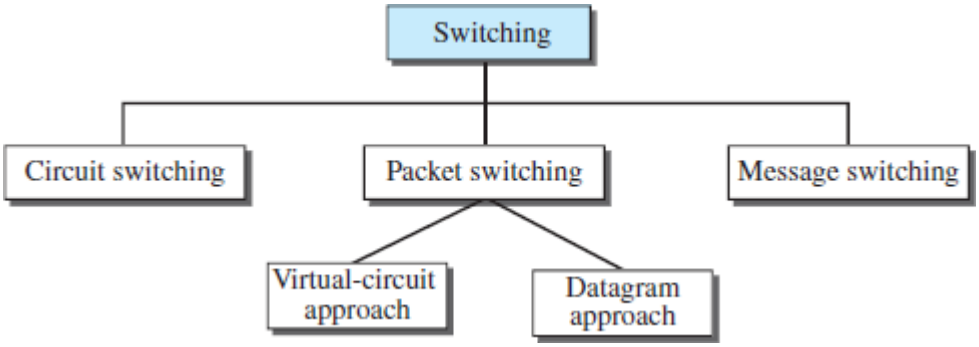
A better solution is switching. A switched network consists of a series of interlinked nodes, called switches. Switches are devices capable of creating temporary connections between two or more devices linked to the switch. In a switched network, some of these nodes are connected to the end systems (computers or telephones, for example). Others are used only for routing.



(An example of switch network)

Three Methods of Switching:

Traditionally, three methods of switching have been discussed: circuit switching, packet switching, and message switching. The first two are commonly used today. The third has been phased out in general communications but still has networking applications. Packet switching can further be divided into two subcategories—virtual circuit approach and datagram approach.



Switching functions at different TCP/IP Layers:

1. Switching at Physical Layer: At the physical layer, we can have only circuit switching. There are no packets exchanged at the physical layer. The switches at the physical layer allow signals to travel in one path or another.
2. Switching at Data-Link Layer: At the data-link layer, we can have packet switching. However, the term packet in this case means frames or *cells*. Packet switching at the data-link layer is normally done using a virtual-circuit approach.
3. Switching at Network Layer :At the network layer, we can have packet switching. In this case, either a virtual-circuit approach or a datagram approach can be used. Currently the Internet uses a datagram approach.
4. Switching at Application Layer: At the application layer, we can have only message switching. The communication at the application layer occurs by exchanging messages. Conceptually, we can say that communication using e-mail is a kind of message-switched communication, but we do not see any network that actually can be called a message-switched network.

CIRCUIT-SWITCHED NETWORKS:

circuit-switched network consists of a set of switches connected by physical links. A connection between two stations is a dedicated path made of one or more links. However, each connection uses only one dedicated channel on each link. Each link is normally divided into n channels by using FDM or TDM.

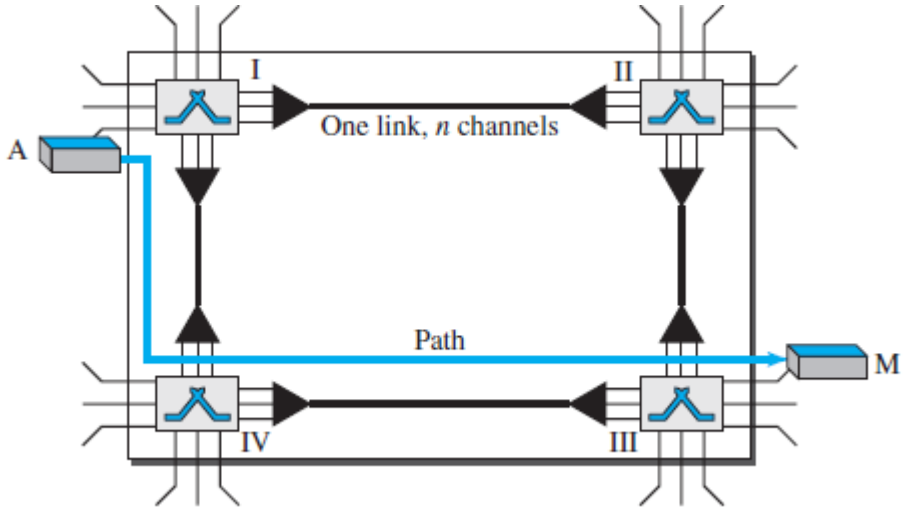
Three Phases:

The actual communication in a circuit-switched network requires three phases: connection setup, data transfer, and connection teardown.

Setup Phase:

Before the two parties (or multiple parties in a conference call) can communicate, a dedicated circuit (combination of channels in links) needs to be established. The end systems are normally connected through dedicated lines to the switches, so connection setup means creating dedicated channels between the switches. For example, as shown in the figure below when system A needs to connect to system M, it sends a setup request that includes the address of system M, to switch I. Switch I finds a channel between itself and switch IV that can be dedicated for this purpose. Switch I then sends the request to switch IV, which finds a dedicated channel between itself and switch III. Switch III informs system M of system A's intention at this time. In the next step to making a connection, an acknowledgment from system M needs to be sent in the opposite direction to system A. Only after system A receives this acknowledgment is the connection established.

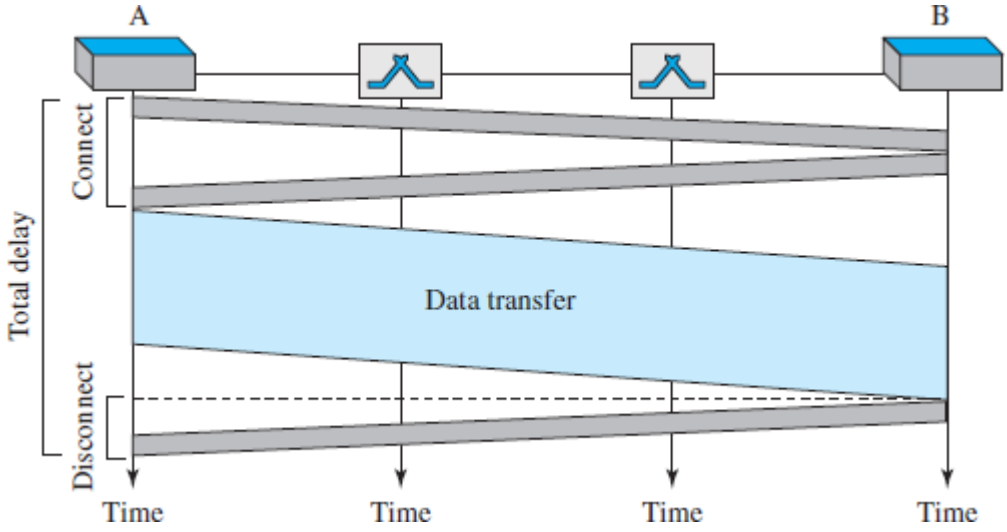
Note that end-to-end addressing is required for creating a connection between the two end systems. These can be, for example, the addresses of the computers assigned by the administrator in a TDM network, or telephone numbers in an FDM network.



Data-Transfer Phase:
After the establishment of the dedicated circuit (channels), the two parties can transfer data.
Teardown Phase:
When one of the parties needs to disconnect, a signal is sent to each switch to release the resources.

Efficiency of circuit switched network:
circuit-switched networks are not as efficient as the other two types of networks because resources are allocated during the entire duration of the connection.These resources are unavailable to other connections. In a telephone network, people normally terminate the communication when they have finished their conversation. However, in computer networks, a computer can be connected to another computer even if there is no activity for a long time. In this case, allowing resources to be dedicated means that other connections are deprived.

Delay:
Although a circuit-switched network normally has low efficiency, the delay in this type of network is minimal. During data transfer the data are not delayed at each switch; the resources are allocated for the duration of the connection.



As above Figure shows, there is no waiting time at each switch. The total delay is due to the time needed to create the connection, transfer data, and disconnect the circuit. The

delay caused by the setup is the sum of four parts: the propagation time of the source computer request (slope of the first gray box), the request signal transfer time (height of the first gray box), the propagation time of the acknowledgment from the destination computer (slope of the second gray box), and the signal transfer time of the acknowledgment (height of the second gray box). The delay due to data transfer is the sum of two parts: the propagation time (slope of the colored box) and data transfer time (height of the colored box), which can be very long. The third box shows the time needed to tear down the circuit. We have shown the case in which the receiver requests disconnection, which creates the maximum delay.

PACKET SWITCHING:

In data communications, we need to send messages from one end system to another. If the message is going to pass through a packet-switched network, it needs to be divided into packets of fixed or variable size. The size of the packet is determined by the network and the governing protocol. In packet switching, there is no resource allocation for a packet. This means that there is no reserved bandwidth on the links, and there is no scheduled processing time for each packet. Resources are allocated on demand. The allocation is done on a firstcome, first-served basis. When a switch receives a packet, no matter what the source or destination is, the packet must wait if there are other packets being processed. As with other systems in our daily life, this lack of reservation may create delay. For example, if we do not have a reservation at a restaurant, we might have to wait.

Note:In a packet-switched network, there is no resource reservation; resources are allocated on demand.

We can have two types of packet-switched networks**: datagram networks and virtualcircuit networks.**

Datagram network or X.25 packet switching network:

In a datagram network, each packet is treated independently of all others. Even if a packet is part of a multipacket transmission, the network treats it as though it existed alone. Packets in this approach are referred to as **datagrams**.

A **datagram** (DG) is, at best, vaguely defined by X.25 and, until it is completely outlined, has very limited usefulness. With a DG, users send small packets of data into the network. Each packet is self-contained and travels through the network independent of other packets of the same message by whatever means available. The network does not acknowledge packets, nor does it guarantee successful transmission. However, if a message will fit into a single packet, a DG is somewhat reliable. This is called a **single-packet-per-segment** protocol.

X.25 Packet Format:

A virtual call is the most efficient service offered for a packet network. There are two packet formats used with virtual calls: a **call request packet** and a **data transfer packet**.

9-4-1 Call request packet. the below Figure shows the field format for a call request packet. The delimiting sequence is 01111110 (an HDLC flag), and the error-detection correction mechanism is CRC-16 with ARQ. The link address field and the control field have little use and, therefore, are seldom used with packet networks. The rest of the fields are defined in sequence.

Format identifier: The format identifier identifies whether the packet is a new call request or a previously established call. The format identifier also identifies the packet numbering sequence (either 0–7 or 0–127).

Logical channel identifier (LCI): The LCI is a 12-bit binary number that identifies the source and destination users for a given virtual call. After a source user has gained access to the network and has identified the destination user, they are assigned an LCI. In subsequent packets, the source and destination addresses are unnecessary; only the LCI is needed. When two users disconnect, the LCI is relinquished and can be reassigned to new users. There are 4096 LCIs available. Therefore, there may be as many as 4096 virtual calls established at any given time.

Packet type: This field is used to identify the function and the content of the packet (new request, call clear, call reset, and so on).

Calling address length: This four-bit field gives the number of digits (in binary) that appear in the calling address field. With four bits, up to 15 digits can be specified.

Called address length: This field is the same as the calling address field except that it identifies the number of digits that appear in the called address field.

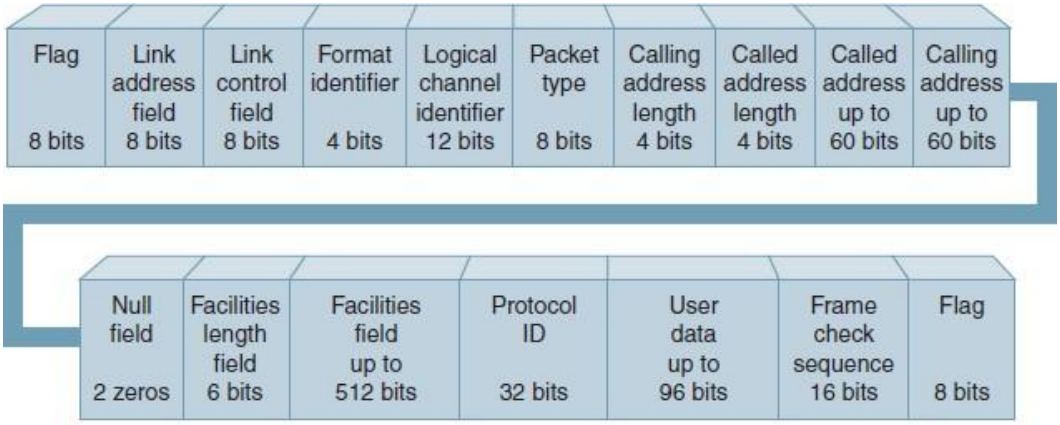
Called address: This field contains the destination address. Up to 15 BCD digits (60 bits) can be assigned to a destination user.

Calling address: This field is the same as the called address field except that it contains up to 15 BCD digits that can be assigned to a source user.

Facilities length field:This field identifies (in binary) the number of eight-bit octets present in the facilities field.

Facilities field: This field contains up to 512 bits of optional network facility information, such as reverse billing information, closed user groups, and whether it is a simplex transmit or simplex receive connection.

Protocol identifier: This 32-bit field is reserved for the subscriber to insert user-level protocol functions such as log-on procedures and user identification practices. User data field. Up to 96 bits of user data can be transmitted with a call request packet. These are unnumbered data that are not confirmed. This field is generally used for user passwords.



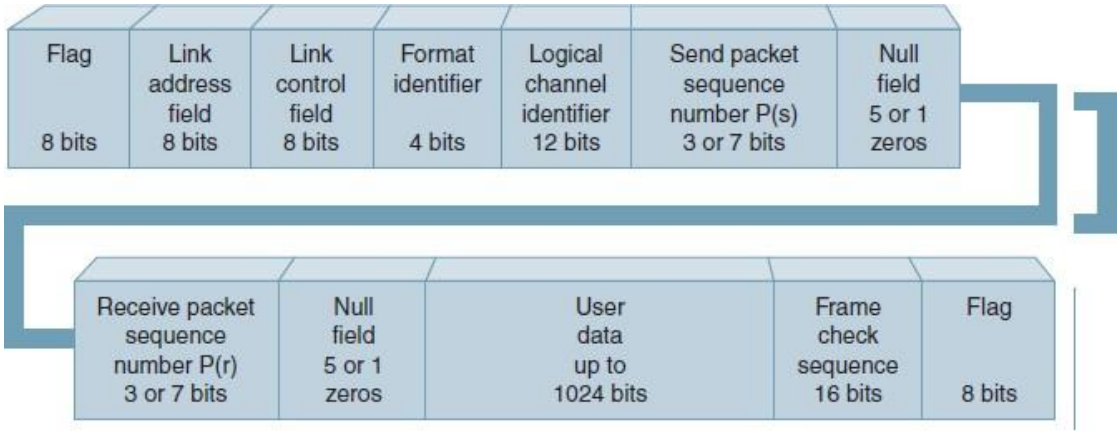
X.25 call request packet format

9-4-2 <u>Data transfer packet:</u>

The below Figure shows the field format for a data transfer packet. A data transfer packet is similar to a call request packet except that a data transfer packet has considerably less overhead and can accommodate a much larger user data field. The data transfer packet contains a send-and-receive packet sequence field that was not included with the call request format. The flag, link address, link control, format identifier, LCI, and FCS fields are identical to those used with the call request packet. The send and receive packet sequence fields are described as follows:

<u>Send packet sequence field:</u> The P(s) field is used in the same manner that the ns and nr sequences are used with SDLC and HDLC. P(s) is analogous to ns, and P(r) is analogous to nr. Each successive data transfer packet is assigned the next P(s) number in sequence. The P(s) can be a 14- or seven-bit binary number and, thus, number packets from either 0–7 or 0–127. The numbering sequence is identified in the format identifier. The send packet field always contains eight bits, and the unused bits are reset.

<u>Receive packet sequence field:</u> P(r) is used to confirm received packets and call for retransmission of packets received in error (ARQ). The I field in a data transfer packet can have considerably more source information than an I field in a call request packet.

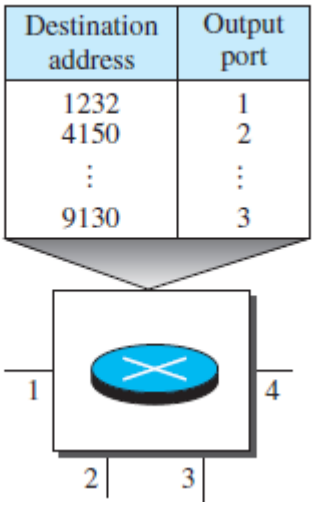


X.25 data transfer packet format

The datagram networks are sometimes referred to as connectionless networks. The

term connectionless here means that the switch (packet switch) does not keep information about the connection state. There are no setup or teardown phases. Each packet is treated the same by a switch regardless of its source or destination

Routing Table:
If there are no setup or teardown phases, how are the packets routed to their destinations in a datagram network? In this type of network, each switch (or packet switch) has a routing table which is based on the destination address. The routing tables are dynamic and are updated periodically. The destination addresses and the corresponding forwarding output ports are recorded in the tables. This is different from the table of a circuitswitched network (discussed later) in which each entry is created when the setup phase is completed and deleted when the teardown phase is over. Figure below shows the routing table for a switch.



A switch in a datagram network uses a routing table that is based on the destination address.

Destination Address:

Every packet in a datagram network carries a header that contains, among other information, the destination address of the packet. When the switch receives the packet, this destination address is examined; the routing table is consulted to find the corresponding port through which the packet should be forwarded.

Congestion:
As Internet can be considered as a Queue of packets, where transmitting nodes are constantly adding packets and some of them (receiving nodes) are removing packets from the queue. So, consider a situation where too many packets are present in this queue (or internet or a part of internet), such that constantly transmitting nodes are pouring packets at a higher rate than receiving nodes are removing them. This degrades the performance, and such a situation is termed as Congestion. Main reason of congestion is more number of packets into the network than it can handle. So, the objective of congestion control can be summarized as to maintain the number of packets in the network below the level at which performance falls off dramatically.

The nature of a Packet switching network can be summarized in following points:

- A network of queues
- At each node, there is a queue of packets for each outgoing channel
- If packet arrival rate exceeds the packet transmission rate, the queue size grows without bound
- When the line for which packets are queuing becomes more than 80% utilized, the queue length grows alarmingly.

When the number of packets dumped into the network is within the carrying capacity, they all are delivered, expect a few that have too be rejected due to transmission errors). And then the number delivered is proportional to the number of packets sent. However, as traffic increases too far, the routers are no longer able to cope, and they begin to lose packets. This tends to make matter worse. At very high traffic, performance collapse completely, and almost no packet is delivered. In the following sections, the causes of congestion, the effects of congestion and various congestion control techniques are discussed in detail.

Causes Of Congestion:

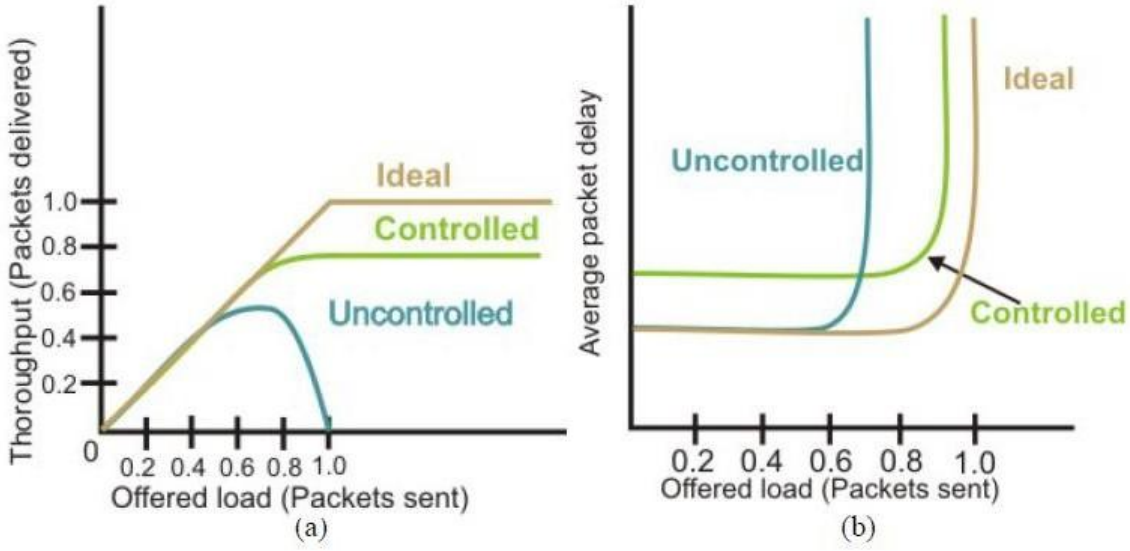
Congestion can occur due to several reasons. For example, if all of a sudden a stream of packets arrive on several input lines and need to be out on the same output line, then a long queue will be build up for that output. If there is insufficient memory to hold these packets, then packets will be lost (dropped). Adding more memory also may not help in certain situations. If router have an infinite amount of memory even then instead of congestion being reduced, it gets worse; because by the time packets gets at the head of the queue, to be dispatched out to the output line, they have already timed-out (repeatedly), and duplicates may also be present. All the packets will be forwarded to next router up to the destination, all the way only increasing the load to the network more and more. Finally when it arrives at the destination, the packet will be discarded, due to time out, so instead of been dropped at any intermediate router (in case memory is restricted) such a packet goes all the way up to the destination, increasing the network load throughout and then finally gets dropped there.

Slow processors also cause Congestion. If the router CPU is slow at performing the task required for them (Queuing buffers, updating tables, reporting any exceptions etc.), queue can build up even if there is excess of line capacity. Similarly, LowBandwidth lines can also cause congestion. Upgrading lines but not changing slow processors, or vice-versa, often helps a little; these can just shift the bottleneck to some other point. The real problem is the mismatch between different parts of the system. Congestion tends to feed upon itself to get even worse. Routers respond to overloading by dropping packets. When these packets contain TCP segments, the segments don't reach their destination, and they are therefore left unacknowledged, which eventually leads to timeout and retransmission.
So, the major cause of congestion is often the bursty nature of traffic. If the hosts could be made to transmit at a uniform rate, then congestion problem will be less common and all other causes will not even led to congestion because other causes just act as an enzyme which boosts up the congestion when the traffic is bursty (i.e., other causes just add on to make the problem more serious, main cause is the bursty traffic).
This means that when a device sends a packet and does not receive an acknowledgment from the receiver, in most the cases it can be assumed that the packets have been dropped by intermediate devices due to congestion. By detecting the rate at which segments are sent and not acknowledged, the source or an intermediate router can infer the level of congestion on the network. In the following section we shall discuss the

ill effects of congestion.

Effects of Congestion:

Congestion affects two vital parameters of the network performance, namely throughput and delay. In simple terms, the throughput can be defined as the percentage utilization of the network capacity. Initially throughput increases linearly with offered load, because utilization of the network increases. However, as the offered load increases beyond certain limit, say 60% of the capacity of the network, the throughput drops. If the offered load increases further, a point is reached when not a single packet is delivered to any destination, which is commonly known as deadlock situation. There are three curves in Fig. 7.5.1(a), the ideal one corresponds to the situation when all the packets introduced are delivered to their destination up to the maximum capacity of the network. The second one corresponds to the situation when there is no congestion control. The third one is the case when some congestion control technique is used. This prevents the throughput collapse, but provides lesser throughput than the ideal condition due to overhead of the congestion control technique. The delay also increases with offered load, as shown in Fig. 7.5.1(b). And no matter what technique is used for congestion control, the delay grows without bound as the load approaches the capacity of the system. It may be noted that initially there is longer delay when congestion control policy is applied. However, the network without any congestion control will saturate at a lower offered load.



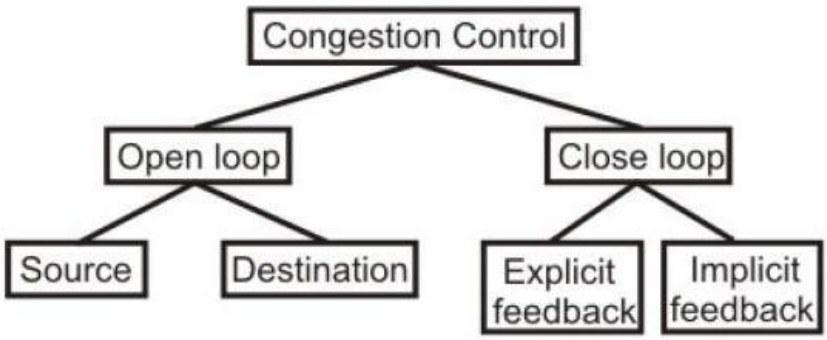
(a) Effect of congestion on throughput (b) Effect of congestion on delay

Congestion Control Techniques:

Congestion control refers to the mechanisms and techniques used to control congestion and keep the traffic below the capacity of the network. Congestion control techniques can be broadly classified two broad categories:

- Open loop: Protocols to prevent or avoid congestion, ensuring that the system (or network under consideration) never enters a Congested State.

- Close loop: Protocols that allow system to enter congested state, detect it, and remove it.



The first category of solutions or protocols attempt to solve the problem by a good design, at first, to make sure that it doesn't occur at all. Once system is up and running midcourse corrections are not made. These solutions are somewhat static in nature, as the policies to control congestion don't change much according to the current state of the system. Such Protocols are also known as Open Loop solutions. These rules or policies include deciding upon when to accept traffic, when to discard it, making scheduling decisions and so on. Main point here is that they make decision without taking into consideration the current state of the network. The open loop algorithms are further divided on the basis of whether these acts on source versus that act upon destination.

The second category is based on the concept of feedback. During operation, some system parameters are measured and feed back to portions of the subnet that can take action to reduce the congestion. This approach can be divided into 3 steps:

- Monitor the system (network) to detect whether the network is congested or not and what's the actual location and devices involved.
- To pass this information to the places where actions can be taken
- Adjust the system operation to correct the problem.

These solutions are known as Closed Loop solutions. Various Metrics can be used to monitor the network for congestion. Some of them are: the average queue length, number of packets that are timed-out, average packet delay, number of packets discarded due to lack of buffer space, etc. A general feedback step would be, say a router, which detects the congestion send special packets to the source (responsible for the congestion) announcing the problem. These extra packets increase the load at that moment of time, but are necessary to bring down the congestion at a later time. Other approaches are also used at times to curtail down the congestion. For example, hosts or routers send out probe packets at regular intervals to explicitly ask about the congestion and source itself regulate its transmission rate, if congestion is detected in the network. This kind of approach is a pro-active one, as source tries to get knowledge about congestion in the network and act accordingly.

Yet another approach may be where instead of sending information back to the source an intermediate router which detects the congestion send the information about the congestion to rest of the network, piggy backed to the outgoing packets. This approach will in no way put an extra load on the network (by not sending any kind of special packet for feedback). Once the congestion has been detected and this information has been passed to a place where the action needed to be done, then there are two basic approaches that can overcome the problem. These are: either to increase the resources or to decrease the load. For example,

separate dial-up lines or alternate links can be used to increase the bandwidth between two points, where congestion occurs. Another example could be to decrease the rate at which a particular sender in transmitting packets out into the network.

The closed loop algorithms can also be divided into two categories, namely explicit feedback and implicit feedback algorithms. In the explicit approach, special packets are sent back to the sources to curtail down the congestion. While in implicit approach, the source itself acts pro-actively and tries to deduce the existence of congestion by making local observations.

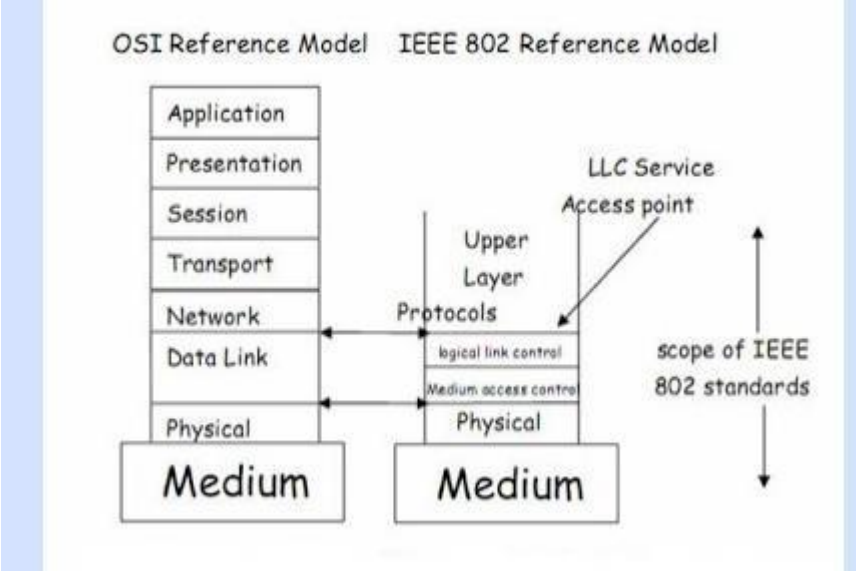
Chapter-6

LAN protocol architecture

The LAN protocol architecture consists of layering of protocols that contribute to the basic functions of a LAN. The standardized LAN protocol architecture encompasses 3 layers. They are Physical layer, Medium Access control layer (MAC), and Logical Link control layers. The physical layer deals with the topology and transmission medium.

IEEE 802 Reference Model:

The protocols of OSI Reference Model can be classified as Network Support Layers(Physical, Data link, Network) and User support layers(Session, Presentation, Application). LAN protocols are concerned with the network support layers i.e., the lower layers. Protocols defined specifically for LAN and MAN transmission address issues related to the transmission of blocks of data over the network. The higher layers of the OSI model are independent of network architecture and are applicable to LAN, MAN and WANs. Therefore LAN protocols are related to the lower layers of the OSI model.



As the figure shows, the lowest layer of the IEEE 802 reference model corresponds to the physical layer of the OSI model. This layer has the following functions.

- Encoding / Decoding of signals.
- Preamble generation / removal.
- Bit transmission reception. It also includes a specification of the transmission medium and topology. The layers above the physical layers are
- Logical Link control layer.
- Medium Access control layer.

The logical link layer's functions are Provide an interface to the higher layers and perform flow control and error control.

1. At the sender, assemble data into a frame with address and error detection fields.
2. At the receiver, disassemble the received frame and perform address recognition and error detection.
3. Govern access to the LAN transmission medium

Of these, the last three functions are treated as a separate layer called, Medium access control layer.

Relationship between the levels:

As the figure shows data from the application layer is passed to the TCP layer along with its header. TCP layer appends its header to the data it received and passes it to the IP layer. It appends its header to the data it received and passes it to the Logical Link Control layer. LLC appends its header creating a Logical Link Control Protocol Data Unit(LLC PDU).This entire LLC protocol data unit is passed on to MAC layer. The MAC layer adds a header at the front of the data and adds a trailer at the back. This is called the MAC Frame.

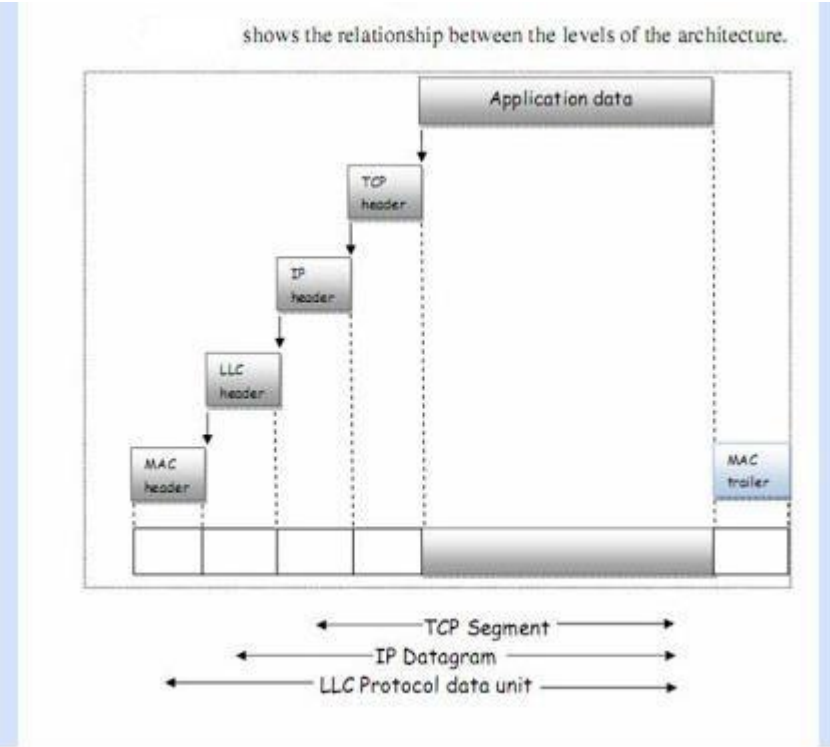


Fig shows Relationship between the levels of the architecture

It is important to note here that the headers are added to the data units to include control information such as, sequence number, source address, destination address etc., In transmission, the control information is necessary for the operation of a protocol.

Logical Link Control (LLC):

LLC is concerned with the transmission of a link level PDU between two stations. There is no need for an intermediate switching node here. LLC has two unique characteristics.

1. There is no primary node involved here.

2. Half of the Link access details are taken care of by the MAC layer.

The LLC user addresses are called Service Access Points (SAPs). Generally LLC users are the higher layer protocols or some network management function.

LLC Services:

The Logical Link Control specifies the following mechanisms

1. for addressing stations across the medium
2. for controlling the exchange of data between two users.

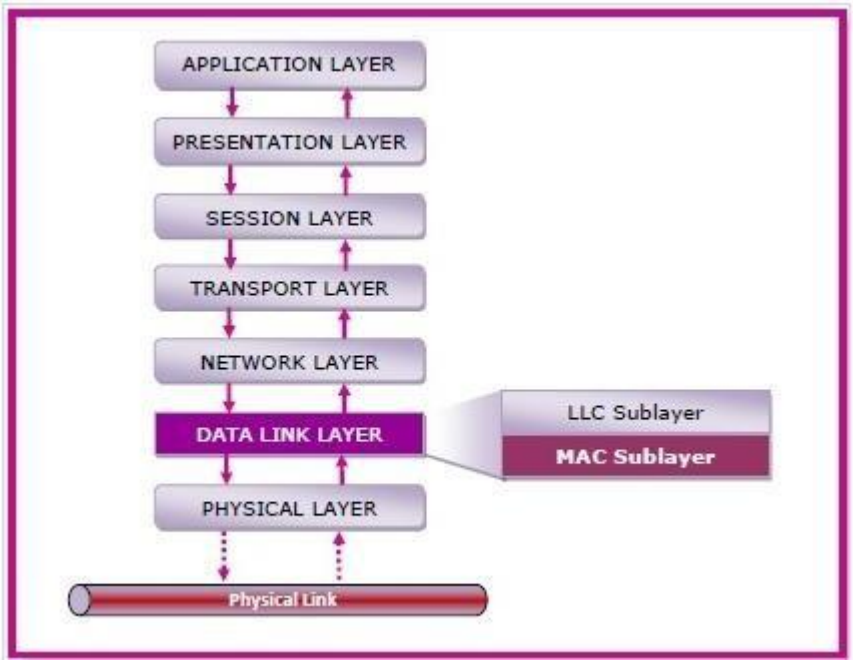
There are three types of services provided by the Logical Link Layer.
They are
1. Connection mode service – A logical connection is set up between two users.
2. Acknowledged connectionless service- No connections are involved. But data grams are acknowledged.
3. Unacknowledged connectionless service- Operates in datagram style. No guarantee for the delivery of data. The service does not include flow control or error control mechanisms.

Medium Access control:

The medium access control (MAC) is a sublayer of the data link layer of the open system interconnections (OSI) reference model for data transmission. It is responsible for flow control and multiplexing for transmission medium. It controls the transmission of data packets via remotely shared channels. It sends data over the network interface card.

MAC Layer in the OSI Model:
The Open System Interconnections (OSI) model is a layered networking framework that conceptualizes how communications should be done between heterogeneous systems. The data link layer is the second lowest layer. It is divided into two sublayers

- The logical link control (LLC) sublayer.
- The medium access control (MAC) sublayer.



Functions of MAC Layer:

- It provides an abstraction of the physical layer to the LLC and upper layers of the OSI network.
- It is responsible for encapsulating frames so that they are suitable for transmission via the physical medium.
- It resolves the addressing of source station as well as the destination station, or groups of destination stations.

- It performs multiple access resolutions when more than one data frame is to be transmitted. It determines the channel access methods for transmission.

- It also performs collision resolution and initiating retransmission in case of collisions.

- It generates the frame check sequences and thus contributes to protection against transmission errors.

MAC Addresses:

MAC address or media access control address is a unique identifier allotted to a network interface controller (NIC) of a device. It is used as a network address for data transmission within a network segment like Ethernet, Wi-Fi, and Bluetooth.

MAC address is assigned to a network adapter at the time of manufacturing. It is hardwired or hard-coded in the network interface card (NIC). A MAC address comprises of six groups of two hexadecimal digits, separated by hyphens, colons, or no separators. An example of a MAC address is 00:0A:89:5B:F0:11.

Network Devices (Hub, Repeater, Bridge, Switch, Router, Gateways and Brouter):

Repeater – A repeater operates at the physical layer. Its job is to regenerate the signal over the same network before the signal becomes too weak or corrupted so as to extend the length to which the signal can be transmitted over the same network. An important point to be noted about repeaters is that they do not amplify the signal. When the signal becomes weak, they copy the signal bit by bit and regenerate it at the original strength. It is a 2 port device.

Hub – A hub is basically a multiport repeater. A hub connects multiple wires coming from different branches, for example, the connector in star topology which connects different stations. Hubs cannot filter data, so data packets are sent to all connected devices. In other words, collision domain of all hosts connected through Hub remains one. Also, they do not have intelligence to find out best path for data packets which leads to inefficiencies and wastage.

Types of Hub:

Active Hub:- These are the hubs which have their own power supply and can clean, boost and relay the signal along with the network. It serves both as a repeater as well as wiring centre. These are used to extend the maximum distance between nodes.

Passive Hub :- These are the hubs which collect wiring from nodes and power supply from active hub. These hubs relay signals onto the network without cleaning and boosting them and can't be used to extend the distance between nodes.

Bridge – A bridge operates at data link layer. A bridge is a repeater, with add on the functionality of filtering content by reading the MAC addresses of source and destination. It is also used for interconnecting two LANs working on the same protocol. It has a single input and single output port, thus making it a 2 port device.

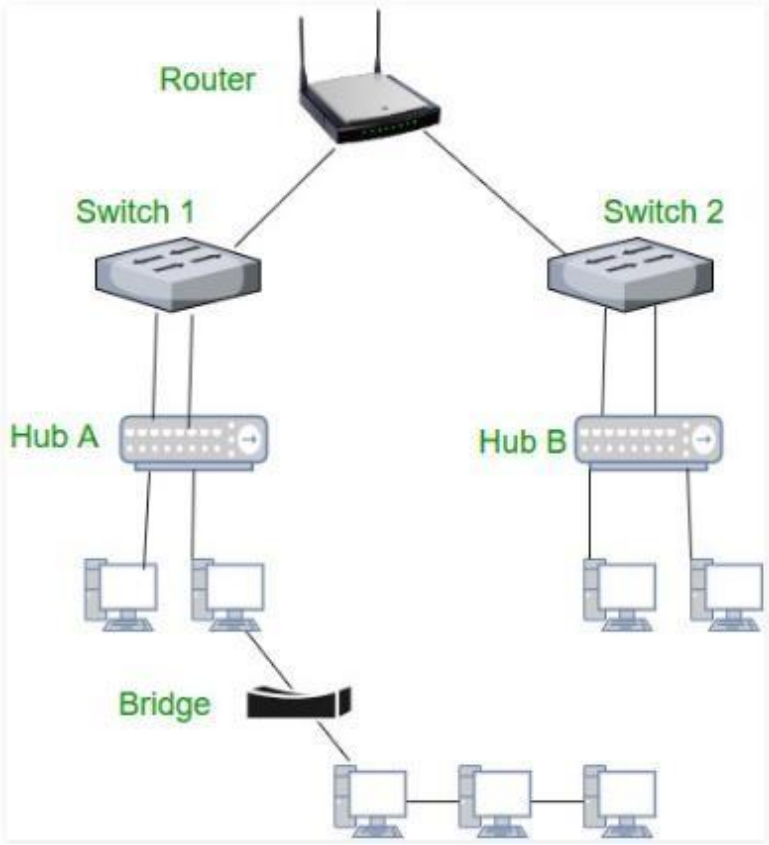
Types of Bridges:

Transparent Bridges:- These are the bridge in which the stations are completely unaware of the bridge's existence i.e. whether or not a bridge is added or deleted from the network, reconfiguration of the stations is unnecessary. These bridges make use of two processes i.e. bridge forwarding and bridge learning.

Source Routing Bridges:- In these bridges, routing operation is performed by source station and the frame specifies which route to follow. The hot can discover frame by sending a special frame called discovery frame, which spreads through the entire network using all possible paths to destination

Switch – A switch is a multiport bridge with a buffer and a design that can boost its efficiency(a large number of ports imply less traffic) and performance. A switch is a data link layer device. The switch can perform error checking before forwarding data, that makes it very efficient as it does not forward packets that have errors and forward good packets selectively to correct port only. In other words, switch divides collision domain of hosts, but broadcast domain remains same.

Routers – A router is a device like a switch that routes data packets based on their IP addresses. Router is mainly a Network Layer device. Routers normally connect LANs and WANs together and have a dynamically updating routing table based on which they make decisions on routing the data packets. Router divide broadcast domains of hosts connected through it.



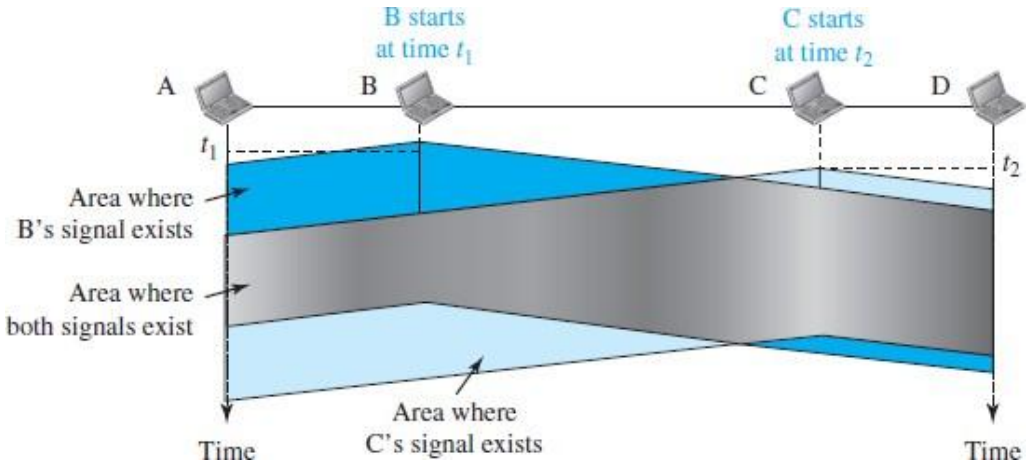
Gateway – A gateway, as the name suggests, is a passage to connect two networks together that may work upon different networking models. They basically work as the messenger agents that take data from one system, interpret it, and transfer it to another system. Gateways are also called protocol converters and can operate at any network layer. Gateways are generally more complex than switch or router.

Brouter – It is also known as bridging router is a device which combines features of both bridge and router. It can work either at data link layer or at network layer. Working as router, it is capable of routing packets across networks and working as bridge, it is capable of filtering local area network traffic.

CSMA:

To minimize the chance of collision and, therefore, increase the performance, the CSMA method was developed. The chance of collision can be reduced if a station senses the medium before trying to use it. Carrier sense multiple access (CSMA) requires that each station first listen to the medium (or check the state of the medium) before sending. In other words, CSMA is based on the principle "sense before transmit" or "listen before talk. CSMA can reduce the possibility of collision, but it cannot eliminate it.

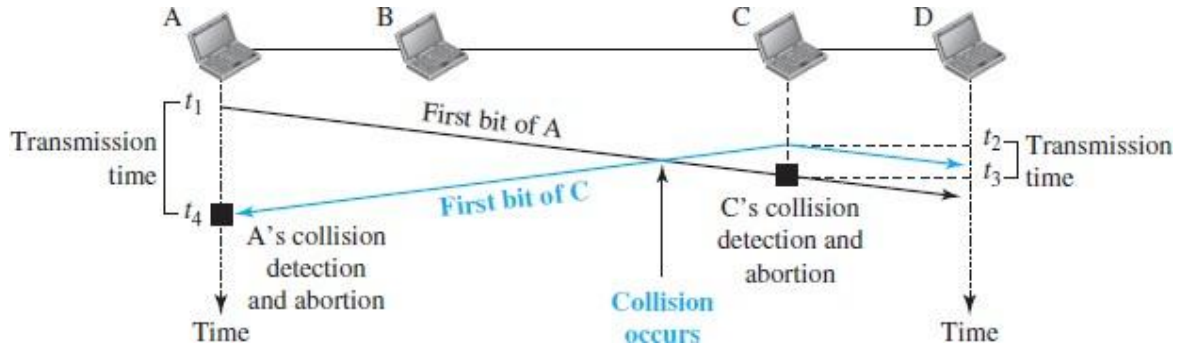
Stations are connected to a shared channel (usually a dedicated medium).
The possibility of collision still exists because of propagation delay; when a station sends a frame, it still takes time (although very short) for the first bit to reach every station and for every station to sense it. In other words, a station may sense the medium and find it idle, only because the first bit sent by another station has not yet been received.



At time t1, station B senses the medium and finds it idle, so it sends a frame. At time t2 (t2 > t1), station C senses the medium and finds it idle because, at this time, the first bits from station B have not reached station C. Station C also sends a frame. The two signals collide and both frames are destroyed.

CSMA/CD:

The CSMA method does not specify the procedure following a collision. Carrier sense multiple access with collision detection (CSMA/CD) augments the algorithm to handle the collision. In this method, a station monitors the medium after it sends a frame to see if the transmission was successful. If so, the station is finished. If, however, there is a collision, the frame is sent again. To better understand CSMA/CD, let us look at the first bits transmitted by the two stations involved in the collision. Although each station continues to send bits in the frame until it detects the collision, we show what happens as the first bits collide. In Figure stations A and C are involved in the collision.



At time t1, station A has executed its persistence procedure and starts sending the bits of its frame. At time t2, station C has not yet sensed the first bit sent by A. Station C executes its persistence procedure and starts sending the bits in its frame, which propagate both to the left and to the right. The collision occurs sometime after time t2. Station C detects a collision at time t3 when it receives the first bit of A's frame. Station C immediately (or after a short time, but we assume immediately) aborts transmission. Station A detects collision at time t4 when it receives the first bit of C's frame; it also immediately aborts transmission. Looking at the figure, we see that A transmits for the duration t4 − t1; C transmits for the duration t3 − t2.

Fiber Channel protocol:

The Fibre Channel Protocol (FCP) is one of the communication protocols designed to carry serial SCSI-3 data over an optical fiber network. The throughput of a Fibre Channel network can provide from 100 MB/s to 1.6 GB/s and the distance can be extended from 500 meters to 10 kilometers. The max number of devices for FC-SW is 16,777,216 or 224.

Because Fibre Channel (FC) is a common data transmission protocol, it is capable of working with many current network protocols and interface or I/O interface protocols, such as:

Network protocols:

- IP

- IEEE 802.2 (MAC)

- Enterprise Systems Connection (ESCON) for Mainframe

- Asynchronous Transfer Mode (ATM)

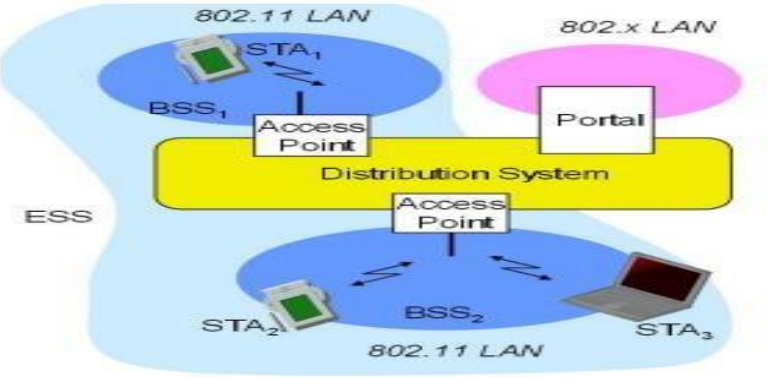
Wireless LAN Technology:

There are two major approaches today for deploying WLAN networks in the enterprise. Both approaches require Wireless 802.11 based Access Points (APs) and some method for managing these network elements. However, the two approaches have some basic philosophical differences which can have a major impact on deployment costs, security and manageability. The first architecture to be presented is the socalled "Centralized" WLAN Architecture. The Centralized Architecture requires one or more servers or special purpose switches to be deployed in conjunction with Wireless Access Points. In the Centralized approach, all wireless traffic is sent through the WLAN switch. Another approach is the "Distributed" Access Point WLAN Architecture. The Distributed Architecture adheres closely to the principles of the IEEE 802.11 standard. In the Distributed approach, APs have built-in WLAN security, layer 2 bridging, and access control features. Depending on the number of APs required, Centralized Management may be required. Distributed AP vendors may provide Centralized Management tools or the APs may be managed by the existing Network Management Infrastructure. The AP is connected directly to the trusted wired infrastructure and "extends" the wired network by providing wireless connections to wireless client devices.

One of the advantages of the Distributed or Wireless Extension approach is that the wireless traffic load is literally distributed across the APs and does not depend on a centralized element to process all of the wireless traffic. In the Centralized approach, loss of the WLAN Switch results in loss of the wireless network, whereas with the Distributed architecture, there is no single point of failure. From a performance point of view, the Distributed Architecture is superior from a performance/efficiency point of view. This is because the Centralized approach requires all wireless packets to be processed by the centralized WLAN Switch whereas in the Distributed Architecture, the packets are handled by the APs and only management traffic needs to go to and from a central point. Centralized Architectures tend to be difficult to scale because each WLAN switch can only handle a limited number of APs.

WLAN ARCHITECTURES

1. Station (STA) terminal with access mechanisms to the wireless medium and radio Contact to the access point.
2. Access Point (or Base Station) station integrated into the wireless LAN and the distribution system.



3. Basic Service Set A BSS is a set of stations that communicate with one another. A BSS does not generally refer to a particular area, due to the uncertainties of electromagnetic propagation. When all of the stations int the BSS are mobile stations and there is no connection to a wired network, the BSS is called independent BSS (IBSS). IBSS is typically short-lived network, with a small number of stations, that is created for a particular purpose. When a BSS includes an access point (AP), the BSS is called infrastructure BSS. When there is a AP, If one mobile station in the BSS must communicate with another mobile station, the communication is sent first to the AP and then from the AP to the other mobile station. This consume twice the bandwidth that the same communication. While this appears to be a significant cost, the benefits provided by the AP far outweigh this cost. One of them is, AP buffers the traffic of mobile while that station is operating in a very low power state
4. Extended Service Set (ESS):- A ESS is a set of infrastructure BSSs, where the APs communicate among themselves to forward traffic from one BSS to another and to facilitate the movement of mobile stations from one BSS to another. The APs perform this communication via an abstract medium called the distribution system (DS). To network equipment outside of the ESS, the ESS and all of its mobile stations appears to be a single MAC-layer network where all stations are physically stationary. Thus, the ESS hides the mobility of the mobile stations from everything outside the ESS.
5. Distribution System:- the distribution system (DS) is the mechanism by which one AP communicates withanother to exchange frames for stations in their BSSs, forward frames to follow mobile stations from one BSS to another, and exchange frames with wired network. Infrastructure wireless LAN is a term often referred to wireless LANs that deploy AP, with the infrastructure being the APs along with wired Ethernet infrastructure that connects APs and router, hub or switch

Chapter-7

TCP/IP Protocol Suite:

The TCP/IP protocol model for internetwork communications was created in the early 1970s and is sometimes referred to as the internet model. This type of model closely matches the structure of a particular protocol suite. The TCP/IP model is a protocol model because it describes the functions that occur at each layer of protocols within the TCP/IP suite. TCP/IP is also used as a reference model. The table shows details about each layer of the OSI model.

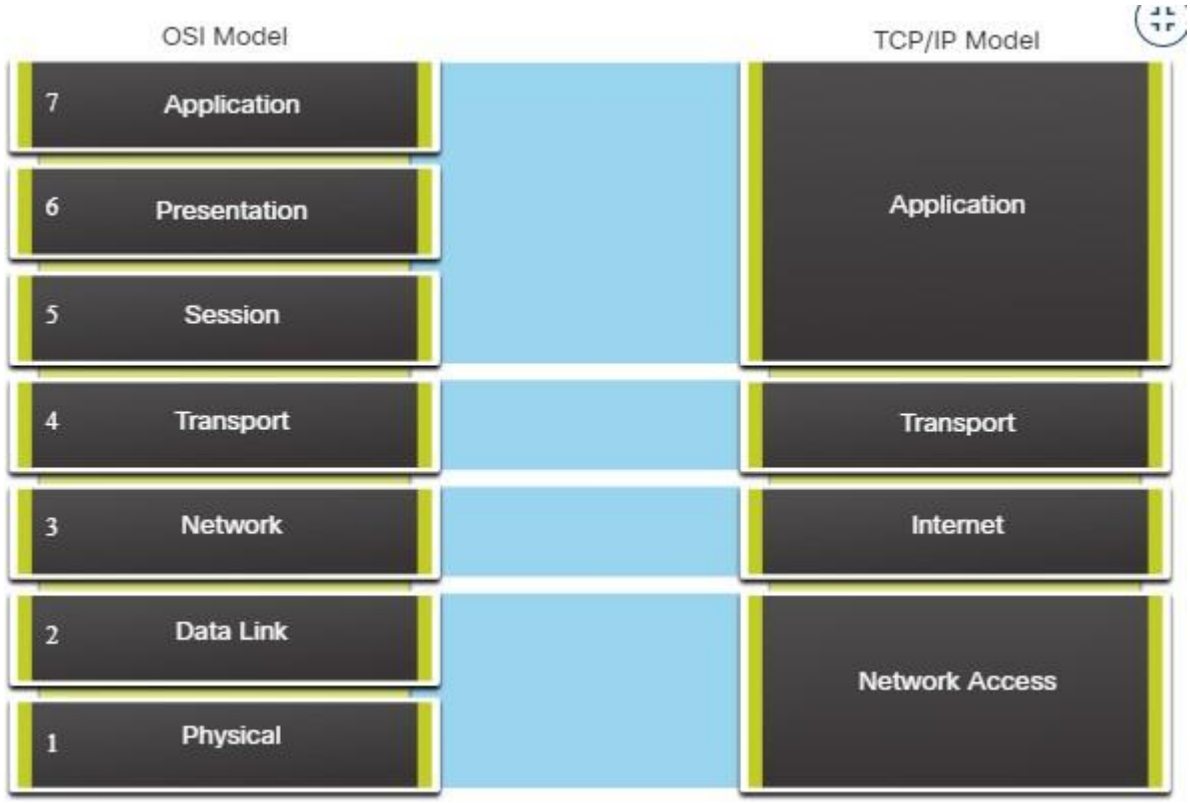
TCP/IP Model Layer	Description
4 - Application	Represents data to the user, plus encoding and dialog control.
3 - Transport	Supports communication between various devices across diverse networks.
2 - Internet	Determines the best path through the network.
1 - Network Access	Controls the hardware devices and media that make up the network.

The protocols that make up the TCP/IP protocol suite can also be described in terms of the OSI reference model. In the OSI model, the network access layer and the application layer of the TCP/IP model are further divided to describe discrete functions that must occur at these layers.

At the network access layer, the TCP/IP protocol suite does not specify which protocols to use when transmitting over a physical medium; it only describes the handoff from the internet layer to the physical network protocols. OSI Layers 1 and 2 discuss the necessary procedures to access the media and the physical means to send data over a network.

The key similarities are in the transport and network layers; however, the two models differ in how they relate to the layers above and below each layer:

OSI Layer 3, the network layer, maps directly to the TCP/IP internet layer. This layer is used to describe protocols that address and route messages through an internetwork.
OSI Layer 4, the transport layer, maps directly to the TCP/IP transport layer. This layer describes general services and functions that provide ordered and reliable delivery of data between source and destination hosts.
The TCP/IP application layer includes several protocols that provide specific functionality to a variety of end user applications. The OSI model Layers 5, 6, and 7 are used as references for application software developers and vendors to produce applications that operate on networks.
Both the TCP/IP and OSI models are commonly used when referring to protocols at various layers. Because the OSI model separates the data link layer from the physical layer, it is commonly used when referring to these lower layers.



Principles of Internetworking:

Internetworking stands for connectivity and communication between two or mornetworks.
Internetwork (internet): a collection of communication networks interconnected by bridges, switches and/or routers.

Requirements for Internetworking:
1. Provide a link between networks. At minimum, a physical and link control connection is needed.

2. Provide for the routing and delivery of data between processes on different networks.

3. Provide an accounting service that keeps track of the use of the various networks and routers and maintains status information.

4. Provide the services just listed without requiring modifications to the networking architecture of constituent networks. This means accommodating the following differences

- Different addressing schemes: e.g., naming (DNS), DHCP.
- Different maximum packet size: e.g., segmentation, ATM cells.
- Different network access mechanisms: e.g., Ethernet, FDDI, ATM.
- Different timeouts: longer with multiple networks.
- Different error recovery services: some networks will have it, others won't. Internetwork error recovery should be independent of individual networks.
- Different status reporting: how and whether this information can be shared.
- Different routing techniques: may depend on fault detection and congestion control techniques. Coordination is needed.
- Different user access control: authorization for use of the network.
- Connection-oriented vs. connectionless

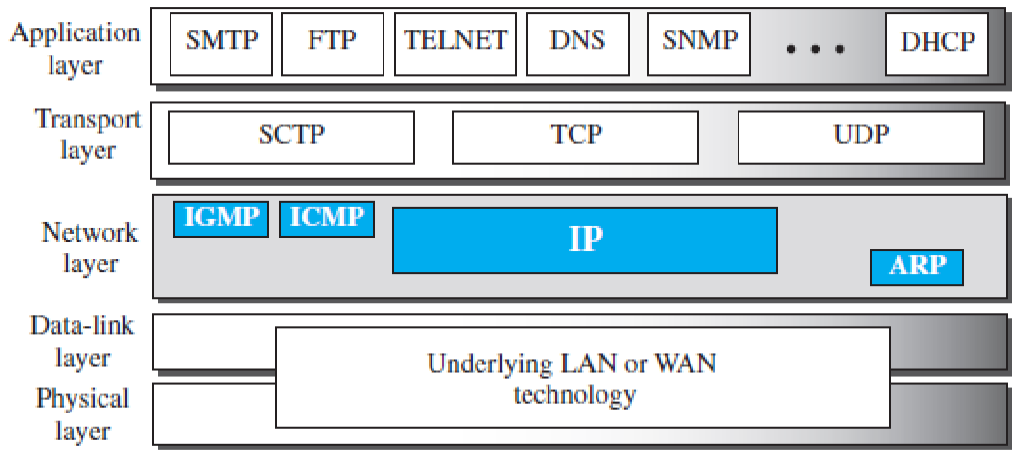
Internet Protocol operations

The network layer, or OSI Layer 3, provides services to allow end devices to exchange data across networks. As shown in the figure, IP version 4 (IPv4) and IP version 6 (IPv6) are the principle network layer communication protocols. Other network layer protocols include routing protocols such as Open Shortest Path First (OSPF) and messaging protocols such as Internet Control Message Protocol (ICMP).

To accomplish end-to-end communications across network boundaries, network layer protocols perform four basic operations:

1. Addressing end devices - End devices must be configured with a unique IP address for identification on the network.
2. Encapsulation - The network layer encapsulates the protocol data unit (PDU) from the transport layer into a packet. The encapsulation process adds IP header information, such as the IP address of the source (sending) and destination (receiving) hosts. The encapsulation process is performed by the source of the IP packet.
3. Routing - The network layer provides services to direct the packets to a destination host on another network. To travel to other networks, the packet must be processed by a router. The role of the router is to select the best path and direct packets toward the destination host in a process known as routing. A packet may cross many routers before reaching the destination host. Each router a packet crosses to reach the destination host is called a hop.
4. De-encapsulation - When the packet arrives at the network layer of the destination host, the host checks the IP header of the packet. If the destination IP address within the header matches its own IP address, the IP header is removed from the packet. After the packet is de-encapsulated by the network layer, the resulting Layer 4 PDU is passed up to the appropriate service at the transport layer. The de-encapsulation process is performed by the destination host of the IP packet.

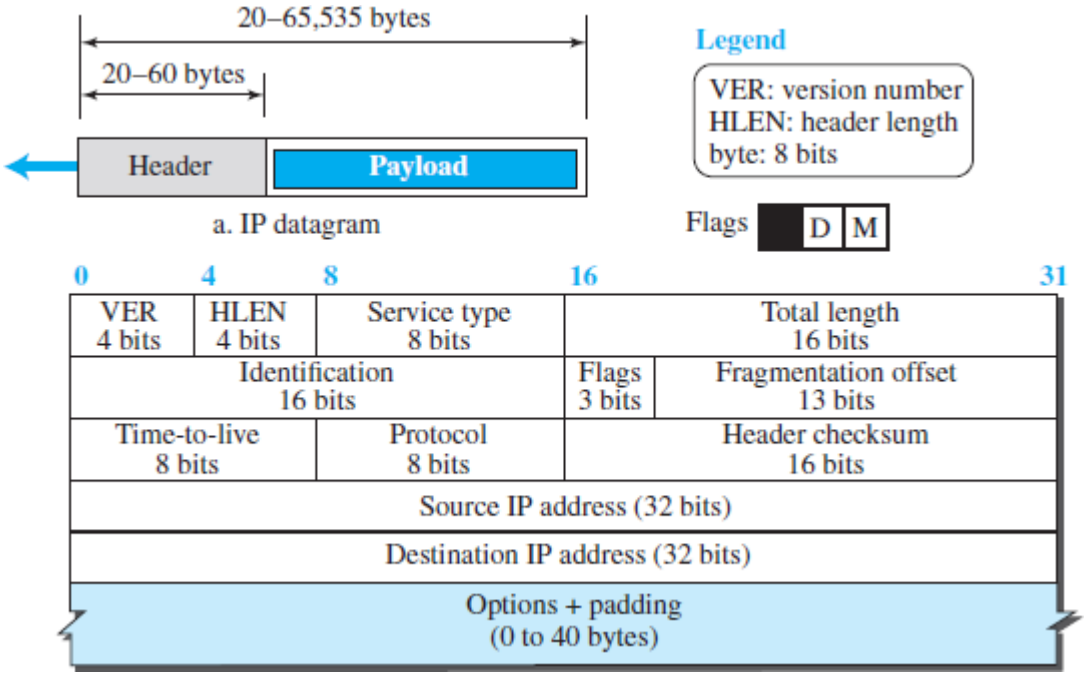
Internet Protocol:



IPv4 is an unreliable datagram protocol—a best-effort delivery service. The term best-effort means that IPv4 packets can be corrupted, be lost, arrive out of order, or be delayed, and may create congestion for the network. If reliability is important, IPv4 must be paired with a reliable transport-layer protocol such as TCP. An example of a more commonly understood best-effort delivery service is the post office. The post office does its best to deliver the regular mail but does not always succeed. If an unregistered letter is lost or damaged, it is up to the sender or would-be recipient to discover this. The post office itself does not keep track of every letter and cannot notify a sender of loss or damage of one. IPv4 is also a connectionless protocol that uses the datagram approach. This means that each datagram is handled independently, and each datagram can follow a different route to the destination. This implies that datagrams sent by the same source to the same destination could arrive out of order. Again, IPv4 relies on a higher-level protocol to take care of all these problems.

Datagram Format

In this section, we begin by discussing the first service provided by IPv4, packetizing. We show how IPv4 defines the format of a packet in which the data coming from the upper layer or other protocols are encapsulated. Packets used by the IP are called datagrams. The below Figure shows the IPv4 datagram format. A datagram is a variable-length packet consisting of two parts: header and payload (data). The header is 20 to 60 bytes in length and contains information essential to routing and delivery. It is customary in TCP/IP to show the header in 4-byte sections



Version Number- The 4-bit version number (VER) field defines the version of the IPv4 protocol, which, obviously, has the value of 4.

Header Length-The 4-bit header length (HLEN) field defines the total length of the datagram header in 4-byte words. The IPv4 datagram has a variable-length header. When a device receives a datagram, it needs to know when the header stops and the data, which is encapsulated in the packet, starts. However, to make the value of the header length (number of bytes) fit in a 4-bit header length, the total length of the header is calculated as 4-byte words. The total length is divided by 4 and the value is inserted in the field. The receiver needs to multiply the value of this field by 4 to find the total length.

Service Type-In the original design of the IP header, this field was referred to as type of service (TOS), which defined how the datagram should be handled. In the late 1990s, IETF redefined the field to provide differentiated services (DiffServ). The use of 4-byte words for the length header isalso logical because the IP header always needs to be aligned in 4-byte boundaries. Total Length. This 16-bit field defines the total length (header plus data) of the IP datagram in bytes. A 16-bit number can define a total length of up to 65,535 (when all bits are 1s). However, the size of the datagram is normally much less than this. This field helps the receiving device to know when the packet has completely arrived. To find the length of the data coming from the upper layer, subtract the header length from the total length. The header length can be found by multiplying the value in the HLEN field by 4.

$$\text{Length of data} = \text{total length} - (\text{HLEN}) \times 4$$

Though a size of 65,535 bytes might seem large, the size of the IPv4 datagram may increase in the near future as the underlying technologies allow even more throughput (greater bandwidth).

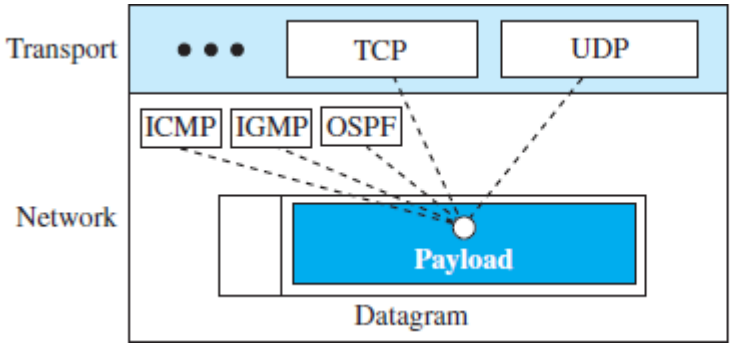
One may ask why we need this field anyway. When a machine (router or host) receives a frame, it drops the header and the trailer, leaving the datagram. Why include an extra field that is not needed? The answer is that in many cases we really do not need the value in this field. However, there are occasions in which the datagram is not the only thing encapsulated in a frame; it may be that padding has been added. For example, the Ethernet protocol has a minimum and maximum restriction on the size of data that can be encapsulated in a frame (46 to 1500 bytes). If the size of an IPv4 datagram is less than 46 bytes, some padding will be added to meet this requirement. In this case, when a machine decapsulates the datagram, it needs to check the total length field to determine how much is really data and how much is padding.

Identification Flags, and Fragmentation Offset:. These three fields are related to the fragmentation of the IP datagram when the size of the datagram is larger than the underlying network can carry. We discuss the contents and importance of these fields when we talk about fragmentation in the next section

Time-to-live. Due to some malfunctioning of routing protocols (discussed later) a datagram may be circulating in the Internet, visiting some networks over and over without reaching the destination. This may create extra traffic in the Internet. The time-to-live (TTL) field is used to control the maximum number of hops (routers) visited by the datagram. When a source host sends the datagram, it stores a number in this field. This value is approximately two times the maximum number of routers between any two hosts. Each router that processes the datagram decrements this number by one. If this value, after being decremented, is zero, the router discards the datagram.

Protocol. In TCP/IP: the data section of a packet, called the payload, carries the whole packet from another protocol. A datagram, for example, can carry a packet belonging to any transport-layer protocol such as UDP or TCP. A datagram can also carry a packet from other protocols that directly use the service of the IP, such as some routing protocols or some auxiliary protocols. The Internet authority has given any protocol that uses the service of IP a unique 8-bit number which is inserted in the protocol field. When the payload is encapsulated in a datagram at the source IP, the corresponding protocol number is inserted in this field; when the datagram arrives at the destination, the value of this field helps to define to which protocol the payload should be delivered. In other words, this field provides multiplexing at the source and demultiplexing at the destination, as shown in below Figure . Note that the protocol fields at the network layer play the same role as the port numbers at

the transport layer). However, we need two port numbers in a transport-layer packet because the port numbers at the source and destination are different, but we need only one protocol field because this value is the same for each protocol no matter whether it is located at the source or the destination.



Some protocol values

ICMP	01
IGMP	02
TCP	06
UDP	17
OSPF	89

Header checksum. IP is not a reliable protocol; it does not check whether the payload carried by a datagram is corrupted during the transmission. IP puts the burden of error checking of the payload on the protocol that owns the payload, such as UDP or TCP. The datagram header, however, is added by IP, and its error-checking is the responsibility of IP. Errors in the IP header can be a disaster. For example, if the destination IP address is corrupted, the packet can be delivered to the wrong host. If the protocol field is corrupted, the payload may be delivered to the wrong protocol. If the fields related to the fragmentation are corrupted, the datagram cannot be reassembled correctly at the destination, and so on. For these reasons, IP adds a header checksum field to check the header, but not the payload. We need to remember that, since the value of some fields, such as TTL, which are related to fragmentation and options, may change from router to router, the checksum needs to be recalculated at each router. As we discussed in Chapter 10, checksum in the Internet normally uses a 16-bit field, which is the complement of the sum of other fields calculated using 1s complement arithmetic.

Source and Destination Addresses. These 32-bit source and destination address fields define the IP address of the source and destination respectively. The source host should know its IP address. The destination IP address is either known by the protocol that uses the service of IP or is provided by the DNS as described in Note that the value of these fields must remain unchanged during the time the IP datagram travels from the source host to the destination host.

Options- A datagram header can have up to 40 bytes of options. Options can be used for network testing and debugging. Although options are not a required part of the IP header, option processing is required of the IP software. This means that all implementations must be able to handle options if they are present in the header. The existence of options in a header creates some burden on the datagram handling; some options can be changed by routers, which forces each router to recalculate the header checksum

Payload. Payload, or data, is the main reason for creating a datagram. Payload is the packet coming from other protocols that use the service of IP. Comparing a datagram to a postal package, payload is the content of the package; the header is only the information written on the package.