❑❑❑

CHAPTER 1

Unit I

Introduction

Syllabus :

Introduction to data communication, Components, Data representation, Data flow, Networks, Network criteria, Physical structures, Network types, Local area network, Wide area network, Switching, The Internet, Accessing the Internet, Standards and administration Internet standards.

1.1 Introduction :

- The communication branch is the oldest branch of the electronics field. Telecommunication means communicating at a distance. A communication system is the means of conveying the information from one place to the other. This information can be of different types such as sound, picture, music, computer data etc.
- The field of communication engineering started developing rapidly in the nineteenth century when the telegraph, telephone and then the radio were invented. The development was still faster in the twentieth century when first the black and white and then colour TVs were brought in use. Then came the age of satellite communication, cable TV, mobile telephones etc.
- In order to understand the subject, it is necessary to understand the basic concepts in communication engineering such as, modulation, noise, demodulation, information theory etc.

1.2 Introduction to Data Communication :

- In this chapter we are going to discuss data communication and networking.
- The aim of data communication and networking is to allow the exchange of data such as audio, text and video between any points in world.
- The transfer of data takes place over a computer network. A network is like a path or a road over which the data travels smoothly from sender to destination.

1.2.1 Definition of Data Communication :

- Before exchanging information, the creaters and the users of data should agree upon how the information should be presented ?
- An information that is presented in such a form is called **data.**
- **Data communication** can be defined as the exchange of data between a source and destination over some kind of transmission medium, such as a co-axial cable, (Wired communication) or air (wireless communication).

1.2.2 Characteristics of Data Communication System :

The three important characteristics of a data communication system are :

1. Delivery 2. Accuracy 3. Timeliness

1. Delivery :

A data communication system (DCS) must deliver data only to the user who is intended to use it and not to any one else.

2. Accuracy :

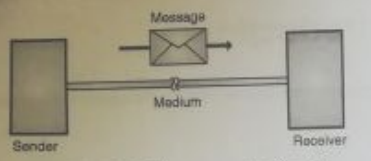
- Due to noise the data may get altered or corrected when it is travelling over a communication medium. Errors will be introduced and the accuracy of the received data is adversely affected.
- The data communication system (DCS) must be designed in such a way that the delivered data is accurate and free from any errors.

3. Timeliness :

- The time delay is unacceptable for the audio and video data, as it introduces errors in the reproduced sound or picture, so the DCS should deliver the data without any time delay.
- Such a data delivery is called as real-time transmission of data.

1.3 Components of Data Communication System :

- If we specifically consider the communication between two computers then the data communication system is as shown in Fig. 1.3.1.
- It has the following five components :
  1. Message
  2. Sender
  3. Medium
  4. Receiver and 5. Protocol



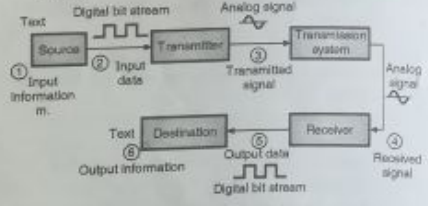
(L-2(a)) **Fig. 1.3.1 : Five components of a data communication system**

Description :

1. **Message :**
   - Message is nothing but information or data which is to be sent from sender to the receiver.
   - A message can be in the form of sound, text, number, pictures, video or combination of them.
2. **Sender :**
   Sender is a device such as a host, video camera, telephone, work station etc. which sends the message over the medium.
3. **Medium :**
   - The message originating from the sender needs a path over which it can travel to the receiver. Such a path is called as the medium or channel.
   - The examples of transmission medium are coaxial cable, twisted pair wire, fiber optic cable, radio waves (used in terrestrial or satellite communication).
4. **Receiver :**
   It is the device which receives the message and reproduces it. A receiver can be in the form of a workstation, telephone handset a TV receiver etc.
5. **Protocol :**
   Protocol is defined as the set of rules agreed by the sender and receiver. There can be different protocols defined for different functions. Protocols govern the exchange of data in true sense.

1.4 Data Communication Model :

- Fig. 1.4.1 shows the simplified data communication model.

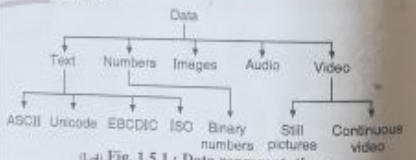


(L-3) **Fig. 1.4.1 : Data communication model**

- Suppose that the source and transmitter are the components of a personal computer. The user of this PC wants to send a message "m" to another PC.
- Then the message "m" will be in the form of a digital bit stream and called as **input data** as shown in Fig. 1.4.1.
- The senders PC is connected to some **transmission medium** such as a local network or a telephone line, by an input/output device (transmitter) such as a modem.
- The input data is applied to the transmitter as a sequence of digital bits on a transmission cable or communication bus.
- The transmitter is connected directly to the medium and converts the incoming digital bit stream into an analog signal suitable for transmission over the communication cable.
- The analog transmitted signal travels on the transmission medium and is subjected to a number of impairments (noise, attenuation etc.), before reaching the receiver.
- Due to these impairments, the received signal may appear completely different from the transmitted signal.
- The receiver tries to estimate the original signal based on the distorted received signal, and the knowledge of transmission medium.
- A sequence of digital bits is produced at the output of the receiver.
- These bits are then sent to the destination which is another PC. The signal "m" represents the output information which is presented to the user and can be seen on the computer screen for display or printed using a printer.

1.5 Data Representation :

- Data can be represented using different forms as shown in Fig. 1.5.1.



(L-4) **Fig. 1.5.1 : Data representation**

1.5.1 Text :

- A binary digit or bit can represent only two symbols as it has only two states "0" or "1".
- But this is not enough for communication between two computers because there we need many more symbols for communication. These symbols are required to represent :
  1. 26 alphabets with capital and small letters.
  2. Numbers from 0 to 9.
  3. Punctuation marks and other symbols.

- Therefore instead of using only single binary bits, a group of bits is used as a code to represent a symbol.
- The number of bits used per code word will be dependent on the total number of symbols to be represented e.g. if 5 bits are used per code word then 2^5 = 32 combination would be possible i.e. 32 distinct code words can be produced.
- If the word length is increased to 8 then the number of combinations would be 2^8 = 256, hence an 8-bit code can represent 256 symbols.
- A set of such code words is called as "code set". The following three code sets are very commonly used for the data representation.

Different codes used are :

1. ASCII (American Standard Code for Information Interchange)
2. EBCDIC (Extended Binary Coded Decimal Interchange code)
3. Baudot code
4. Extended ASCII
5. Unicode 6. ISO.

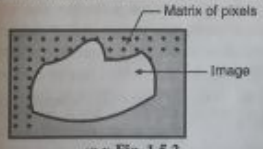
ASCII code is a 7-bit code whereas EBCDIC is an 8-bit code. ASCII code is more commonly used world wide while EBCDIC is used primarily in large IBM computers.

1.5.2 Numbers :

- The numbers are represented using bit patterns. The code such as ASCII or EBCDIC is not used for this purpose.
- This will simplify the mathematical operations on numbers to a great extent.

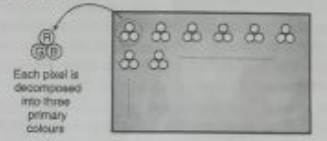
1.5.3 Images :

- Image is another form of data. They also are represented by bit patterns but with a different mechanism.
- The basic principle is as follows. An image is divided into a matrix of pixels (pixel means picture element).
- Each pixel is in the form of a small dot as shown in Fig. 1.5.2. We can use a large number of pixels for better resolution.
- Higher resolution ensures better quality of picture.



(G-8) **Fig. 1.5.2**

- After dividing the image into pixels, each pixel is represented by a unique bit pattern.
- For black and white image one 1 bit per pattern is sufficient to represent a pixel. 1 will represent white and 0 will represent black.
- If the gray shades are to be included then we can use 2 bit pattern to represent each pixel. Then 00 is black, 01 is dark gray, 10 is light gray and 11 represents white.
- If a coloured image is to be represented, then each coloured pixel is decomposed into three primary colours namely Red, Green and Blue (RGB), as shown in Fig. 1.5.3.

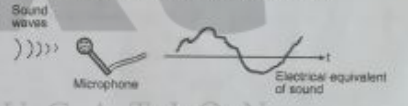


(G-4) **Fig. 1.5.3 : Pixels for coloured images**

- The intensity of each colour at each pixel is then converted into an 8 bit pattern. Thus colour image representation needs more number of bits. Thus the intensity information corresponding to each pixel is to be converted into a unique binary bit pattern.

1.5.4 Audio :

- Audio means sound. It is continuous with respect to time and not discrete like the text or numbers.



(G-5) **Fig. 1.5.4**

- A microphone is used to convert sound into an equivalent electrical signal. The audio signal should be converted into a digital signal before transmitting it over the medium.

1.5.5 Video :

- The principle behind video or moving pictures is the technique called **animation.** This technique is used in cartoons and motion pictures or films. It says - if we show a set of pictures at a sufficiently rapid pace, the human eye due to persistence of vision, feels that the picture is in continuous motion.
- If 24 pictures are shown in one second, our eye senses a continuous motion without any flicker effect.
- So if pictures of successive incremental actions stored in a disc or memory are retrieved at 24 pictures / second then we can play a video on the computer.
- A video camera captures images continuously at the rate of 24 pictures per second. These images can then be converted into binary form.

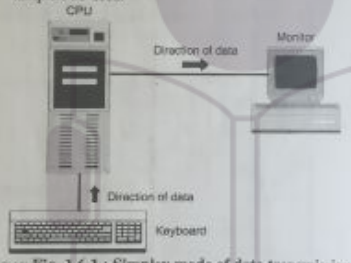
1.6 Types of Communication : Simplex, Half Duplex, Full Duplex (Data Flow) :

Based on whether the given communication system communicates only in one direction only or in both the directions, the communication systems are classified as :

- Simplex systems.
- Half duplex systems.
- Full duplex systems.

1.6.1 Simplex Systems :

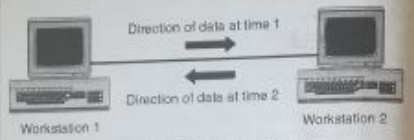
- In these systems the information is communicated in only one direction. For example the radio or TV broadcasting systems can only transmit. They cannot receive.
- In data communication system the simplex communication takes place as shown in Fig. 1.6.1.
- The communication from CPU to monitor or keyboard to CPU is unidirectional.
- Keyboard and traditional monitors are examples of simplex devices.



(G-10) **Fig. 1.6.1 : Simplex mode of data transmission**

1.6.2 Half Duplex Systems :

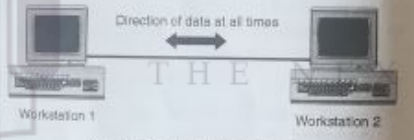
- These systems are bi-directional, i.e. they can transmit as well as receive but not simultaneously.
- At a time these systems can either transmit or receive, for example a transreceiver or walky talky set. Thus the direction of communication will keep changing itself.
- A data communication system working in the half duplex mode is shown in Fig. 1.6.2.
- Each station can transmit and receive, but not at the same time. When one device is ending the other one is receiving and vice versa.
- In half duplex transmission, the entire capacity of the channel is utilized by the transmitting (sending) system.



(G-11) **Fig. 1.6.2 : Half duplex system**

1.6.3 Full Duplex Systems :

- These are truly bi-directional systems as they allow the communication to take place in both the directions simultaneously.
- These systems can transmit as well as receive simultaneously, for example the telephone systems.
- A full duplex data communication system is shown in Fig. 1.6.3. Each station can transmit and receive simultaneously.
- In full duplex mode, signals going in either direction share the full capacity of link.
- The link may contain two physically separate transmission paths one for sending and another for receiving.
- Otherwise the capacity of channel is divided between signals travelling in both directions.



(G-12) **Fig. 1.6.3 : Full duplex mode**

1.7 Computer Networks :

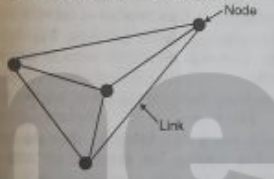
Network :

- Network is a broad term similar to "system". Network is a communication system which supports many users.
- In relation with the computers we can say that a "computer network" is a system which allows communication among the computers connected in the network.
- There are various ways of interconnecting the computers.

Protocol :

- For successful communication to occur, it is not enough for the "sender" to simply transmit the message and "assume" that the "receiver" will receive it properly.

- There are certain rules that must be followed to ensure proper communication.
- A set of such rules is known as a "protocol" of the data communication system.
- Many different protocols are used in the modern data communication system.
- The interconnection of one station to many stations is called as networking.
- A network is any interconnection of two or more stations that wish to communicate.
- **Node :** Each station in a communication network is called as a node. The nodes are connected in different way to each other to form a network.
- One of such networks is shown in Fig. 1.7.1.
- Many other forms of interconnections are possible. The most familiar network is the telephone system. It is the largest and most sophisticated network of all.



(G-13) **Fig. 1.7.1 : A simple communication network**

1.7.1 Introduction to Computer Networks :

- During 20th century the most important technology has been the information gathering, its processing and distribution.
- The computers and communication fields have been merged together and their merger has had a deep impact on the manner in which computer systems are organized.
- The old model in which a single computer used to serve all the computational needs of an organization has been replaced by a new one in which a large number of separate but interconnected computers do the job.
- Such systems are called as **computer networks.**
- Two computers are said to be interconnected if they interchange information. The connection between the separate computers can be done via a copper wire, fiber optics, microwaves or communication satellite.

Distributed system :

- A system with one control unit (master computer) and many slaves, or a large computer with remote printers and terminals is not called a **computer network, it is called a Distributed System.**
- In distributed system the existence of multiple autonomous computers is not visible to the user.
- With a computer network, the user has to conciously log onto a machine, submit jobs remotely, move files around etc. in short handle all the network management personally.
- With a distributed system nothing of this needs to be done explicitly, it all happens automatically because the system takes care of it without the users knowledge.
- Basically a distributed system is a software system built on top of a network. The software gives it a high degree of cohesiveness, (homoginity) and transparency to the system.

1.7.2 Computer Network Criteria :

- Network is a broad term similar to system. Network is a communication system which supports many users.
- In context with the computers we can say that, a "computer network" is a system which allows communication among the computers connected in the network.
- A network must be able to meet certain criteria. The most important of them are :
  1. Performance 2. Reliability 3. Security

Performance :

- Performance can be measured in different ways. We can measure it in terms of transit time and response time.
- **Transit time** is defined as the time required for a message to travel from one device to the other.
- **Response time :** It is the time elapsed between the instant of enquiry and the instant of giving response.
- The other factors deciding the performance are as follows :
  1. Number of users.
  2. Type of transmission medium.
  3. The hardware used. 4. The software used.

Reliability :

- The network reliability is important because it decides the frequency at which network failure takes place.
- It also decides the time taken by the network to recover and its robustness in the catastrophe.

Security :

The network security refers to protection of data from the unauthorized user or access. It also includes the data protection against damage and recovering it in the events of data losses.

1.8 Physical Structures :

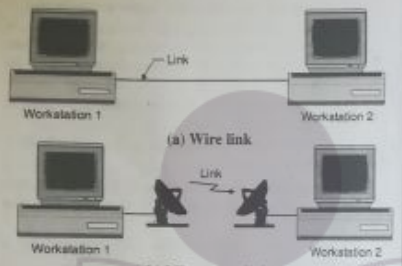
Two characteristics which can be used as the basis of distinguishing various data link configurations are topology and whether the link is half duplex or full duplex.

1.8.1 Type of Connections (Topology) :

- In a network two or more devices are connected to each other through connecting links.
- There are two possible ways to connect the devices. They are follows :
  1. Point to point connection
  2. Multipoint connection.

1.8.2 Point-to-Point Connection :

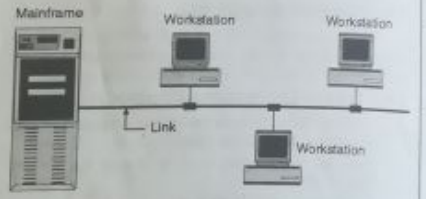
- A point to point connection provides a dedicated link between two devices as shown in Fig. 1.8.1. The meaning of the word dedicated is that the entire capacity of the link is reserved for transmission between these two devices only.
- It is possible to connect the two devices by means of a pair of wires (see Fig. 1.8.1(a)) or using a microwave or satellite link (wireless link) as shown in Fig. 1.8.1(b).



(G-8) **Fig. 1.8.1 : Point to point connection**

1.8.3 Multipoint Connection :

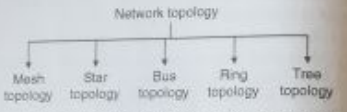
- A multipoint connection is also called as a multidrop connection.
- In such a connection more than two devices are connected to and share a single link as shown in Fig. 1.8.2.
- In the multipoint connection the channel capacity is shared among multiple users. If many devices share the link simultaneously, it is called spatically shared connection.
- But if users share it turn by turn then it is time sharing connection.



(G-9) **Fig. 1.8.2 : Multipoint configuration**

1.9 Network Topology Types :

- The word physical network topology is used to explain the manner in which a network is physically connected.
- Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.
- The geometric representation of such a relationship of links and nodes is known as the topology of that network.
- The five basic network topologies are as shown in Fig. 1.9.1.

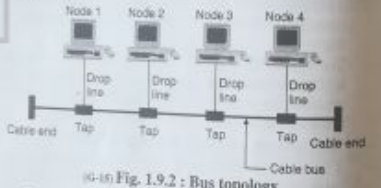


(G-14) **Fig. 1.9.1 : Classification of network topology**

- While selecting one of the above five topologies we have to consider the relative status of the device to be linked.
- These topologies can be classified into two types :
  1. Peer to peer 2. Primary – secondary
- Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.
- In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

1.9.1 Bus Topology :

- The bus topology is usually used when a network under consideration is small, simple or temporary as shown in Fig. 1.9.2.



(G-15) **Fig. 1.9.2 : Bus topology**

- On a typical bus network a simple cable is used without additional electronics to amplify the signal or pass it along from computer to computer. Therefore the bus is a passive topology.
- When one computer sends a signal on the cable; all the computers on the network receive the information. However only the one with the address that matches with the destination address stored in the message accepts the information while all the others reject the message.

- The speed of the bus topology is slow because only one computer can send a message at a time. A computer must wait until the bus is free before it can transmit.
- The bus topology requires a proper termination at both the ends of the cable in order to avoid reflections.
- Since the bus is a passive topology, the electrical signal from a transmitting computer is free to travel over the entire length of the cable.
- Without termination when the signal reaches the end of the cable, it returns back and travels back on the cable.
- The transmitted waves and reflected waves, if they are in phase add and if they are out of phase cancel.
- Thus addition and cancellation of wave results in a standing wave.
- The standing waves can distort the normal signals which are travelling along the cable. This can be avoided by terminating the bus on both ends in 50 Ω load impedance.
- The terminators absorb the electrical energy and avoid reflections.

Characteristics of the bus topology :

Following are some of the important characteristics of the bus topology :

1. This is a multipoint configuration. There are more than two devices connected to the medium and they are capable of transmitting on the medium. Hence the Medium Access Control (MAC) is essential for the bus topology.
2. The signal strength of the transmitted signal should be adequately high so as to meet the minimum signal strength requirements of the receiver.
3. Adequate Signal to Noise Ratio (SNR) should be maintained for better quality reception.
4. The signal should not be too strong. This is necessary to avoid the overloading of transmitter and hence the possibility of signal distortion.
5. This is called as signal balancing which is not an easy task at all. Specially the signal balancing becomes increasingly difficult with increase in the number of stations.

Transmission media for bus LANs :

We can use the following transmission media for the bus LANs :

1. Twisted pair 2. Baseband co-axial cable
3. Broadband co-axial cable 4. Optical fibre

Advantages of bus topology :

1. The bus topology is easy to understand, install, and use for small networks.
2. The cabling cost is less as the bus topology requires a small length of cable to connect the computers.
3. The bus topology is easy to expand by joining two cables with a BNC barrel connector.
4. In the expansion of a bus topology repeaters can be used to boost the signal and increase the distance.

Disadvantages of bus topology :

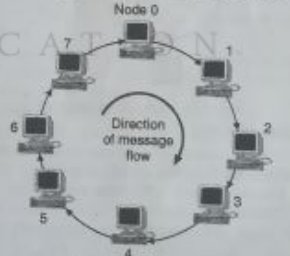
1. Heavy network traffic slows down the bus speed. In bus topology only one computer can transmit and other have to wait till their turn comes and there is no co-ordination between computers for reservation of transmitting time slot.
2. The BNC connectors used for expansion of the bus attenuates the signal considerably.
3. A cable break or loose BNC connector will cause reflections and bring down the whole network causing all network activity to stop.

Note :

- Ethernet 10 base 2 also known as thinnet is an inexpensive network based on the bus topology.
- A bus network behaves erratically if it is not terminated or improperly terminated.
- Token bus networks are defined by the IEEE 802.4 standard.

1.9.2 Ring Topology :

- In a ring topology, each computer is connected to the next computer, with the last one connected to the first as shown in Fig. 1.9.3.
- Rings are used in high-performance networks where large bandwidth is necessary e.g. time sensitive features such as video and audio.
- Every computer is connected to the next computer in the ring and each retransmits what it receives from the previous computer hence the ring is an active network.



(G-16) **Fig. 1.9.3 : Ring topology**

- The messages flow around the ring in one direction. There is no termination because there is no end to the ring.
- Some ring networks do token passing. A short message called a token is passed around the ring until a computer wishes to send information to another computer.
- That computer modifies the token, adds an electronic address and data and sends it around the ring.

- Each computer in sequence receives the token and the information and passes them to the next computer until either the electronic address matches the address of a computer or the token returns to its origin.
- The receiving computer returns a message to the originator indicating that the message has been received.
- The sending computer then creates another token and places it on the network, allowing another station to capture the token and begin transmitting.
- The token circulates until a station is ready to send and capture the token. Faster networks circulate several tokens at once.
- Some ring networks have two counter-rotating rings that help them recover from network faults.

Characteristics of ring LANs :

- The basic ring LAN is shown in Fig. 1.9.4, which shows that along with the nodes A, B, C, D equal number of repeaters are used and that the transmission is unidirectional.
- The data is travels in a sequential manner around the ring. Each repeater will receive regenerate and retransmit this data bit.

(G-17) **Fig. 1.9.4 : Ring topology**

Functions of a ring :

- A ring can operate as a communication network if it performs the following three functions :
  1. Data insertion
  2. Data reception
  3. Data removal
- The above mentioned functions are actually provided by the repeaters.
- Each repeater also acts as the device attachment point, so that the function of data insertion can be accomplished.
- Data is transmitted in the form of packets.
- Each packet consists of a destination address field. As this packet passes through a repeater, the destination address field is copied by the repeater.
- If the destination address field corresponds to the address of a device then that repeater copies the remaining contents of packet as well.
- Data insertion and reception can be done easily by the repeaters but the task of removing data removal is more difficult on a ring.
- As the ring is a closed loop, a packet will circulate on it indefinitely if it is not removed.
- A packet can be removed by the addressed repeater or each packet can be removed by the transmitting repeater itself after the packet has made one trip around the ring.
- The second approach is more desirable.

Problems faced in the ring topology :

1. If any link breaks or if any repeater fails then the entire network will be disabled.
2. To install a new repeater for supporting a new device, it is necessary to have the identification of two nearby, topologically adjacent repeaters.
3. It is necessary to take preventive measures to deal with the time jitter.
4. Due to the closed nature of the ring topology it is necessary to remove the circulating packets.

These problems except for the last one can be rectified by refinements of the ring topology.

Advantages of ring topology :

1. Every computer gets an equal access to the token.
2. There are no standing waves produced.

Disadvantages of ring topology :

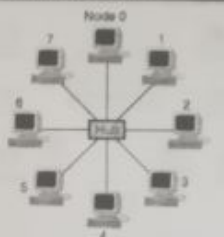
1. Failure of one computer on the ring can affect the whole network.
2. It is difficult to trouble short the ring.
3. Adding or removing the computers disturbs the network activity.

Note : Token ring networks are defined by the IEEE 802.5 standard.
Fibre Distributed Data Interface (FDDI) is a fast fibre-optic network based on the ring topology.

1.9.3 Star Topology :

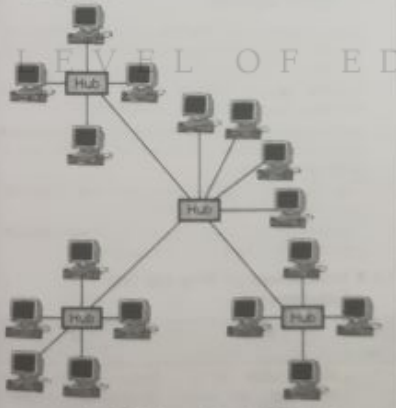
In a star topology all the computers are connected via cables to a central location where they are all connected by a device called a hub as shown in Fig. 1.9.5. There is no direct connections among the computers. All the connections are made via the central hub.

- Stars are used in concentrated networks, where the endpoints are directly reachable from a central location; when network expansion is expected and when the greater reliability of a star topology is needed.
- Each computer on a star network communicates with a central hub. The hub then resends the message to all the computers in a broadcast star network. It will resend the message only to the destination computer in a switched star network.



(G-18) **Fig. 1.9.5 : Star topology**

- The hub in a broadcast star network can be active or passive. An active hub generates the electrical signal and sends it to all the computers connected to it.
- This type of hub is usually called a multiport repeater. Active hubs require external power supply.
- A passive hub is a wiring panel or punch down block which acts as a connection point. It does not amplify or regenerate the signal. Passive hubs do not require electrical power supply.
- Several types of cables can be used to implement a star network. A hybrid hub can use different types of cable in the same star network.
- A star network can be expanded by placing another star hub as shown in Fig. 1.9.6.
- This arrangement allows several more computers or hubs to be connected to that hub. This creates a hybrid star network.



(G-19) **Fig. 1.9.6 : Expansion of star topology**

1.9.4 STAR LANs :

- In the star type LANs, the Unshielded Twisted Pair (UTP) is used as the transmission medium.
- This is because the unshielded twisted pair is a telephone wire which is available in each and every office building. The other advantages of using twisted wires are as follows :
  1. No no additional installation cost is required for the installation of LAN.
  2. Since the telephone wires cover the entire building it is possible to spread the network in every part of each building.
- The basic star topology is as shown in Fig. 1.9.7. This is called as a single level star topology.
- As shown in Fig. 1.9.7, the central element of the star topology is an active element called hub or repeater.
- Each station (A, B, C, ...) is connected to the hub with the help of two links one for transmitting and the other for reception of the data.

(G-20) **Fig. 1.9.7 : Single level star topology**

- When a single station transmits, the hub repeats the signal and sends it to each station.
- Typically the length of each link is 100 m. If the twisted pair is used and the length may increase upto 500 m if the optical fibre is used as transmission medium.
- It is important to note that if two stations transmit simultaneously, then there will be a collision between their transmitted signals.

Disadvantages of star topology :

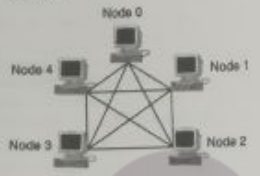
1. If the central hub fails, the whole network fails to operate.
2. Many star networks require a device at the central point to rebroadcast or switch the network traffic.
3. The cabling cost is more since cables must be pulled from all computers to the central hub.

Note : Ethernet 10 base T is a popular network based on the star topology.
Intelligent hubs with microprocessor that implement features in addition to repeating network. Signals provide for centralized monitoring and management of the network. It is the most flexible and the easiest to diagnose when there is a network fault.

1.9.5 Mesh Topology :

In a mesh topology every device is physically connected to every other device with a point to point dedicated link as shown in Fig. 1.9.8.

- The term dedicated means that the link carries data only between two devices connected on it.
- A fully connected mesh network therefore has n(n–1)/2 physical cables to connect n devices. To accommodate that many links every device on the network must have n–1 input/output ports.
- So too many cables are required to be used for the mesh topology.



(G-21) **Fig. 1.9.8 : Mesh topology**

Advantages :

1. The use of dedicated links guarantees that each connection can carry its own data reliably.
2. A mesh topology is robust because the failure of any one computer does not bring down the entire network.
3. It provides security and privacy because every message sent travels along a dedicated line.
4. Point to point links make fault diagnose easy.

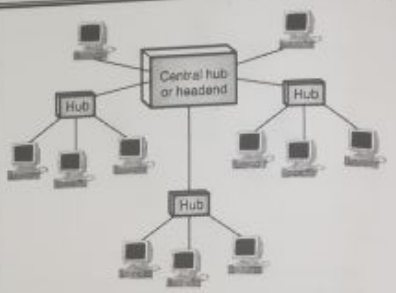
Disadvantages :

1. Since every computer must be connected to every other computer installation and reconfiguration is difficult.
2. Cabling cost is more.
3. The hardware required to connect each link input/output and cable is expensive.

Note : Mesh topology is usually implemented as a backbone connecting the main computers of a hybrid network that can include several other topologies.

1.9.6 Tree Topology :

- A tree topology is a variation of a star. As in a star, nodes in a tree are connected to a central hub that controls the entire network.
- However, every computer is not plugged into the central hub. Most of them are connected to a secondary hub which in turn is connected to the central hub as shown in Fig. 1.9.9.
- The central hub in the tree is an active hub which contains repeater. The repeater amplify the signal and increase the distance a signal can travel.
- The secondary hubs may be active or passive. A passive hub provides a simple physical connection between the attached devices.



(G-22) **Fig. 1.9.9 : Tree topology**

Advantages :

1. It allows more devices to be attached to a single hub and can therefore increase the distance of a signal can travel between devices.
2. It allows the network to isolate and attach priorities to the communications from different computers.

Disadvantages :

1. If the central hub fails the system breaks down.
2. The cabling cost is more.

Note : The advantages and disadvantages of a tree topology are generally the same as those of a star.

1.9.7 Logical Topology :

- Logical topology describes the manner in which the stations are logically connected to each other for the purpose of data unit exchange.
- Physical topology discussed earlier can be different from the logical topology, of the network.
- As an example consider the bus topology. The bus acts as a central controller. It receives data and forwards it to the various nodes.
- Thus the stations have a logical connection to the bus which acts as a centralized controller.
- Therefore the logical topology of a bus is star topology. eventhough the physical topology is bus.

1.9.8 Comparison of Ring and Star Topologies :

Sr. No.	Ring	Star
1.	Media failure on uni-directional or single loop ring causes complete network failure.	Media faults are automatically isolated to the failed segment.
2.	Relatively difficult to reconfigure.	Relatively easy to configure.

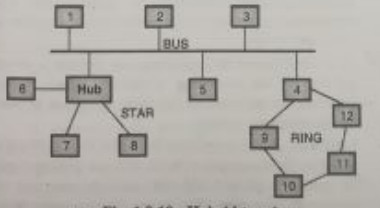
Sr. No.	Ring	Star
3.	It is difficult to troubleshoot.	Easy to troubleshoot.
4.	The failure of one computer can affect the whole network.	The failure of single computer or cable doesn't bring the network down.
5.	No computer has a monopoly over the network.	Failure of the central hub causes the whole network failure.
6.	Adding and removing computers disrupts the network.	Adding and removing the computers is relatively easier.

1.9.9 Comparison of Bus and Star Topologies :

Sr. No.	Bus	Star
1.	Uses a cable as bus or backbone to connect all nodes.	Uses a central hub to connect the nodes to each other.
2.	Baseband or broadband coaxial cable is used.	Twisted pair, coaxial cables or optical fiber cables are used.
3.	If a part of bus fails, the whole network fails.	Failure of the central hub will make the entire network collapse.
4.	Adding an new node is difficult.	Adding and removing a node is relatively easy.
5.	Fault diagnosis is relatively difficult.	Fault diagnosis is easy.

1.9.10 Hybrid Topology :

- We have discussed various basic topologies such as bus, ring, mesh, star etc.
- Hybrid topology is the one which makes use of two or more basic topologies mentioned above, together.
- There are different ways in which a hybrid network is created. Fig. 1.9.10 shows the hybrid topology in which bus, star and ring topologies are used simultaneously.



(G-23) **Fig. 1.9.10 : Hybrid topology**

- In Fig. 1.9.10, the nodes 1, 2, 3, 4 and 5 are connected in the bus topology, node 6, 7 and 8 form a star and the nodes 4, 9, 10, 11, 12 are arranged in a ring topology.
- The practical networks generally make use of hybrid topology. Many complex networks can be reduced to some form of hybrid topology.
- The hybrid topology which is to be used for a particular application depends on the requirements of that application.

1.9.11 Comparison of Star Bus and Star Ring Topologies :

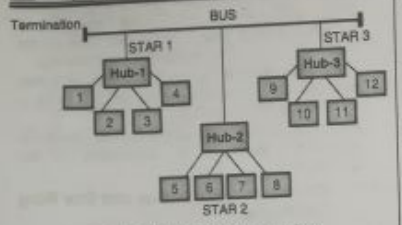
Sr. No.	Parameter	Star bus	Star ring
1.	Topology	Hybrid. It is a combination of star and bus topologies.	Hybrid. It is a combination of star and ring topologies.
2.	Configuration	Fig. A	Fig. B
3.	Peculiarity	Many stars are connected to a bus.	Many rings are connected in star.
4.	Applications	Suitable for large networks.	Suitable for interconnection of many small networks.

Fig. A

(G-24) **Fig. B**

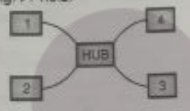
Ex. 1.9.1 : Draw the star bus topology connecting three star networks consisting of four computers.

Soln. : Fig. P. 1.9.1 shows the required star bus topology.



(G-25) **Fig. P. 1.9.1 : Star-bus topology**

Ex. 1.9.2 : Identify the network topology shown in Fig. P. 1.9.2.



(G-26) **Fig. P. 1.9.2**

Soln. : The given network uses the star topology.

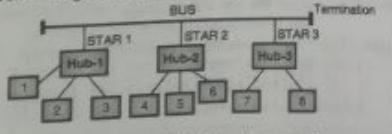
Ex. 1.9.3 : Identify the topology shown in Fig. P. 1.9.3. Comment on whether this topology is active or passive.

(G-27) **Fig. P. 1.9.3**

Soln. : This is the ring topology. The ring topology is an active network.

Ex. 1.9.4 : Draw with neat-labelled sketch of star-bus topology connecting 3 star networks having 3 computers in 2 stars and 2 computers in one star.

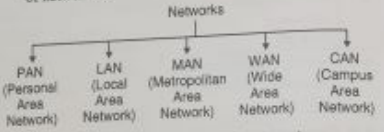
Soln. : Fig. P. 1.9.4 shows the required star-bus topology.



(G-28) **Fig. P. 1.9.4 : Star-bus network**

1.10 Network Classification by their Geography (Categories of Networks) :

- Computer network can be classified based on the geographical area they cover, i.e. the area over which the network is spread.
- Such a classification is shown in Fig. 1.10.1.
- In this section, we will discuss the following categories of networks :



(G-1400) **Fig. 1.10.1 : Network categories**

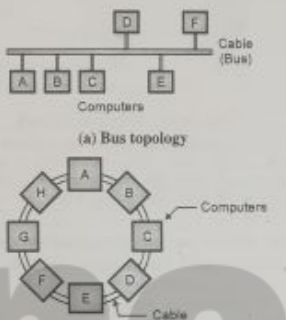
1.10.1 Local Area Networks (LAN) :

- The Local Area Network (LAN) is a network which is designed to operate over a small physical area such as an office, factory or a group of buildings. LANs are very widely used in a variety of applications.
- LANs are easy to design and troubleshoot. The personal computers and workstations in the offices are interconnected via LAN.
- The exchange of information and sharing of resources becomes easy because of LAN.
- In LAN all the machines are connected to a single cable. Different types of topologies such as Bus, Ring, Star, Tree etc. are used for LANs.
- LAN uses a layered architecture and they are capable of operating at hundreds of Mbits/sec.
- A Local Area Network (LAN) is usually a privately owned and links the devices in a single office, building or campus of upto a few kilometres in size as shown in Fig. 1.10.1.
- Depending on the needs of an organisation and the type of technology used, a LAN can be as simple as a few computers and a printer at home or it can contain many computers in a company and include voice, sound and video peripherals.
- LANs are widely used to allow resources to be shared between personal computers or workstations. The resources to be shared can be hardware like a printer or softwares or data.
- In a LAN one of the computer can become a server serving all the remaining computers called clients. Software can be stored on the server and it can be used by the remaining clients.
- LAN's are also distinguished from MAN's and WAN's based on the transmission media they use and topology. In general a given LAN will use only one type of transmission medium. The most common networking topologies used are bus, ring and star.

- The data rates for LAN can now range from 10 Mbps to 16 Gbps.

LAN topologies :

Various topologies are possible for the broadcast LANs such as bus topology or ring topology as shown in Fig. 1.10.2.



(G-32) **Fig. 1.10.2 : LAN topologies**

Static and Dynamic broadcast networks :

- The broadcast networks are further classified into two types namely :
  1. Static networks and
  2. Dynamic networks.
- This classification is based on how the common channel is allocated.
- In static allocation, each machine is allowed to broadcast only in its allotted time slot.
- But static allocation wastes the channel capacity when a machine does not want to transmit in its allotted time slot.
- Hence most of the systems try to allocate the channel dynamically i.e. on demand.

LAN components :

Some of the important LAN components are as follows :

1. Workstations.
2. File servers.
3. Gateway.
4. Network interfacing unit.
5. Active and passive hubs.
6. LAN cables or communication channels.

Workstation :

Workstation refers to the individual, single computer.

A communication capability is added to enable it for networking.

File server :

File server is a computer that allows the sharing of data, software and hardware resources by running special softwares.

Gateway :

It assists the transfer of data from one LAN to the other LAN.

Network Interfacing Unit (NIU) :

It is a unit which consists of hardware as well as software. It uses microprocessor to control the access and communication in a network.

LAN cables or communication channel :

A cable is used for connecting the computers in a LAN. The communication from one computer to others takes place over the cables. So cables are called communication channels. The twisted pair, coaxial cables or optical fiber cables are used in LANs.

Advantages of LAN :

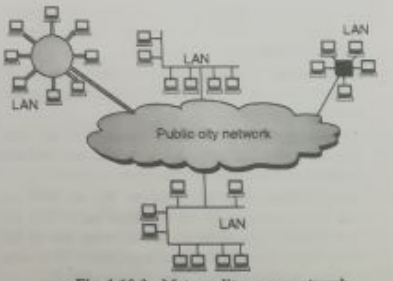
1. High reliability. Failure of individual computers does not affect the entire LAN.
2. It is possible to add a new computer easily.
3. The transmission of data is at a very high rate.
4. Sharing of peripheral devices such as printer is possible.

Applications of LAN :

1. File transfer and file access.
2. Personal computing.
3. Office automation.
4. Distributed computing.
5. Word and text processing.
6. Document distribution.
7. Remote access to database.
8. Electronic message handling.

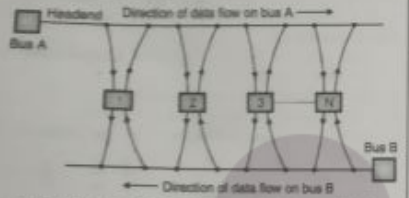
1.10.2 Metropolitan Area Network (MAN) :

- A MAN is basically a bigger version of a LAN and normally uses similar technology. It is designed to extend over a larger area such as an entire city.
- The MAN can be in the form of a single network such as a cable network or it can be a combination of multiple LANs as shown in Fig. 1.10.3.
- A MAN may be wholly owned and operated by a private company or it may be a service provided by a public company, such as a local telephone company (telco).



(G-33) **Fig. 1.10.3 : Metropolitan area network**

- A MAN is distinguished by the IEEE 802.6 standard or it is also known as Distributed Queue Dual Bus (DQDB).
- The DQDB consists of two unidirectional cables (buses) to which all the computers are connected as shown in Fig. 1.10.4.
- Each bus has a device which initiates the transmission activity called as the head-end.
- Traffic that is destined for a computer to the right of the sender uses the upper bus and to the left uses the lower bus as shown in Fig. 1.10.4.



(G-34)**Fig. 1.10.4 : Distributed queue dual bus architecture (DQDB)**

1.10.3 Wide Area Network (WAN) :

- When a network spans a large distance or when the computers to be connected to each other are at widely separated locations a local area network cannot be used.
- For such situations a Wide Area Network (WAN) must be installed. The communication between different users of "WAN" is established using leased telephone lines or satellite links and similar channels.
- It is cheaper and more efficient to use the phone network for the links.
- Most wide area networks are used for transferring large blocks of data between its users. As the data is from existing records or files, the exact time taken for this data transfer is not a critical parameter.
- An example of WAN is an airline reservation system. Terminals are located all over the country through which the reservations can be made.
- It is important to note here that all the terminals use the same centralized common data provided by the central reservation computer.
- Because of the large distances involved in the wide area networks, the propagation delays and variable signal travel times are major problems.
- Therefore most wide area networks are not used for time critical applications. As explained earlier they are more suitable for transfer of data from one user to the other which is not a time critical application. Wide area networks are basically packet switching networks.

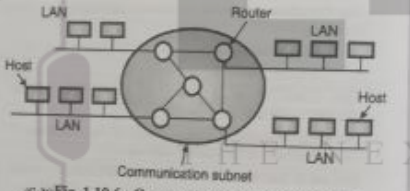
(G-35)**Fig. 1.10.5 : Wide area network**

- A WAN provides long distance transmission of data, voice image and video information over large geographical areas that may comprise a country, a continent or even the whole world as shown in Fig. 1.10.5.

Host : Host is a large computer. It can provide services to many computers. The services provided are :

1. Providing computing capabilities
2. Providing access to database.

- WAN contains a collection of machines used for running user (i.e. application) programs. All the machines called hosts are connected by a communication subnet as shown in Fig. 1.10.6.



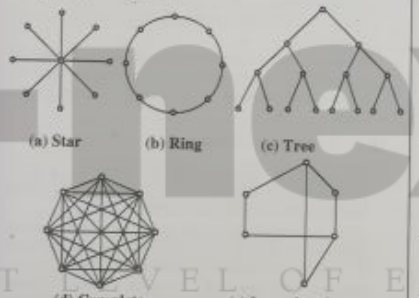
(G-36)**Fig. 1.10.6 : Communication subnet and hosts**

- The function of the subnet is to carry messages from host to host. The subnet consists of two important components; transmission lines and switching elements.
- Transmission lines move bits from one machine to another. The switching elements are specialised computers used to connect two or more transmission lines. When data arrive on an incoming line, the switching element has to choose an outgoing line on which it is to be forwarded.
- The switching elements are either called as packet switching nodes, intermediate systems, data switching exchanges or routers.
- When a packet is sent from one router to another via one or more intermediate routers, the packet is received at intermediate router. It is stored in the routers until the required output line is free and then forwarded. A subnet using this principle is called a point to point, store-forward or packet switched subnet.

- WAN's may use public, leased or private communication devices, and can spread over a wide geographical area. A WAN that is wholly owned and used by a single company is often called as an enterprise network.
- In most WANs the network contains a large number of cables or telephone lines each one connecting a pair of routers.
- If two routers which are not connected to each other via a cable want to communicate, then they have to do it indirectly via other routers.

Router interconnection topologies :

- Fig. 1.10.7 shows some of the possible router interconnection topologies in a point to point subnet.
- The LANs have a symmetric topology while WANs have irregular topologies.
- The WANs can also be formed using satellite or ground radio system. Satellite networks are inherently broadcast type so they are useful when the broadcast property is important.



(G-38)**Fig. 1.10.7 : Router interconnection topologies**

1.10.4 Comparison of LAN, WAN and MAN :

Sr. No.	Parameter	LAN	WAN	MAN
1.	Ownership of network	Private	Private or public	Private or public
2.	Geographical Area covered	Small	Very large (states or countries)	Moderate (city)
3.	Design and maintenance	Easy	Not easy	Not easy
4.	Communication medium	Coaxial cable	PSTN or satellite links	Coaxial cables, PSTN, optical fiber cables, wireless.
5.	Data rates (speed)	High	Low	Moderate
6.	Mode of communication	Each station can transmit and receive	Each station cannot transmit	Each station can transmit or receive.
7.	Principle	Operates on the principle of broadcasting	Switching	Both
8.	Propagation delay	Short	Long	Moderate
9.	Bandwidth	Low	High	Moderate

1.11 Internetworking :

- Earlier we have discussed the basic concepts in computer networking. An another very important concept is **Internetworking** which deals with connecting many computer networks together.
- A network of computer networks is known as **internetwork** or simply an **Internet** (note that i is small). The best example of an internetwork is Internet (note that I is capital).
- The difference between Networking and Internetworking may be states as follows : In networking all the devices (hosts) involved are compatible with each other. But in internetworking this may not be true.
- The networks that are being connected to each other through internetworking may or may not be compatible with each other, in many respects.
- For example the networks which are to be connected to form an internetwork can be an Ethernet, a token ring LAN and a WAN. All these networks are entirely different from each other, in terms of their topologies, signaling, transmission mechanism, wiring etc.
- Therefore in internetworking it is a challenge to handle these incompatibilities and bring all these networks together.
- The Internet is the biggest network of computer networks. The TCP/IP protocol is the backbone of Internet.

1.11.1 Why Internetworking ?

- Different networks such as LANs, MANs and WANs are designed with a specific task or application. So these networks won't have the same technology (hardware and protocols used).
- So the computers connected in the same network only can communicate to each others. It is not possible for a computer to communicate with some other computer outside its own network.

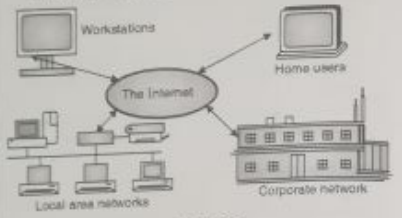
- For example an employee would be unable to communicate with the other computer connected to a printer or it would not possible to access a file on a computer which is on some other network and so on.
- This affected the productivity to a large extent in 1970's. So the concept of **universal service** came into existence.
- The simple meaning of universal service was that there was no dependence on the underlying physical technology or existence of separate physical networks.
- People wanted a single computer network to exist the way a telephone network exists. People should be able to use resources such as a printer or a file on any other computer without any hurdles.
- For this to become a reality it was necessary to connect all the computer networks together. This is why internetworking is essential.

1.11.2 The Problems in Internetworking :

- When internetworking is to be done, it must be remembered that the organizations involved have invested a lot of money on the infrastructure, cabling etc. of their existing network that they would like to reuse it when becoming the member of the internetwork.
- But it is not that simple to form a network merely by interconnecting wires from two networks. The problems are due to the incompatibilities in the electrical as well as the software aspects.
- The packet sizes used by different networks will be different, the methods of acknowledgement or error detection etc. can also be totally different. There could be many more differences between them other than these.
- Hence any two networks can not be connected to each other just by connecting a wire between them to form an internetwork.

1.12 Internet :

- This was the next step of ARPANET and NSFNET.
- Sometime in mid 1980s, people started thinking about the interconnection of many networks and the Internet came into existence.
- It started growing exponentially with all the types of existing networks getting connected to Internet.
- This was possible due to the use of TCP/IP reference model and TCP/IP protocol stack.
- The Internet is a globally existing network of networks consisting of a huge number of computers located in all the parts of the world.
- Already billions of people share it and the number is increasing every day.



(G-43) **Fig. 1.12.1**

- All the computers connected to the internet are part of this huge network. Networking is interconnection of computers. Generally the networking topologies used for networking are star, bus, ring, loop etc.
- When a limited number of computers are to be interconnected, the local area network (LAN) is used. But in the Internet the interconnection is achieved even via satellites.
- The power that a computer gets due to the Internet is amazing. It is now possible to send and receive the data within a few seconds of time from any part of the world.

1.12.1 Evolution of Internet :

- The origin of Internet can be traced to a U.S. Department of Defence (DoD) organization called Advanced Research Projects Agency (ARPA).
- ARPA developed a four node packet switching network called ARPANET in 1969. The network was intended to support the military research on fault tolerant computer networks. DoD wanted to ensure a reliable data transfer in the event of a nuclear war, even if parts of the network has been destroyed.
- As the years pass by, the network grew in size and by January 1983, the ARPANET no longer remained an experimental network but its control was passed over to Defence Communications Agency (DCA).
- The network then became available for the academic research, government employee and contractors.
- It was not until 1986, however that the dramatic growth of the Internet began. At this time, the National Science Foundation (NSF) which formed the National Science Foundation Network (NSFNET) linked five of its regional super-computer centres together to provide a national high speed backbone network across the United States.
- The world's fastest and most powerful computers were made available to a academic and scientific community.

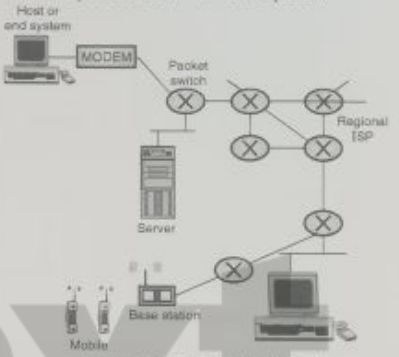
- In 1990 the ARPANET was officially decommissioned. Some of the major networks contributing to the growth of the internet are as follows :
  1. ARPANET
  2. USENET (User's network)
  3. CSNET (Computer science network)
  4. BITNET (Because it's time network)
  5. NSFNET (National science foundation network)
  6. WWW (World wide web)
  7. NREN (National research and education network)
  8. Intranet
- It is said that the Internet is growing at a rate of 20% per month. The data speeds have gone up considerably, which makes the access even more fast.
- The event which started as a military assistance program is now largely a private enterprise.

1.12.2 Net Structure :

- To understand how the Internet works, imagine complex network of roads. It consists of superhighways, highways, to the small roads on the countryside, all connected to each other in one way or the other. The nature of the internet is much similar to this.
- The high speed and powerful super computers connected to the Internet is the backbone of the system.
- The data that is being moved by these supercomputers is a few million bits per second.
- The regional networks are then connected to these supercomputers. The small computer network and individual users are then connected to this regional network.
- The data and information available on the Internet is increasing every day because every user can contribute to it.
- Due to this the entire net will never crash down. Even if a part of it closes down the remaining network keeps working.
- There is no central computer to control the internet. There is no central body which governs the Internet structure, though volunteer groups of individuals have set some standards for Internet technologies. No particular person, group or organization owns the Internet.

1.12.3 Parts of the Internet :

- Now we will discuss the basic hardware and software components of the Internet. Refer Fig. 1.12.2.
- As shown in Fig. 1.12.2 the Internet is made up of hosts, packet switches, servers, regional ISPs (Internet Service Providers), base stations, mobile units and different types of communication links.
- Now a days some nontraditional Internet end systems such as TV, cell phones, etc are also being used. We will call all the traditional and non-traditional devices as **hosts** or **end systems.**
- The end systems are interconnected by **communication links.** Different types of communication links are as follows :
  1. Coaxial cables
  2. Copper wires
  3. Optical fiber cable
  4. Radio spectrum



(G-44) **Fig. 1.12.2 : Parts of the Internet**

Packet switches :

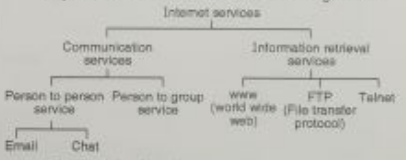
- The end systems are not directly connected to each other via communication links. Instead they are interconnected through the switching devices called as the **packet switches.** The two most popular packet switches are **routers** and **link-layer switches.**

1.12.4 Services on the Internet :

The services that are available on the Internet can be classified into two categories :

1. Communication services
2. Information retrieval services

They can be further classified as shown in Fig. 1.12.3.



(G-45) **Fig. 1.12.3 : Classification of services on Internet**

- The person to person service includes E-Mail. This is the most popular service and the most of the Internet traffic corresponds to the E-mail.
- The person to group service includes on line discussions between a user and many other participants from around the globe. A user can choose his topic of discussion from a list of more than 25, 000 topics.

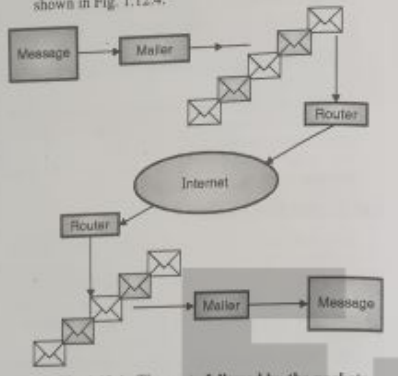
- The fast spreading popularity of the Internet is basically due to the information retrieval services. These services as shown in Fig. 1.12.3 include :
  1. The world wide web (www), which is world's largest and ever-growing data base.
  2. FTP or file transfer protocol, which allows remote accessing to the files which contain programs, technical handouts, reports etc.
  3. Telnet or remote login to commercial services.

1.12.5 A Service Description :

- The Internet allows distributed applications running on its end systems to exchange data with each other, eg web surfing, audio / video streaming, internet telephony, P2P, etc.
- The Internet provides two services to its distributed applications :
  1. A connection oriented reliable service.
  2. Connectionless unreliable service.
- Typically a distributed application uses one of these services (but not both).

1.12.6 Internet Protocols :

- The Internet protocol is like any other communication protocol is a set of rules which will govern every possible communication over the internet. Since the development of the ARPANET, TCP/IP together has emerged as the controlling body.
- It is being used in computers of not only in the U.S. but all over the world for all the types and sizes of computers. It has become the language of the Internet.
- TCP/IP are two protocols : Transmission control protocol and Internet protocol. These two protocols describe the movement of data between the host computers on Internet.
- The protocol however is a suite of many other protocols which provide for reliable communications across the Internet and the web.
- In the TCP/IP protocol suite, there are various layers, with each layer being responsible for different facets of communication.
- The Internet Protocol (IP) and Transmission Control Protocol (TCP) are together known as TCP/IP protocol. TCP/IP offers a simple naming and addressing scheme whereby different resources on Internet can be easily located.
- Information on Internet is carried in "packets". The IP protocol is used to put a message into a "packet". Each packet has the address of the sender and the recipient's address. These addresses are known as the IP addresses.
- Using the TCP protocol, a single large message is divided into a sequence of packets and each is put into an IP packet. The packets are passed from one network to another until they reach their destination.
- At the destination the TCP software reassembles the packets into a complete message. It is not necessary for all the packets in a single message to take the same route each time it is sent.
- The Internet is thus a "packet switched" network. The route followed by the packets of information are as shown in Fig. 1.12.4.



(G-46) **Fig. 1.12.4 : The route followed by the packets**

- The packets as shown in Fig. 1.12.4 are read by the routers and then are sent down to the destination address.

1.12.7 Internet Address :

- With so many computers and users using the Internet, it would be impossible to differentiate them without a proper addressing system.
- This addressing system assigns names and numbers to identify the computers on the Internet.
- The names are called as "domain names" and the numbers are called as "IP addresses". Every computer on the Internet will have both a domain name and an IP address.

An IP address :

- Each computer on the Internet is identified by its unique IP address. An IP address is made of four numbers and each one should be less than 256.
- Thus each set of numbers can have values from 1 to 255. An example of an IP address is
  202.64.254.10
- The numbers between the dots are called as "octets". The leftmost octet i.e. "202" represents the largest network and the rightmost octet "10" describes the specific machine which is being addressed.
- Although the IP addressing scheme can identify the computers on Internet, the numbers are impossible to remember. Therefore a method of recognizing the computers by names was deviced.

Domain name addressing :

- A smaller network making up the internet and having many computers within it is called a "domain".
- The domain may represent a type of organization or a geographical location. For example, all the computers on the Internet which belong to the educational institutions form a part of the "educational" or "edu" domain.
- Commonly used domain types are as follows :

Domain name	Description
com	A company or commercial organization
edu	Educational Institutes.
gov	Government organization
mil	Military sites
net	Network resources or Internet service provider.
org	Non profit or non commercial organizations.

- In order to accommodate different countries all over the world, a two letter abbreviation for every country in the world is added at the end of the domain name.
  e.g. techpub@pn2.vsnl.net.in
- Here the last two letters i.e. "in" represent "India". Commonly used abbreviations as country suffix are as follows :

Abbreviation	Country name
au	Australia
fr	France
ca	Canada
uk	United Kingdom
in	India

Take again the same example :

techpub@pn2.vsnl.net.in

Where reading from right to left,
in - India
net - Network Provider
vsnl - Organization Name, in this case it is VSNL
pn2 - Server or computer name
techpub - Users name

- The actual Internet protocol however will recognize only the IP number. The domain naming convention is introduced only for the user's convenience.
- Therefore there must be some mechanism to convert the domain names to the corresponding IP addresses, and vice versa.
- This is done by an electronic server called "domain name server" (DNS). It is an electronic directory which is maintained on almost all the host computers.

1.12.8 Internet Service Providers (ISPs) :

- Internet Service Providers (ISPs) all over the world offer various options and packages to the general public for internet access.
- With the privatization of the Internet services, various ISPs, are offering their services in India now. The major players currently are :
  1. Videsh Sanchar Nigam Limited (VSNL).
  2. Mahanagar Telephone Nigam Limited (MTNL).
  3. Satyam Infoway Limited (Satyam online).
  4. Bharti BT Internet Pvt. Limited (Mantra online).
  5. Reliance, Tata Indicom and many others.

1.12.9 Who Owns the Internet :

- The Internet is a decentralized network and no one runs or owns it. There is no organization or agency that controls the activity on the Internet.
- However there are a few organizations which co-ordinate and guide the technical parts of Internet. They are :
  1. Internet Engineering Task Force (IETF).
  2. Internet Research Task Force (IRTF).
  3. Internet Architecture Board (IAB).
- The Internet's communication protocols are developed and maintained by the IETF. The protocols are the methods by which computers on the Internet are connected.
- The long term research problem which may be critical to the Internet are looked into by the IRTF.
- Any major changes coming from the IETF are ratified by the IAB.

1.13 Accessing the Internet :

- Now we can access the Internet in a number of different ways. Earlier the only way to access the Internet was to obtain a telephone connection and then open an account with an ISP (Internet Service Provider). This type of Internet access is known as the basic access method and it is as shown in Fig. 1.13.1. So in this system the ISP will allow the user to access the Internet.

(G-1435) **Fig. 1.13.1 : Basic Internet access method**

- In modern era, the basic framework of Internet access has not changed considerably. But the ways of getting connected to the ISP from user's premises have changed to a great extent.
- The different possible ways are as follows :
  1. Dial up access 2. Cable modems

3. ADCL 4. Leased lines
5. ISDN

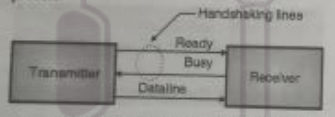
- One important thing to be noted is that all such methods still use the services of an ISP. They only provide different ways for the user to get connected to the ISP and Internet.

1.14 Protocols and Standards :

- Protocol and standards are the two frequently used words in data communication.
- Let us define them first and then explain them.

1.14.1 Protocols :

- In order to communicate successfully using serial digital data, some rules and procedures should be agreed upon at the sending and receiving ends of the system.
- Such rules and procedures are called as protocols. Different types of protocols are used in data communications.
- In data communications a message is more than one character. A group of characters forms a block.
- When a message is sent it is broken up into blocks. To identify a block one or more special characters are transmitted before and after the block.
- Some of the characters at the beginning and end of each block are used for "handshaking" purpose.
- Fig. 1.14.1 demonstrates the basic handshaking process.



(U-qu) **Fig. 1.14.1 : The handshaking process**

- The transmitter may send a "Ready" signal to the receiver indicating that it wants to send a character.
- The receiver identifies this signal and communicates its status on the "busy" line to the transmitter.
- If the receiver is busy then it is indicated by the receiver by sending a character on the busy line.
- The transmitter will send the data only when the receiver is not busy and the transmitter becomes ready.

1.14.2 Important Elements of a Protocol :

Some of the important elements of a protocol are :

1. Syntax 2. Semantics 3. Timing

1. Syntax :

Syntax means the structure or format of data. That means the order in which the data is presented.

2. Semantics :

- Semantics means the meaning of each section of bits, or how to interprete a particular pattern.
- It also tells us about what action is to be taken based on the interpretation.

3. Timing :

This aspect refers to two characters, namely the instant of sending the data and the speed at which it is to be sent.

1.14.3 Standards :

Data communication standards are classified into two categories :

1. De facto standards 2. De jure standards

De facto :

- The meaning of this word is "by fact" or "by convention".
- These are the standards that have not been approved by an organized body, but have been adopted as standards through widespread use.
- These standards are often established originally by the manufacturers.

De jure :

- The meaning of this word is "by law" or "by regulation".
- These are the standards that have been approved by an officially recognized body.

1.15 Standard Organizations for Data Communication :

Standards are developed by the standards creation committees, forums and government regulatory agencies.

Standard creation committees :

Some of the standard creation committees are :

1. International organization for standardization (ISO).
2. International Telecommunication Union - Telecommunication standards.
3. American National Standards Institute (ANSI).
4. Institute of Electrical and Electronics Engineers (IEEE).
5. Electronic Industries Association (EIA).

Regulatory agencies :

Federal communications commission (FCC) is the government regulator body in U.S. for all communication technology.

Standard organizations for data communications :

Some of the standard organizations for data communication are as follows :

1. **International Standard Organization (ISO) :**
   - ISO is the international organization for standardization. It creates sets of rules and standards for graphics, document exchange etc.
   - ISO endorses and coordinates the work of the other standard organizations.
2. **Consultative Committee for International Telephony And Telegraphy (CCITT) :**
   - The CCITT is now a standard organization for the United Nations.
   - Many government authorities and representatives are members of CCITT.
   - CCITT develops the recommended sets of rules and standards for telephone and telegraph communications.
   - CCITT has developed three sets of specifications as follows :
     1. V series for MODEM interfacing
     2. X series for data communication
     3. Q series for Integrated Services Digital Network (ISDN).
3. **American National Standards Institute (ANSI) :**
   ANSI is the official standard agency for United States.
4. **Institute of Electrical and Electronics Engineers (IEEE) :**
   IEEE is a U.S. based professional organization of electronics, computer and communications engineers.
5. **Electronic Industries Association (EIA) :**
   - EIA is a U.S. organization. It establishes and recommends industrial standards.
   - EIA has developed the RS (Recommended Standard) series of standards for data and telecommunications.
6. **Standards Council of Canada (SCC) :**
   SCC is the official standards agency for Canada. It has similar responsibilities to those of ANSI.

1.16 Internet Standards :

- An **Internet Standard** is defined as the specification which is used by and adhered by all the uses of the Internet.
- A specification is tested thoroughly before it is allowed to become the Internet standard.
- The procedure for a specification to become the Internet standard is as follows :
  1. A specification begins as an Internet draft.
  2. It remains an Internet draft i.e. a working document for 6 months.
  3. This draft can be then published as **Request For Comment (RFC).**
  4. Each such RFC is edited, given a number and made available to all the interested parties.
  5. RFCs are categorized as per their maturity levels.

Review Questions

- Q. 1 Explain simplex, half duplex and full duplex communication with examples.
- Q. 2 What is the difference between broadcast and point to point networks ?
- Q. 3 What is meant by internet work ?
- Q. 4 Write a short note on MAN.
- Q. 5 Write a short note on WAN.
- Q. 6 Compare LAN, WAN and MAN.
- Q. 7 Name the different network topology types.
- Q. 8 Explain the basic concepts of bus topology with the help of suitable diagram.
- Q. 9 State the important characteristics of bus topology.
- Q. 10 Name the transmission media used for bus LANs.
- Q. 11 State advantages and disadvantages of bus topology.
- Q. 12 Write a note on : Ring topology.
- Q. 13 What are the problems faced by the ring topology ?
- Q. 14 State the advantages and disadvantages of ring topology.
- Q. 15 Write a short note on star topology.
- Q. 16 What is the difference between single level star topology and two level star topology ?
- Q. 17 State the advantages and disadvantages of star topology.
- Q. 18 Write a short note on Mesh topology.
- Q. 19 State advantages and disadvantages of mesh topology.
- Q. 20 Write a short note on tree topology.
- Q. 21 Compare Ring and Bus.
- Q. 22 Compare Star and Ring.
- Q. 23 What are the important blocks of a data communication system ?
- Q. 24 What are the different methods of representing the data ?
- Q. 25 Explain the net structure of the Internet.
- Q. 26 State various services provided by the Internet.
- Q. 27 Name the Internet protocols.
- Q. 28 Write a note on : Internet address.
- Q. 29 State and explain various applications of Internet.

❑❑❑

CHAPTER 2

Unit I

Network Models

Syllabus :

Network models, Protocol layering, Scenarios, Principle of protocol layering, Logical connections, TCP/IP protocol suite, Layered architecture, Layers in TCP/IP protocol suite, Encapsulation and Decapsulation, Addressing, Multiplexing and Demultiplexing, Detailed introduction to physical layer, Detailed introduction to data link layer, Detailed introduction to the network layer, Detailed introduction to the transport layer, Detailed introduction to application layer.

2.1 Network Models :

- This chapter is the base for the remaining book. In this chapter the idea of network model has been discussed first and then the TCP/IP protocol suite has been discussed in detail.
- In order to define the computer network operations, two models have been derived. They are as follows :
  1. TCP/IP protocol suite.
  2. OSI model.
- The International Standards Organisation (ISO) covers all aspects of network communication in the Open Systems Interconnection (OSI) model.
- An OSI model is a layered framework for the design of network systems that allows for communication across all types of computer systems.
- The purpose of each layer is to offer certain services to the higher layers.
- Layer n on one machine (source) will communicate with layer n on another machine (destination).
- The rules and conventions used in this communication are collectively known as the layer n protocol.
- Basically a protocol is an agreement between the two communicating machines about how the communication link should be established, maintained and released.

2.2 Protocol Layering :

Protocol :

A protocol in data communication and networking is designed to define certain rules which are to be followed by both sender and the receiver and all the intermediate devices, so as to make the communication effective.

Protocol layering :

- For a simple type of communication, we need to use only one simple protocol.
- However for a complex type of communication, we need to divide the tasks among various layers. At each layer we need to use a protocol to carry out a specific task. This is known as **protocol layering.**

2.2.1 Scenarios :

Need of protocol layering :

In order to understand the need of protocol layering, let us develop two simple scenarios as follows :

1. First scenario :

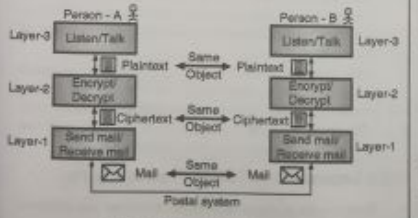
- In the first scenario, the communication between the source and destination is very very simple. Therefore only **one layer** will be sufficient to carry it out successfully.
- Assume that A and B are neighbours staying next to each other in the same building. They speak the same language and can talk face to face very easily and frequently.
- Therefore the communication between A and B can take place in one layer as shown in Fig. 2.2.1(a).

(G-2062) **Fig. 2.2.1(a) : A single layer protocol**

- Even in the simple single layer scenario, a set of rules must be followed. The set of rules which should be followed by both A and B are as follows :
  1. Both A and B should greet each other.
  2. They must choose proper words for communication.
  3. If A is speaking, B should remain silent and listen to A and vice versa.
  4. Both know that the communication should be bidirectional (dialog) and not unidirectional (monolog).
  5. They should say goodbye while leaving.

2. Second scenario :

- Now let us discuss the second scenario, in which person A has been offered a high-level position in his company and therefore he needs to relocate himself to company's another branch which is located in a city which is far away from person B.
- But A and B being very good friends wish to continue their communication about an innovative project to start a new business after their retirement.
- They choose the conventional mail through post office as their way of communication. But they do not want to reveal their ideas to anyone in case their mails are intercepted. Therefore both of them agree upon using the technique of **encryption and decryption.**
- Thus the sender encrypts the letter so that any intruder won't be able to read and understand the contents of the letter.
- Only the receiver knows how to decrypt it. So he will decrypt the received letter and understand its contents.
- From this discussion we conclude that the communication between A and B takes place in three layers as shown in Fig. 2.2.1(b). Let us assume that both persons A and B have three different machines or robots to perform the tasks specified at each layer.
- Refer Fig. 2.2.1(b) and imagine that person A sends the first letter to B. For this A talks to the robot at layer-3 as if it is person B.
- The layer-3 robot (or machine) listens to what A says and converts it into the **plaintext** i.e. a letter written in English. This letter is then sent to the robot or machine present at layer-2.
- This **plaintext** is encrypted by the machine at layer-2 to create the **ciphertext** which is sent to the machine present at layer-1.
- The robot/machine present at layer-1 will take the ciphertext, puts it in an envelope, write the addresses of sender and receiver over it and mails the envelope.



(G-2063) **Fig. 2.2.1(b) : A three layer protocol**

- At **person B's place,** the letter from the mailbox is picked up by the robot/machine at layer-1, the letter in the **ciphertext** is taken out of the envelope and gives it to the machine/robot at layer-2.
- The machine at layer-2, decrypts the ciphertext to obtain the **plaintext** and hands it over to the machine at layer-3.
- Finally the machine/robot present at layer-3 reads the plaintext as if person A is talking to person B.

Advantages of protocol layering :

1. It allows us to divide a complex task into many simpler tasks.
2. It allows us to separate the services from the implementation.
3. In practice, the communication does not always take place directly between the two end systems (A and B) but there are intermediate systems which need only some layers. Without the protocol layering the intermediate systems will be as complex as the two end system, thus making the entire system very complex and expensive.
4. It simplifies the design process as the function of each layer is well defined.
5. It provides flexibility to modify and develop network services.
6. Addition of new services and management of network infrastructure becomes easy.

Disadvantages of protocol layering :

1. We lose the touch with reality.
2. Sometimes the protocol layering can result in poor performance of protocol.

2.2.2 Principles of Protocol Layering :

There are two different principles of protocol layering. We will discuss them one by one.

1. First principle :

- According to the first principle, in order to have a successful bidirectional communication, each layer should be able to perform two opposite tasks one in each direction.
- For example layer-1 performs send and receive mail functions or layer-2 performs the encryption and decryption and so on.

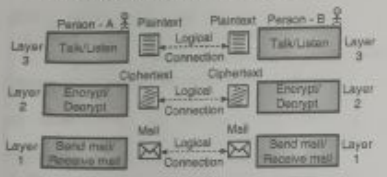
2. Second principle :

- According to the second principle, in protocol layering the two objects under each layer at both the ends should be the same.
- For example in Fig. 2.2.1(b), the object under the second layer at A as well as B is cipher text.

2.2.3 Logical Connections :

- We can think of the logical connection between each layer as shown in Fig. 2.2.2. This is after following the two principles of protocol layering.

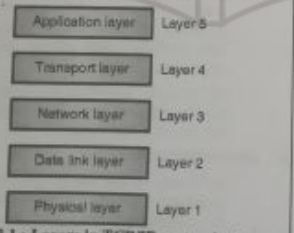
- Fig. 2.2.2 shows that there is a logical (imaginary) connection from a layer at A to the corresponding layer at B.
- The logical connection between each layer implies that there is a layer to layer communication.
- Due to logical connections, persons A and B can think that it is possible to send the object created from layer to the corresponding layer at the other end.



(G-2064) **Fig. 2.2.2 : Concept of logical connections between the peer layers**

2.3 TCP/IP Protocol Suite :

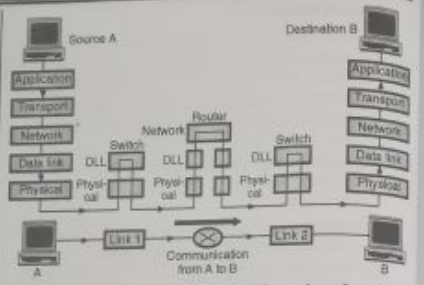
- After discussing about the concept of protocol layering and about the logical communication taking place between layers, now it is time to introduce the **TCP/IP protocol suite.**
- TCP/IP is the short form of two important protocols namely Transmission Control Protocol/Internet Protocol.
- A **protocol suite** is defined as the set of protocols organized in different layers. The TCP/IP protocol suite is used in Internet today.
- TCP/IP is a hierarchical protocol suite means that each upper layer protocol receives support and services from either one or more lower level protocols.
- In the original TCP/IP protocol suite, there were four software layers built upon the hardware. But today's TCP/IP protocol suite uses a five layer model as shown in Fig. 2.3.1.



(G-2065) **Fig. 2.3.1 : Layers in TCP/IP protocol suite**

2.3.1 Layered Architecture :

- In order to understand how the communication takes place between various layers of TCP/IP protocol suite, we have considered a small internetwork consisting of three LANs (links) with all LANs connected to each other via a router as shown in Fig. 2.3.2.



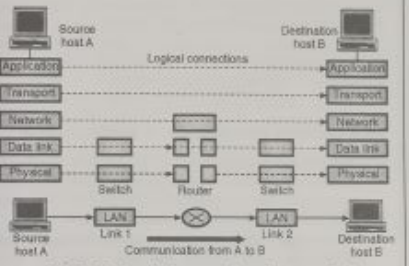
(G-2176) **Fig. 2.3.2 : Communication through an Internet**

- In Fig. 2.3.2, there are two computers A and B communicating with each other and three more devices namely : the link layer switch in link-1, the router and the link layer switch in link-2.
- Computer A is called as the **source host** and computer B is called as the **destination host.**
- Each device in the Internet has a specific role to play, depending on which each device uses a set of layers as shown in Fig. 2.3.2.
- All the five layers are involved in communication for the source and destination hosts A and B respectively.
- At the source host, a message is created at the application layer and then it is sent in down the layers in order to physically send it to the destination host.
- At the destination host this message is received at the physical layer and then it is delivered to the application layer via the other layers between the physical and application layers.
- At the router, as shown in Fig. 2.3.2 only three layers of TCP/IP protocol suite are needed to be involved. Thus a router does not need the transport or application layers when it is being used only for routing.
- The router is connected to multiple links. At each link we use a switch which involves only two layers of the TCP/IP protocol suite as shown in Fig. 2.3.2.
- However note that the link layer and physical layer protocols used by each link can be completely different.
- Thus the router may have to receive a packet from link-1 based on one pair of protocol and may have to deliver a packet to link-2 based on a totally different pair of protocols.
- Now consider a switch in Fig. 2.3.2 which shows that it has two different connections. But both of them belong to the same link. Therefore two different protocol pairs will not be involved. A switch has to deal with only one pair of DLL and physical layer protocols.

2.3.2 Layers in the TCP/IP Protocol Suite :

- Now we are going to discuss the functions and duties of various layers in the TCP/IP protocol suite.

- In this section, we will think about the logical connections between various layers, so as to clearly understand the duties of each layer.
- The logical connections in a simple internetwork have been shown in Fig. 2.3.3.



(G-2177) **Fig. 2.3.3 : Logical connections between the layers of TCP/IP suite**

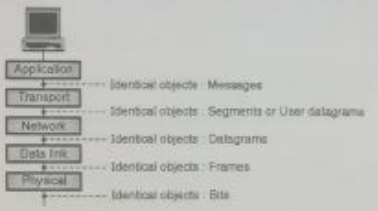
- Each layer has some specific duties and we can use the logical connections to think about them easily.
- From Fig. 2.3.3 it is clear that the network, transport and application layers have an **end-to-end** duty. But the data link and physical layers have the **hop to hop** duty. (Hop is a host or router).
- In this way the upper three layers have a **domain of duty** of the entire Internet while the lower two have a domain of duty of only link.

Data unit created by every layer :

- We can think about the logical connections in a different way i.e. in terms of the **data unit** created by each layer.
- The names of data units (packets) created by different layers are as follows :

Layer	Data unit	Layer	Data unit
Application	Message	Datalink	Frame
Transport	Segment	Physical	Bits
Network	Datagram		

- The data unit (packet) created by the top three layers, should not be changed by a router or a link layer switch.
- However the data unit created at the lower two levels can be changed only by the router. The link layer switches cannot modify it.
- The second principle that we discussed for the protocol layering has been shown in Fig. 2.3.4. Note that the objects shown below each layer related to each device are identical.



(G-2178) **Fig. 2.3.4 : Identical objects in the TCP/IP suite**

2.4 Detailed Description of Each Layer :

In this section we are going to discuss the duties of various layers in TCP/IP.

2.4.1 Detailed Introduction to Physical Layer :

- Physical layer is the lowest layer in the TCP/IP protocol suite. The communication at the physical layer level is still **logical** because of the presence of a hidden layer (transmission media) under the physical layer.
- The **primary responsibility** of the physical layer is to carry the individual bits present in a frame across the link.
- The transmission media (wired or wireless) is used for connecting two devices to each other. Here it is important to understand that the transmission media does not actually carry the bits.
- Instead it carries the electrical or optical signals which represents the bits which are to be carried from one device to the other.
- That means the bits received in a frame from the data link layer are transformed into an electrical or optical signal and sent over the transmission media.
- Still we consider **bit** as the data unit for communication between physical layers of two communicating devices.
- For the transformation of bits to signal, several physical layer protocols are available.

Following are the functions of the physical layer :

1. To define the type of encoding i.e. how 0's and 1's are changed to signals.
2. To define the transmission rate i.e. the number of bits transmitted per second.
3. To deal with the synchronization of the transmitter and receiver.
4. To deal with network connection types, including multipoint and point-to-point connections.
5. To deal with physical topologies i.e. bus, star, ring, or mesh.

6. To deal with the media bandwidth i.e. baseband and broadband transmission.
7. Multiplexing which deals with combining several data channels into one.
8. To define the characteristics between the device and the transmission medium.
9. To define the transmission mode between two devices i.e. whether it should be simplex, half duplex or full duplex.

Note : Passive hubs, simple active hubs, terminators, couplers, cable and cabling, connectors, repeaters, multiplexers, transmitters, receivers, transceivers are associated with the physical layer.

2.4.2 Detailed Introduction to Data Link Layer :

- An internetwork consists of many LANs and WANs, connected to each other by routers.
- While travelling from source to destination a datagram has to travel through many overlapping sets of links.
- It is the responsibility of router to choose the best possible link for a datagram to travel.
- When a router does so, it is the responsibility of the data link layer to take the datagram across the link.
- The said link can be anything such as a wired LAN, a wireless LAN, or a link layer switch etc. Every type of link will use different types of protocols. The data link layer should be able to handle all the different types of protocols and move the packet through the link.
- The data link layer receives a datagram from the network layer and encapsulates it into a packet called as **frame.**
- There are no specific data link layer protocols defined by the TCP/IP suite. Instead it supports all the standard protocols that can carry the datagram successfully over the link.
- The services provided by each data link layer protocol are different.

Following are the functions of data link layer :

1. Framing :

The bits received from the network layer are divided into another type of data units called frames at the data link layer.

2. Flow control :

It provides a flow control mechanism to avoid a fast transmitter from over-running a slow receiver by buffering the extra bits.

3. Physical addressing :

It adds a header to the frame which consists of the physical address of the sender and / or receiver of that frame.

4. Error control :

A trailer is added at the end of the frame in order to achieve error control. It also uses a mechanism to prevent duplication of frames.

5. Access control :

- The data link layer protocol perform an important function of determining which device has control over the link at any given time, when two or more devices are connected to the same link.
- The Institution of Electrical and Electronics Engineers (IEEE) felt the need to define the data link layer in more details, so they split it into two sub-layers :
  1. Logical Link Control (LLC).
  2. Media Access Control (MAC).

2.4.3 Detailed Introduction to Network Layer :

- The primary responsibility of the network layer is to create a connection between the source and destination computers. The communication at the network layer level is called as host to host communication.
- The several routers present between the source and destination hosts choose the best route for each travelling packet.
- Therefore the two responsibilities of the network layer are : host to host communication and routing of the packet through the possible routers.
- The main protocol in the network layer of the Internet is IP (Internet Protocol). The format of the packet (datagram) at network layer is decided by IP.
- The routing of datagrams from their source to destination is also the responsibility of IP. It achieves this by making each router forward the datagrams to the next router in its path towards the destination.
- IP is a **connectionless** protocol. It does not provide services like **flow control, error control** or even the **congestion control.**
- Therefore it is dependent on the transport layer in case if an application needs these services.
- The routing protocols included in the network layer are of unicast (one-to-one) and multicast (one-to-many) nature.
- These routing protocols have a responsibility of creating the forwarding tables for the routers to help them in the process of routing.
- There are some auxiliary protocols at the network layer, that are designed to assist IP in its delivery and routing tasks. The examples of such protocols are ICMP, IGMP, DHCP, ARP etc.
- The functioning of these protocols is as follows :

Sr. No.	Protocol	Function
1.	ICMP	To help IP report problems when routing a packet
2.	IGMP	Helps IP in multitasking
3.	DHCP	To help IP to get the network layer address for a host.
4.	ARP	Helps IP to find the link layer address of a host or router.

Functions of the network layer :

1. It translates logical network address into physical machine addresses i.e. the numbers used as destination IDs in the physical network cards.
2. It determines the quality of service by deciding the priority of message and the route a message will take if there are several ways a message can get to its destination.
3. It breaks the larger packets into smaller packets if the packet is larger than the largest data frame the data link will accept.
4. It is concerned with the circuit, message or packet switching.
5. It provides connection oriented services, including network layer flow control, network layer error control and packet sequence control.
6. Routers and gateways operate in the network layer.

2.4.4 Detailed Introduction to Transport Layer :

- The primary responsibility of the transport layer is also to provide an end to end connection.
- At the source host, the application layer sends a message to the transport layer which **encapsulates** it into a transport layer packet (which is also called as a **segment** or **user datagram**) and sends it through the logical connection (which is imaginary) to the transport layer of the destination host.
- In short the transport layer takes message from the application layer of source host and via the transport layer at the destination host delivers the message to the application layer at the destination.
- For the Internet applications, there are number of transport layer protocols designed to give specific service to various application programs.
- The main protocol in the transport layer is TCP (Transmission Control Protocol) which is a connection oriented protocol.
- The main task of TCP is to establish a logical connection between the transport layers of the source and destination hosts before actually transferring the data.
- Being connection oriented, the TCP is a reliable protocol which provides the following services to an application layer program :
  1. Flow control 2. Error control and
  3. Congestion control
- The other commonly used transport layer protocol is UDP (User Datagram Protocol). This is a connectionless protocol. Therefore it does not need to create any logical connection before transmitting the user datagrams.
- The UDP treats each datagram as a totally in dependent packet with absolutely no relation with the previous or next datagrams.
- UDP is a very simple protocol as compared to TCP. It does not provide flow control, error control or congestion control.
- UDP is an attractive protocol for certain application program specially for those who want to send small messages or those who do not afford retransmission of a packet if the packet is corrupted or lost.
- For new emerging applications in the field of multimedia, a new transport layer protocol has been designed which is called as SCTP (Stream Control Transmission Protocol).

Functions of transport layer :

The transport layer performs the following functions :

1. It divides each message into packets at the source and re-assembles then at the destination.
2. The transport layer header H4 includes a service point address to deliver a specific process from source to a specific process at the destination.
3. The transport layer is capable of either connectionless or connection-oriented transfer of data.
4. It performs end to end flow control. Flow control is an important function of the transport layer.
5. It makes sure that the entire message arrives at the receiving transport layer without error.

2.4.5 Detailed Introduction to Application Layer :

- The logical connection between the application layers of source and destination hosts is **end-to-end** type.
- The communication between the application layers of source and destination hosts takes place through all the layers.
- The application layer communication is between **two processes.** A process is nothing but a program running at the application layer.
- Thus the primary responsibility of the application layer is the **process to process communication.**
- There are many predefined protocols at the application layer in the Internet. Some of these protocols are HTTP, WWW, SMTP, FTP, TELNET, SNMP etc. These protocols and their functions are shown in Table 2.4.1.

Table 2.4.1

Sr. No.	Protocol	Function
1.	HTTP	As tool to access World Wide Web i.e. WWW.
2.	SMTP	It is the main protocol used in e-mail service.
3.	FTP	To transfer files from one host to the other.
4.	TELNET	To access a website remotely.
5.	SNMP	To manage the Internet.

Sr. No.	Protocol	Function
6.	DNS	To find the network layer address of a computer.
7.	IGMP	To collect the membership in a group.

The application layer performs the following functions :

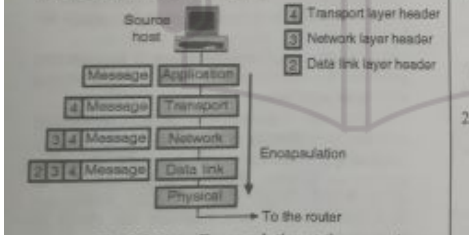
1. The application layer allows the creation of a virtual terminal which is the software version of a physical terminal. The user can log on to the remote host due to this arrangement.
2. The application layer provides File Transfer Access and Management (FTAM) which allows a user to access, retrieve, manage or control files in a remote computer.
3. It creates a basis for forwarding and storage of e-mails.

2.5 Encapsulation and Decapsulation :

- The encapsulation / decapsulation is one of the most important concepts in the protocol layering in Internet.
- This concept applied to a small Internet has been illustrated in Fig. 2.5.1.
- In this figure, the layers of data link switches have not been shown because encapsulation or decapsulation does not take place in the data link layer switches.
- In Fig. 2.5.1, the encapsulation takes place at the source host, decapsulation takes place at the destination host while both encapsulation and decapsulation takes place at the router.

2.5.1 Encapsulation at the Source Host :

Refer Fig. 2.5.1(a) to understand the process of encapsulation at the source host.



(G-2066) **Fig. 2.5.1(a) : Encapsulation at the source host**

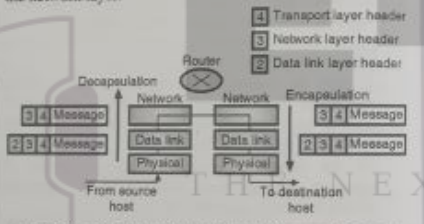
1. The data to be exchanged at the application layer is called as **message.** Normally a message does not contain any header or trailer. This message is passed on to the transport layer.
2. The transport layer takes this message which is also called as the **payload** and adds a transport layer header to it to produce the **segment** of **user datagram.** It is then passed on to the network layer. The transport layer header consists of the identifiers of the application programs at the source and destination and some additional information needed for the flow control, error control and congestion control.
3. The packet from transport layer is accepted by the network layer as its payload and adds its own header to it to produce a **datagram** as shown in Fig. 2.5.1(a). The network layer header contains the source and destination host's addresses and some additional information needed for checking errors in the header. This network layer packet (datagram) is then passed on to the data link layer.
4. The packet from the network layer is taken by the data link layer as its payload and adds its own header to it to produce a **frame** (packet at the data link layer). The link layer header contains the link layer addresses of the host or the next hop i.e. the router. This **frame** is then passed on to the physical layer for transmission.

2.5.2 Decapsulation and Encapsulation at the Router :

Refer Fig. 2.5.1(b) which illustrates the processes of decapsulation and then encapsulation occurring at a router connected to two or more links.

1. Decapsulation :

The router receives a set of bits at its input port. When these bits are delivered to the data link layer at the router, it decapsulates the datagram from the frame as passes it on to the network layer.

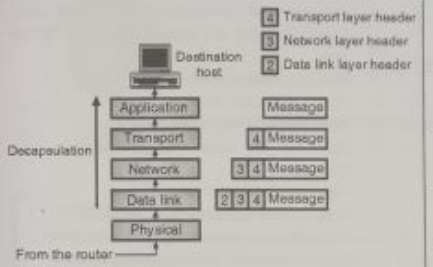


(G-2067) **Fig. 2.5.1(b) : Decapsulation / Encapsulation at the router**

2. The router checks the header for source and destination addresses. Then it refers to its forwarding tables and finds out the next hop to which this datagram is to be forwarded. Note here that the network layer at the router should not change the contents of the datagram unless fragmentation of the datagram is needed. Fragmentation is done if the datagram is too large in size. After all this, the datagram is passed to the data link layer.
3. At the data link layer, the datagram received from the network layer is encapsulated again into a frame and the frame is passed on to the physical layer which transmits it to the destination host.

2.5.3 Decapsulation at the Destination Host :

- At the destination host, only the decapsulation process is carried out at each layer, as shown in Fig. 2.5.1(c).

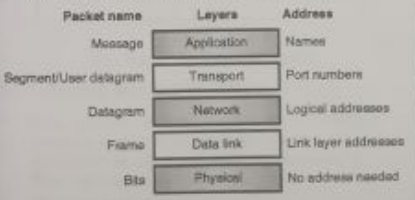


(G-2068) **Fig. 2.5.1(c) : Decapsulation at the destination host**

- At each layer, the payload is removed from the packet and the payload is delivered to the higher layer, by removing the headers at each stage.
- Finally after removing all the headers, the message is delivered to the application layer.
- It is important to note that the **error checking** is involved in the process of decapsulation at the destination host.

2.6 Addressing :

- Addressing is another important concept related to the protocol layering in the Internet.
- There is a logical connection between the pair of layers as discussed earlier. For any communication to take place between a source and a destination, two addresses namely source address and destination address are needed.
- Thus we will need four pairs of such addresses corresponding to the data link, network, transport and application layers.
- There is no need of addresses at the physical layer because communication at the physical layer takes place in bits which can not have an address.
- Fig. 2.6.1 shows the addressing at each layer.



(G-2069) **Fig. 2.6.1 : Addressing in TCP/IP protocol suite**

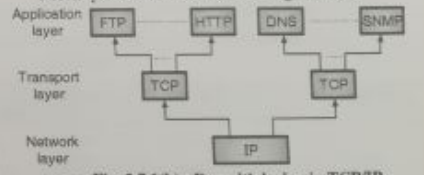
- Fig. 2.6.1 also shows the relationship between various layers, the addresses used in each layer and the name of the packet at each layer.
- We generally use the **names** to define the **site address** which provides the required services. For example **techmaxbook.com,** at the application layer. It is also possible to use the email address such as **jayantkatre@gmail.com.**
- The addresses at the transport layer are called as **port numbers.** These define the programs at the application layer of source and destination.
- There are several application layer programs running at a time. Port numbers are the local addresses which are used to distinguish between these programs.
- The addresses at the network layer are global in nature because the whole Internet is the scope of these addresses.
- The connection of a device to the Internet is uniquely defined by a network layer address.
- The addresses at the data link layer are called as the **MAC addresses.** These are the locally defined addresses. Each host or router in a network such as LAN or WAN always has a MAC address.

2.7 Multiplexing and Demultiplexing :

- In TCP/IP protocol, many protocols are being used at the same layer. Therefore multiplexing is needed at the source and demultiplexing is needed at the destination.
- In the process of **multiplexing** as shown in Fig. 2.7.1(a), a protocol at one layer in TCP/IP can encapsulate a packet (one at a time) from several protocols corresponding to the next higher layer in TCP/IP suite.

(G-2070) **Fig. 2.7.1(a) : Multiplexing in TCP/IP**

- In the process of demultiplexing, a protocol will decapsulate and deliver a packet one at a time to several protocols belonging to the next higher layer in TCP/IP protocol suite as shown in Fig. 2.7.1(b).



(G-2071) **Fig. 2.7.1(b) : Demultiplexing in TCP/IP**

- As shown in Fig. 2.7.1(a), at the transport layer two protocols TCP and UDP are capable of multiplexing the messages coming from various protocols at the application layer.

- Next the segments from TCP or user datagrams from UDP are accepted and multiplexed by IP at the network layer.
- IP can also multiplex the packets from some other protocols such as ICMP or IGMP etc.
- The frames at the data link layer level can carry the payload coming from the network layer protocols such as IP or ARP etc.

Review Questions

Q. 1 State the names of two network models.
Q. 2 Define the word protocol.
Q. 3 What is protocol layering ?
Q. 4 Explain the concept of logical connections.
Q. 5 Draw the layers of TCP/IP suite.
Q. 6 Explain the layered architecture of TCP/IP suite.
Q. 7 Explain in detail the physical layer in TCP/IP suite.
Q. 8 Explain in detail the data link layer in TCP/IP suite.
Q. 9 Explain in detail the network layer in TCP/IP suite.
Q. 10 Explain in detail the transport layer in TCP/IP suite.
Q. 11 Explain in detail the application layer in TCP/IP suite.
Q. 12 Name any three network layer protocols.
Q. 13 Write a short note on : IP.
Q. 14 State various functions of network layer.
Q. 15 State the two most important transport layer protocols.
Q. 16 State various duties of transport layer.
Q. 17 State any four application layer protocols.
Q. 18 Explain the concept of encapsulation in TCP/IP.
Q. 19 Explain the concept of decapsulation in TCP/IP.
Q. 20 Write a note on following in TCP/IP suite :
1. Addressing.
2. Multiplexing and demultiplexing.

❑❑❑

CHAPTER 3
Unit I

Data and Signals

Syllabus :

Data and signals, Analog and digital data, Analog and digital signals, Sine wave phase, Wavelength, Time and frequency domains, Composite signals, Bandwidth, Digital signal, Bit rate, Bit length, Transmission of digital signals, Transmission impairments, Attenuation, Distortion, Noise, Data rate limits, Performance, Bandwidth, Throughput, Latency (Delay).

3.1 Analog and Digital Signals :

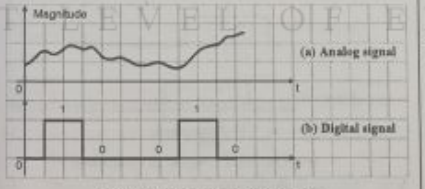
- Signals can be of two types :
1. Analog signals. 2. Digital signals.

1. Analog signal :

It is the signal in which the signal magnitude varies in a smooth fashion without any break with respect to time, as shown in Fig. 3.1.1(a).

2. Digital signal :

It is the signal in which the signal magnitudes has a constant level for some period of time, then it changes suddenly to another constant level as shown in Fig. 3.1.1(b). The examples of digital signal are binary signal, hexadecimal signal etc.



(L-24) **Fig. 3.1.1 : Types of signals**

3.1.1 Analog and Digital Data :

- Data are the entities which convey meaning, or information such as temperature, pressure etc. Signals are electric or electromagnetic representation of data. Thus signal is the representation of data.
- Data can be of two types :
1. Analog data 2. Digital data.

1. Analog data :

Analog data is the type of data that varies continuously (smoothly) with respect to time. Voice and video are the best examples of analog data. The other examples are temperature, pressure etc.

2. Digital data :

Digital data is the type of data that can take on discrete values i.e. it is discrete in nature. The examples of digital data are text and integers.

3.1.2 Sources of Digital Signal :

- The digital signals can be obtained directly from the computers. All the data used by the computers is digital.
- We can also use an A to D converter (Analog to digital converter) so as to convert analog signals into digital signals.

3.1.3 Advantages of Digital Signals :

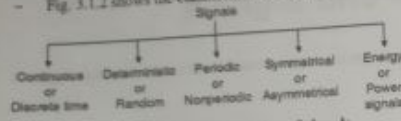
1. Digital signals can be processed and transmitted more efficiently and reliably than analog signals.
2. It is possible to store the digital data.
3. Play back or further processing of the digital data is possible.
4. The effect of "noise" (unwanted voltage fluctuations) is less. So digital data does not get corrupt.
5. It is possible to separate signal and noise and use repeaters between the transmitter and receiver.
6. Use of microprocessor and digital systems is possible.

3.1.4 Comparison of Digital and Analog Signals :

Sr. No.	Parameter	Analog signals	Digital signals
1.	Number of values	Infinite	Finite (2, 8, 16 etc.)
2.	Nature	Continuous	Discrete
3.	Sources	Signal generators, transducers etc.	Computers, A to D converters
4.	Examples	Sinewave, triangular wave	Binary signal

3.1.5 Classification of Signals :

- Fig. 3.1.2 shows the classification of signals.



(G-1289) **Fig. 3.1.2 : Classification of signals**

- Out of these we will concentrate only on periodic or non periodic signals.

3.1.6 Periodic and Non-periodic Signals :

Periodic signal :

- A signal which repeats itself after a fixed time period is called as a periodic signal. The periodicity of a signal can be defined mathematically as follows :

$x(t) = x(t + T_0)$: Condition of periodicity...(3.1.1)

Where T_0 is called as the period of signal x (t), in other words, signal x (t) repeats itself after a period of T_0 sec.

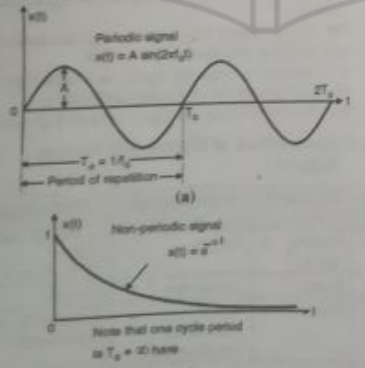
- Examples of periodic signals are sine wave, cosine wave, square wave etc. Fig. 3.1.3(a) shows a sine wave which is periodic because it repeats itself after a period T_0.

Non-periodic signal :

- A signal which does not repeat itself after a fixed time period or does not repeat at all is called as a non-periodic or aperiodic signal.
- The non-periodic signals do not satisfy the condition of periodicity stated in Equation (3.1.1).

$\therefore$ For a non-periodic signal $x(t) \neq x(t + T_0)$ ...(3.1.2)

- Sometimes it is said that an aperiodic signal has a period $T_0 = \infty$. Fig. 3.1.3(b) shows a decaying exponential signal.

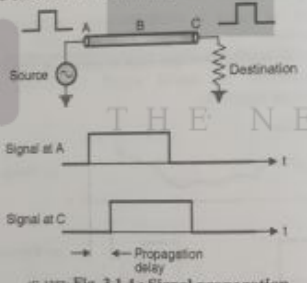


(G-1291) **Fig. 3.1.3 : Periodic and non-periodic signals**

- This exponential signal is non-periodic but it is deterministic because we can mathematically express it as $x(t) = e^{-at}$.

3.1.7 Signal Propagation :

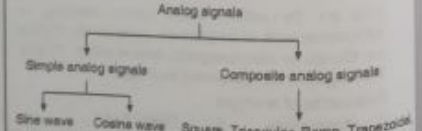
- Refer Fig. 3.1.4 which shows a signal source, a communication channel and the destination or signal receiver.
- The signal containing the data information is in the electric form. It is applied at point A of the conducting medium.
- The electrons in the conducting medium will transfer the charge to the adjacent electrons and the signal at point A gets transferred to B and then to C which is the receiving point.
- The shape of the signal at the receiver (point C) is almost same as that at the source (point A), but the signal reaches point C after a finite delay called **propagation delay.**
- The signal producing source in Fig. 3.1.4 can be a person talking on the phone or a computer producing a data signal or a video camera producing a video signal etc.
- Thus if we apply a signal at one end of the conducting medium then eventually this signal gets propagated to the other end of the medium with some time delay.



(G-1282) **Fig. 3.1.4 : Signal propagation**

3.2 Analog Signals :

We can classify the analog signals as follows :



(L-770) **Fig. 3.2.1 : Classification of analog signals**

3.2.1 Simple Analog Signal :

- It is the analog signal which cannot be decomposed into simpler signals. So this is the most basic analog signal which can be used as basic building block to build other composite signals.
- Examples of simple analog signal are sine and cosine waves.

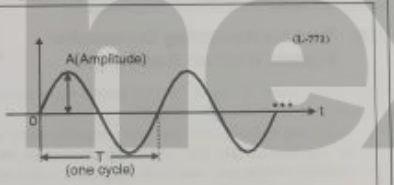
3.2.2 Composite Analog Signal :

- A composed analog signal is made of multiple sine or cosine waves of different amplitudes added to / subtracted from each other.
- Examples of composite analog signals are square wave, triangular wave, ramp, trapezoidal signal, pulse etc.

3.2.3 Sinewaves :

- It is the most basic type of periodic analog signal. That means it is the simple analog signal.
- Table 3.2.1 shows the graphical representation along with some of its characteristics.

Table 3.2.1 : Sinewave and its characteristics

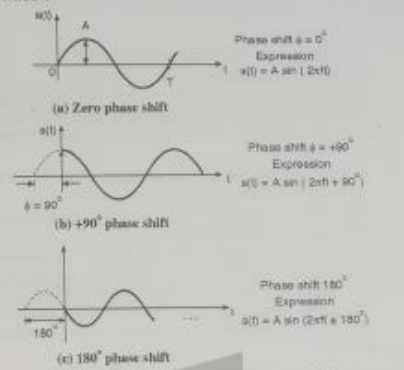


Mathematical expression	:	$s(t) = A \sin(2\pi ft + \phi)$
Peak amplitude	:	A
Frequency	:	f Hz = 1 / T
Period	:	T = (1/f) sec.
Phase	:	ϕ radians

Important points about frequency and time period :

1. Frequency and time period (T) are reciprocals of each other.
2. Frequency can be defined as rate of change of magnitude with respect to time. Change of signal magnitude in a shorter time span corresponds to high frequency.
3. Change of signal magnitude over a long time span corresponds to low frequency.
4. The signal which does not change at all has zero frequency. Example of such a signal is dc signal.
5. The signal that changes instantaneously has an infinite frequency. Example is delta signal or an impulse.

Phase :



(L-772) **Fig. 3.2.2 : Concept of phase shift**

- The letter ϕ in the mathematical expression for a sinewave denotes phase of the sinewave.
- It describes the position of the waveform with respect to time zero (i.e. t = 0). This is explained in Fig. 3.2.2.

Wavelength :

- Wavelength is another important characteristics of a signal travelling through a transmission medium.
- Wavelength relates the frequency of a sinewave with the speed of propagation as follows :

$$\text{Wavelength } \lambda = \frac{\text{Propagation speed}}{\text{Frequency}}$$

$$= \text{Propagation speed} \times \text{Period}$$

$$\therefore \lambda = \frac{C}{f} = C \times T \qquad ...(3.2.1)$$

- Wavelength of a signal is dependent on the medium. In data communication the wavelength is used to describe the transmission of light in an optical fiber.
- Wavelength is the distance travelled by a signal or an electromagnetic wave during the time period of one cycle of the signal.
- In vaccum the speed of light (c) is 3×10^8 m/s but it does not remain same in a cable. Hence the wavelength of the same signal will be different in air and on a cable.

3.3 Time and Frequency Domain :

It is possible to display a signal in either time domain or in frequency domain.

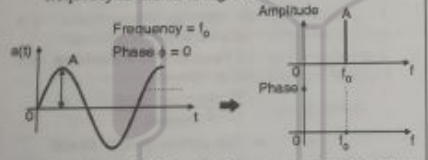
3.3.1 Time Domain Display of Analog Signals :

- It is the graph of signal magnitude with respect to time. That means on the x-axis we plot the time t and on the y-axis the magnitude is plotted. (Refer Fig. 3.3.1(a)).

- A time domain display tell us about the following :
  1. Shape of the signal
  2. Its frequency
  3. Type of the signal (periodic or nonperiodic).
  4. One cycle period.
- But we cannot know anything about what frequency components are present and in what proportion they have been mixed in order to obtain the particular shape of the signal.
- All this information can be obtained from the line spectrum of a signal.
- The phase and frequency cannot be explicitly measured from the time domain plot.
- In case of a composite waveform, which is composed of multiple sinewaves of different amplitudes, frequencies and phases, the time domain plot does not reveal any information about the relationship between amplitude, frequencies and phases of the various sinewaves.

3.3.2 Frequency Domain Display of a Signal :

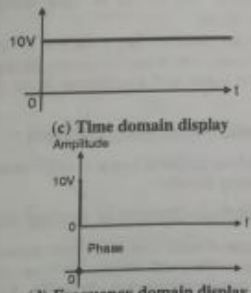
- It is the graph of signal amplitude (peak or sometimes rms), plotted on Y-axis with respect to frequency plotted on the x-axis.
- It is also possible to plot phase ϕ with respect to frequency as shown in Fig. 3.3.1.



(a) Time domain display (b) Frequency domain display

(L-773) **Fig. 3.3.1**

- Now refer Fig. 3.3.1(c) which shows a dc signal with amplitude 10 V. The frequency and phase both will be zero. The frequency domain display of this signal is shown in Fig. 3.3.1(d).



(c) Time domain display

(d) Frequency domain display

(L-774) **Fig. 3.3.1 : DC signal**

Conclusions :

1. In the frequency domain we can show two characteristics of the signal using one vertical line. First is the amplitude and second is the frequency of the signal.
2. We can also show the phase angle.
3. The analog signal is best represented in the frequency domain.

3.4 Composite Signals :

- Sine and cosine waves are the basic and purest signals but they are not useful in data communication.
- To use the sinewave for data communication, we have to change one or more of its characteristics, such as amplitude in accordance with the data to be transmitted.
- But then the sinewave does not remain the simple sinewave, instead it becomes a **composite signal** which is made up of many simple sinewaves.

Thus when we change amplitude, frequency or phase of a single frequency signal, it gets transformed into a **composite signal** made of many frequencies.

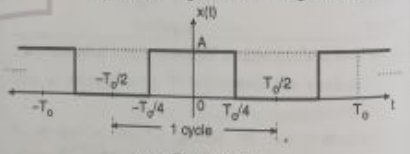
3.5 Tool for Analyzing Composite Signals (Fourier Analysis) :

- Fourier series is one of the well known tools used for the analysis of composite signals.
- According to fourier analysis, any composite signal can be represented as sum of simple sinewaves with different frequencies, amplitudes on phases as follows :

$s(t) = a_0 + a_1 \sin(2\pi f_1 t + \phi_1) + a_2 \sin(2\pi f_2 t + \phi_2) + \ldots$

where a_0, a_1, a_2...are called as the Fourier coefficients.

Ex. 3.5.1 : Obtain the quadrature fourier series for the square wave signal shown in Fig. P. 3.5.1.



(G-819) **Fig. P. 3.5.1**

Soln. :

- Let us obtain the quadrature fourier series for the given square wave. The quadrature fourier series is given by the following expression :

$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n \omega_0 t) + \sum_{n=1}^{\infty} b_n \sin(n \omega_0 t) \quad ...(1)$$

- Substituting $\omega_0 = 2\pi f_0 = \frac{2\pi}{T_0}$ into Equation (1) we get,

$$\therefore \quad x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left[\frac{2\pi nt}{T_0}\right] + \sum_{n=1}^{\infty} b_n \sin\left[\frac{2\pi nt}{T_0}\right] \quad ...(2)$$

- To find the fourier coefficients, we must consider one complete cycle of x (t) for integration.
- Here we will consider one cycle from $t = -\frac{T_0}{2}$ to $t = \frac{T_0}{2}$. Let us obtain the fourier coefficients now.

1. To obtain the value of a_0 :

$$a_0 = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x(t)\,dt = \frac{1}{T_0}\int_{-T_0/4}^{T_0/4} x(t)\,dt$$

.....As x (t) exists from $-T_0/4$ to $T_0/4$ only

$$= \frac{1}{T_0}\int_{-T_0/4}^{T_0/4} A\,dt \quad .....\text{As } x(t) = A \text{ from } -T_0/4 \text{ to } T_0/4$$

$$\therefore \quad a_0 = \frac{A}{T_0}[t]_{-T_0/4}^{T_0/4} = \frac{A}{T_0} \times \frac{T_0}{2} = \frac{A}{2} \quad ...(3)$$

2. To obtain the value of a_n : a_n is given by,

$$a_n = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2} x(t) \cdot \cos\left[\frac{2\pi nt}{T_0}\right] dt$$

$$= \frac{2}{T_0}\int_{-T_0/4}^{T_0/4} A \cos\left[\frac{2\pi nt}{T_0}\right] dt$$

$$= \frac{2A}{T_0} \cdot \frac{1}{\frac{2\pi n}{T_0}} \left[\sin\left(\frac{2\pi nt}{T_0}\right)\right]_{-T_0/4}^{T_0/4}$$

$$= \frac{A}{\pi n}\left[\sin\left(\frac{2\pi n\, T_0/4}{T_0}\right) - \sin\left(\frac{-2\pi n\, T_0/4}{T_0}\right)\right]$$

$$= \frac{A}{\pi n}\left[\sin\left(\frac{n\pi}{2}\right) + \sin\left(\frac{n\pi}{2}\right)\right] = \frac{2A}{\sin}\left[\frac{\pi nt}{T_0}\right]$$

$$\therefore \quad a_n = \frac{2A}{n\pi}\sin\left(\frac{n\pi}{2}\right) \quad ...(4)$$

3. To obtain the value of b_n : b_n is given by,

$$b_n = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2} x(t) \sin\left[\frac{2\pi nt}{T_0}\right] dt$$

$$= \frac{2}{T_0}\int_{-T_0/4}^{T_0/4} A \sin\left[\frac{2\pi nt}{T_0}\right] dt$$

$$= \frac{-2A}{T_0} \cdot \frac{1}{\frac{2\pi n}{T_0}} = \left[\cos\left(\frac{2\pi nt}{T_0}\right)\right]_{-T_0/4}^{T_0/4}$$

$$= \frac{-A}{\pi n}\left[\cos\left(\frac{2\pi n\, T_0/4}{T_0}\right) - \cos\left(\frac{2\pi n\, T_0/4}{T_0}\right)\right]$$

$$\therefore \quad b_n = \frac{-A}{\pi n}[\cos(n\pi/2) - \cos(n\pi/2)]$$

$$\therefore \quad b_n = 0 \quad ...(5)$$

4. Substitute the values of a_0, a_n and b_n in the Equation (2) we get the quadrature fourier series for x (y) as :

$$x(t) = \frac{A}{2} + \sum_{n=1}^{\infty} \frac{2A}{n\pi}\sin\left(\frac{n\pi}{2}\right)\cos n\,\omega_0 t \quad ...(6)$$

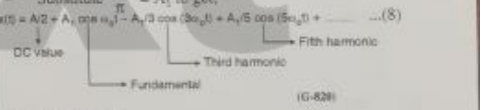
Explanation : Equation (6) can be expanded by opening the summation sign as follows :

$$x(t) = \frac{A}{2} + \frac{2A}{n\pi}\sin\left(\frac{\pi}{2}\right)\cos\omega_0 t + \frac{2A}{2\pi}\sin\pi\cos 2\omega_0 t + \frac{2A}{3\pi}\sin\frac{3\pi}{2}\cos 3\omega_0 t + \ldots$$

$$= \frac{A}{2} + \frac{2A}{n\pi}\cos(\omega_0 t) - \frac{2A}{3\pi}\cos(3\omega_0 t) + \frac{2A}{5\pi}\cos(5\omega_0 t) + \ldots \quad ...(7)$$

- Substitute $\frac{2A}{\pi} = A_1$ to get,

$$x(t) = A/2 + A_1 \cos\omega_0 t - A_1/3 \cos(3\omega_0 t) + A_1/5 \cos(5\omega_0 t) + \ldots \quad ...(8)$$



(G-828)

Conclusions :

- Equation (8) shows that a square wave consists of a dc component, fundamental and only the odd harmonic components such as third, fifth, seventh ... etc.
- The even harmonic components are automatically cancelled.
- The amplitude of harmonic components decreases with increase in their order.

3.5.1 Fundamental Component :

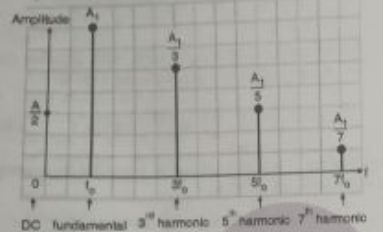
The composite signal x(t) has a frequency of f_0. Now refer to Equation (8) in Ex. 3.5.1. The second term $A_1 \cos\omega_0 t$ has the same frequency as that of composite signal. It is called as fundamental frequency or fundamental component.

3.5.2 Harmonics :

- The other sinewaves (cosine terms in Equation (8)) having frequencies which are integral multiples of fundamental frequency are called as harmonics.
- The sinewave having a frequency $3f_0$ is called as third harmonic, then the sinewave with frequency $5f_0$ is fifth harmonic and so on.

.3 Frequency Spectrum :

- The frequency domain description of a signal is called as frequency spectrum.
- A square wave is made of fundamental and odd harmonics (third, fifth, seventh etc). The frequency spectrum of a square wave is shown in Fig. 3.5.1.



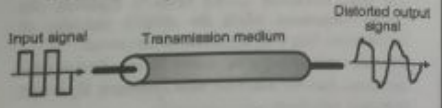
(G-821) **Fig. 3.5.1 : Frequency spectrum of a square wave**

3.6 Composite Signal and Transmission Medium :

- The data is generally in the form of pulses and pulse is a composite signal which contains many frequencies.
- Note that the peculiar shape of a pulse is due to the sum of specific frequencies at specific amplitudes and phases.
- If there is any change in the amplitudes or phases of these frequency components, then the shape of the pulse will not remain the same.

3.6.1 Medium :

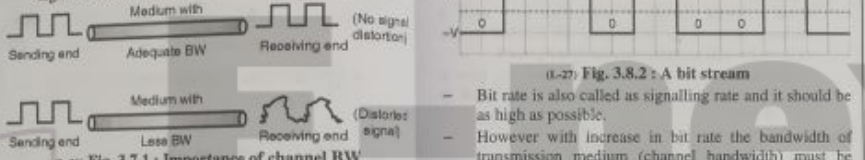
- The signal always travels over some medium from sender to destination.
- The medium can be a coaxial cable or optical fiber etc. A medium does not pass all frequencies equally due to its inadequate frequency spectrum.
- It may pass some frequencies and weaken or block the other frequencies.
- Hence when a composite signal is passed over such a transmission medium, at the receiving end we get a wave, having a different shape as shown in Fig. 3.6.1.
- To avoid the signal distortion, the medium should pass all the frequencies present at the input without any change.
- But no medium is perfect and so some signal distortion is bound to take place.



(G-82) **Fig. 3.6.1 : Signal distortion on a transmission medium**

3.7 Channel Bandwidth :

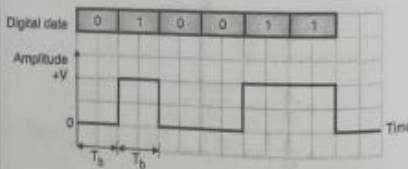
- The range of frequencies that contain the information is called as the bandwidth. But the term channel bandwidth is used to describe the range of frequencies required to transmit the desired information.
- For example the amplitude modulation (AM) system needs a channel bandwidth of 10 kHz to transmit a signal of 5 kHz bandwidth.
- But the single sideband system (SSB) needs only 5 kHz channel bandwidth to transmit the same signal.
- All the efforts should be made to reduce the required channel bandwidth so that we can fit in more number of channels in the same available EM spectrum.
- Bandwidth of a medium (also called as channel bandwidth) is defined as the maximum frequency it can allow to pass through it without attenuating it and without distorting the shape of the signal.
- If the medium has less bandwidth than required, then signal distortion will take place as shown in Fig. 3.7.1.



(L-30) **Fig. 3.7.1 : Importance of channel BW**

3.8 Digital Signals :

- The input data which is either analog or digital can also be represented by a digital signal.
- A digital signal is a discrete time signal having finite number of amplitudes. For example see the digital signal shown in Fig. 3.8.1.
- A 0 is represented by zero volts and a 1 by some positive voltage.



(L-26) **Fig. 3.8.1 : Digital signal**

3.8.1 Bit Interval (T_b) :

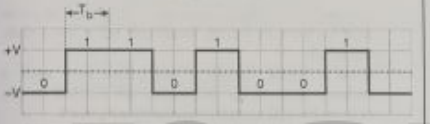
- The bit interval is the time corresponding to one single bit (0 or 1).
- As shown in Fig. 3.8.2, time corresponding to a 0 or a 1 is T_b hence it is the bit interval or bit length.

3.8.2 Bit Rate (Data Rate) :

- Bit rate is defined as the number of bits transmitted or sent in one second. It is expressed in bits per second (bps).
- Relation between bit rate and bit interval is as follows :

$$\text{Bit rate} = \frac{1}{\text{Bit interval}}$$

- Bit rate is also called as signalling rate and is defined as the number of bits which can be transmitted in a second.
- If the bit duration is "T_b" then bit rate will be $1 / T_b$. Look at Fig. 3.8.2, you will see that the bit duration is necessarily equal to the pulse duration.
- In Fig. 3.8.2 the first pulse is of two bit duration.

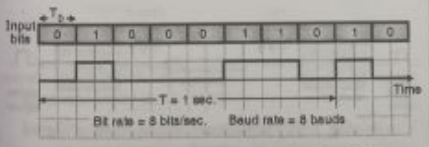


(L-27) **Fig. 3.8.2 : A bit stream**

- Bit rate is also called as signalling rate and it should be as high as possible.
- However with increase in bit rate the bandwidth of transmission medium (channel bandwidth) must be increased, in order to ensure that the signal is received without any distortion.

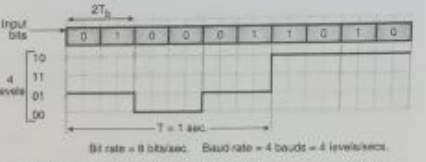
3.8.3 Bauds (or Baud Rate) :

- Baud is the unit of signalling speed or modulation rate or the rate of symbol transmission.
- It indicates the rate at which a signal level changes over a given period of time.
- When binary bits are transmitted as an electrical signal with two levels "0" and "1" the bit rate and the modulation rate i.e. baud rate are same. This is as shown in Fig. 3.8.3(a).
- Note that for a two level signal (binary signal) the bit rate and bauds are equal.
- Now consider Fig. 3.8.3(b) where four different levels are used to represent the data.



(L-28) **Fig. 3.8.3(a) : Baud rate for two level modulation**

- Each level is being represented by a combination of two bits i.e. 00 or 01 etc.
- The bit rate is therefore not equal to the baud rate.
- The bit rate is 8 bits/sec. but baud rate is only 4 bauds as there are 4-levels per second.



(L-29) **Fig. 3.8.3(b) : Baud rate for a four level modulation**

Ex. 3.8.1 : For a binary PCM system, the number of bits per transmitted word is 8 and the sampling frequency f_s = 8 kHz. Calculate the bit rate and baud rate.

Soln. :

Given : $N = 8$, $f_s = 8$ kHz

$$\text{Bit rate} = N \times f_s = 8 \times 8 \text{ kHz}$$
$$= 64 \text{ k bits/sec.}$$
$$\text{Baud rate} = \text{Bit rate} = 64 \text{ kHz}$$
(as transmission is binary)

Ex. 3.8.2 : For the same data in the previous example, calculate the bit rate and baud rate if a QPSK system is used.

Soln. :

In the QPSK system, two successive bits are clubbed together to form one message. Hence one symbol corresponds to 2 bit duration.

$$\therefore \text{Baud rate} = \frac{1}{2} \times \text{bit rate} = 32 \text{ k bits/sec.}$$

Bit rate does not change.

Ex. 3.8.3 : A system sends a signal that can assume 8 different voltage levels. It sends 400 of these signals per second. What are the baud and bit rates ?

Soln. :

1. As the signal assumes 8 different voltage levels we need 3 bit digital signal to have 8 different combinations. Hence the number of bits per voltage level is 3. Let each voltage level represent one symbol.
   $\therefore$ Number of bits / symbol = 3
2. The system sends 400 signal / sec. Hence the number of symbols transmitted per second is also 400.
   $\therefore$ Symbol rate = Number of symbols / sec.
   = 400 symbols / sec.
3. The baud rate is defined as the number of symbols per second.
   $\therefore$ Baud rate = Symbol rate.
   $\therefore$ Baud rate = 400 symbols / sec. ...Ans.
4. We are using 3 bit to represent each symbol.
   So bit rate = 3 × symbol rate = 3 × 400
   = 1200 bits / sec. ...Ans.

Ex. 3.8.4 : A system sends a signal that can assume 4 different voltage levels. It sends 200 of such signals per second. What is the baud rate ?

Soln. :

This example is similar to Ex. 3.8.3. A system has 4 different voltage levels and it sends 200 of such signals/sec. Hence number of voltage levels transmitted will be 200/sec. Therefore the symbol rate is 200 symbols / sec.

∴ Baud rate = 200 symbols / sec. ...Ans.

Ex. 3.8.5 : A system sends a signal that can assume two different voltage level. It sends 100 of signal per second, what is baud rate ?

Soln. :

1. As the signal assumes 2 different voltage levels we need. 1 bit digital signal to have 2 different combinations. Hence the number of bits per voltage is 1. Let each voltage level represent one symbol.
   ∴ Number of bits / symbol = 1
2. The system sends 100 signals / sec. Hence the number of symbols transmitted per second is also 100.
   ∴ Baud rate = 100 symbols / sec.

3.8.4 Bit Length :

- The bit length for a digital signal is similar to the term wavelength for an analog signal.
- Bit length of a digital signal is defined as the distance corresponding to one bit on the transmission medium. It is measured in meters or cm.

∴ Bit length = Propagation speed $\times$ Bit duration = $\frac{\text{meters}}{\text{sec}} \times \text{sec}$

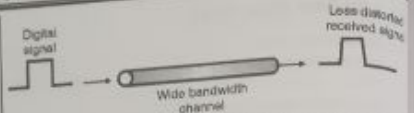
3.9 Transmission of Digital Signals :

The digital signals can be transmitted from one point to the other using one of the following two approaches :

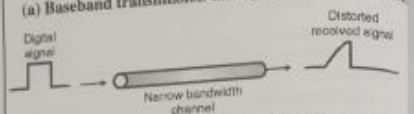
1. Baseband transmission
2. Bandpass transmission (with modulation)

3.9.1 Baseband Transmission :

- A baseband digital signal is the original signal without any modulation. In the modulation process the baseband digital signal is converted into analog signal.
- Baseband transmission is the transmission of baseband digital signal.
- A baseband signal occupies bandwidth from 0 to f_1 Hz.
- Hence baseband transmission requires the use of low pass channel. This low pass channel can have a narrow bandwidth or wide bandwidth.
- If we send the digital signals over the low pass channel with a small bandwidth (telephone cable) then same frequency components in the digital signal get blocked and the shape of the received signal will be badly distorted as shown in Fig. 3.9.1(b).



(a) Baseband transmission through a wideband channel



(b) Baseband transmission through a narrowband channel

(G-1301) **Fig. 3.9.1**

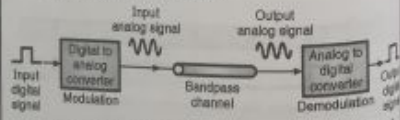
- A practically available medium such as coaxial cable does not have an infinite bandwidth but it has a wide bandwidth.
- When a digital signals is transmitted over such a medium, some of the frequencies are blocked by the medium but still enough frequencies are passed.
- So the digital signal at the receiving end will have a different shape with a small distortion as shown in Fig. 3.9.1(a).

Conclusion :

Baseband transmission of digital signals over a low pass channel without waveform distortion is possible if and only if the channel has a wide or infinite bandwidth.

3.9.2 Broadband Transmission (With Modulation) :

- A baseband signal is passed through a D to A converter to obtain an equivalent analog signal. This is modulation and the modulated analog signal is called a broadband signal.
- Transmission of a broadband signal is known as broadband transmission of digital signal. The spectrum of a broadband signal extends from f_1 to f_2 so it is a bandpass spectrum.
- We have to use a bandpass channel to carry the transmission. Fig. 3.9.2 shows the modulation process and broadband transmission.



(G-1302) **Fig. 3.9.2 : Broadband transmission of digital signal**

- Note that the output digital signal is distortion free.
- We can send the digital signals over a bandlimited medium. The best example of this is the data transmission over telephone cables in internet.

- The most important question here is that what should be minimum bandwidth of the medium (B Hz) if we want to transmit a signal of n bps.
- The answer to this question will be given when we study the Nyquist theorem and Shannon capacity.

3.9.3 Relation between Required Bandwidth and Bit Rate :

- In computer communication we have to send as many bits as possible per second for fast data transfer.
- That means the bit rate should be as high as possible. But increase in bit rate has an undesired side effect.
- The signal bandwidth and the required bandwidth of the medium (channel bandwidth) increase with increase in bit rate. If we double the bit rate then the required channel bandwidth needs to be doubled.
- Thus bit rate and bandwidth are proportional to each other.
- The general relation between required bandwidth (B) and bit rate (n) is as follows

$$B \geq \frac{n}{2} \text{ or } n \leq 2B.$$

- Thus over a medium having a bandwidth of 4 kHz we can send a digital signal with a bit rate upto 8 kbps.
- In practice the maximum bit rate can be more than 30 kbps using the traditional MODEMS.

3.10 Some Important Definitions :

Channel Capacity (C) :

- The channel capacity is defined as the maximum data rate at which the digital data can be transmitted over the channel reliably.
- The various other concepts related to channel capacity are as follows :
  1. Data rate
  2. Bandwidth
  3. Noise
  4. Error rate

Data Rate :

- It is defined as the number of bits transmitted by the transmitter per second. It indicates how fast a signal can be transmitted reliably over the given medium.
- This capability depends on the following factors :
  1. The amount of energy put into transmitting each signal.
  2. Distance to be travelled.
  3. Noise.
  4. Channel bandwidth

Channel bandwidth :

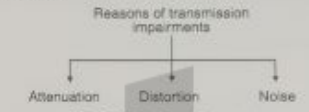
- The bandwidth of the communication medium should be large enough to transmit the digital signal reliably.
- An inadequate bandwidth will distort the signal and introduce errors into the received signal.

Noise : This is the average level of noise over the communication path.

Error rate : It is defined as the rate at which errors occur in the received (or detected) signal.

3.11 Transmission Impairments :

- In any communication system, the received signal is never identical to the transmitted one due to some transmission impairments.
- The quality of analog signals will deteriorate due to the transmission impairment whereas errors get introduced if the signal is digital.
- The most important reasons for the impairments are as follows :
  1. Attenuation and attenuation distortion
  2. Delay distortion
  3. Noise.



(G-46) **Fig. 3.11.1 : Impairment types**

3.11.1 Attenuation :

- The strength of a signal (signal energy) decrease with increase in distance travelled over a medium. This is known as attenuation.
- When any signal travels over a medium or channel, it loses some of its energy in the form of heat in the resistance offered to the signal by the medium.
- Attenuation is expressed in decibels as,

$$\text{Attenuation (dB)} = 10 \log_{10} \frac{P_{out}}{P_{in}} \quad ...(3.11.1)$$

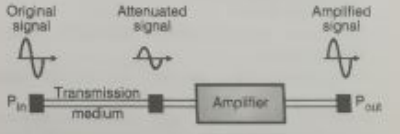
Where P_{in} = Power at the sending end

P_{out} = Power at the receiving end.

- Attenuation decides the signal strength and hence the signal to noise ratio and the quality of received signal.

Remedy to reduce attenuation :

We can introduce amplifiers to compensate for any loss of signal as shown in Fig. 3.11.2. Note that amplification and attenuation are exactly opposite to each other.



(G-87) **Fig. 3.11.2 : Concept of attenuation**

Ex. 3.11.1 : Calculate the attenuation in dB if input power is three times higher than the received power.

Soln. :

$$P_{in} = 3 P_{out}$$

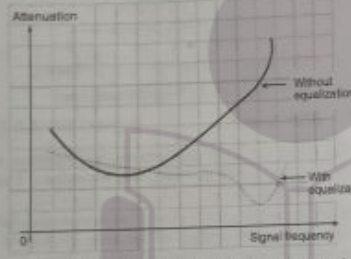
$$\therefore \frac{P_{out}}{P_{in}} = \frac{1}{3}$$

$$\therefore \text{Attenuation in dB} = 10 \log_{10} \frac{P_{out}}{P_{in}} = 10 \log_{10} (1/3)$$

$$= -4.8 \text{ dB} \quad \text{...Ans.}$$

Effect of frequency on attenuation :

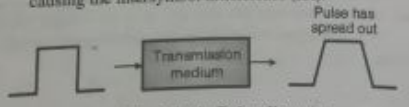
- Attenuation increases with increase in signal frequency.
- The SNR can be improved by using amplifiers or repeaters to boost the signal strength.
- The effect of frequency on attenuation is important if the signal is analog. This effect can be reduced by using equalizers. Another way to do this is to use an amplifier that amplifies the higher frequencies more than lower frequencies.
- The effect of equalization on attenuation is illustrated in Fig. 3.11.3.



(G-88) **Fig. 3.11.3 : Effect of equalizer on attenuation**

3.11.2 Delay Distortion :

- This problem is particularly present in the wired media.
- This distortion is caused due to a property which states that the velocity of propagation of a signal through a medium varies with frequency.
- The velocity is maximum near the center frequency and reduces as the signal frequency deviates away from the center frequency on both sides of frequency spectrum.
- Hence various frequency components of a signal arrive at the receiver at different instants of time resulting in phase shifts between different frequencies.
- This distorts the signal and the distortion is called as delay distortion.
- Delay distortion is particularly important for digital data.
- Due to delay distortion a digital pulse transmitted over a medium tends to spread out as shown in Fig. 3.11.4 causing the intersymbol interference (ISI).

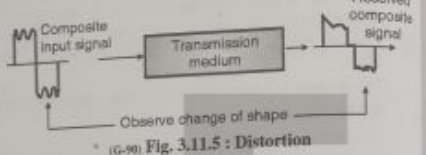


(G-89) **Fig. 3.11.4 : Delay distortion**

- The Intersymbol interference (ISI) put a major limitation on the maximum bit rate over a transmission channel.
- Delay distortion can be reduced by using equalizers.

3.11.3 Harmonic Distortion :

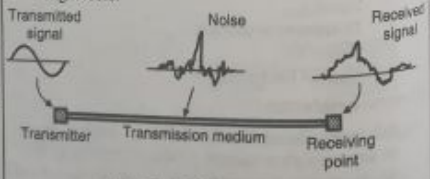
- Another meaning of distortion is change in shape of the signal. The shape of the received signal can be substantially different than the transmitted signal.
- If the medium is not perfect, then all the frequency components present at the input will not be equally attenuated and will not be proportionally delayed.
- Hence the shape of the received signal is different from that of the transmitted signal as shown in Fig. 3.11.5.



(G-90) **Fig. 3.11.5 : Distortion**

3.11.4 Noise :

- When the data travels over a transmission medium, noise gets added to it.
- Noise is a major limiting factor in communication system performance.
- Noise can be categorized into four types as follows :
  1. Thermal noise 2. Intermodulation noise
  3. Crosstalk 4. Impulse noise
- Thermal noise is due to the random motion of electrons in a wire. But the noise induced into a wire is generated by the sources such as electric motors and various appliances.
- Crosstalk is the interference caused by one wire to the other one. Here one wire radiates its signal by acting as a sending antenna and the other one acts as a receiving antenna to induce voltage into it.
- Impulse noise is in the form of a high energy spike. It is basically a pulse of short duration which comes from power lines, lightning etc.
- The effect of noise on a signal has been illustrated in Fig. 3.11.6.



(L-719) **Fig. 3.11.6 : Effect of noise**

Thermal Noise :

- It is due to thermal agitation of electrons.

- Thermal noise is present in all electronic devices and the transmission media.
- Thermal noise is proportional to temperature and it is uniformly distributed across the frequency spectrum. So it is called as white noise.
- We cannot eliminate the thermal noise completely. So it puts a limit on the performance of the communication system.
- The thermal noise power is given by,

$$N = kTB \text{ Watts} \quad \text{...(3.11.2)}$$

where k = Boltzman's constant
T = Absolute temperature °K
B = Bandwidth

- It can be expressed in dBW as :

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

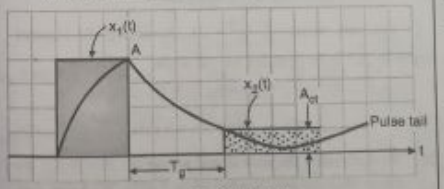
- Thermal noise can be reduced by reducing temperature or bandwidth.

Intermodulation Noise :

- If signals at different frequencies are transmitted simultaneously on a common transmission medium (e.g. FDM) then it results in intermodulation noise.
- The intermodulation produces frequency components at sum and difference frequencies of the frequencies travelling on the medium.
- Suppose signals of frequencies f_1 and f_2 are travelling simultaneously over the same medium, then due to intermodulation produces frequency components $(f_1 + f_2)$ and $(f_1 - f_2)$, which are called intermodulation noise.
- This is similar to a mixer. So whenever there is any non- linearity in the transmitter or channel, the intermodulation noise is produced.

Crosstalk :

- Crosstalk basically means interference between the adjacent telephone channels.
- It is the unwanted coupling of information from one channel to the other adjacent channels. The guard time (T_g) is the time spacing introduced between the adjacent telephone channels represented by $x_1(t)$ and $x_2(t)$ in Fig. 3.11.7.
- A signal without baseband filtering also has crosstalk if the pulse tail or postcursor overlaps into the next time slot of the frame as shown in Fig. 3.11.7.

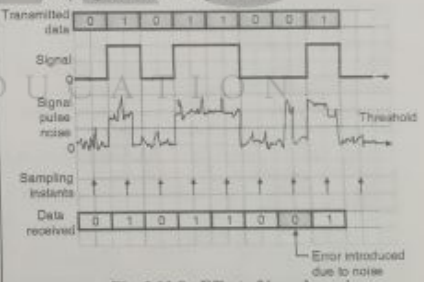


(G-91) **Fig. 3.11.7 : Crosstalk**

- Pulse overlap shown in Fig. 3.11.7 can be controlled by using "guard time" (T_g) between the pulses. The guard time (T_g) is analogous to the guard bands between the channels in the FDM systems.
- In practice crosstalk can be experienced while using a telephone. We can hear another conversation due to crosstalk.
- It can occur due to electrical coupling between nearby twisted pairs or sometimes even in the coaxial cables.
- Typically the magnitude of crosstalk is almost same as that of the thermal noise.

Impulse Noise :

- All the types of noise discussed so far are reasonably predictable and their magnitude is relatively constant.
- So it is a bit easier for the system engineers to deal with them.
- But impulse noise is completely different from the other types. It is non continuous, by nature and it is made of noise spikes of short duration of high amplitude. These noise pulses occure irregularly and they are extremely unpredictable.
- Impulse noise gets generated due to many reasons such as external electromagnetic disturbances, lightning etc.
- Impulse noise does not affect the quality of analog signal to a great extent but it affect the digital data badly because it introduces errors in the digital data as shown in Fig. 3.11.8.
- It is possible to recover the original data by means of sampling as shown in Fig. 3.11.8.



(G-92) **Fig. 3.11.8 : Effect of impulse noise**

3.11.5 Signal to Noise Ratio :

- The signal to noise ratio (SNR) is defined as :

$$SNR = \frac{\text{Average Signal Power}}{\text{Average Noise Power}}$$

- SNR can be used to find the theoretical bit rate limit of a given communication medium. SNR is the ratio of the desired portion (signal) and the undesired portion (noise) in the transmitted or received waveform. Its value should be as high as possible.
- SNR is a ratio of two powers. So it is often defined in decibels (dB).

$$[SNR]_{dB} = 10 \log_{10} SNR$$

Note :

1. A high SNR indicates that the signal is less degraded due to noise.
2. A low SNR indicates that the signal is heavily degraded due to the noise.

3.12 Data Rate Limits :

- In data communication a large data is required to be transferred from one place to the other.
- It is necessary to transfer it as quickly as possible. In other words the data rate in bits per second over a channel should be as high as possible.
- The data rate is decided by the following factors :
  1. The maximum bandwidth.
  2. The signal level.
  3. The noise presented by the channel.
- Two theorems were developed to calculate the data rate and we can use them on the basis of the type of channel as follows :
  1. A noiseless channel : Nyquist theorem
  2. A noisy channel : Shannon's theorem

3.12.1 Noiseless Channel : Nyquist Bit Rate :

- As we know a transmission channel is a medium over which the electrical signals from a transmitter travel to the receiver. Two important characteristics of a transmission channel are :
  1. Signal to Noise ratio (SNR) and
  2. Channel bandwidth.
- These two characteristics will ultimately decide the maximum capacity of a channel to carry information.
- Nyquist and Shannon worked on finding the maximum channel capacity of a bandlimited channel.
- Nyquist's theorem states that if the bandwidth of a transmission channel is "B" which carries a signal having "L" number of levels, then the maximum data rate "R" on this channel is given by.

$$R = 2 B \log_2 L \quad ...(3.12.1)$$

- As maximum data rate for reliable transmission is defined as channel capacity C, the above expression gets modified as :

$$C = 2 B \log_2 L \quad ...(3.12.2)$$

- This expression indicates that the data rate can be increased by increasing the number of different signal elements (L).

3.12.2 Noisy Channel : Shannon's Channel Capacity :

- A noiseless channel is not possible in the real world, so Shannon introduced a theorem called Shannon's capacity theorem to determine the highest possible data rate on the noisy channel.
- We have seen the Nyquist bandwidth earlier in this chapter.
- Shannon extended Nyquist's work. He included the effect of noise present on the transmission channel.
- According to Shannon's theorem, if (S/N) is the signal to noise ratio then the maximum data rate is given by

$$C = R_{max} = B \log_2 \left[1 + \frac{S}{N}\right] \text{ bits/sec} \quad ...(3.12.3)$$

- Shannon's theorem puts a limit on the maximum number of levels for a given (S/N) ratio and bandwidth. This expression shows that the maximum data rate for a communication channel is dependent on the channel bandwidth B and signal to noise ratio (S/N).
- It is important to note that the Shannon's formula does not indicate the signal level. It says no matter how much is the signal level, it is not possible to achieve a data rate (R) which is greater than the capacity of the channel (C).

Importance of channel bandwidth :

- Bandwidth of the communication channel should be higher than the bandwidth of the signal that is to be transmitted over it.
- This is essential in order to preserve the shape of the signal being transmitted.
- If the channel bandwidth is less than the signal bandwidth then the signal shape will be distorted when it travels over this channel.

3.12.3 Solved Examples :

Ex. 3.12.1 : A channel has a bandwidth of 5 kHz and a signal to noise power ratio 63. Determine the bandwidth needed if the S/N power ratio is reduced to 31. What will be the signal power required if the channel bandwidth is reduced to 3 kHz ?

Soln. :

1. To determine the channel capacity :

It is given that B = 5 kHz and $\frac{S}{N} = 63$. Hence using the Shannon Hartley theorem the channel capacity is given by,

$$C = B \log_2 \left[1 + \frac{S}{N}\right] = 5 \times 10^3 \log_2 [1 + 63]$$

$$\therefore C = 30 \times 10^3 \text{ bits/sec.} \quad ...(1)$$

2. To determine the new bandwidth :

The new value of $\frac{S}{N} = 31$. Assuming the channel capacity "C" to be constant we can write,

$$30 \times 10^3 = B \log_2 [1 + 31]$$

$$\therefore B = \frac{30 \times 10^3}{5} = 6 \text{ kHz} \quad ...(2)$$

3. To determine the new signal power :

Given that the new bandwidth is 3 kHz. We know that noise power $N = N_0 B$.

Let the noise power corresponding to a bandwidth of 6 kHz be $N_1 = 6 N_0$ and the noise power corresponding to the new bandwidth of 3 kHz be $N_2 = 3 N_0$.

$$\therefore \frac{N_1}{N_2} = \frac{6 N_0}{3 N_0} = 2 \quad ...(3)$$

The old signal to noise ratio $= \frac{S_1}{N_1} = 31$

$$\therefore S_1 = 31 N_1 \quad ...(4)$$

The new signal to noise ratio $= \frac{S_2}{N_2}$. We do not know its value, hence let us find it out.

$$30 \times 10^3 = 3 \times 10^3 \log_2 \left(1 + \frac{S_2}{N_2}\right)$$

$$\therefore \frac{S_2}{N_2} = 1023 \quad ...(5)$$

$$\therefore S_2 = 1023 N_2$$

But from Equation (3), $N_2 = \frac{N_1}{2}$, substituting we get,

$$\therefore S_2 = 1023 \frac{N_1}{2} \quad ...(6)$$

Dividing Equation (6) by Equation (4) we get,

$$\therefore \frac{S_2}{S_1} = \frac{1023 N_1}{2 \times 31 N_1} = 16.5$$

$$\therefore S_2 = 16.5 S_1 \quad \text{...Ans.}$$

Thus if the bandwidth is reduced by 50% then the signal power must be increased 16.5 times i.e. 1650% to get the same capacity.

Ex. 3.12.2 : Calculate the maximum bit rate for a channel having bandwidth 3100 Hz and S/N ratio 20 dB.

Soln. :

Given : B = 3100 Hz

$$\frac{S}{N} = 20 \text{ dB.}$$

But 20 dB = 10 log (S/N)

$$\therefore S/N = 100$$

The maximum bit rate is given by,

$$R_{max} = B \log_2 \left[1 + \frac{S}{N}\right]$$

$$= 3100 \log_2 [1 + 100]$$

$$= \frac{3100 \log_{10} 101}{\log_{10} 2}$$

$$= 20{,}640 \text{ bits/sec.} \quad \text{...Ans.}$$

Ex. 3.12.3 : Calculate the maximum bit rate for a channel having bandwidth 3100 Hz and S/N ratio 10 dB.

Soln. :

Given : B = 3100 Hz

$$\left(\frac{S}{N}\right)_{dB} = 10$$

$$\therefore 10 = 10 \log_{10} \left(\frac{S}{N}\right)$$

$$\therefore \frac{S}{N} = 10$$

$$\therefore \text{Maximum bit rate} = R_{max} = B \log_2 \left[1 + \frac{S}{N}\right]$$

$$= 3100 \log_2 (1 + 10)$$

$$= \frac{3100 \log_{10} 11}{\log_{10} 2}$$

$$= 10{,}724 \text{ bits/s} \quad \text{...Ans.}$$

Ex. 3.12.4 : Calculate the maximum bit rate for a channel having bandwidth 1600 Hz if :
(a) S/N ratio is 0 dB (b) S/N ratio is 20 dB.

Soln. :

Given : B = 1600 Hz

(a) R_{max} for S/N = 0 dB :

$$\left(\frac{S}{N}\right)_{dB} = 10 \log_{10} \left(\frac{S}{N}\right)$$

$$\therefore \frac{S}{N} = 1$$

$$\therefore R_{max} = B \log_2 (1 + S/N) = 1600 \log_2 (1 + 1)$$

$$= 1600 \text{ bits/sec} \quad \text{...Ans.}$$

(b) R_{max} for S/N = 20 dB :

$$\left(\frac{S}{N}\right)_{dB} = 10 \log_{10} \left(\frac{S}{N}\right)$$

$$\therefore 20 = 10 \log_{10} (S/N)$$

$$\therefore \frac{S}{N} = 100$$

$$\therefore R_{max} = B \log_2 \left(1 + \frac{S}{N}\right)$$

$$= 1600 \log_2 (101) = \frac{1600 \log_{10} (101)}{\log_{10} (2)}$$

$$R_{max} = 10{,}654 \text{ bits/sec} \quad \text{...Ans.}$$

Using both the limits :

In practice we have to use both the methods to calculate the required bandwidth and signal level. Consider the following example for the same.

Ex. 3.12.5 : The bandwidth of a channel is 2 MHz and its signal to noise ratio is 63. Calculate the appropriate bit rate and signal level.

Soln. :

Step 1 : Calculate the upper limit using Shannon's theorem :

$$C = B \log_2 \left(1 + \frac{S}{N}\right)$$

$$= 2 \times 10^6 \log_2 (1 + 63)$$

$$= 12 \text{ M bits / sec.}$$

This is the upper limit. For ensuring a better performance we select a somewhat lower value say 8 Mbps.

Step 2 : Calculate number of signal levels using Nyquist theorem :

$$C = 2B \log_2 L$$

$$\therefore 8 \times 10^6 = 2 \times 2 \times 10^6 \log_2 L$$

$$\therefore 2 = \log_2 L$$

$$\therefore L = 4 \quad \text{...Ans.}$$

Ex. 3.12.6 : Calculate the maximum bit rate of channel having bandwidth 1200 Hz if :
1. S/N ratio is 0 dB 2. S/N ratio is 20 dB

Soln. :

Given : B = 1200 Hz.

1. R_{max} for S/N = 0 dB

$$\left(\frac{S}{N}\right)_{dB} = 10 \log_{10}\left(\frac{S}{N}\right)$$

$$\frac{S}{N} = 1$$

$$\therefore R_{max} = B \log_2\left(1+\frac{S}{N}\right) = 1200 \log_2 (1+1)$$

$$= 1200 \text{ bits/sec} \quad \text{...Ans.}$$

2. R_{max} for S/N = 20 dB

$$\left(\frac{S}{N}\right)_{dB} = 10 \log_{10}\left(\frac{S}{N}\right)$$

$$20 = 10 \log_{10}\left(\frac{S}{N}\right)$$

$$\therefore \frac{S}{N} = 100$$

$$\therefore R_{max} = B \log_2\left(1+\frac{S}{N}\right)$$

$$= 1200 \log_2 (101) = 1200 \frac{\log_{10}(101)}{\log_{10}(2)}$$

$$= 7990 \text{ bits/sec} \quad \text{...Ans.}$$

Maximum bit rate = 1200 bits/sec for 0 dB
Maximum bit rate = 7990 bits/sec for 20 dB

Ex. 3.12.7 : Find the number of coding or symbol levels if C = 31000 bits/s and B is 3100 Hz.

Soln. : C is the channel capacity while B is the bandwidth.

According to Shannon's theorem,

$$C = B \log_2\left(1+\frac{S}{N}\right)$$

where S/N is the signal to noise ratio.

$$\therefore 31000 = 3100 \log_2\left(1+\frac{S}{N}\right)$$

$$\therefore \log_2\left(1+\frac{S}{N}\right) = 10$$

$$\therefore \frac{S}{N} = 1023 \text{ or } 30\text{dB}$$

$$\left(\frac{S}{N}\right) \text{dB} = 1.8 + 6 \text{ NdB}$$

$$\therefore 30 = 1.8 + 6N$$

where N = Number of bits per word.

$$\therefore N = 4.72 = 5$$

Number of symbol levels $Q = 2^N = 2^5 = 32$ ...Ans.

Ex. 3.12.8 : Calculate the channel capacity for a noiseless channel having bandwidth = 5 kHz and SNR = 0 using appropriate formula.

Soln. :

Given : B = 5 kHz, $\frac{S}{N} = 0$

Find : Channel capacity.

1. To determine channel capacity :

$$C = B \log_2\left[1+\frac{S}{N}\right] = 5 \times 10^3 \log_2 [1+0]$$

$$C = 0 \text{ bits/sec.} \quad \text{...Ans.}$$

Ex. 3.12.9 : An analog signal has a bit rate of 8000 bps and a baud rate of 1000 baud. How many data elements are carried by each signal element ? How many signal elements do we need ?

Soln. :

Given : Bit rate (n) = 8000 bps, Baud rate = 1000 baud.

To find : 1. Data elements carried by each signal element (R)
2. Total signal elements (L)

Step 1 : Calculate R :

Bit rate = Numbers of data elements per signal × Baud rate

$$\therefore \text{Number of data elements per signal (R)} = \frac{\text{Bit rate (n)}}{\text{baud rate}}$$

$$\therefore R = \frac{8000}{1000} = 8 \text{ bits/ baud} \quad \text{...Ans.}$$

Step 2 : Calculate L :

Total signal elements $(L) = 2^R = 2^8 = 256$ ...Ans.

Ex. 3.12.10 : State and explain the Nyquist theorem and Shannon capacity and solve the following example :
Calculate the maximum bit rate for noiseless channel with a bandwidth of 3000 Hz transmitting a signal with two signal levels.

Soln. : For Nyquist theorem and Shannon capacity refer sections 3.12.1 and 3.12.2.

Given : B = 3000 Hz, L = 2

To find : Maximum bit rate

Maximum bit rate (R) $= 2B \log_2 L$

$$= 2 \times 3000 \log_2 (2)$$

$$= 2 \times 3000 \times \left(\frac{\log_{10}(2)}{\log_{10} 2}\right)$$

$$\therefore R = 6000 \text{ Bits/sec.} \quad \text{...Ans.}$$

3.13 Performance :

- The performance of a data communication network can be measured with the help of the following parameters :
  1. Bandwidth 3. Jitter.
  2. Delay 4. Throughput
- We will discuss them one by one in the following sections.

3.14 Bandwidth :

- Bandwidth is a very important characteristics of a network, which can be used for measuring the network performance.
- Bandwidth can have two different values :
  1. BW in hertz and 2. BW in bits per second.

BW in Hz :

It is the range of frequencies present in a composite signal. It can also be defined as a range of frequencies that a channel can pass through without much attenuation.

BW in bits per second :

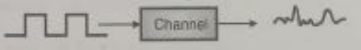
We can also define bandwidth as the number of bits per second (bps) that a channel or network can transmit. For example the BW of Fast Ethernet is 100 Mbps i.e. that network can transmit 100 Mbps.

Relationship :

- There is a clear relationship between the bandwidth in Hz and BW in bps. With increase in BW in Hz, there is an increase in bps bandwidth.
- The relation between them depends on whether baseband transmission is being used or transmission with modulation is being used.

3.14.1 Signal and Channel Bandwidths :

- We can define two different bandwidths :
  1. Signal bandwidth 2. Channel bandwidth
- Signal bandwidth B_s is defined as the range of frequencies contained in the signal.
- Whereas the bandwidth of a channel B_c is the range of frequencies that is passed by a channel.
- If the bandwidth of the input signal is larger than the channel bandwidth, then the output of the channel will not contain all the frequencies of the input signal.
- Fig. 3.14.1 shows a typical digital signal at the input and the output is of different shape than input, if the channel BW is less than signal BW.



(G-698) **Fig. 3.14.1 : Effect of $B_c < B_s$**

- If the signaling rate of input signal is increased, then the channel bandwidth has to be increased so as to pass the signal without any change in shape of the signal.
- The maximum rate at which pulses can be transmitted through the channel is given by,

$$r_{max} = 2W \text{ pulses/sec.}$$

where, $W = 1/2\tau$ and τ = Smallest pulse width of input signal.

- The bandwidth is also important in deciding the channel capacity C of a transmission system.
- **The channel capacity C** of a transmission system is the maximum rate at which bits can be transferred reliably.
- The relation between C and channel bandwidth B_c is given by,

$$C = B_c \log_2\left(1+\frac{S}{N}\right) \text{ bits/sec.} \quad \text{...(3.14.1)}$$

- The channel capacity should be as high as possible and to increase C. We have to increase the channel bandwidth B_c.
- For a telephone channel B_c = 3.4 kHz. If SNR = 10,000 then the channel capacity is given by,

$$C = 3400 \log_2 (1 + 10000) = 45,200 \text{ bits/sec.}$$

- That means we can transmit at a maximum rate of 45.2 kbits per second on the telephone channel.

3.15 Throughput (T) :

- The throughput is a parameter that is used to know the speed of data transmission over a network.
- The throughput of a system is defined as the actual rate at which the information is sent over the channel. It is measured in bits/second or frames/second.
- Throughput is a measure of performance of any network.
- The definitions of bandwidth and throughput appear to be the same but they are not. They are different.
- If B is the bandwidth and T is the throughput of a network then T is always less than and at the most equal to B.
- Thus bandwidth is the theoretical measurement while throughput is the actual measurement of how fast data can be sent.

Ex. 3.15.1 : A network has a bandwidth of 20 Mbps. It can pass 15,000 frames per minute and each frame contains 10,000 bits. Calculate the throughput. Comment on the result.

Soln. :

Given : $B = 20 \times 10^6$ bps, Number of frames = 15000 per min, Number of bits per frame = 10,000.

Throughput T = Number of frames per sec. × Number of bits per frame

$$= \frac{15000}{60} \times 10,000 = 3.5 \text{ Mbps.} \quad \text{... Ans.}$$

Comment :

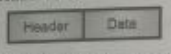
The value of B = 20 Mbps shows the potential of the network and T = 3.5 Mbps shows the actual capability.

3.16 Latency (Delay) :

- The latency or delay is defined as the time required for an entire message to reach its destination from the instant at which the first bit was sent out from the source.
- Latency is the sum of four delay components viz :
  1. Processing delay 2. Queueing delay
  3. Transmission delay 4. Propagation delay.
- The sum of all these delays amounts to the total delay or latency.

1. Processing delay :

- A packet consists of a header and a data field as shown in Fig. 3.16.1. The header contains the destination address.
- The time required to examine the packet header and deciding the direction in which the packet is to be sent is a part of the processing delay.
- The processing delay may also include some other factors such as the time required to check the errors.
- The processing delay is of the order of few microseconds or less.



(G-6981(k)) **Fig. 3.16.1 : Format of a packet**

2. Queueing delay :

- At the queue, the packets experience a queueing delay, when they wait to get transmit on the links.
- The queueing delay depends on the number of packets arrived earlier in the queue.
- With no waiting packets in the queue, the queueing delay will be equal to zero.
- Queueing delays can be of the order of microseconds or milliseconds in practice.

3. Transmission delay :

- The packets are transmitted on the first come first served basis. So a particular packet can get transmitted only after all the earlier packets are transmitted.
- Transmission delay is also called as store and forward delay. It is the time required to push (transmit) all the packet bits into the link.
- Typically, the transmission delay is of the order of microseconds or milliseconds in practice.

$$\text{Transmission delay} = \frac{\text{Message size}}{\text{Bandwidth}}$$

4. Propagation delay :

- The time required for the packet bits to reach from the beginning of the link to the desired router is called as propagation delay.
- The signals travel at a speed which is slightly less than that of light.
- So propagation delay is the ratio of the distance to be travelled by the signal to the speed of propagation.
- Practically the propagation delays are of the order of few milliseconds.

$$\text{Propagation delay} = \frac{\text{Distance}}{\text{Propagation speed}}$$

3.17 Jitter :

Definition :

- The delay introduced by the data communication networks is not constant. It varies packet to packet. The jitter measures the variability in packet delays and it is measured in terms of the difference of the minimum delay and maximum value of delay.
- Jitter is defined as the variation in delay for the packets belonging to the same flow.
- The real time audio and video cannot tolerate jitter on the other hand the jitter does not matter if the packets are carrying any data information contained in a file.
- For the audio and video transmission if the packets take 20 msec to 30 msec (delay) to reach the destination, it does not matter, provided that the delay remains constant.
- The quality of sound or video will be hampered if the delays associated with different packets have different values.

Jitter control :

- When a packet arrives at a router, the router will check to see whether the packet is behind or ahead and by what time.
- This information is stored in the packet and updated at every hop.
- If the packet is ahead of the schedule (early) then the router will hold it for a slightly longer time and if the packet is behind the schedule (late), then the router will try to send it out as quickly as possible.
- This will help in keeping the average delay per packet constant and will avoid time jitter.

3.17.1 Difference between Delay and Jitter :

End to end delay :

- End to end delay is the time required for the signal to travel from transmitter to receiver.
- This delay is due to the time required for buffering, queueing, switch and routing.
- This time delay remains the same for all the types of packets in the same flow.

Jitter :

Jitter is defined as the variation in delay for the packets belonging to the same flow. This is the difference between end to end delay and jitter.

Review Questions

Q. 1 What are the different methods of representing the data ?

Q. 2 Explain the importance of frequency spectrum in analysis of a signal.

Q. 3 Define the term signal bandwidth with the help of frequency spectrum.

Q. 4 Define analog and digital signals and compare them.

Q. 5 Define signal bandwidth and channel bandwidth.

Q. 6 Define : Bit interval, Bit rate, Baud rate.

Q. 7 What is Shannon's channel capacity ?

Q. 8 State and explain the term Nyquist's bandwidth.

Q. 9 What is the significance of channel capacity ?

Q. 10 Explain the effect of channel bandwidth.

Q. 11 What is meant by impulse noise ?

Q. 12 Write a short note on crosstalk and guard time.

Q. 13 What is delay distortion and what is its effect ?

Q. 14 State and explain various transmission impairments.

Q. 15 Define bandwidth in Hz and in bps.

Q. 16 What is throughput ? How is it different from bandwidth ?

Q. 17 Define latency and its four components.

Q. 18 Explain the concept of bandwidth delay product.

□□□

CHAPTER 4

Unit II

Digital and Analog Transmission

Syllabus :

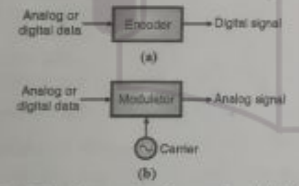
Digital transmission, Digital to digital conversion, Line coding, Line coding schemes, Analog to digital conversion, Pulse Code Modulation (PCM), Transmission modes, Parallel transmission, Serial transmission, Analog transmission, Digital to analog conversion, Aspects of digital to analog conversion, Amplitude shift keying, Frequency shift keying, Phase shift keying, Analog to analog conversion, Amplitude modulation, Frequency modulation, Phase modulation.

4.1 Digital Transmission :

- In computer networks, information is sent from one point to the other. We cannot send this information as it is. Instead it has to be converted into either a digital signal or an analog signal, for transmission.
- Use of digital signals for transmission has many advantages. In this chapter various schemes have been discussed for analog and digital transmission.

4.1.1 Encoding and Modulation :

- It is possible to encode any type of data into any type of signal as shown in Fig. 4.1.1.



(L-25) **Fig. 4.1.1 : Conversion from analog / digital data to analog / digital signal**

- Fig. 4.1.1(a) illustrates the concept of **digital signalling** in which the input data (analog or digital) is encoded into a digital signal.
- Fig. 4.1.1(b) illustrates the concept of **analog signalling** in which the analog / digital source is used for modulating a continuous time carrier signal to produce an analog signal called modulated signal.

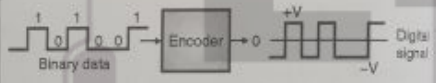
Encoding Types :

There are four different possible transformations as follows :

1. Digital data, digital signal.
2. Analog data, digital signal.
3. Digital data, analog signal.
4. Analog data, analog signal.

4.2 Digital to Digital Conversion :

- In this type of encoding, the digital data which is normally binary in nature is converted into a sequence of discrete, discontinuous voltage pulses (digital signal).

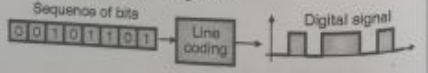


(L-254) **Fig. 4.2.1 : Digital to digital conversion**

- The digital data at the input of the encoder may not be suitable for transmission over a longer distance. Hence it is converted into the digital signal which is more suitable for long distance communication.
- The digital signals at the output of the encoder are known as the line codes.

4.2.1 Definition of Line Coding :

- The line coding is defined as the process of converting binary data, a sequence of bits to a digital signal.
- The digital data such as text, numbers, graphical images, audio and video are stored in computer memory in the form of sequences of bits.
- Line coding converts these sequences into digital signals as shown in Fig. 4.2.2.



(L-255) **Fig. 4.2.2**

4.2.2 Some Important Characteristics of Line Coding :

Some of the important characteristics of line coding are :

1. Signal level and data level
2. Pulse rate and bit rate
3. DC component
4. Self synchronization

4.2.3 Signal Level and Data Level :

- A digital such as binary, octal, hex signal has a limited number of values. However all of them are not used to represent the data. Instead some of them only are used for representing the data.
- The remaining values are used for some other purpose.

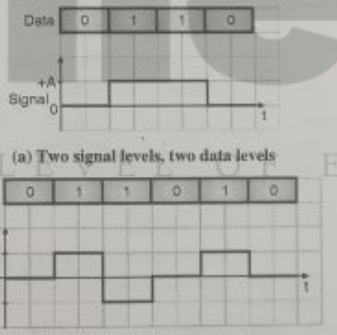
Signal levels :

The number of values allowed in a particular signal is defined as the number of signal levels. Refer Fig. 4.2.3 for better understanding.

Data levels :

The number of values to represent data is called as the number of data levels. The binary data has two values 0 and 1.

- Fig. 4.2.3 gives you a clear idea about signal levels and data levels.
- Fig. 4.2.3 (a) has two signal levels (0 and A) and two data levels (0 and 1) whereas Fig. 4.2.3(b) has three signal levels (0, + A and – A) and two data levels.



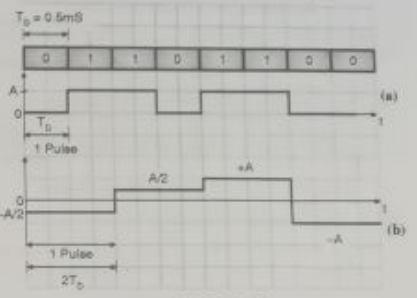
(L-256) **Fig. 4.2.3**

4.2.4 Pulse Rate and Bit Rate :

- **Pulse Rate** is defined as the number of pulses per second and a pulse is defined as the minimum amount of time required to transmit a symbol.
- **Bit Rate** is defined as the number of bits per second. It one pulse corresponds to one bit then the pulse rate is equal to the bit rate. But if a pulse carries more than 1 bit then the pulse rate is lower than bit rate.
- The relation between bit rate and pulse rate is as follows :

$$\text{Bit rate} = \text{Pulse rate} \times \log_2 L \quad \ldots(4.2.1)$$

Ex. 4.2.1 : For the signals shown in Fig. P. 4.2.1(a) and (b) calculate the bit rate and pulse rate.



(L-257) **Fig. P. 4.2.1**

Soln. :

- The data has two levels and the bit duration $T_b = 0.5$ mS.
- Refer Fig. P. 4.2.1(a) which shows that there are two signal levels (0 or A)

$$\therefore \text{ Pulse rate} = \frac{1}{0.5 \times 10^{-3}} = 2000 \text{ pulses / sec.}$$

$$\text{and Bit rate} = 2000 \times \log_2 2 = 2000 \text{ bps}$$

- Refer Fig. P. 4.2.1(b). Now one pulse corresponds to a duration of 2 T_b.

So pulse duration = $2 T_b = 1$ mS.

$$\therefore \text{ Pulse rate} = \frac{1}{1 \times 10^{-3}} = 1000 \text{ pulses / sec.}$$

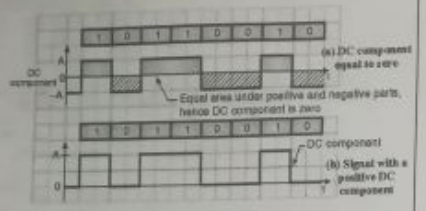
$$\text{And Bit rate} = \text{Pulse rate} \times \log_2 L$$

$$\text{Here } L = 4 \text{ levels}$$

$$\text{Bit rate} = 1000 \times \log_2 4 = 2000 \text{ bps.}$$

4.2.5 DC Component :

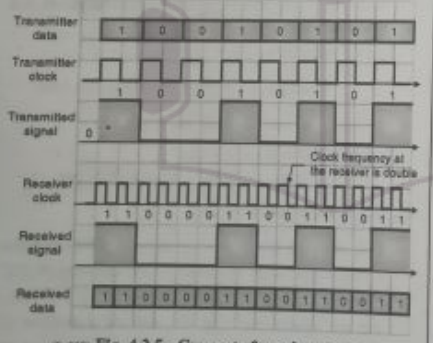
- Over one cycle period of a waveform, if all the positive voltages are cancelled by negative voltages then the DC component of the waveform is zero. (See Fig. 4.2.4(a)).
- But the waveform of Fig. 4.2.4(b) has a positive dc component because the instantaneous voltage can be either zero or positive.
- In line coding, the signal with a non-zero dc component is treated as a distorted one and it can create errors in the received signal.
- The signals with a dc component cannot pass through a transformer. Hence the signals with zero dc component are preferred.



(L-258) **Fig. 4.2.4 : Concept of DC component**

4.2.6 Self Synchronization :

- If the receiver's bit intervals correspond exactly to the sender's bit intervals, only then it is possible to receive (recognize) a signal correctly.
- The clock frequency of the transmitter and receiver should be the same.
- If the clock frequency at the receiver is slower or faster than the bit intervals are not matched and the received signal is different that the transmitted one.
- Fig. 4.2.5 illustrates the effect of change in clock frequency. The receiver clock frequency is twice that of the transmitter frequency. So received data is totally different than the transmitted one.



(L-259) **Fig. 4.2.5 : Concept of synchronization**

- Such thing would not happen if the receiver clock is **synchronized** with the transmitter clock.
- To achieve this, the transmitted digital signal includes the timing information. This will force the self synchronization.
- For achieving synchronization, the transmitted signal should cross the zero frequently. So if the transmitted signal consists of long trains of 0's or 1's, then the synchronization is affected.

4.2.7 Built in Error Detection :

- A line code should have a built in error detecti capability, so that at least some if not all the err which have been introduced during the transit can detected. Some encoding schemes discussed in chapter have the error detection capability.

4.2.8 Immunity to Noise Interference :

- Noise immunity is another desirable property of a li code. The noise interference should not be allowed induce errors in the line codes.

4.2.9 Complexity :

- A line code should be as simple as possible. A li coding scheme which uses four signal levels is mo complex than that using two signal levels.

4.2.10 Bandwidth :

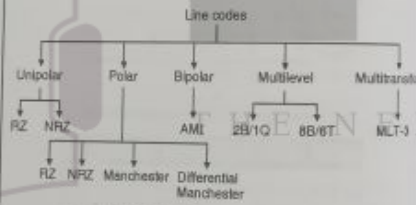
- Most digital signals that we come across have a fin bandwidth. This is not the absolute bandwidth but it the effective bandwidth.
- The bandwidth is proportional to the signal rate (bu rate). The minimum bandwidth is given by,

$$B_{min} = C \times N \times \frac{1}{r}$$

Where C = Case factor, N = Data rate bps.

4.3 Classification of Line Codes :

- Fig. 4.3.1 shows the classification of line codes.



(L-260) **Fig. 4.3.1 : Classification of line codes**

- The line codes are basically divided into thre categories :
  1. Unipolar codes
  2. Polar codes
  3. Bipolar codes

1. Unipolar codes :

- Unipolar codes use only one voltage level othe than zero.
- So the encoded signal will have either + A volt value or 0.
- These codes are very simple and primitive and ar not used now a days.

2. Polar codes :

- Polar coding using two voltage levels other tha zero such as + A/2 and – A/2 volts.
- This will bring the dc level for some codes to zer which is a desired characteristics.

3. Bipolar codes :

- Bipolar coding uses three voltage levels positive, negative and zero which is similar to polar codes.
- But here the zero level is always used for representing the "0" in the data stream at the input.

4.4 Unipolar Line Codes :

4.4.1 Unipolar RZ Format :

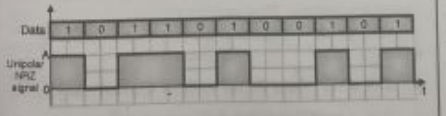
- The return to zero (RZ) unipolar format is as shown in Fig. 4.4.1.
- In this format each "0" is represented by an off pulse (0) and each "1" by an on pulse with amplitude A and a duration of $T_b/2$, followed by a return to zero level.
- Therefore this is called as return to zero (RZ) format. As the voltage level is either + A or zero, this is a unipolar format. (Unipolar means only one polarity).

(L-262) **Fig. 4.4.1 : Unipolar RZ format**

- Due to the unipolar nature, the unipolar RZ format has a nonzero dc value. The dc value does not contain any information.

4.4.2 Unipolar NRZ Format :

- A non-return to zero (NRZ) format is as shown in Fig. 4.4.2.
- In this format a logic "1" is represented by a pulse of full bit duration T_b and amplitude + A while a logic "0" is represented by an off pulse or zero amplitude.
- During the on time, the pulse does not return to zero after half bit period. Therefore the name NRZ format.
- As the pulses have either + A or 0 amplitude it is called as a unipolar format.
- Internal computer waveforms are usually of this type. Due to the unipolar nature, the unipolar NRZ format also will have a nonzero average (dc) value which does not carry any information.
- Due to longer pulse duration, the NRZ pulses carry more "energy" than the RZ pulses. But they need synchronization at the receiver as there is no separation between the adjacent pulses.

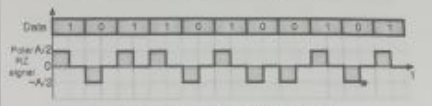


(L-263) **Fig. 4.4.2 : Unipolar NRZ format**

4.5 Polar Line Codes :

4.5.1 Polar RZ Format :

- The disadvantage of the two unipolar formats discussed earlier is that they result in a dc component that does not carry any information and wastes power.

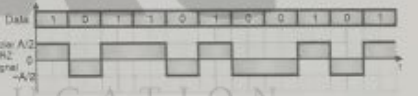


(L-264) **Fig. 4.5.1 : Polar RZ format**

- The polar RZ format is as shown in Fig. 4.5.1. It shows that opposite polarity pulses of amplitude ± "A/2" are used to represent logic "1" and "0".
- Therefore it is called as a "polar" format. As the pulses return to zero after half the bit duration "$T_b/2$" this format is a RZ format.

4.5.2 Polar NRZ Format :

- In the polar NRZ format, as shown in Fig. 4.5.2 a pulse of amplitude "+ A/2" of duration T_b is used to represent a logic "1" and a pulse of amplitude "– A/2" of the same duration represents a logic "0".
- Unlike the unipolar waveform, a polar waveform has no dc component if the 0s and 1s in the input data occur in equal proportion.

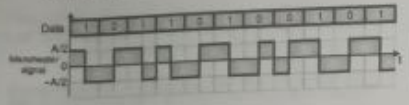


(L-265) **Fig. 4.5.2 : Polar NRZ format**

4.5.3 Split Phase Manchester Format :

- The split phase Manchester format is as shown in Fig. 4.5.3.
- In this format, symbol "1" is represented by transmitting a positive pulse of "+ A/2" amplitude for one half of the symbol duration, followed by a negative pulse of amplitude "– A/2" for remaining half of the symbol duration.
- For symbol "0" these two pulses are, transmitted in reverse order.
- This waveform does not have any dc component.
- The Manchester format has a built in synchronization capability as it crosses zero at regular intervals. But this capability is attained at the expense of a bandwidth requirement of twice that of the NRZ unipolar, polar and bipolar formats.

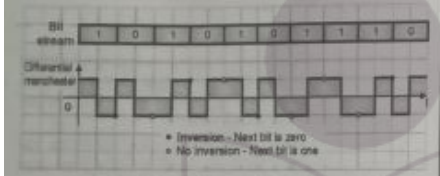
– Local area networks (LAN) such as Ethernet and Cheapernet are increasingly using the Manchester code for signal transmission over the network.



(L-266) **Fig. 4.5.3 : Split phase manchester format**

4.5.4 Differential Manchester Code :

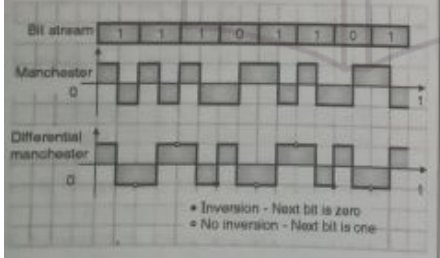
In this code there is always a transition in the middle of a bit interval. The binary zero has an additional transition at the beginning of the bit interval. This is as shown in Fig. 4.5.4 as inversion. There will be no additional transition (inversion) if the next bit is a 1.



(L-266(a)) **Fig. 4.5.4 : Differential Manchester Coding**

Ex. 4.5.1 : Show the Manchester and differential Manchester encoding pattern for the bit stream 11101101.

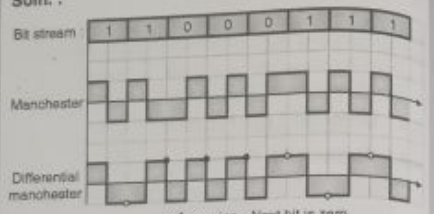
Soln. :



(L-287) **Fig. P. 4.5.1**

Ex. 4.5.2 Show Manchester and differential Manchester encoding pattern for the bit stream 11000111.

Soln. :



(L-288) **Fig. P. 4.5.2**

4.6 Bipolar Line Codes :

4.6.1 Bipolar NRZ Format (AMI) :

– The bipolar NRZ format is as shown in Fig. 4.6.1. Here the successive "1s" are represented by pulses with alternating polarity, and no pulse is transmitted for a logic "0".

– Note that in this representation there are three levels : + A, 0 and – A.

– Therefore this is also known as "pseudoternary or alternative mark inversion (AMI)" format.

(L-268) **Fig. 4.6.1 : Bipolar NRZ format (AMI)**

– An attractive feature of the bipolar format is the absence of a dc component eventhough the input binary data may contain long strings of "0s" and "1s".

– Moreover the bipolar format eliminates ambiguity that may arise because of polarity inversion during the course of transmissions. (This problem is observed in the switched telephone networks).

– This is the reason why the bipolar NRZ format is used in the PCM-TDM T_1 system for digital telephony.

– The absence of dc component allows the use of transformers for coupling.

4.7 Some Other Line Codes :

Some other line codes have been created for some special applications. They are :

1. 2 B1Q 2. Polar Quaternary 3. MLT - 3
4. Biphase M 5. Biphase S

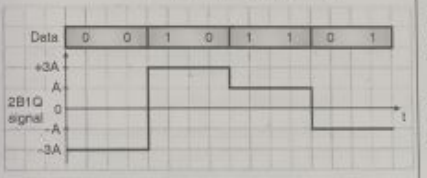
4.7.1 2 B1Q (Two Binary, One Quaternary) :

– The 2 B1Q code uses four voltage levels – 3A, – A, + 3A and + A in order to represent the digital input sequence.

– The input data bits are divided into groups of two bits each and each group is represented by one level as shown in Fig. 4.7.1.

Table 4.7.1

Message	Level assigned
0 0	– 3A Volts
0 1	– A Volts
1 0	+ 3A Volts
1 1	+ A Volts

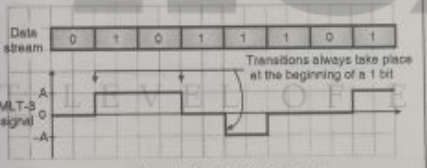


(L-269) **Fig. 4.7.1 : 2 B1Q signal**

4.7.2 MLT-3 (Multi Line Transmission, Three Level) :

– The MLT-3 is similar to NRZ-I but it uses three levels of signals (+ A, 0, – A).

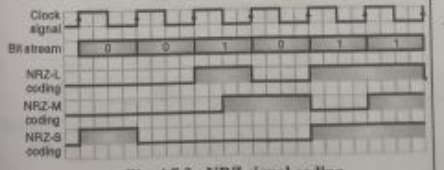
– The signal will change its value from one level to the next at the beginning of a 1 bit and there is no change in the signal value at the beginning of a 0 bit, as shown in Fig. 4.7.2.



(L-271) **Fig. 4.7.2 : MLT-3 signal**

4.7.3 NRZ-L (Non Return to Zero Level) :

– In NRZ-L coding a bit 0 or 1 is represented by a voltage level which remains constant during the bit duration. A binary "1" is represented by a high level and "0" is represented by a low level.



(L-272) **Fig. 4.7.3 : NRZ signal coding**

4.7.4 NRZ-M (Non Return to Zero Mark) :

As shown in Fig. 4.7.3, in NRZ-M the waveform changes its level when the binary digit is "1". The waveform does not change its level when the binary digit is "0".

4.7.5 NRZ-S (Non Return to Zero Space) :

As shown in Fig. 4.7.3, in NRZ-S the waveform changes its level when the binary digit is "0". The waveform does not change its level when the binary digit is "1".

The other line codes are RZ (return to zero), AMI, manchester etc.

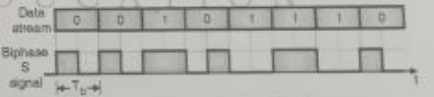
4.7.6 Biphase-M Code :

In this code, there is always a transition at the beginning of every bit interval. If the bit is a binary 1 then the coded signal returns to zero in the middle of the bit duration as shown in Fig. 4.7.4.

(L-273) **Fig. 4.7.4 : Biphase-M coding**

4.7.7 Biphase-S Code :

– In this code also there is always a transition at the beginning of every bit interval. If the data bit is a binary 0 then the coded signal returns to zero in the middle of the bit duration as shown in Fig. 4.7.5. No such return transition takes place if the bit is 1.



(L-274) **Fig. 4.7.5 : Biphase-S coding**

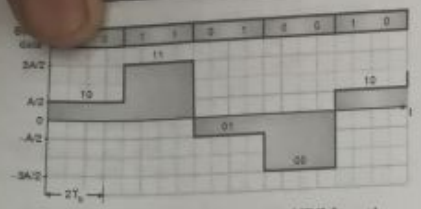
4.7.8 Polar Quaternary NRZ Format :

– Fig. 4.7.6 shows quaternary NRZ format derived by grouping the message bits in the blocks of two and using four amplitude levels to represent the four possible combinations 00, 01, 10 and 11.

– To these four combinations, four amplitude levels are assigned, as shown in Table 4.7.2.

Table 4.7.2

Message combination	$x(t) = a_k$
0 0	– 3 A/2
0 1	– A/2
1 0	A/2
1 1	3 A/2



(L-275) **Fig. 4.7.6 : Polar quaternary NRZ format**

- In the waveforms of Fig. 4.7.6, the first combination of two bits is "10" hence it is represented by a level "A/2".
- The second combination is "11" hence it is represented by a level of "3A/2". Thus here for a message of two bits only one pulse of duration $D = 2T_b$ is transmitted.

$$\therefore D = 2T_b \quad ...(4.7.1)$$

The signaling rate : $r = \frac{1}{2T_b}$ messages/sec ...(4.7.2)

- If there are "M" levels obtained from the combination of "k" bits (here M = 4 and k = 2) then :

$$M = 2^k \quad ...(4.7.3)$$

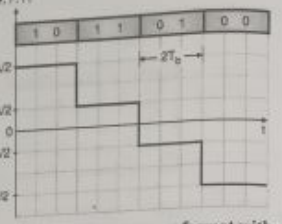
4.7.9 Gray Code :

- There is another scheme used for coding the quaternary format. It is called as the gray code.
- The gray coding scheme is illustrated in the following table. The adjacent bits are arranged in such a way that they differ by only one bit.

Table 4.7.3 : Gray encoding

Message combination	x (t)
0 0	– 3 A/2
0 1	– A /2
1 1	A/2
1 0	3 A/2

- The polar quaternary format with gray coding is shown in Fig. 4.7.7.



(L-276) **Fig. 4.7.7 : Polar quaternary format with gray coding**

4.7.10 Differential Encoding :

- The differential encoding format is shown in Fig. 4.7.8
- The advantage of differential encoding is that it is immune from the polarity inversion-ambiguity problem.
- Differential encoding starts with an arbitrary initial bit
- Corresponding to every 0 at the input, the differential encoding format make a transition from + A to – A or – A to + A whereas no transition takes place corresponding to a logic 1 at the input.

(L-277) **Fig. 4.7.8 : Differential encoding format**

- The original binary information can be recovered by sampling the received wave and comparing the polarities of the adjacent samples.

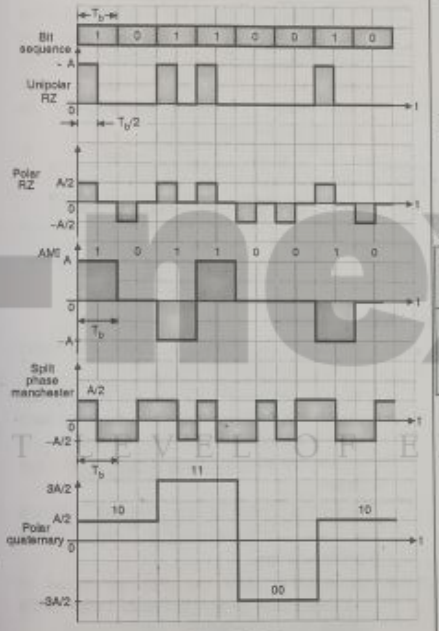
4.7.11 Comparison of Line Codes :

Sr. No.	Parameter	Polar RZ	Polar NRZ	AMI	Manchester	Polar Quaternary NRZ
1.	Transmission of DC component	Yes	Yes	No	No	Possible
2.	Signaling rate	$1/T_b$	$1/T_b$	$1/T_b$	$1/T_b$	$1/2T_b$
3.	Noise immunity	Low	Low	High	High	High
4.	Synchronizing capability	Poor	Poor	Very good	Very good	Poor
5.	Bandwidth required	$1/T_b$	$1/2T_b$	$1/2T_b$	$1/T_b$	$1/2T_b$
6.	Crosstalk	High	High	Low	Low	Low

Ex. 4.7.1 : Consider that the bit sequence given below is to be transmitted. Bit sequence = 10110010. Draw the resulting waveform if the sequence is transmitted using :
1. Unipolar RZ 2. Polar RZ
3. AMI
4. Split Phase Manchester
5. M-ary where M = 4. (Polar quaternary)

Soln. :

The required waveforms are as shown in Fig. P. 4.7.1.

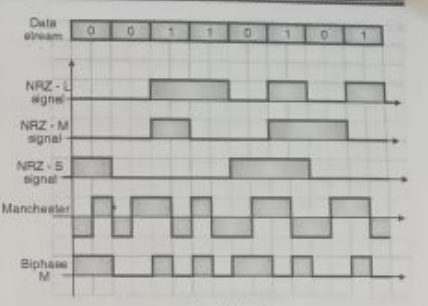


(L-279) **Fig. P. 4.7.1**

Ex. 4.7.2 : Sketch the signal waveforms when 0011 0101 is transmitted in the following signal codes :
1. NRZ-L
2. NRZ-M
3. NRZ-S
4. Manchester code 5. Biphase-M.

Soln. :

The required waveforms are as shown in Fig. P. 4.7.2.



(L-280) **Fig. P. 4.7.2**

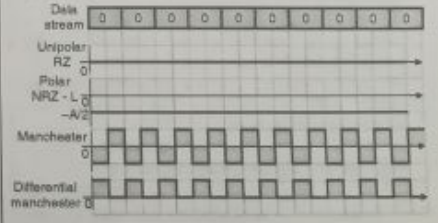
Ex. 4.7.3 : How many amplitude levels are there for each of the following methods ?
1. NRZ-L 2. NRZ-S 3. Manchester
4. RZ 5. Differential Manchester
6. NRZ-M.

Soln. :

Name of the format	NRZ-L	NRZ-S	Manchester	RZ	NRZ-M
Number of amplitude levels	2	2	2	2	2

Ex. 4.7.4 : Assume a data stream is made of ten 0's. Encode this stream using the following schemes. How many changes can you find in each scheme ?
(a) Unipolar RZ (b) Polar NRZ-L
(c) RZ (d) Manchester
(e) Differential Manchester

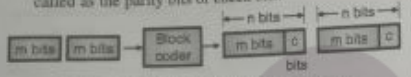
Soln. : The required waveforms are as shown in Fig. P. 4.7.4.



(L-281) **Fig. P. 4.7.4**

4.7.12 Block Coding :

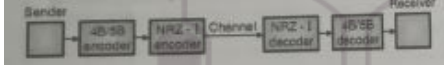
- In order to ensure better error performance and synchronization, it is necessary to introduce redundancy.
- Block coding is one of the ways to introduce redundancy and enhance the performance of line coding.
- In general block coding is a technique that changes a block of m bits into a block of n bits with n > m. Therefore block coding is referred to as mB/nB coding technique.
- The additional bits added to the original m bits are called as the parity bits or check bits.



m = message bits, c = parity bits or check bits

(L-282) **Fig. 4.7.9 : Concept of block coding**

- The examples of block codes are 4B/5B encoding in which a 4 bit code is converted into a 5 bit code or an 8B/10B code in which the 8 bit codes are converted into 10 bit codes.
- Fig. 4.7.10 shows the block diagram of a system using 4B/5B encoding.



(L-283) **Fig. 4.7.10 : 4B/5B system**

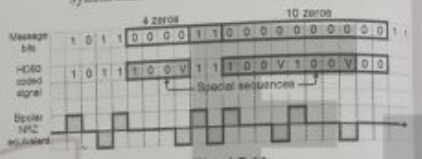
4.7.13 Scrambling :

- The Biphase schemes that are suitable for the dedicated links in a LAN prove to be unsuitable for long distance communication as they require large bandwidths.
- The combination of block coding and NRZ line coding also proves to be unsuitable due to the presence of DC component.
- The AMI scheme has smaller BW and no DC component but a long sequence of 0s can create problems for its synchronization.
- For long distance communication, a technique is needed to be found which has small BW, no DC component and no synchronization problems.
- One of the solutions is to use **scrambling.** We can modify AMI for scrambling. Two commonly used scrambling techniques are :

1. B8ZS 2. HDB3 encoding

1. High density bipolar (HDB) signalling :

- In case of the bipolar NRZ or AMI signal, the transmitted signal is equal to zero when a binary "0" is to be transmitted. This is true even for the unipolar RZ and unipolar NRZ signals.
- The absence of transmitted signal can cause problems in synchronization at the receiver, if long sequence of binary "0"s are being transmitted.
- This problem can be solved by adding (transmitting) pulses when long strings of 0's exceeding a number n are being transmitted. This type of coding is called as High Density Bipolar coding. It is denoted by HDBN. Here N = 1, 2, 3 ..... . The most widely used HDB format is with N = 3 i.e. HDB3.
- In the string of message bits when (N + 1) or more number of zeros occur, they are replaced by special binary sequences of (N + 1) length. As shown in Fig. 4.7.11, these sequences contain some binary 1's which are necessary for synchronization at the receiver.

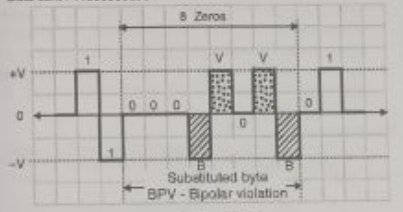


(L-284) **Fig. 4.7.11**

2. B8ZS line code :

- The long form of B8ZS is bipolar with 8-zero substitution.
- We have discussed about the line codes in this chapter. We know that in order to have synchronization between the transmitter and receiver, the line code needs to cross the zero line frequently.
- As per U.S. T_1 standard, not more than 15 0's can be sent in succession to ensure proper synchronization.
- In order to solve the problems related to synchronization a new line code called B8ZS (Binary 8-zeros suppression) was developed.
- Whenever eight successive 0's are detected, the implementation of this line code will automatically insert a special 8 bit sequence containing a bipolar violation.
- This can be easily detected and corrected by the CSU/DSU (channel service unit/digital service unit).
- Refer Fig. 4.7.12 to get a clear idea about the B8ZS line code.
- The violations (BPV in Fig. 4.7.12), will distinguish a byte substituted for all 0's from a normal byte which contains 1's.

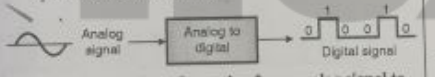
- The B8ZS does not allow more than 8 - consecutive 0's and the bipolar violation pattern uniquely identifies the eight 0's.
- Note that the voltage levels in the "violating byte", has a zero average (dc) value.



(L-285) **Fig. 4.7.12**

4.8 Analog to Digital Conversion :

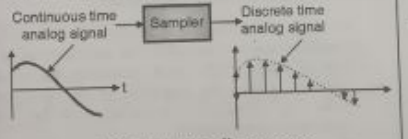
- The process of converting the analog data to digital signal is known as digitisation.
- This process is essential in all the digital communication systems such as pulse code modulation (PCM) or delta modulation (D.M).
- In order to carry out this transformation, one has to follow a sequence of operations such as sampling, quantization and encoding.



(L-218) **Fig. 4.8.1 : Transformation from analog signal to digital signal**

4.9 Sampling Process :

- In the pulse modulation and digital modulation systems, the signal to be transmitted must be in the discrete time form.
- If the message signal is coming from a digital source (e.g. a digital computer) then it is in the proper form for a digital communication system.



(L-156) **Fig. 4.9.1 : Sampling process**

- But this is not always the case. The message signal can be analog in nature (e.g. speech or video signal).
- In such a case it has to be first converted into a discrete time signal. We use the "sampling process" to do this.
- Thus using the sampling process we convert a continuous time signal into a discrete time signal.
- For the sampling process to be of practical utility it is necessary to choose the sampling rate properly. The sampling process should satisfy the following requirements :
  1. Sampled signal should represent the original signal faithfully.
  2. We should be able to reconstruct the original signal from its sampled version.
- Fig. 4.9.1 summarizes the sampling process.
- Thus sampling is the process of converting a continuous analog signal to a discrete analog signal and the sampled signal is the discrete time representation of the original analog signal.

4.9.1 Sampling Theorem for Low Pass Signals :

- In order to represent the original message signal "faithfully" (without loss of information), it is necessary to take as many samples of the original signal as possible.
- Higher the number of samples, closer is the representation.
- The number of samples depends on the "sampling rate" and the maximum frequency of the signal to be sampled.
- Sampling theorem was introduced to the communication theory in 1949 by Shannon. Therefore this theorem is also called as **"Shannon's sampling theorem".**
- The statement of sampling theorem in time domain, for the bandlimited signals of finite energy is as follows :

Statement :

1. If a finite energy signal x (t) contains no frequencies higher than "W" Hz (i.e. it is a band limited signal) then it is completely determined by specifying its values at the instants of time which are spaced (1/2W) seconds apart.
2. If a finite energy signal x (t) contains no frequency components higher than "W" Hz then it may be completely recovered from its samples which are spaced (1/2W) seconds apart.

- Combined statement of sampling theorem : A continuous time signal x (t) can be completely represented in its sampled form and recovered back from the sampled form if the sampling frequency $f_s \geq 2W$ where "W" is the maximum frequency of the continuous time signal x (t).

Nyquist Rate :

The minimum sampling rate of 2W samples per second for a signal x(t) having maximum frequency of W Hz is called as Nyquist rate.

$$\therefore \text{Nyquist rate} = f_{s(min)} = 2W \text{ Hz} \qquad ...(4.9.1)$$

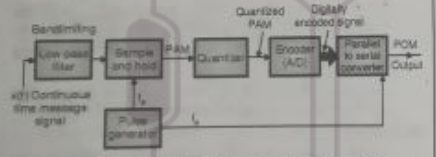
$f_s = 2f_m$

4.10 Pulse Code Modulation (PCM) :

- PCM is a type of pulse modulation like PAM, PWM or PPM but there is an important difference between them. PAM, PWM or PPM are "analog" pulse modulation systems whereas PCM is a "digital" pulse modulation system.
- That means the PCM output is in the coded digital form. It is in the form of digital pulses of constant amplitude, width and position.
- The information is transmitted in the form of "code words". A PCM system consists of a PCM encoder (transmitter) and a PCM decoder (receiver).
- The essential operations in the PCM transmitter are sampling, quantizing and encoding.
- All these operations are usually performed in the same circuit called as analog-to-digital (A to D) converter.
- It should be understood that the PCM is not modulation in the conventional sense.
- Because in modulation, one of the characteristics of the carrier is varied in proportion with the amplitude of the modulating signal. Nothing of that sort happen in PCM.

4.10.1 PCM Transmitter (Encoder) :

- Block diagram of the PCM transmitter is as shown in Fig. 4.10.1.



(L-221) **Fig. 4.10.1 : PCM transmitter (Encoder)**

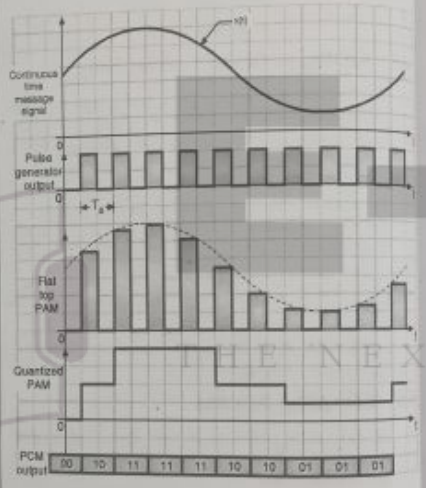
Operation of PCM transmitter :

Operation of the PCM transmitter is as follows :

- The analog signal x (t) is passed through a bandlimiting low pass filter, which has a cut-off frequency $f_c = W$ Hz. This will ensure that x (t) will not have any frequency component higher than "W". This will eliminate the possibility of aliasing.
- The band limited analog signal is then applied to a sample and hold circuit where it is sampled at adequately high sampling rate. Output of sample and hold block is a flat topped PAM signal.
- These samples are then subjected to the operation called "Quantization" in the "Quantizer". Quantization process is the process of approximation as will be explained later on. The quantization is used to reduce the effect of noise. The combined effect of sampling and quantization produces the quantized PAM at the quantizer output.
- The quantized PAM pulses are applied to an encoder which is basically an A to D converter. Each quantized level is converted into an N bit digital word by the A to D converter. The value of N can be 8, 16, 32, 64 etc.
- The encoder output is converted into a stream of pulses by the parallel to serial converter block. Thus at the PCM transmitter output we get a train of digital pulses.
- A pulse generator produces a train of rectangular pulses with each pulse of duration "τ" seconds. The frequency of this signal is "f_s" Hz. This signal acts as a sampling signal for the sample and hold block. The same signal acts as "clock" signal for the parallel to serial converter. The frequency "f_s" is adjusted to satisfy the Nyquist criteria.

Waveforms :

The waveforms at various points in the PCM transmitter are as shown in Fig. 4.10.2.

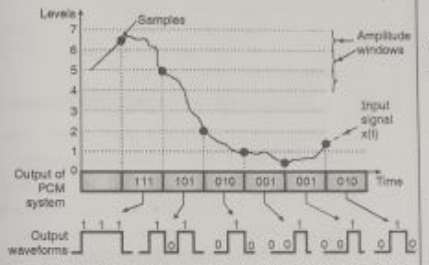


(L-222) **Fig. 4.10.2 : Waveforms at different points in PCM transmitter**

4.10.2 Shape of the PCM Signal :

- Fig. 4.10.3 shows input to and output of a PCM system. It is important to understand that the output is in the form of binary codes. Each transmitted binary code represents a particular amplitude of the input signal. Hence the "information" is contained in the "code" which is being transmitted.
- The range of input signal magnitudes is divided into 8 equal levels. Each level is denoted by a three bit digital word between 000 and 111.
- Input signal x (t) is sampled. If the sample is in the 5th window of amplitude then a digital word 101 is transmitted. If the sample is in the 2nd - window then the transmitted word is 010 and so on.

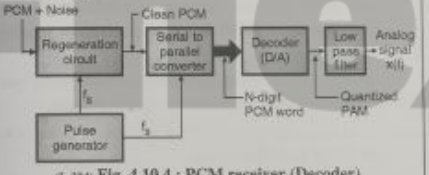
- In this example we have converted the amplitudes into 3 bit codes, but in practice the number of bits per word can be as high as 8, 9 or 10.



(L-223) **Fig. 4.10.3 : Input and output waveforms of a PCM system**

4.10.3 PCM Receiver (Decoder) :

- Fig. 4.10.4 shows the block diagram of a PCM receiver.



(L-224) **Fig. 4.10.4 : PCM receiver (Decoder)**

Operation of PCM receiver :

- A PCM signal contaminated with noise is available at the receiver input.
- The regeneration circuit at the receiver will separate the PCM pulses from noise and will reconstruct the original PCM signal.
- The pulse generator has to operate in synchronization with that at the transmitter. Thus at the regeneration circuit output we get a "clean" PCM signal.
- The reconstruction of PCM signal is possible due to the digital nature of PCM signal. The reconstructed PCM signal is then passed through a serial to parallel converter.
- Output of this block is then applied to a decoder.
- The decoder is a D to A converter which performs exactly the opposite operation of the encoder.
- The decoder output is the sequence of a quantized multilevel pulses. The quantized PAM signal is thus obtained, at the output of the decoder.
- This quantized PAM signal is passed through a low pass filter to recover the analog signal, x (t).
- The low pass filter is called as the reconstruction filter and its cut off frequency is equal to the message bandwidth W.

4.10.4 Quantization Process :

- Quantization is a process of approximation or rounding off. The sampled signal in PCM transmitted is applied to the quantizer block.
- Quantizer converts the sampled signal into an approximate quantized signal which consists of only a finite number of predecided voltage levels.
- Each sampled value at the input of the quantizer is approximated or rounded off to the nearest standard predecided voltage level.
- These standard levels are known as the "quantization levels". Refer to Fig. 4.10.5 to understand the process of quantization.
- The quantization process takes place as follows :
- The input signal x (t) is assumed to have a peak to peak swing of V_L to V_H volts. This entire voltage range has been divided into "Q" equal intervals each of size "s".
- "s" is called as the step size and its value is given as,

$$s = \frac{V_H - V_L}{Q} \qquad \ldots(4.10.1)$$

In Fig. 4.10.5, the value of Q = 8

- At the center of these ranges, the quantization levels q_0, q_1,...q_7 are placed. Thus the number of quantization levels is Q = 8. The quantization levels are also called as decision thresholds.
- x_q (t) represents the quantized version of x (t). We obtain x_q (t) at the output of the quantizer.
- When x (t) is in the range Δ_0, then corresponding to any value of x (t), the quantizer output will be equal to "q_0".
- Similarly for all the values of x (t) in the range Δ_1, the quantizer output is constant equal to "q_1".
- Thus in each range from Δ_0 to Δ_7, the signal x (t) is rounded off to the nearest quantization level and the quantized signal is produced.
- The quantized signal x_q (t) is thus an approximation of x (t). The difference between them is called as **quantization error or quantization noise.**
- This error should be as small as possible.
- To minimize the quantization error we need to reduce the step size "s" by increasing the number of quantization levels Q.

(L-229) **Fig. 4.10.5 : Process of quantization**

Why is quantization required ?

- If we do not use the quantizer block in the PCM transmitter, then we will have to convert each and every sampled value into a unique digital word.
- This will need a large number of bits per word (N). This will increase the bit rate and hence the bandwidth requirement of the channel.
- To avoid this, if we use a quantizer with only 256 quantization levels then all the sampled values will be finally approximated into only 256 distinct voltage levels.
- So we need only 8 bits per word to represent each quantized sampled value.
- Thus the number of bits per word can be reduced. This will eventually reduce the bit rate and bandwidth requirement.

4.10.5 Quantization Error or Quantization Noise ∈ :

- The difference between the instantaneous values of the quantized signal and input signal is called as quantization error or quantization noise.

$$\in = x_q(t) - x(t) \qquad ...(4.10.2)$$

- The quantization error is shown by shaded portions of the waveform in Fig. 4.10.5.
- The maximum value of quantization error is $\pm s / 2$ where s is step size.
- Therefore to reduce the quantization error we have to reduce the step size by increasing the number of quantization levels i.e. Q.
- The mean square value of the quantization is given by,

$$\text{Mean square value of quantization error} = \frac{s^2}{12} \qquad ...(4.10.3)$$

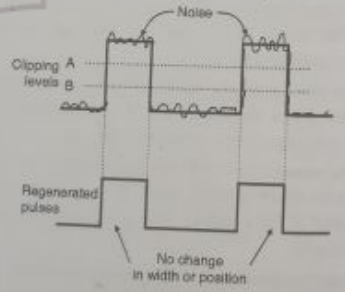
- The relation between the number of quantization levels Q and the number of bits per word (N) in the transmitted signal can be found as follows :
- Because each quantized level is to be converted into a unique N bit digital word, assuming a binary coded output signal,
- The number of quantization levels Q = Number of combinations of bits/word.

$$\therefore \quad Q = 2^N \qquad ...(4.10.4)$$

- Thus if N = 4 i.e. 4 bits per word then the number of quantization levels will be 2^4 i.e.16.

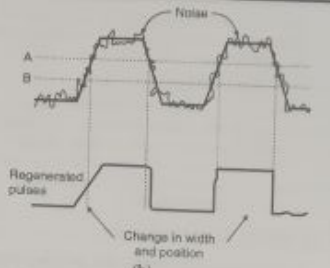
4.10.6 Effect of Noise on the PCM System :

- Look at the two Figs. 4.10.6(a) and 4.10.6(b) which illustrate the effect of noise on the transmitted pulses.
- Consider Fig. 4.10.6(a) first. Due to the noise superimposed on the pulses, only the PAM system will be affected.
- However the PWM, PPM and PCM systems will remain unaffected. The regeneration of the pulses is achieved by using a clipper circuit with reference levels A and B.
- Now consider Fig. 4.10.6(b). Here the sides of the transmitted pulse are not perfectly vertical. In practice the transmitted pulses usually have slightly sloping sides (edges).
- As the noise is superimposed on them, the width and the position of the regenerated pulses is changed.
- Now this is going to distort the information contents in the PWM and PPM signals.
- But PCM is still unaffected as it does not contain any information in the width or the position of the pulses.
- Thus PCM has much better noise immunity as compared to PAM, PWM and PPM systems.



(a)

Fig. 4.10.6 (Contd...)



(b)

(L-229) **Fig. 4.10.6 : Effect of noise on PCM**

4.11 Advantages, Disadvantages, Applications and Modifications in PCM :

- The PCM is considered to be the best modulation scheme to transmit the voice and video signals.
- All the advantages of PCM are due to the fact that it uses coded pulses for the transmission of information.

4.11.1 Advantages of PCM :

1. Very high noise immunity (Noise does not affect the information content).
2. Due to digital nature of the signal, repeaters can be placed between the transmitter and the receivers. The repeaters actually regenerate the received PCM signal. This is not possible in analog systems. Repeaters further reduce the effect of noise.
3. It is possible to store the PCM signal due to its digital nature.
4. It is possible to use various coding techniques so that only the desired receiver can decode the received signal. This makes the communication secure.
5. The increased channel bandwidth requirement for PCM is balanced by the improved SNR. This is the due to the fact that PCM obeys an exponential law.
6. There is a **uniform format** used for the transmission of different types of base band signals. Hence it is easy to integrate all these signals together and send them on the common network.
7. It is easy to drop or reinsert the message sources in a PCM-TDM system.

4.11.2 Disadvantages of PCM :

1. The encoding, decoding and quantizing circuitry of PCM is complex.
2. PCM requires a large bandwidth as compared to the other systems.

4.11.3 Applications of PCM :

Some of the applications of PCM are as follows :

1. In telephony (with the advent of fibre optic cables).
2. In the space communication, space craft transmits signals to earth. Here the transmitted power is very low (10 to 15W) and the distances are huge (a few million km). Still due to the high noise immunity, only PCM systems can be used in such applications.

4.11.4 Modifications in PCM :

- Even though PCM is complex, it is possible to implement it using the VLSI technology.
- Due to the improvements in VLSI technology, the use of PCM for digital transmission of analog signals is going to increase.
- But if the simplicity is more important than the performance, then one should use the **Delta Modulation** in place of PCM.
- The requirement of large channel bandwidth for PCM is not a real problem now, due to the availability of wideband communication channels.
- As the problem of limitation on bandwidth has been solved, it has become possible to use the communication satellites and optical fiber communication.
- It is possible to remove the **redundancy** in PCM by using the **data compression techniques.** This will reduce the bit rate of transmitted data without any significant loss of quality in the contents.
- This will increase the complexity of PCM further.

Why is PCM not used for broadcasting ?

- In radio broadcasting a relatively large signal to noise ratio (typically of the order of 60 dB) is required.
- To get this level of $(S/N)_0$ the PCM with $b > 8$ is required, where $b = B_T / W$ i.e. ratio of transmission bandwidth to baseband bandwidth.
- However we can obtain the same performance with an FM system with $b = 6$ and with much simpler transmitter and receiver circuits.
- So higher bandwidth requirement and complicated circuitry are the disadvantages of PCM which does not make it suitable for the radio, TV broadcasting applications.

4.12 Data Transmission Modes :

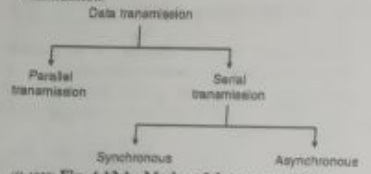
- Data transmission means the movement of data which is in the form of bits between two or more digital devices. The data transmission takes place over some physical medium from one computer to the other.
- There are two ways of transmitting the digital data. They are :

  1. Parallel transmission 2. Serial transmission

4.12.1 Transmission Mode :

- Various modes of data transmission are shown in Fig. 4.12.1.

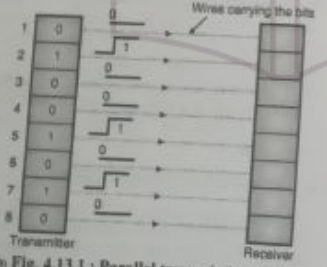
- As seen from Fig. 4.12.1, serial transmission and parallel transmission are the two basic types of transmission. The serial transmission is the most preferred mode of data transmission.
- The serial transmission is further classified into two types namely synchronous and asynchronous transmission.



(G-1039) **Fig. 4.12.1 : Modes of data transmission**

4.13 Parallel Transmission :

- In parallel transmission of data, all the bits of a byte are transmitted simultaneously on separate wires as shown in Fig. 4.13.1.
- This type of transmission requires multiple wires for interconnecting the two devices.
- Parallel transmission is possible practically only if the two devices are close to each other due to the length and the number of wires required.
- For example parallel transmission takes place between a computer and its primer.
- Fig. 4.13.1 shows the parallel transmission of an 8-bit digital data.
- This will require eight wires for connection between a transmitter and a receiver.
- With increase in the number of receivers, the number of wires will increase to an unmanageable number.



(G-1040) **Fig. 4.13.1 : Parallel transmission of data**

4.13.1 Advantages of Parallel Transmission :

The advantage of parallel transmission is that all the data bits will be transmitted simultaneously. Therefore the time required for the transmission of an N-bit word is only one clock cycle.

The serial transmission will require N number of clock cycles for the transmission of same word.

3. Due to this the clock frequency can be kept low without affecting the speed of operation. For serial transmission the clock frequency has to be high.

4.13.2 Disadvantages :

To transmit an N-bit word, we need N number of wires. With increase in the number of users, the number of wires increase and it becomes impossible to handle them. The serial transmission uses only one wire, for connecting the transmitter to the receiver. Hence practically the serial transmission is always preferred.

4.14 Serial Transmission :

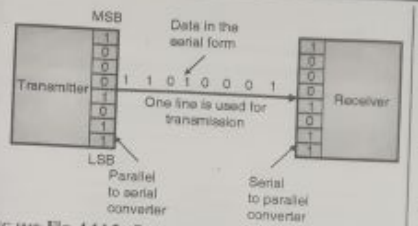
- In serial transmission, the bits of a byte are serially transmitted one by one as shown in Fig. 4.14.1.
- The byte to be transmitted is first stored in a shift register. Then these bits are shifted from MSB to LSB bit by bit in synchronization with the clock. Bits are shifted right [see Fig. 4.14.1] by one position per clock cycle.
- The bit which falls out of the shift register is transmitted. Hence LSB is transmitted first and MSB is the last bit getting transmitted.
- For serial transmission only one wire is needed between the transmitter and the receiver. Hence serial transmission is preferred for long distance data communication. This is the advantage of serial transmission over parallel transmission.

(G-1041) **Fig. 4.14.1 : Serial transmission**

- The serial transmission has a serious drawback. As only one bit is transmitted per clock cycle, it requires a time corresponding to 8-clock cycles to transmit one byte. (The parallel transmission needs only one clock cycle to transmit a byte). The time can be reduced by increasing the clock frequency.

4.14.1 Practical Serial Transmission System :

- Fig. 4.14.2 shows the practical serial transmission system. The transmitter and receiver both are computers.
- Since the communication within a computer is parallel, it is necessary to convert the parallel data into a serial one at the transmitter.
- At the receiver, the serial to parallel conversion is required to be performed as shown in Fig. 4.14.2.



(G-1042) **Fig. 4.14.2 : Practical serial transmission system**

Advantages of serial transmission :

1. Only one wire is required to be used.
2. Reduction in cost due to less number of conductors.

Disadvantages :

1. The speed of data transfer is low as only one bit is sent at a time.
2. To increase the speed of data transfer, it is necessary to increase the clock frequency.

Application :

It is used for computer to computer communication, specially long distance communication.

4.14.2 Comparison of Serial and Parallel Transmission :

Sr. No.	Parameter	Parallel transmission	Serial transmission
1.	Number of wires required to transmit N bits.	N wires	1 wire
2.	Number of bits transmitted simultaneously	N bits	1 bit
3.	Speed of data transfer	Fast	Slow
4.	Cost	Higher due to more number of conductors	Low, since only one wire is used.
5.	Application	Short distance communication such as computer to printer communication	Long distance computer to computer communication.

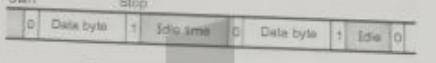
4.14.3 Types of Serial Transmission :

There are three types of serial transmission. They are :

1. Synchronous data transmission.
2. Asynchronous data transmission
3. Isosynchronous transmission

4.15 Asynchronous Transmission :

- In asynchronous transmission, the transmitter can begin the transmission of data bytes at any instant of time.
- Only one byte is sent at a time. After sending one byte the next byte can be sent after an arbitrary time delay as shown in Fig. 4.15.1.
- The transmitter and receiver can operate at different clock frequencies. There is no synchronization between them on this account.
- As the data transmission can commence at any instant, it becomes difficult for the receiver to understand the instant at which the byte has been transmitted.
- To help the receiver to receive the data bytes "start" and "stop" bits are used alongwith each data byte as shown in Fig. 4.15.1. The start bit is always "0" and stop bit is always "1".



(G-1043) **Fig. 4.15.1 : Asynchronous transmission**

- The idle time in between the adjacent data bytes is not constant. The idle time is also called as the gaps between the data bytes.
- In the asynchronous transmission the timing of the signal is not important, instead information is received and translated by the patterns which are agreed upon by the sender and receiver.
- As long as these patterns are being followed, the receiver can retrieve the information without any problem.

Why is it called asynchronous ?

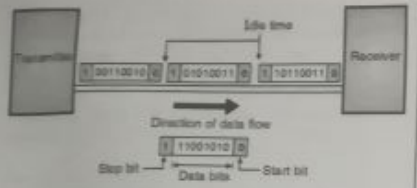
- This mechanism is called as asynchronous because at the byte level the sender and receiver do not have to be synchronized.
- However within each byte, the receiver should still be synchronized with the incoming bit stream.
- This means that some synchronization is required only for the duration of single byte.

Response to the start and stop bits :

- When the receiver detects a start bit, it will set a timer and begins counting bits as they come in.
- After "n" bits, the receiver searches for the stop bit.
- As soon as it detects the stop bit, it will wait until it detects the next start bit.
- So the meaning of asynchronous is actually asynchronous at the byte level but the bits are still synchronized. So their durations are same.

4.15.1 Block Diagram of Asynchronous Transmission :

- Fig. 4.15.2 shows the block diagram of asynchronous transmission.
- The start bits are 0 and stop bits are 1, as shown in Fig. 4.15.2.



(G-1046) **Fig. 4.15.2 : Asynchronous transmission**

The use of start and stop bits and the gaps (idle time) between adjacent data units will make the asynchronous transmission slow. This is the major disadvantage of using the start and stop bits.

4.15.2 Disadvantages of Asynchronous Transmission :

1. Additional bits called start and stop bits are required to be used.
2. It is difficult to determine the sampling instants hence the timing error can take place.
3. The start/stop bits and idle time makes the asynchronous transmission slow.

4.15.3 Advantages of Asynchronous Transmission :

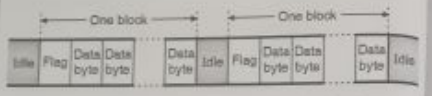
1. Synchronization between the transmitter and receiver is not necessary.
2. It is possible to transmit signals from the sources having different bit rates.
3. The transmission can commence as soon as the data byte to be transmitted becomes available.
4. This mode of transmission is easy to implement.
5. It is a cheap scheme.
6. It is an effective scheme.

4.15.4 Application of Asynchronous Transmission :

The connection of a keyboard to a computer is an example of asynchronous transmission.

4.16 Synchronous Transmission :

- Synchronous transmission is carried out under the control of a common master clock. Here the bits which are being transmitted are synchronized to a reference clock.
- No start and stop bits are used instead the bytes are transmitted as a block in a continuous stream of bits as shown in Fig. 4.16.1. There is an inter block idle time which is filled with idle characters.
- The receivers operates at exactly the same clock frequency as that of transmitter as both are synchronized with each other.
- This is essential for error free reception of data. Flag is a sequence of fixed number of bits which is prefixed to each block as shown in Fig. 4.16.1. Flag is useful in identifying the beginning of a block.

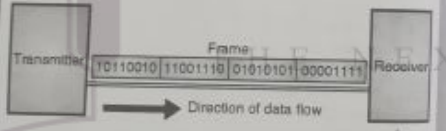


(G-1045) **Fig. 4.16.1 : Synchronous transmission**

- In the synchronous transmission the bit stream to be transmitted is combined into longer "frames". A frame would contain more than one bytes.
- There is no gap between the successive frames. The receiver separates the bit stream into bytes for the purpose of decoding.
- Start and stop bits are not used. Instead bits are transmitted serially one after the other.
- The grouping of these bits is responsibility of the receiver.

4.16.1 Block Diagram of Synchronous Transmission :

- Fig. 4.16.2 shows the block schematic of synchronous transmission. Note the absence of gaps and start stop bits.



(G-1046) **Fig. 4.16.2 : Synchronous transmission**

4.16.2 Advantages :

1. The speed of transmission is much higher than that of asynchronous transmission. This is due to the absence of gaps between the data units and absence of start stop bits.
2. Start and stop bits are not needed any more.
3. Timing errors are reduced due to synchronization.

4.16.3 Disadvantages :

1. The timing is very important. The accuracy of the received data is dependent entirely on the ability of the receiver to count the received bits accurately.
2. The transmitter and receiver have to operate at the same clock frequency. This requires proper synchronization which makes the system complicated.

4.16.4 Application of Synchronous Transmission :

The synchronous transmission, due to its high speed is used for the data exchange from one computer to the other.

4.16.5 Synchronization :

The byte synchronization is achieved at the data link layer level for the synchronous transmission between computers.

4.16.6 Comparison of Synchronous and Asynchronous Transmission :

Table 4.16.1

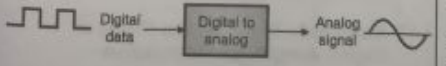
Sr. No.	Parameter	Asynchronous transmission	Synchronous transmission
1.	Synchronization	Not needed	Needed
2.	Start and Stop bits	Used	Not used
3.	Gaps between data blocks	Present	Absent
4.	Speed	Low	High
5.	Application	Communication between a computer and keyboard.	Communication between two computers.

4.17 Isochronous :

- This is the third type of serial transmission. In the real time streaming of audio and video, the time delay introduced during the transmission must remain constant. An uneven time delay would introduce distortion.
- For such applications we cannot use the synchronous serial transmission successfully. The isosynchronous transmission can be used for such applications.

4.18 Digital to Analog Conversion :

- In the process of D to A conversion the digital data at the input is converted into an analog signals. These analog signals are transmitted over the transmission medium.
- The most familiar application of D to A conversion is for transmitting digital data through the public telephone network.
- The D to A conversion is done by the modems to convert the digital data from the computers into the analog signals that are sent on the telephone lines for the Internet.



(L-791) **Fig. 4.18.1 : Digital data to analog signal**

4.18.1 Aspects of Digital to Analog Conversion :

- The two most important aspects related to D to A conversion are as stated below :
  1. Data element versus signal element
  2. Data rate versus signal rate.

1. Data element versus signal element :

- We may define data element as the smallest piece of information that can be exchanged and as we know it is a "bit".
- The signal element is classically defined as the smallest unit of a signal that is constant. This definition is true in the digital context. But the signal will be analog here hence the nature of the signal element is slightly different than that for the digital transmission.

2. Data rate versus signal rate :

- We have defined the data rate (bit rate) and signal rate (baud rate) earlier.
- The relation between them is as follows :

$$\text{Signal rate } S = N \times \frac{1}{r} \text{ baud}$$

Where N = Data rate (bps)
r = Number of data elements in one signal element

- The value of "r" in the analog transmission is as follows :

$$r = \log_2 L$$

Where L is the type of signal element not the level.

Note :
- Bit rate = Number of bits per second.
- Baud rate = Number of signal elements or symbols per second.
- In analog communication, of digital data the bit rate (bps) is always greater than or equal to the baud rate (bauds).

Ex. 4.18.1 : An analog signal carries 4 bits per signal element. If number of signal elements sent per second is 500 calculate the bit rate.

Soln. :

Given : r = 4, S = 500, N = ?

We know that $S = N \times \frac{1}{r}$

$\therefore 500 = N \times \frac{1}{4}$

$\therefore N = 2000$ bps ...Ans.

Bandwidth :

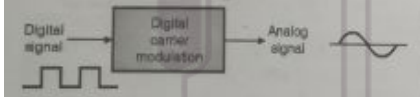
The bandwidth requirement for analog transmission of digital data is proportional to the signal rate i.e. the baud rate. But this is not true for FSK system, this has been discussed later.

Carrier signal :

- In the D to A conversion, at the sending end a high frequency signal which acts as the base signal for transmission of information is produced. This signal is known as the carrier signal or carrier frequency.
- The input digital signal (which is the information signal) will change one of the characteristics of this carrier such as amplitude, frequency or phase.
- This type of modification or modulation is known as **shift keying**. Depending on which parameter of the carrier is being modified we get Amplitude Shift Keying (ASK), frequency shift keying (FSK) or Phase Shift Keying (PSK).

4.18.2 Need of Digital Continuous Wave Modulation :

- PCM converts analog message signal into a digital signal. Now we will learn some techniques which convert the digital message signal into an analog signal and then transmits it.
- Such modulation schemes are called as digital carrier modulation schemes.
- This type of digital to analog conversion is essential when the digital message signal is to be sent over a bandlimited channel such as the telephone line.
- The best application of digital carrier modulation is MODEM.
- The modem will modulate the digital data signal from the DTE (computer) into an analog signal.
- This analog signal is then transmitted on the telephone lines.

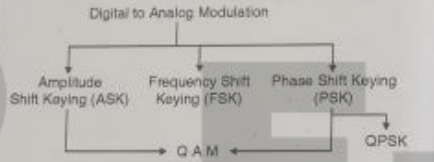


(L-61) **Fig. 4.18.2 : Digital carrier modulation**

- The question is why can't we send the digital signal as it is on the telephone lines ? Why should we modulate it ?
- Here is the answer for it. The digital data consists of binary 0s and 1s, therefore the waveform changes its value abruptly from high to low or low to high.
- In order to carry such a signal without any distortion being introduced, the communication medium needs to have a large bandwidth.
- Unfortunately the telephone lines do not have high bandwidth. Therefore we have to convert the digital signal first into an analog signal which needs lower bandwidth by means of the modulation process.

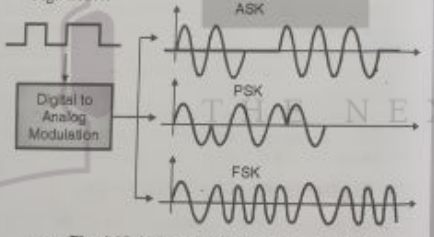
4.18.3 Types of Digital Carrier Modulation :

- There are three basic types of modulation techniques for the transmission of digital signals.
- These methods are based on the three characteristics of a sinusoidal signal; amplitude, frequency and phase. The corresponding modulation methods are then called as :
  1. Amplitude shift keying (ASK)
  2. Frequency shift keying (FSK)
  3. Phase shift keying (PSK)
  4. Quadrature phase shift keying (QPSK) or 4-psk.
  5. Quadrature amplitude modulation (QAM).
- QPSK is a multilevel modulation in which four phase shifts are used for representing four different symbols.
- At high bit rates, a combination of ASK and PSK is employed in order to minimize the errors in the received data.
- This method is known as "Quadrature Amplitude Modulation (QAM)". Let us discuss these methods one by one.
- Fig. 4.18.3 shows the classification of digital to analog modulation systems.



(L-62) **Fig. 4.18.3 : Types of digital to analog modulation**

- Digital to analog modulation is demonstrated in Fig. 4.18.4.



(L-63) **Fig. 4.18.4 : Digital to analog modulation**

4.18.4 Advantages and Disadvantages of CW Modulation :

1. The advantage of CW modulation techniques such as ASK, PSK, FSK etc. used for transmission of data is that we can use the telephone lines for transmission of high speed data. Due to the use of CW modulation the BW requirement is reduced.
2. The disadvantage of CW modulation is we need to use a MODEM alongwith every computer. This makes the system costly and complex.

4.19 Amplitude Shift Keying (ASK) or Digital Amplitude Modulation :

Definition :

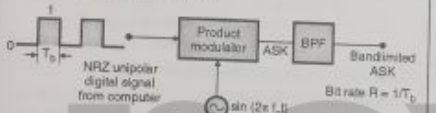
ASK is the digital carrier modulation in which the amplitude of the sinusoidal carrier will take one of the two predetermined values in response to 0 or 1 value of digital input signal.

Generation and waveforms :

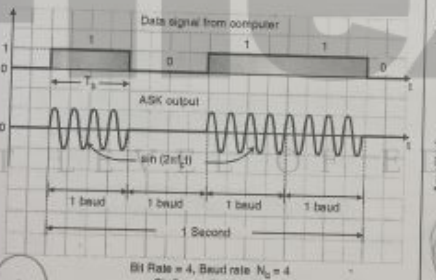
- Amplitude shift keying (ASK) is the simplest type of digital CW modulation. Here the carrier is a sinewave of frequency f_c. We can represent the carrier signal mathematically as follows :

$$e_c = \sin(2\pi f_c t) \quad ...(4.19.1)$$

- The digital signal from the computer is a unipolar NRZ signal which acts as the modulating signal. The ASK modulator is nothing but a multiplier followed by a band pass filter as shown in Fig. 4.19.1(a).
- Due to the multiplication, the ASK output will be present only when a binary "1" is to be transmitted.
- The ASK output corresponding to a binary "0" is zero as shown in Fig. 4.19.1(b).



(L-64) **Fig. 4.19.1(a) : ASK generator**



(L-65) **Fig. 4.19.1(b) : ASK waveforms**

- From the waveforms of Fig. 4.19.1(b) we can conclude that the carrier is transmitted when a binary 1 is to be sent and no carrier is transmitted when a binary 0 is to be sent.
- The ASK signal can be mathematically expressed as follows :

$$V_{ASK}(t) = d \sin(2\pi f_c t) \quad ...(4.19.2)$$

where d = Data bit which can take values 1 or 0.

$$\therefore V_{ASK}(t) = \sin(2\pi f_c t) \text{ when } d = 1 \quad \text{and} \quad V_{ASK}(t) = 0 \text{ when } d = 0 \quad ...(4.19.3)$$

4.19.1 Baud Rate (N_b) :

- For ASK we use 1 bit (0 or 1) to represent one symbol. So the rate of symbol transmission i.e. the baud rate will be same as the bit rate.
- N_b will be same as bit rate R as shown in Fig. 4.19.1(b).

$$\therefore \text{Baud rate} = \text{Bit rate} \qquad N_b = f_b$$

4.19.2 Transmission Bandwidth of the ASK Signal :

- The bandwidth of ASK signal is dependent on the bit rate f_b. Where bit rate $f_b = 1/T_b$ as shown in Fig. 4.19.1(a). For a bit rate of "f_b" bits/sec. the maximum bandwidth required for an ASK signal is

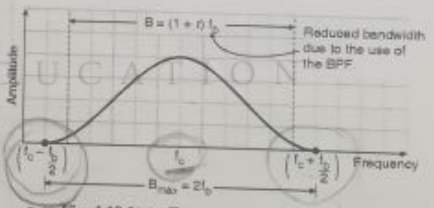
$$BW_{(max)} = f_b \text{ Hz} \quad ...(4.19.4)$$

- The frequency spectrum of an ASK signal is shown in Fig. 4.19.1(c) which shows that the spectrum consists of the carrier frequency f_c with upper and lower sidebands.
- The transmission bandwidth BW of the ASK signal can be restricted by using a filter. The restricted value of bandwidth is given as :

$$BW = (1 + r) f_b \quad ...(4.19.5)$$

where "r" is a factor related to the filter characteristics and its value lies between 0 and 1.

- f_c is the carrier frequency i.e. frequency of the sine wave being transmitted.



(L-66) **Fig. 4.19.1(c) : Frequency spectrum of an ASK signal**

4.19.3 Bandwidth of ASK in Terms of Baud Rate :

- For ASK, as shown in Fig. 4.19.1(b), the baud rate = Bit rate.
- The bandwidth of ASK in terms of bit rate is given by,

$$BW = f_c + \left(\frac{f_b}{2}\right) - \left[f_c - \left(\frac{f_b}{2}\right)\right] = f_b$$

Where $f_b = \frac{1}{T_b}$ = Bit rate and T_b = One bit interval.

- Since bit rate and baud rate are equal for ASK, the expression for bandwidth is given by

$$BW = f_c + \frac{N_b}{2} - \left[f_c - \frac{N_b}{2}\right]$$

Where N_b = Baud rate = f_b i.e. Bit rate

$\therefore$ Bandwidth = N_b

- But practically the bandwidth requirement of ASK is given by,

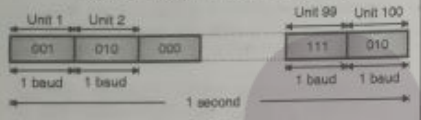
$$BW = (1 + d) \times N_b$$

where d is related to modulation process.

Ex. 4.19.1 : An analog signal carries 3 bits in each signal unit. If the transmitter send 100 such units per second, calculate the bit rate and baud rate.

Soln. :

Given : Number of bits per signal unit = 3,
Number of signal units per second = 100



(L-67) **Fig. P. 4.19.1**

1. Bit rate = Number of bits per second = Number of bits/unit × Number of units / sec.
   = 3 × 100 = 300 bits/sec. ...Ans.
2. Baud rate = Number of signal units per second
   = 100 bauds/sec. ...Ans.

Ex. 4.19.2 : An ASK transmitter transmits 5000 bits per second. Calculate the minimum bandwidth. Assume the transmission to be half duplex.

Soln. :

Given : Bit rate = 5000 bps, Mode : Half duplex

To find : Minimum bandwidth

- The minimum bandwidth of ASK is given by,

$$BW_{(min)} = N_b$$

- For ASK, baud rate N_b is equal to bit rate

$\therefore$ $BW_{(min)}$ = 5000 Hz ...Ans.

Ex. 4.19.3 : The bandwidth of an ASK system is 3000 Hz. Calculate the bit rate and baud rate.

Soln. :

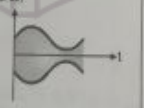
1. For an ASK system, the baud rate is equal to the bandwidth,
   $\therefore$ Baud rate N_b = 3000 bauds per sec. ...Ans.
2. For ASK, bit rate is equal to baud rate.
   $\therefore$ Bit rate = 3000 bits/sec. ...Ans.

4.19.4 Merits and Demerits of ASK :

The advantage of using ASK is its simplicity. It is easy to generate and detect. However its disadvantage is that it is very sensitive to noise, therefore it finds limited application in data transmission. It is used at very low bit rates, upto 100 bits/sec.

4.19.5 Comparison of AM and ASK :

Table 4.19.1 : Comparison of AM and ASK

Sr. No.	Parameter	AM	ASK
1.	Variable characteristics of the carrier.	Amplitude	Amplitude
2.	Nature of modulating signal.	Modulating signal is analog	Modulating signal is digital
3.	Modulated signal shape.	(L-68) 	(L-68) 
4.	Variation in the carrier amplitude.	Continuous variation in accordance with the amplitude of modulating signal.	Carrier ON or OFF depending on whether a 1 or 0 is to be transmitted.
5.	Number of sidebands produced.	Two	Two
6.	Bandwidth	$2 f_m$	$(1 + r) f_b$
7.	Noise immunity	Poor	Poor
8.	Application	Radio broadcasting	Data transmission at low bit rate
9.	Detection method	Envelope	Envelope

4.19.6 Multilevel ASK :

- In the above discussion, there were only two amplitude levels of the input digital data involved. But we can have multilevel ASK in which the digital data can have multiple amplitude levels (e.g. 4, 8, 16 levels etc.). This is possible if we consider a group of consecutive data bits as one symbol.
- 4 levels correspond to 2 bits at a time. Similarly 8 and 16 levels correspond to 3 and 4 bits at a time. Thus multiple bits/symbol would result in multilevel ASK.
- This principle is actually not implemented with pure ASK but it is implemented with QAM.

Ex. 4.19.4 : Given bandwidth of 10,000 Hz (100 to 11000 Hz), draw the full duplex ASK diagram of the system. Find the carriers and the bandwidths in each direction. Assume there is no gap between the bands in the two directions.

Soln. :

Given : System : ASK mode : Full duplex
BW = 10,000 Hz No gap between bands

Step 1 : Calculate BW in each direction :

For a full duplex ASK, the bandwidth in each direction is given by ,

$$BW = \frac{10,000}{2} = 5000 \text{ Hz.}$$

Step 2 : Carrier frequencies in each direction :

- The frequency bands allocated for backward and forward transmission are : 1000 Hz to 6000 Hz and 6000 Hz to 11000 Hz respectively.
- The carrier frequencies are placed at the center of each band.

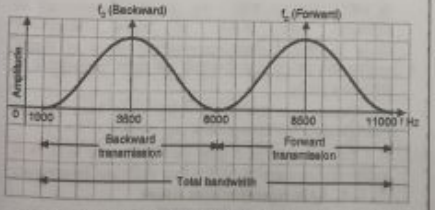
$$\therefore f_c \text{ (backward)} = 1000 + \frac{5000}{2} = 3500 \text{ Hz}$$

and

$$f_c \text{ (forward)} = 6000 + \frac{5000}{2} = 8500 \text{ Hz}$$

Step 3 : Draw the frequency spectrum :

- The required frequency spectrum is shown in Fig. P. 4.19.4.



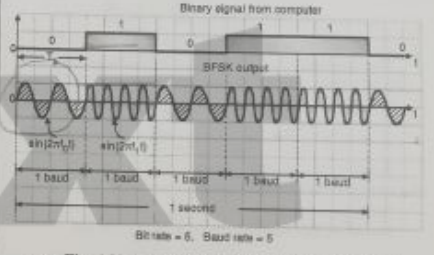
(L-70) **Fig. P. 4.19.4 : Frequency spectrum of a duplex ASK system**

4.20 Frequency Shift Keying (FSK) :

Definition and waveforms :

- In "frequency shift keying (FSK)", the frequency of a sinusoidal carrier is shifted between two discrete values, in response to the value (0 or 1) of the digital input signal.
- One of these frequencies (f_1) represents a binary "1" and the other value (f_0) represents a binary "0".
- The representation of digital data using FSK is as shown in Fig. 4.20.1(b). Note that there is no change in the amplitude of the carrier.

(L-785) **Fig. 4.20.1(a) : FSK generation**



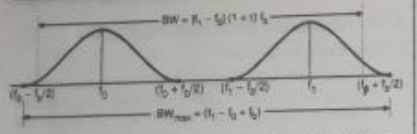
(L-72) **Fig. 4.20.1(b) : Representation of digital signal using FSK**

4.20.1 FSK Generation :

- Refer to the FSK generator shown in Fig. 4.20.1(a). It is basically a voltage controlled oscillator (VCO) which produce sinewaves at frequencies f_1 and f_0, respectively.
- Corresponding to binary 0 input, the VCO produces a sinewave of frequency f_0 whereas corresponding to binary 1 input, the VCO produces a sinewave of frequency f_1. ($f_1 > f_0$).
- Thus we obtain the binary FSK (BFSK) signal at the output of VCO corresponding to the input digital data bits.

4.20.2 Frequency Spectrum of Binary FSK Signal :

- The FSK signal can be considered to be containing two ASK signals, with the carrier frequencies f_1 and f_0.
- Therefore the frequency spectrum of the FSK signal is as shown in Fig. 4.20.1(c) which is identical to that of ASK for each of the two frequencies.



(L-73(b)) **Fig. 4.20.1(c) : Frequency spectrum of a binary FSK signal**

4.20.3 Bandwidth of FSK Signal :

- The bandwidth of FSK signal is dependent on the pulse width T_b or bit rate $f_b = 1/T_b$ and the separation between the frequencies f_0 and f_1, as shown in Fig. 4.20.1(c).
- The maximum bandwidth of FSK system is given by,

$$BW_{max} = \left(f_1 + \frac{f_b}{2}\right) - \left(f_0 - \frac{f_b}{2}\right)$$
$$= (f_1 - f_0 + f_b) \quad ...(4.20.1)$$

- The bandwidth can be restricted by using a bandpass filter after the VCO in the FSK generator. The restricted bandwidth is given as :

$$BW = |f_1 - f_0| + (1 + r)\frac{f_b}{2} \quad ...(4.20.2)$$

Where "r" is the factor related to the filter characteristics and its value lies between 0 and 1.

- The separation between f_1 and f_0 is kept at least $2\ f_b/3$. Substitute this value in Equation (4.20.2) to get

$$BW_{min} = \frac{2}{3}f_b + f_b = \frac{5f_b}{3} \quad ...(4.20.3)$$

- This shows that FSK requires larger bandwidth than ASK and PSK (to be discussed next).

4.20.4 Bandwidth for FSK in terms of Baud Rate :

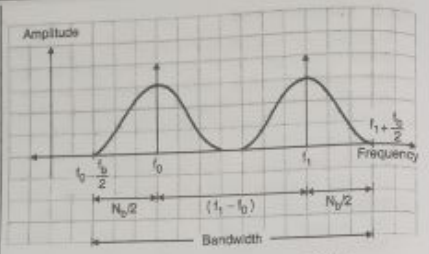
- For FSK also bit rate is equal to baud rate. This is due to the fact that each data bit at the input is treated as a separate symbol.
- We can imagine the FSK spectrum to be a combination of two ASK spectrums centered at frequencies f_1 and f_0 as shown in Fig. 4.20.1(d).
- From Fig. 4.20.1(d) the expression for bandwidth is given by

$$BW = \frac{N_b}{2} + (f_1 - f_0) + \frac{N_b}{2}$$
$$= (f_1 - f_0) + N_b \quad ...(4.20.4)$$

Where N_b = Baud rate = Bit rate = f_b

- Minimum bandwidth will correspond to the situation in which $(f_1 - f_0) = N_b$

$$\therefore BW_{(min)} = N_b + N_b = 2N_b = 2f_b \quad ...(4.20.5)$$



(L-74) **Fig. 4.20.1(d) : Spectrum of FSK**

Ex. 4.20.1 : Calculate the bandwidth of an FSK system in which, the transmission takes place at 4000 bits per second rate and the frequency difference between the two carriers is 3000 Hz.

Soln. :

Given : $(f_1 - f_0) = 3000$ Hz, Bit rate = 4000 bps.

To find : Bandwidth

- Bandwidth = $(f_1 - f_0) + N_b$
- But N_b = baud rate = bit rate = 4000

$\therefore$ BW = 3000 + 4000 = 7000 Hz **...Ans.**

Ex. 4.20.2 : For a half duplex FSK transmission, the bandwidth of medium is 8000 Hz. If the frequency difference between the two carriers is 4000 Hz calculate the maximum bit rate.

Soln. :

Given : FSK, half duplex, BW = 8000 Hz, $f_1 - f_0 = 4000$ Hz.

To find : Maximum bit rate

1. $BW = (f_1 - f_0) + N_b$

$\therefore 8000 = 4000 + N_b$

$\therefore N_b = 4000$ bauds/sec.

2. For FSK system baud rate is equal to bit rate.

$\therefore$ Bit rate = 4000 bits per second **...Ans.**

4.20.5 Multilevel FSK (MFSK) :

- In BFSK (Binary FSK) we use two frequencies to represent two levels of signal amplitude (0 or 1). But if there are multiple levels of signal amplitude, then it is possible to use more frequencies to represent them.
- As discussed in multilevel ASK, even here we use multiple data bits to represent one symbol.
- If 2 bits are used at a time then there will be $2^2 = 4$ levels. So we have to assign four different frequencies (f_0, f_1, f_2, f_3) to represent these levels.
- The bandwidth requirement of MFSK is higher than that of BFSK.

4.20.6 Advantages of FSK :

1. FSK is relatively easy to implement.
2. It has better noise immunity than ASK. Therefore the probability of error free reception of data is high.

4.20.7 Disadvantages of FSK :

1. The major disadvantage is its high bandwidth requirement as discussed earlier.
2. Therefore FSK is extensively used in low speed modems having bit rates below 1200 bits/sec.
3. The FSK is not preferred for the high speed modems because with increase in speed, the bit rate increases.
4. This increases the channel bandwidth required to transmit the FSK signal.
5. As the telephone lines have a very low bandwidth, it is not possible to satisfy the bandwidth requirement of FSK at higher speed. Therefore PSK is preferred only for the low speed modems.

4.21 Phase Shift Keying (PSK) :

Definition and waveforms :

- Phase shift keying (PSK) is the most efficient of the three modulation methods.
- Therefore it is used for high bit rates. In PSK, phase of the sinusoidal carrier is changed according to the data bit to be transmitted.
- Fig. 4.21.1(a) shows the simplest form of PSK called Binary PSK (BPSK). The carrier phase is changed between 0° and 180° by the bipolar digital signal. A bipolar NRZ signal is used to represent the digital data from the DTE.
- The BPSK signal can be represented mathematically as :

$V_{BPSK}(t) = \sin(2\pi f_c t)$ when binary "0" is to be represented

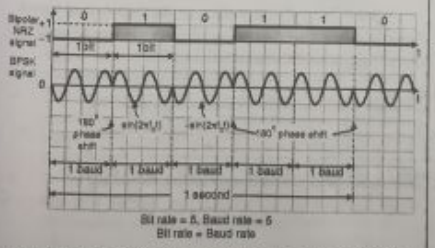
and $V_{BPSK}(t) = -\sin(2\pi f_c t)$

$= \sin(2\pi f_c t + \pi)$ when binary "1" is to be represented.

- Combining the two conditions we can write

$$V_{BPSK}(t) = d \sin(2\pi f_c t) \quad ...(4.21.1)$$

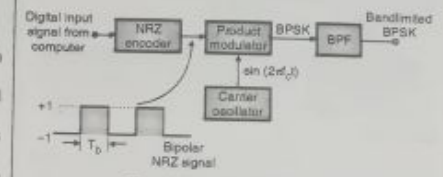
where $d = \pm 1$



(L-80) **Fig. 4.21.1(a) : Binary phase shift keying (BPSK)**

4.21.1 BPSK Generation :

- The BPSK generation takes place as shown in Fig. 4.21.1(b).



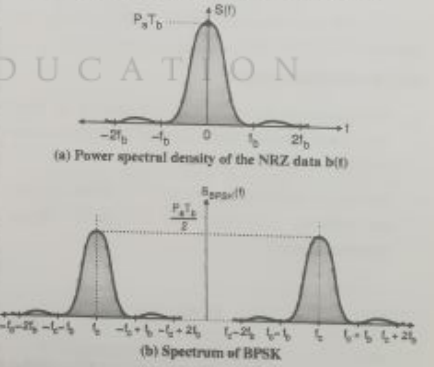
(L-81) **Fig. 4.21.1(b) : BPSK generation**

- The binary data signal (0s and 1s) is converted into a NRZ bipolar signal by an NRZ encoder, which is then applied to a multiplier (balanced modulator). The other input to the multiplier is the carrier signal ($2\pi f_c t$).
- The data bits 0s and 1s are converted into a bipolar NRZ signal "d" as shown in the following table.

Digital signal	Bipolar NRZ signal	BPSK output
Binary 0	d = 1	$V_{BPSK}(t) = \sin(2\pi f_c t)$
Binary 1	d = – 1	$V_{BPSK}(t) = -\sin(2\pi f_c t)$

4.21.2 Spectrum of BPSK :

- The spectrum of BPSK is as shown in Fig. 4.21.2.



(a) Power spectral density of the NRZ data b(t)

(b) Spectrum of BPSK

(L-88) **Fig. 4.21.2 : Spectrum of BPSK**

4.21.3 Bandwidth of BPSK :

- From the frequency spectrum of BPSK signal, shown in Fig. 4.21.2(b), we can come to a conclusion that the bandwidth of a BPSK signal is given by,

BW = Highest frequency – Lowest frequency in main lobe = $(f_c + f_b) - (f_c - f_b)$

$\therefore$ BW = $2 f_b$ ...(4.21.2)

where $f_b = 1 / T_b$

- Thus the minimum bandwidth of BPSK signal is equal to twice the highest frequency contained in the baseband signal.

Baud rate :

In BPSK also each digit (0 or 1) of the input digital data represents a symbol. Hence symbol rate is equal to bit rate.

$\therefore$ Baud rate N_b = Bit rate f_b

$\therefore$ BW = $2 N_b$

4.21.4 Advantages of BPSK :

1. BPSK has a bandwidth which is lower than that of a BFSK signal.
2. BPSK has the best performance of all the systems in presence of noise. It gives the minimum possibility of error.
3. BPSK has a very good noise immunity.

4.21.5 Disadvantage of BPSK :

The only disadvantage of BPSK is that generation and detection of BPSK is not easy. It is quite complicated.

4.21.6 Applications :

- Phase shift keying is the most efficient of the three modulation methods and it is used for high bit rates even higher than 1800 bits/sec.
- Due to low bandwidth requirement the BPSK modems are preferred over the FSK modems, at higher operating speeds.

4.21.7 Comparison of Binary Modulation Systems :

Sr. No.	Parameter	Binary ASK	Binary FSK	Binary PSK
1.	Variable characteristic.	Amplitude	Frequency	Phase
2.	Bandwidth (Hz)	2R	$\lvert f_1 - f_0 \rvert$ + (1 + r) R	(1+ r) R
3.	Noise immunity.	Low	High	High
4.	Error probability	High	Low	Low
5.	Performance in presence of noise.	Poor	Better than ASK	Better than FSK
6.	Complexity	Simple	Moderately complex	Very complex
7.	Bit rate	Suitable upto 100 bits/sec.	Suitable upto about 1200 bits/sec.	Suitable for high bit rates.
8.	Detection method.	Envelope	Envelope	Coherent

Ex. 4.21.1 : If the data bit sequence consists of the following string of bits, what will be the nature of waveform transmitted by BPSK transmitter ?

The data bit sequence is 1 0 1 1 1 0 1 0.

Soln. :

The BPSK signal can also be expressed in terms of cosine wave as :

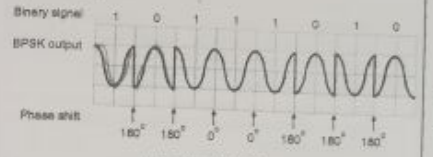
$$V_{BPSK}(t) = \sqrt{2 P_s}\, b(t) \cos \omega_c t$$

where $b(t) = \pm 1$ depending on the digital input signal. Table P. 4.21.1 lists the values of b (t) and the transmitted signal V_{BPSK} for different bit intervals.

Table P. 4.21.1

Binary signal	1	0	1	1	1	0	1	0
b (t)	+ 1	– 1	+ 1	+ 1	+ 1	– 1	+ 1	– 1
V_{BPSK} (t)	$\cos \omega_c t$	$-\cos \omega_c t$	$\cos \omega_c t$	$-\cos \omega_c t$	$\cos \omega_c t$	$-\cos \omega_c t$	$\cos \omega_c t$	$-\cos \omega_c t$

The transmitted BPSK signal is as shown in Fig. P. 4.21.1.



(E-89) **Fig. P. 4.21.1**

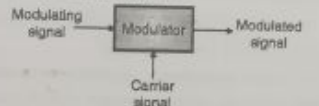
4.22 Analog to Analog Conversion :

- In some applications we have to transform analog data such as voice, video etc. into analog signal.
- This process is known as modulation. The analog data at the input is called as modulating signal. It modulates a high frequency sinusoidal signal called carrier to produce another analog signal called modulated signal.

(E-31) **Fig. 4.22.1 : Transformation from analog data to analog signals**

4.23 Modulation :

- In the **Modulation** process, two signals are used namely the **modulating signal** and the **carrier**.
- The modulating signal is nothing but the baseband signal or information signal while carrier is a high frequency sinusoidal signal.



(D-8) **Fig. 4.23.1 : Modulation**

- In the modulation process some parameter of the carrier wave (such as amplitude, frequency or phase) is varied in proportion with the modulating signal.
- The result of this process is called as the modulated signal. This modulated signal is then transmitted by the transmitter over a communication channel or medium.
- The receiver will **"Demodulate"** the received modulated signal and get the original information signal back. Thus demodulation is exactly opposite to modulation.
- In the process of modulation, the carrier wave actually acts as a **carrier** which carries the information signal (modulating signal) from the transmitter to receiver.
- This is similar to a situation in which a person travels in his car or on his bike from one place to the other. The person can be viewed as the modulating signal and the car or bike as the carrier as shown in Fig. 4.23.2.

(D-9) **Fig. 4.23.2 : Concept of modulation**

4.23.1 Need of Modulation :

- A question may be asked as, when the baseband signals can be transmitted directly why to use the modulation ?
- The answer is that the baseband transmission has many limitations which can be overcome using modulation. It is as explained below.
- In the process of modulation, the baseband signal is "translated" i.e. shifted from low frequency side to high frequency side of the frequency spectrum.
- This frequency shift is proportional to the frequency of carrier. The modulation process has the following advantages :

Advantages of (Reasons for) modulation :

1. Reduction in the height of antenna
2. Avoids mixing of signals
3. Increases the range of communication
4. Multiplexing becomes possible
5. Improves quality of reception.

Reduction in height of antenna :

- For transmission of radio signals, the antenna height must be a multiple of (λ/4). Here λ is the wavelength. $\lambda = c/f$ where c is velocity of light and f is the frequency of the signal to be transmitted.
- The minimum antenna height required to transmit a baseband signal of f = 10 kHz is calculated as follows :

$$\text{Minimum antenna height} = \frac{\lambda}{4} = \frac{c}{4f}$$

$$= \frac{3 \times 10^8}{4 \times 10 \times 10^3}$$

$$= 7500 \text{ meters i.e. } 7.5 \text{ km}$$

The antenna of this height is practically impossible to install.

- Now consider a modulated signal at f = 1 MHz. The minimum antenna height is given by,

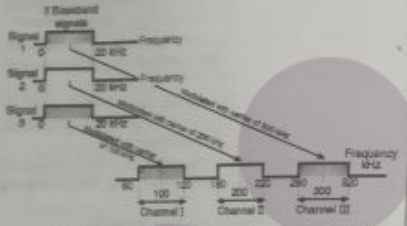
$$\text{Minimum antenna height} = \frac{\lambda}{4} = \frac{c}{4f} = \frac{3 \times 10^8}{4 \times 10 \times 10^6}$$

= 75 meters

- This antenna can be easily installed practically. Thus modulation reduces the height of the antenna.

Avoids mixing of signals :

- If the baseband sound signals are transmitted without using the modulation by more than one transmitter, then all the transmitted signal by multiple transmitters will be in the same frequency range i.e. 0 to 20 kHz.
- Therefore the signals from different stations get mixed together and a receiver cannot separate them from each other.
- So if each baseband sound signal is used to modulate a different carrier which corresponds to a different station then they will occupy different slots in the frequency spectrum (different channels).
- This is as shown in Fig. 4.23.3. Thus modulation avoids mixing of signals.



(D-18) **Fig. 4.23.3 : Modulation avoids mixing of signals**

Increases the range of communication :

- The frequency of baseband signals is low, and the low frequency signals can not travel a long distance when they are transmitted. They get attenuated (suppressed) quickly.
- The attenuation reduces with increase in frequency of the transmitted signals, and they travel longer distance.
- The modulation process increases the frequency of the signal. Hence it increases the range of communication.

Multiplexing becomes possible :

- Multiplexing is a process in which two or more signals can be transmitted over the same communication channel simultaneously.
- This is possible only with modulation. The multiplexing allows the same channel to be used by many signals.
- So many TV channels can use the same frequency range, without getting mixed with each other. OR different frequency signals can be transmitted at the same time.

Improves quality of reception : With frequency modulation (FM), and the digital communication techniques like PCM, the effect of noise is reduced to a great extent. This improves quality of reception.

4.23.2 Demodulation or Detection :

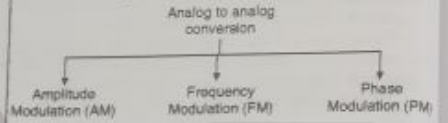
- The modulated signals are transmitted by the transmitter via air medium or wire medium. These signals then reach the receivers by travelling over the communication medium.
- At the receiver, the original information signal is separated from the carrier. This process is called as demodulation or detection. Detection is exactly the opposite process of modulation.

Frequency translation :

The frequency up conversion at the transmitter and frequency down conversion at the receiver together is called as frequency translation.

4.23.3 Types of Analog to Analog Conversion :

The three basic types of analog to analog conversion are shown in Fig. 4.23.4. They are AM, FM and PM.



(L-780) **Fig. 4.23.4 : Types of analog to analog conversion**

4.24 Amplitude Modulation (AM) :

Definition :

Amplitude modulation (AM) or Amplitude Modulation with Full Carrier (AM-FC) is the process of changing the amplitude of a high frequency sinusoidal carrier signal in proportion with the instantaneous value of modulating signal.

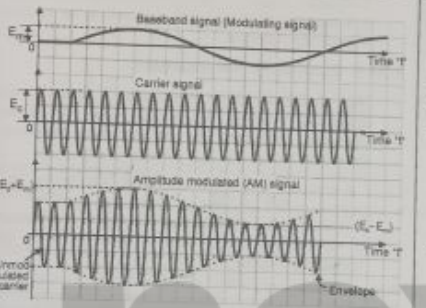
- Fig. 4.24.1 shows the amplitude modulated wave when the modulating signal is a sinusoidal signal.

Observations :

1. The frequency of the sinusoidal carrier is much higher than that of the modulating signal.
2. In AM the instantaneous amplitude of the sinusoidal high frequency carrier is changed in proportion to the instantaneous amplitude of the modulating signal. This is the principle of AM.
3. The time domain display of AM signal is as shown in Fig. 4.24.1. This AM signal is transmitted by a transmitter. The information in the AM signal is contained in the amplitude variations of the carrier or the envelope shown by dotted lines in Fig. 4.24.1.
4. Note that the frequency and phase of the carrier remain constant.

5. AM is used in the applications such as radio transmission, TV transmission etc.

Note : The modulating signal in practice may or may not be purely sinusoidal. Most of the times it will have a complex shape.



(D-12) **Fig. 4.24.1 : AM waveform for sinusoidal modulating signal**

4.24.1 Mathematical Representation of an AM Wave :

Expression of AM wave :

- Let the modulating signal be sinusoidal and be represented as,

$$e_m = E_m \cos \omega_m t \quad ...(4.24.1)$$

where "e_m" is the instantaneous amplitude of the modulating signal, E_m is the peak amplitude, $\omega_m = 2\pi f_m$ and f_m = Frequency of the modulating signal.

- Let the carrier signal also be sinusoidal at a much higher frequency than that of the modulating signal. The instantaneous carrier signal e_c is given by,

$$e_c = E_c \cos \omega_c t \quad ...(4.24.2)$$

where E_c = Peak carrier amplitude,

f_c = Carrier frequency and $\omega_c = 2 \pi f_c$.

- The AM wave is expressed by the following expression,

$$e_{AM} = A \cos (2 \pi f_c t) \quad ...(4.24.3)$$

where A = Envelope of AM wave

- Where A represents the instantaneous value of the envelope. The modulating signal either adds or gets subtracted from the peak carrier amplitude E_c as shown in Fig. 4.24.1. Hence we can represent the instantaneous value of envelope as,

$$A = E_c + e_m = E_c + E_m \cos (2 \pi f_m t) \quad ...(4.24.4)$$

- Hence the AM wave is given by,

$$e_{AM} = A \cos (2 \pi f_c t)$$

$$= [E_c + E_m \cos (2 \pi f_m t)] \cos (2 \pi f_c t)$$

$$\therefore \quad e_{AM} = E_c \left[1 + \frac{E_m}{E_c} \cos (2 \pi f_m t) \right] \cos (2 \pi f_c t)$$

- Let $m = E_m / E_c$ be the modulation index.

$$\therefore \quad e_{AM} = E_c [1 + m \cos (2 \pi f_m t)] \cos (2 \pi f_c t) \quad ...(4.24.5)$$

This expression represents the time domain representation of an AM signal.

Note : It is not necessary to always consider the cosine waves to obtain the mathematical expression. We can even use the sine waves to obtain the mathematical expression for AM.

4.24.2 Modulation Index or Modulation Factor :

- In AM wave the modulation index (m) is defined as the ratio of amplitudes of the modulating and carrier waves as follows :

$$m = \frac{E_m}{E_c} \quad ...(4.24.6)$$

- When $E_m \leq E_c$ the modulation index "m" has values between 0 and 1 and no distortion is introduced in the AM wave. But if $E_m > E_c$ then m is greater than 1. This will distort the shape of AM signal. The distortion is called as "over modulation."
- The modulation index is also called as modulation factor, modulation coefficient or degree of modulation. However if modulation index is expressed as percentage it is called as "percentage modulation."

$$\therefore \% \text{ Modulation} = \frac{E_m}{E_c} \times 100 \quad ...(4.24.7)$$

- Note that "m" is a dimensionless quantity.

4.24.3 Frequency Spectrum of the AM Wave (Frequency Domain Description) :

- The frequency spectrum is a graph of amplitude on Y axis versus frequency on X axis. The frequency spectrum of AM wave tells us about which frequency components are present in the AM wave and what are their amplitudes. So consider the equation for AM wave.

$$e_{AM} = (E_c + E_m \cos \omega_m t) \cos \omega_c t$$

$$E_c = \left[1 + \frac{E_m}{E_c} \cos \omega_m t \right] \cos \omega_c t$$

- As per the definition of the modulation index, $m = E_m / E_c$.

$$\therefore \quad e_{AM} = E_c (1 + m \cos \omega_m t) \cos \omega_c t \quad ...(4.24.8)$$

Simplifying we get,

$$e_{AM} = E_c \cos \omega_c t + m E_c \cos \omega_m t \cos \omega_c t \quad ...(4.24.9)$$

- For the second term in the above expression use the following standard identity :

$$2 \cos A \cos B = \cos (A + B) + \cos (A - B)$$

– Therefore Equation (4.24.9) gets simplified as follows :

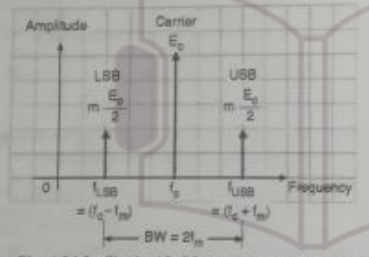
$$e_{AM} = \underbrace{E_c \cos \omega_c t}_{\text{Carrier}} + \underbrace{\frac{m E_c}{2} \cos (\omega_c + \omega_m) t}_{\text{Upper sideband}} + \underbrace{\frac{m E_c}{2} \cos (\omega_c - \omega_m) t}_{\text{Lower sideband}} \quad ...(4.24.10)$$

Observations :

The expression for the AM wave shows that it consists of three terms :

1. First term is nothing else but the unmodulated carrier signal.
2. The second term is a sinusoidal signal at frequency ($f_c + f_m$). This is called as the upper sideband (USB). Its amplitude is $\frac{m E_c}{2}$.
3. The third term represents a sinusoidal signal at frequency ($f_c - f_m$). It is called as the lower sideband (LSB). Its amplitude is $\frac{m E_c}{2}$.

– Hence the frequency spectrum of an A.M. wave is as shown in Fig. 4.24.2. Note that it is a single sided spectrum i.e. the spectrum plotted for only the positive values of frequency.



(D-37) **Fig. 4.24.2 : Single sided frequency spectrum of AM wave**

4.24.4 Concept of Sidebands :

– The AM wave consists of three frequency components (i.e. three sinewaves of different frequencies) namely, the carrier, the lower sideband and upper sideband.

– The lower sideband (LSB) is a sinusoidal component which has a frequency of ($f_c - f_m$) and an amplitude of ($mE_c / 2$).

– The upper sideband (USB) is another sinusoidal component which has a frequency of ($f_c + f_m$) and an amplitude of ($mE_c / 2$).

– The carrier has a frequency f_c and an amplitude of E_c.

Information content in AM wave :

– Note that the amplitude of LSB and USB is ($mE_c / 2$) i.e. it is directly proportional to modulation index m. But the amplitude of carrier (E_c) does not depend on the modulation index.

– This indicates that only the sidebands contain all the information to be conveyed from the transmitter to receiver. Also note that both the sidebands contain identical information.

– The carrier does not contain any information.

– So in order to recover back the transmitted information, it is essential to recover the sidebands at the receiver without any distortion.

4.24.5 Bandwidth Requirement :

– The bandwidth of the AM signal is equal to the difference between the highest and the lowest frequency component in the frequency spectrum. Therefore :

$$BW = f_{USB} - f_{LSB} = (f_c + f_m) - (f_c - f_m)$$
$$BW = 2f_m \quad ...(4.24.11)$$

– This shows that the minimum bandwidth requirement of the DSBFC AM system is equal to twice the modulating frequency (f_m).

Solved Examples :

Ex. 4.24.1 : A modulating signal 10 sin ($2\pi \times 10^3$ t) is used to modulate a carrier signal 20 sin ($2\pi \times 10^4$ t). Find the modulation index, percentage modulation, frequencies of the sidebands components and their amplitudes. What is the bandwidth of the modulated signal ? Also draw the spectrum of the AM wave.

Soln. :

– The modulating signal $e_m = 10 \sin (2\pi \times 10^3 t)$. So comparing this with the expression

$$e_m = E_m \sin (2\pi f_m t) \text{ we get,}$$
$$E_m = 10 \text{ Volts}, f_m = 1 \times 10^3 \text{ Hz} = 1 \text{ kHz}$$

– The carrier signal $e_c = 20 \sin (2\pi \times 10^4 t)$.

Comparing this with the expression $e_c = E_c \sin (2\pi f_c t)$ we get,

$$E_c = 20 \text{ Volts}, f_c = 1 \times 10^4 \text{ Hz} = 10 \text{ kHz}$$

Step 1 : Modulation index and percentage modulation :

$$m = \frac{E_m}{E_c} = \frac{10}{20} = 0.5 \text{ and } \% \text{ modulation}$$
$$= 0.5 \times 100 = 50\%.$$

Step 2 : Frequencies of sideband components :

1. Upper sideband $f_{USB} = f_c + f_m = (10 + 1) = 11$ kHz
2. Lower sideband $f_{LSB} = f_c - f_m = (10 - 1) = 9$ kHz

Step 3 : Amplitudes of sidebands :

The amplitudes of upper as well as the lower sideband is given by,

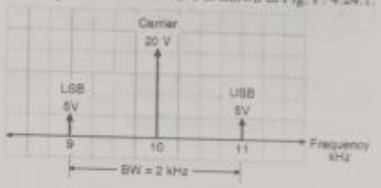
$$\text{Amplitude of each sideband} = \frac{m E_c}{2} = \frac{0.5 \times 20}{2} = 5 \text{ Volts}$$

Step 4 : Bandwidth :

$$\text{Bandwidth} = 2 f_m = 2 \times 1 = 2 \text{ kHz}$$

Step 5 : Spectrum :

The spectrum of AM wave is shown in Fig. P. 4.24.1.



(D-703) **Fig. P. 4.24.1 : Spectrum of the AM wave**

Ex. 4.24.2 : In AM, modulating signal frequency is 10 kHz and carrier frequency is 1 MHz. Determine the resultant frequency components.

Soln. :

Given : $f_m = 10$ kHz, $f_c = 1$ MHz

The resultant frequency components will include the carrier, upper sideband and lower sideband.

1. Carrier frequency, $f_c = 1$ MHz
2. Upper sideband, $(f_c + f_m) = 1000$ kHz + 10 kHz = 1010 kHz
3. Lower sideband, $(f_c - f_m) = 1000$ kHz – 10 kHz = 990 kHz

4.24.6 Effects of Modulation Index on the A.M. Wave :

Depending on the value of percentage modulation (m) the AM wave can be classified into two categories :

1. Linear modulation 2. Overmodulation

Linear modulation :

– If $m \leq 1$ or if the percentage modulation is less than 100% then the type of amplitude modulation is linear amplitude modulation.

– The waveforms of AM waves with linear modulation are in Figs. P. 4.24.3(a) and (b) respectively (Refer Ex. 4.24.3)

Overmodulation :

– If $m > 1$ i.e. if the percentage modulation is greater than 100% then the type of amplitude modulation is called as overmodulation.

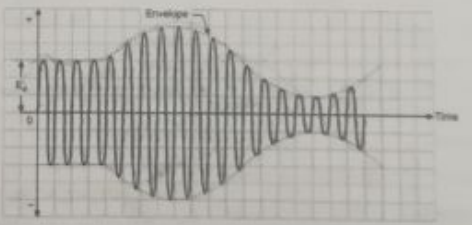
– For $m > 1$ the envelope can sometimes reverse the phase as shown in Fig. P. 4.24.3(d) in Ex. 4.24.3.

– Overmodulation introduces envelope distortion. Hence it should be avoided.

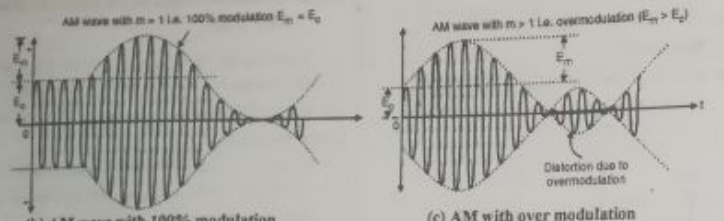
Ex. 4.24.3 : Draw the AM waveforms for less than 100 %, with 100 %, more than 100 % and with 0 % percentage modulation. Assume that the modulating signal is a pure sine wave.

Soln. :

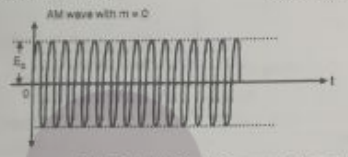
The required waveforms are shown in Figs. P. 4.24.3(a), (b), (c) and (d) respectively.



(D-704) **Fig. P. 4.24.3(a) : AM wave for percentage modulation less than 100 %**



(b) AM wave with 100% modulation (c) AM with over modulation



(B-768) **Fig. P. 4.24.3(d) : AM wave with m = 0**

Ex. 4.24.4 : An audio frequency signal 10 sin 2 π × 500 t is used to amplitude modulate a carrier of 50 sin 2 π × 10^5 t. Calculate :

1. Modulation index
2. Sideband frequencies
3. Amplitude of each sideband frequencies
4. Bandwidth requirement
5. Total power delivered to a load of 600 Ω.

Soln. :

1. The modulating signal $e_m = 10 \sin (2\pi \times 500\, t)$

Comparing it with standard modulating signal given by,

$e_m = E_m \sin (2\pi f_m t)$ we get,

$E_m = 10$ V, $f_m = 500$ Hz

2. The carrier signal $e_c = 50 \sin (2\pi \times 10^5\, t)$

Comparing it with the standard carrier signal given by,

$e_c = E_c \sin (2\pi f_c t)$ we get,

$E_c = 50$ V, $f_c = 1 \times 10^5$ Hz = 100 kHz

Step 1 : Modulation index :

$$m = \frac{E_m}{E_c} = \frac{10}{50} = 0.2$$

Step 2 : Sideband frequencies :

1. $f_{USB} = f_c + f_m = 100.5$ kHz
2. $f_{LSB} = f_c - f_m = 99.5$ kHz

Step 3 : Amplitude of sidebands :

$$\text{Amplitude of sidebands} = \frac{m E_c}{2} = \frac{0.2 \times 50}{2} = 5 \text{ V}$$

Step 4 : Bandwidth :

$BW = 2 f_m = 2 \times 500$ Hz = 1 kHz

Step 5 : Power delivered to load :

$$\text{Carrier power, } P_c = \frac{(E_c/\sqrt{2})^2}{R_L} = \frac{E_c^2}{2 R_L} = \frac{(50)^2}{2 \times 600} = 2.0833 \text{W}$$

$$\text{Total power, } P_t = P_c \left[1 + \frac{m^2}{2}\right] = 2.0833 \left[1 + \frac{(0.2)^2}{2}\right] = 2.125 \text{ W}$$

Ex. 4.24.5 : A carrier wave of frequency 1 MHz and peak value 10 V is amplitude modulated by a 5 kHz sine wave of amplitude 6 V. Determine the modulation index and draw spectrum.

Soln. :

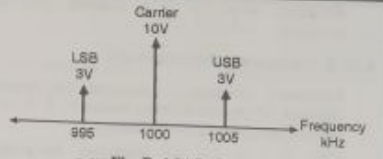
Given : $f_c = 1$ MHz, $E_c = 10$ V, $f_m = 5$ kHz, $E_m = 6$ V

Step 1 : Modulation index :

$$m = \frac{E_m}{E_c} = \frac{6}{10} = 0.6 \quad \text{...Ans.}$$

Step 2 : Spectrum :

1. $f_{USB} = f_c + f_m = 1000$ kHz + 5 kHz = 1005 kHz
2. $f_{LSB} = f_c - f_m = 1000$ kHz − 5 kHz = 995 kHz
3. Amplitude of each sideband $\frac{m E_c}{2} = \frac{0.6 \times 10}{2}$ = 3 Volts
4. Spectrum is shown in Fig. P. 4.24.5.



(L-788) **Fig. P. 4.24.5 : Spectrum**

Ex. 4.24.6 : For given data, find modulation index, frequencies of the sideband components and their amplitudes and plot the frequency spectrum of Amplitude modulated wave :

Modulating signal $e_m = 5 \cos 2\pi 10^3 t$

Carrier signal $e_c = 10 \cos 2\pi 10^4 t$

Soln. :

Given : Modulating signal $e_m = 5 \cos 2\pi 10^3 t$

Carrier signal $e_c = 10 \cos 2\pi 10^4 t$

To find : 1. Modulation index
2. Sideband frequencies and amplitudes
3. Frequency spectrum.

1. Modulation index (m) :

From the expressions of e_m and e_c we get,

$E_m = 5$ V, $f_m = 1000$ Hz, $E_c = 10$ V, $f_c = 10$ kHz.

$$\therefore \quad m = \frac{E_m}{E_c} = \frac{5}{10} = 0.5 \quad \text{...Ans.}$$

2. Sideband frequencies :

$f_{USB} = f_c + f_m = 10 + 1 = 11$ kHz ...Ans.

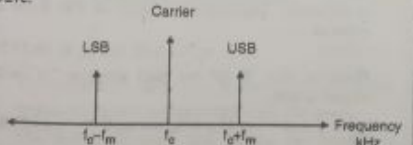
$f_{LSB} = f_c - f_m = 10 - 1 = 9$ kHz ...Ans.

3. Amplitude of sidebands :

$$\text{Amplitude of both the sidebands} = \frac{m E_c}{2} = \frac{0.5 \times 10}{2} = 2.5 \text{ V} \quad \text{...Ans.}$$

4. Frequency spectrum :

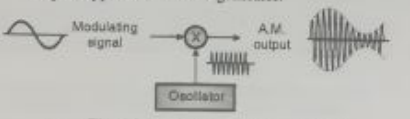
Fig. P. 4.24.6 shows the frequency spectrum of AM wave.



(L-789) **Fig. P. 4.24.6**

4.24.7 Generation of AM Wave :

- Fig. 4.24.3 shows the principle of generation of AM waves. The oscillator produces a sinusoidal carrier of desired high frequency.
- This carrier and the modulating signal are applied to a multiplier circuit. At the output of the multiplier we get the A.M. wave as shown. Thus a multiplier is the simplest type of A.M. wave generator.



(L-781) **Fig. 4.24.3 : Generation of AM wave**

4.25 Advantages, Disadvantages and Applications of AM :

4.25.1 Disadvantages of AM (DSBFC) :

The AM signal is also called as "Double Sideband Full Carrier (DSBFC)" signal. The three main disadvantages of this technique are :

1. Power wastage takes place (Carrier does not contain any information).
2. AM needs larger bandwidth.
3. AM wave gets affected due to noise.

4.25.2 Advantages of AM :

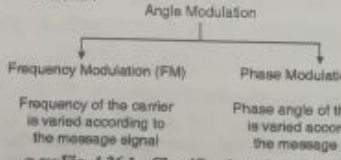
1. AM transmitters are less complex.
2. AM receivers are simple, detection is easy.
3. AM receivers are cost efficient. Hence even a common person can afford to buy it.
4. AM waves can travel a longer distance.
5. Low bandwidth.

4.25.3 Applications of AM :

1. Radio broadcasting.
2. Picture transmission in a TV system.

4.26 Angle Modulation : Basic Concepts :

- There is another method of modulating a sinusoidal carrier namely the angle modulation. In angle modulation either frequency or phase of the carrier is varied in proportion with the message signal amplitude, but the carrier amplitude remains constant.
- Thus angle modulation systems can be classified as follows :



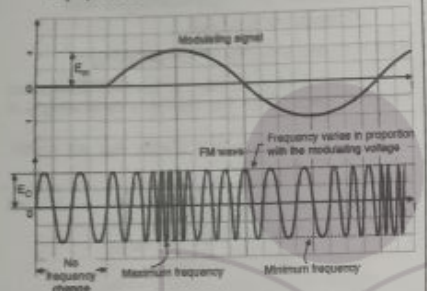
(B-768) **Fig. 4.26.1 : Classification of angle modulation**

4.27 Frequency Modulation (FM) :

- In sinusoidal Frequency Modulation (FM), the modulating signal $x(t) = E_m \cos (2\pi f_m t)$ is a pure

sinusoidal signal. The carrier signal c (t) is also a sinewave at much higher frequency.

- FM is a system of modulation in which the instantaneous frequency of the carrier is varied in proportion with the amplitude of the modulating signal. The amplitude of the carrier signal remain constant. Thus the information is conveyed via frequency changes.
- FM was first practically tried in 1936 as an alternative to AM. As will be shown later on, FM transmission is more resistant to noise than A.M. The time domain display of FM wave is as shown in the Fig. 4.27.1.



(D-760) **Fig. 4.27.1 : Time domain display of FM wave**

- The amount by which the carrier frequency deviates from its unmodulated value is called as **"deviation"**. The deviation (δ) is made proportional to the instantaneous value of modulating voltage.
- The rate at which these frequency variations or oscillations takes place in the FM wave is equal to the modulating frequency (f_m).
- The amplitude of the FM wave always remains constant. This is the biggest advantage of FM.

4.27.1 Important Definitions in Frequency Modulation :

- For the F.M. wave the modulating signal x (t) be a sinusoidal signal of amplitude E_m and frequency f_m.

$$\therefore \quad x(t) = E_m \cos(2\pi f_m t) \qquad ...(4.27.1)$$

- The unmodulated carrier is represented by the expression,

$$e_c = A \sin(\omega_c t + \phi) \qquad ...(4.27.1(a))$$

Instantaneous frequency of an FM wave :

In FM, the frequency f of the FM wave varies in accordance with the modulating voltage. The instantaneous frequency of the FM wave is denoted by f_i (t) and is given by,

$$f_i(t) = f_c[1 + k_f x(t)] = f_c[1 + k_f \cdot E_m \cos(2\pi f_m t)] = f_c + \delta \cos(2\pi f_m t) \qquad ...(4.27.2)$$

Where $\delta = k_f E_m f_c$ and it is called as **frequency deviation,** where k_f is a constant with units Hz / Volts.

4.27.2 Frequency Deviation (δ) :

- **Frequency deviation** δ represents the maximum departure of the instantaneous frequency f_i (t) of the FM wave from the carrier frequency f_c.
- Since $\delta = k_f E_m$, the **frequency deviation** is proportional to the amplitude of modulating voltage (E_m) and it is independent of the modulating frequency f_m.

Maximum and minimum frequency of FM wave :

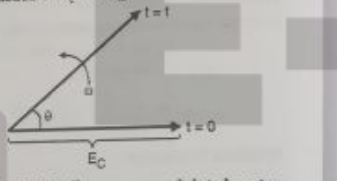
The maximum frequency of FM wave is,

$$f_{max} = f_c + \delta \qquad ...(4.27.3)$$

The minimum frequency of a FM wave is $f_{min} = (f_c - \delta)$.

4.27.3 Mathematical Expression for F.M. :

- We know that the FM wave is a sinewave having a constant amplitude and a variable instantaneous frequency. As the instantaneous frequency is changing continuously, the angular velocity "ω" of an FM wave is the function of ω_c and ω_m.



(D-154) **Fig. 4.27.2 : Frequency modulated vector**

- Therefore the FM wave is represented by,

$$e_{FM} = s(t) = E_c \sin[F(\omega_c, \omega_m)] \qquad ...(4.27.4)$$

$$= E_c \sin\theta(t) \qquad ...(4.27.5)$$

where $\theta(t) = F(\omega_c, \omega_m)$ ...(4.27.6)

- As shown in Fig. 4.27.2, $E_c \sin\theta(t)$ is a rotating vector. If "E_c" is rotating at a constant velocity "ω" then we could have written that $\theta(t) = \omega t$. But in FM this velocity is not constant. In fact it is changing continuously. The angular velocity of FM wave is given as,

$$\omega = \omega_c [1 + k_f E_m \cos\omega_m t] \qquad ...(4.27.7)$$

- Hence to find "θ (t)" we must integrate "ω" with respect to time.

$$\therefore \theta(t) = \int \omega \, dt = \int \omega_c [1 + k_f E_m \cos\omega_m t] \, dt \qquad ...(4.27.8)$$

$$\therefore \theta(t) = \omega_c \int [1 + k_f E_m \cos\omega_m t] \, dt$$

$$= \omega_c \left[t + \frac{k_f E_m \sin\omega_m t}{\omega_m} \right] = \omega_c t + \frac{k_f E_m \omega_c \sin\omega_m t}{\omega_m}$$

$$\therefore \theta(t) = \omega_c t + \frac{k_f E_m f_c \sin\omega_m t}{f_m} \qquad ...(4.27.9)$$

- As per the definition, $\delta = k_f E_m f_c$

$$\therefore \quad \theta(t) = \omega_c t + \frac{\delta \sin\omega_m t}{f_m} \qquad ...(4.27.10)$$

- Substitute this value of θ (t) in Equation (4.27.5) to get the equation for the FM wave as,

$$e_{FM} = s(t) = E_c \sin\left[\omega_c t + \frac{\delta}{f_m} \sin\omega_m t\right] \qquad ...(4.27.11)$$

- But $\frac{\delta}{f_m} = m_f$ i.e. the modulation index of FM wave.

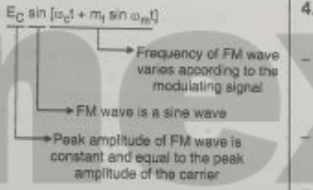
Hence the equation for FM wave is given as,

$$e_{FM} = E_c \sin\{\omega_c t + m_f \sin\omega_m t\} \qquad ...(4.27.12)$$

- This is the expression for a FM wave, where m_f represents the modulation index.

Meaning of mathematical representation :

- The mathematical expression for a FM wave is as follows :



- The amplitude of FM wave is constant and equal to the amplitude of the carrier i.e. E_c.
- FM wave is sinusoidal i.e. it has a shape of sine or cosine wave.
- The frequency of FM wave is not constant. It varies continuously, above and below the carrier frequency f_c.

4.27.4 Modulation Index of FM :

- The modulation index of an FM wave is defined as :

$$m_f = \frac{\text{Maximum frequency deviation}}{\text{Modulating frequency}} \qquad ...(4.27.13)$$

$$\therefore \quad m_f = \frac{\delta}{f_m} \qquad ...(4.27.14)$$

- The modulation index (m_f) is very important in FM because it decides the bandwidth of the FM wave.
- The modulation index also decides the number of sidebands having significant amplitudes.
- In AM the maximum value of the modulation index m is 1. But for FM the modulation index can be greater than 1. The modulation index m_f is measured in radians.

4.27.5 Deviation Ratio :

In FM broadcasting the maximum value of deviation is limited to 75 kHz. The maximum modulating frequency is also limited to 15 kHz. The modulation index corresponding to the maximum deviation and maximum modulating frequency is called as the "deviation ratio".

$$\text{Deviation ratio} = \frac{\text{Maximum deviation}}{\text{Maximum modulating frequency}} \qquad ...(4.27.15)$$

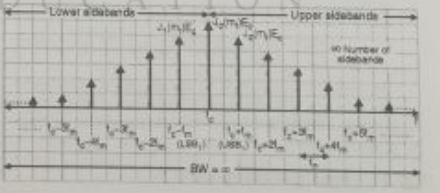
4.27.6 Percentage Modulation of FM Wave :

The percent modulation is defined as the ratio of the actual frequency deviation produced by the modulating signal to the maximum allowable frequency deviation.

$$\therefore \quad \% \text{ Modulation} = \frac{\text{Actual frequency deviation}}{\text{Maximum allowed deviation}} \qquad ...(4.27.16)$$

4.27.7 Frequency Spectrum of FM Wave (Frequency Domain Representation) :

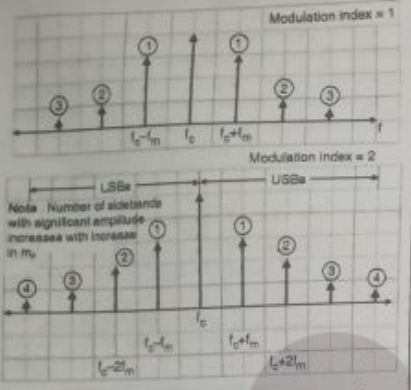
- Frequency domain representation of FM wave is a graph of amplitude plotted on y axis versus the frequency plotted on the x axis.
- To represent the FM wave in the frequency domain, consider the equation of FM wave again.
- The expression for the FM wave is not simple. It is complex since it is sine of sine function. The only way to solve this equation is by using the Bessel functions.



(D-713) **Fig. 4.27.3 : Ideal frequency spectrum of FM wave**

4.27.8 Effect of Modulation Index on the Frequency Spectrum of FM :

- As the amplitude of modulating signal varies, the frequency deviation will change. The number of sidebands produced and their amplitudes will change.



(D-167) **Fig. 4.27.4 : Effect of modulation index on the significant number of sidebands**

- Fig. 4.27.4 illustrates the effect of modulation index on the frequency spectrum of FM.
- Higher the value of m_f, more will be the number of sidebands having significant amplitudes as shown in Fig. 4.27.4.

4.27.9 Bandwidth Requirement of FM :

- Bandwidth of an FM wave is defined as the frequency difference between the highest pair of sidebands.
- Ideally the bandwidth of FM is infinite, because its spectrum consists of infinite number of upper and lower sidebands.
- But practically the bandwidth depends on the number of significant sidebands.
- The number of sidebands having significant amplitudes will increase with increase in the value of modulation index m_f. Hence the bandwidth increases with increase in the value of m_f.

4.27.10 Practical Bandwidth :

- Theoretically the bandwidth of the FM wave is infinite. But practically it is calculated based on how many sidebands have significant amplitude.
- The simplest method to calculate the bandwidth is as follows :

$$BW = 2 f_m \times \text{Number of significant sidebands} \quad ...(4.27.17)$$

- With increase in modulation index, the number of significant sidebands increase. This will increase the bandwidth. The bandwidth of FM is higher than that of AM.

Carson's Rule :

- The second method to find the practical bandwidth is a rule of thumb (Carson's rule). It states that the bandwidth of FM wave is equal to twice the sum of the deviation and the highest modulating frequency.

$$BW = 2 [\delta + f_{m(max)}] \quad ...(4.27.18)$$

- The Carson's rule gives correct results if the modulation index is greater than 6.

Ex. 4.27.1 : Define the term "percent modulation" and determine the percent modulation for an FM wave with a frequency deviation of 10 kHz if the maximum deviation allowed is 25 kHz.

Soln. :

The percent modulation is defined as the ratio of the actual frequency deviation produced by the modulating signal to the maximum allowable frequency deviation.

$$\therefore \quad \% \text{ Modulation} = \frac{\text{Actual frequency deviation}}{\text{Maximum allowed deviation}} \quad ...(1)$$

In the example it is given that $\delta = 10$ kHz and $\Delta f_{max} = 25$ kHz

$$\therefore \quad \% \text{ Modulation} = \frac{10 \text{ kHz}}{25 \text{ kHz}} = 40\%$$

Ex. 4.27.2 : In an F.M. system, if the maximum value of deviation is 75 kHz and the maximum modulating frequency is 10 kHz calculate the deviation ratio and bandwidth of the system using Carson's rule.

Soln. :

Given : $\delta_{max} = 75$ kHz, $f_{m(max)} = 10$ kHz

1. Deviation ratio $D = \frac{\delta_{max}}{f_{m(max)}} = \frac{75 \text{ kHz}}{10 \text{ kHz}} = 7.5$ **...Ans.**
2. System bandwidth $B = 2 [\delta_{max} + f_{m(max)}] = 2 (75 + 10) = 170$ kHz **...Ans.**

4.27.11 Generation of FM Wave :

- Fig. 4.27.5 shows the block schematic of a simple FM modulator. It is a simple Voltage Controlled Oscillator (VCO). The output frequency of voltage controlled oscillator is proportional to the control voltage applied to it.
- The modulating signal is applied at its input. This signal acts as the control voltage. The VCO output frequency varies in proportion with the modulating signal instantaneous value. Thus we get an FM wave at the output of the VCO.

(L-782) **Fig. 4.27.5 : Generation of FM**

4.28 Advantages, Disadvantages and Applications of FM :

4.28.1 Advantages of FM :

1. Improved noise immunity.
2. Low power is required to be transmitted to obtain the same quality of received signal at the receiver.
3. F.M. transmission covers larger area with the same amount of transmitted power.
4. Transmitted power remains constant.
5. All the transmitted power is useful.

4.28.2 Disadvantages of FM :

1. Very large bandwidth is required.
2. Since the space wave propagation is used, the radius of transmission is limited by the line of sight.
3. FM transmission and reception equipments are complex.

4.28.3 Applications of FM :

Some of the applications of FM are :

1. Radio broadcasting (Vividh Bharti, Radio Mirchi).
2. Sound broadcasting in T.V.
3. Satellite communication.
4. Police wireless.
5. Point to point communication.

Ex. 4.28.1 : What is the bandwidth required for FM in which the modulating frequency is 2 kHz and maximum deviation is 10 kHz. Assume highest needed sidebands are 8.

Soln. :

$f_m = 2$ kHz, $\delta = 10$ kHz

Method I : Bandwidth = $2 f_m \times$ Number of significant sidebands

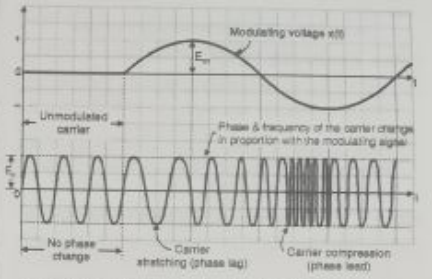
$= 2 \times 2 \text{ kHz} \times 8 = 32$ kHz

Method II : Bandwidth = $2 [\delta + f_{m(max)}] = 2 [10 + 2]$

$= 24$ kHz

4.29 Phase Modulation (PM) :

- Phase modulation is very similar to the frequency modulation. The only difference is that the phase of the carrier is varied instead of varying the frequency. The amplitude of the carrier remains constant.
- As shown in Fig. 4.29.1, as the modulating signal goes positive, the amount of phase lag increases with the amplitude of the modulating signal. The effect of this is that the carrier signal is stretched or its frequency is reduced.
- When the modulating signal goes negative, the phase shift becomes leading. This causes the carrier wave to be effectively compressed. The effect of this is as if the carrier frequency is increased.
- Thus phase modulation is always associated with frequency modulation and vice versa.
- Note that the P.M. wave of Fig. 4.29.1 is the same as the F.M. wave produced by dx(t) / dt i.e. the derivative of x (t) with respect to time.



(L-45) **Fig. 4.29.1 : Time domain display of PM wave**

- So in Fig. 4.29.1 we have plotted the derivative of x (t) which is original x (t) shifted by 90°.
- From the discussion it is clear that the difference between F.M. and P.M. waves can be made only by comparing with the original modulating wave.

4.29.1 Mathematical Representation of Phase Modulation (PM) :

- The phase modulation is another type of angle modulation. PM and FM are closely related. It is possible to obtain FM from PM, using the method called "Armstrong method".
- The PM wave is obtained by varying the phase angle ϕ of a carrier in proportion with the amplitude of the modulating voltage.

If the carrier voltage is expressed as,

$$e_c = A \sin (\omega_c t + \phi) \quad ...(4.29.1)$$

- Then the PM wave can be expressed as,

$$e_{PM} = A \sin (\omega_c t + \phi_m \sin \omega_m t) \quad ...(4.29.2)$$

- Here ϕ_m = Maximum phase change corresponding to the maximum amplitude of the modulating signal. For the sake of uniformity let us modify the Equation (4.29.2) as,

$$e_{PM} = A \sin [\omega_c t + m_p \sin \omega_m t] \quad ...(4.29.3)$$

Where $m_p = \phi_m$ = Modulation index of PM.

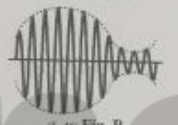
- The FM and PM waves look identical when their modulation index are identical. However if we change the modulating frequency f_m then m_f will change but there is no change in the value of m_p.

4.29.2 Generation of PM :

- Fig. 4.29.2 shows the scheme for the generation of PM wave. The modulating signal x (t) is applied to a differentiator.
- Then the differentiated modulating signal dx (t)/dt is applied to the VCO to produce the PM wave.

Sr. No.	FM	AM
13.	FM wave : **Fig. A**	AM wave : **Fig. B**
14.	**Applications** : Radio, TV broadcasting, police wireless, point to point communications.	**Applications** : Radio and TV broadcasting.

(L-46) **Fig. A**



(L-46) **Fig. B**

Review Questions

Q. 1 What is line coding ? Why line codes are essential ?

Q. 2 What are the disadvantages of RZ codes ?

Q. 3 State the various requirements of a line code.

Q. 4 Why is serial transmission preferred over the parallel transmission ?

Q. 5 Why is the speed of asynchronous transmission is low ?

Q. 6 What is the function of start and stop bits in asynchronous ?

Q. 7 In _____ type transmission, the bits are transmitted simultaneously.

Ans. : Parallel

Q. 8 In the asynchronous transmission, the time gap between successive bytes is ____.

Ans. : Variable

Q. 9 Write a short note on : Manchester coding.

Q. 10 Explain the term parallel transmission.

Q. 11 State the advantages, disadvantages and applications of parallel transmission.

Q. 12 Explain the serial transmission mode.

Q. 13 Explain the asynchronous transmission.

Q. 14 State advantages, disadvantages and application of asynchronous transmission.

Q. 15 Explain synchronous transmission.

Q. 16 Compare synchronous and asynchronous transmission.

Q. 17 Define pulse modulation. Give the types of pulse modulation.

Q. 18 Define Nyquist rate and Nyquist interval ?

Q. 19 What is quantizing noise ?

Q. 20 State the applications of PCM signals ?

Q. 21 Explain the slope overload distortion. How can it be minimized ?

Q. 22 What is granular noise ?

Q. 23 How is the "information" transmitted in a PCM system ?

Q. 24 What is quantization ?

Q. 25 What is quantization error ? What is its maximum value ?

Q. 26 How to reduce the quantization error ?

Q. 27 State and explain sampling theorem.

Q. 28 Draw and explain the block diagram for generation of PCM signal.

Q. 29 Represent ASK mathematically.

Q. 30 State the bandwidth requirement of ASK system.

Q. 31 What is the maximum B.W. of BPSK system ?

Q. 32 Draw the BPSK signal for the following binary signal.
1 0 1 1 1 0 1 0

Q. 33 Express QPSK mathematically.

Q. 34 How many phases are transmitted in QPSK ?

Q. 35 What are the advantages of QPSK system ?

Q. 36 What is the type of demodulation used for QPSK ?

Q. 37 State the expression for BFSK.

Q. 38 How is a message transmitted in BFSK ?

Q. 39 What is the BW of BFSK ?

Q. 40 What is the BW requirement of QAM ?

Q. 41 Why is QPSK superior to BPSK ?

Q. 42 State merits and demerits of BASK.

Q. 43 What type of receiver is used for the BPSK detection ?

Q. 44 Compare ASK and FSK.

Q. 45 Draw the waveforms for FSK and PSK modulation.

Q. 46 Explain the QPSK modulation scheme with constellation diagram.

Q. 47 What is ASK ? Draw its waveform ?

(L-784) **Fig. 4.29.2 : Generation of PM wave**

4.29.3 Bandwidth of PM :

- The formula for bandwidth of PM is same as that for FM. But the actual bandwidth of PM is less than that for FM.
- The bandwidth of a PM signal can be calculated from the maximum modulating frequency and the maximum amplitude of the modulating signal.

4.29.4 Comparison of FM and PM Systems :

Sr. No.	FM	PM
1.	$s(t) = E_c \sin [\omega_c t + m_f \sin \omega_m t]$	$s(t) = E_c \sin [\omega_c t + m_p \sin \omega_m t]$
2.	Frequency deviation is proportional to modulating voltage.	Phase deviation is proportional to the modulating voltage.
3.	Associated with the change in f_c there is some phase change.	Associated with the changes in phase there is some change in f_c
4.	m_f is proportional to the modulating voltage as well as the modulating frequency f_m.	m_p is proportional only to the modulating voltage.
5.	It is possible to receive FM on a PM receiver.	It is possible to receive PM on a FM receiver.
6.	Noise immunity is better than AM and PM.	Noise immunity is better than AM but worse than FM.
7.	Amplitude of the FM wave is constant.	Amplitude of the PM wave is constant.
8.	Signal to noise ratio is better than that of PM.	Signal to noise ratio is inferior to that in FM.
9.	FM is widely used.	PM is used in some mobile systems.
10.	In FM the frequency deviation is proportional to the modulating voltage only.	In PM the frequency deviation is proportional to both the modulating voltage and modulating frequency.

4.29.5 Comparison of FM and AM Systems :

Sr. No.	FM	AM
1.	Amplitude of FM wave is constant. It is independent of the modulation index.	Amplitude of AM wave will change with the modulating voltage.
2.	Hence transmitted power remains constant. It is independent of m_f.	Transmitted power is dependent on the modulation index.
3.	All the transmitted power is useful.	Carrier power and one sideband power are useless.
4.	FM receivers are immune to noise.	AM receivers are not immune to noise.
5.	It is possible to decrease noise further by increasing deviation.	This feature is absent in AM.
6.	Bandwidth = 2 [$\delta + f_m$]. The bandwidth depends on modulation index.	BW = $2f_m$. It is not dependent on the modulation index.
7.	BW is large. Hence wide channel is required.	BW is much less than FM.
8.	Space wave is used for propagation. So radius of transmission is limited to line of sight.	Ground wave and sky wave propagation is used. Therefore larger area is covered than FM
9.	Hence it is possible to operate several transmitters on same frequency.	Not possible to operate more channels on the same frequency.
10.	FM transmission and reception equipment are more complex.	AM equipments are less complex.
11.	The number of sidebands having significant amplitudes depends on modulation index m_f.	Number of sidebands in AM will be constant and equal to 2.
12.	The information is contained in the frequency variation of the carrier.	The information is contained in the amplitude variation of the carrier.

Q. 48 Draw the block diagram of binary PSK system and explain with signal space diagram.

Q. 49 Write an expression for the BFSK and explain the spectrum of BFSK.

Q. 50 Draw the BFSK waveform to represent the following bit stream.

00101110.

Q. 51 Explain clearly the difference between phase modulation and frequency modulation.

Q. 52 Explain the direct method of FM generation (reactance modulator).

Q. 53 Justify FM is called a constant B.W. system.

Q. 54 Compare and contrast : Frequency modulation and phase modulation.

Q. 55 Define FM and draw the necessary waveforms to explain it.

Q. 56 Derive an equation for FM wave.

Q. 57 Compare AM and FM.

Q. 58 Explain the generation of FM wave.

Q. 59 Write short note on : Frequency spectrum of FM wave.

Q. 60 Compare AM with FM with special reference to power requirements signed to noise ratio and bandwidth required.

Q. 61 Derive the formula for instantaneous value of an FM voltage and define modulation index.

Q. 62 What is angle modulation ?

□□□

CHAPTER 5

Unit II

Multiplexing

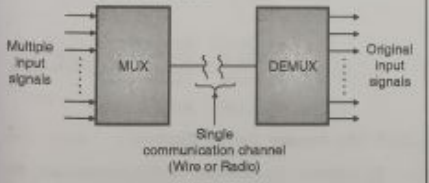
Syllabus :
Multiplexing, Frequency division multiplexing, Wavelength division multiplexing, Time division multiplexing.

5.1 Introduction to Multiplexing :

- Multiplexing is the process of simultaneously transmitting two or more individual signals over a single communication channel.
- Due to multiplexing it is possible to increase the number of communication channels so that more information can be transmitted.
- The typical applications of multiplexing are in telemetry and telephony or in the satellite communication.

5.2 Concept of Multiplexing and Demultiplexing :

- The concept of a simple multiplexer is illustrated in Fig. 5.2.1.
- The multiplexer receives a large number of different input signals.
- Multiplexer has only one output which is connected to the single communication channel.
- The multiplexer combines all input signals into a single composite signal and transmits it over the communication medium.

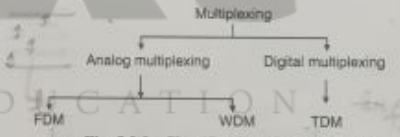


(L-105) **Fig. 5.2.1 : Concept of multiplexing**

- Sometimes the composite signal is used for modulating a carrier before transmission.
- At the receiving end, of communication link, a demultiplexer is used to separate out the signals into their original form.
- The operation of demultiplexer is exactly opposite to that of a multiplexer. Demultiplexing is the process which is exactly opposite to that of multiplexing.

5.2.1 Types of Multiplexing :

- There are three basic types of multiplexing. They are :
  1. Frequency Division Multiplexing (FDM).
  2. Time Division Multiplexing (TDM).
  3. Wavelength Division Multiplexing (WDM).
- The multiplexing techniques can be broadly classified into two categories namely analog and digital.
- Analog multiplexing can be either FDM or WDM and digital multiplexing is TDM.
- Fig. 5.2.2 shows the classification of multiplexing techniques.



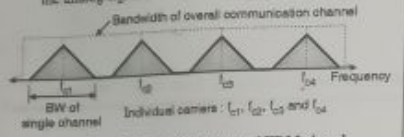
(L-106) **Fig. 5.2.2 : Classification of multiplexing techniques**

- Generally the FDM and WDM systems are used to deal with the analog information whereas the TDM systems are used to handle the digital information.
- In FDM many signals are transmitted simultaneously where each signal occupies a different frequency slot within a common bandwidth.
- In TDM the signals are not transmitted at a time, instead they are transmitted in different time slots.

5.3 Frequency Division Multiplexing (FDM) :

- The operation of FDM is based on sharing the available bandwidth of a communication channel among the signals to be transmitted.
- That means many signals are transmitted simultaneously with each signal occupying a different frequency slot within the total available bandwidth.
- Each signal to be transmitted modulates a different carrier. The modulation can be AM, SSB, FM or PM.

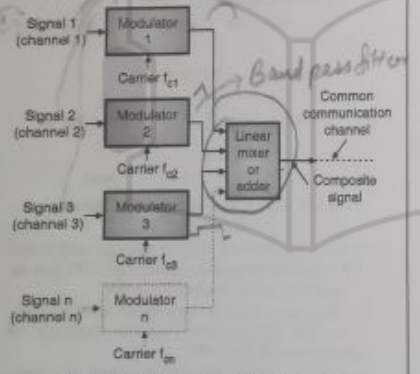
- The modulated signals are then added together to form a composite signal which is transmitted over a single channel.
- The spectrum of composite FDM signal is shown in Fig. 5.3.1(a).
- Generally the FDM systems are used for multiplexing the analog signals.



(L-107) **Fig. 5.3.1(a) : Spectrum of FDM signal**

5.3.1 FDM Transmitter (Multiplexing Process) :

- Fig. 5.3.1(b) shows the block diagram of an FDM transmitter. The signals which are to be multiplexed will each modulate a separate carrier.
- The type of modulation can be AM, SSB, FM or PM.
- The modulated signals are then added together to form a complex signal which is transmitted over a single channel.



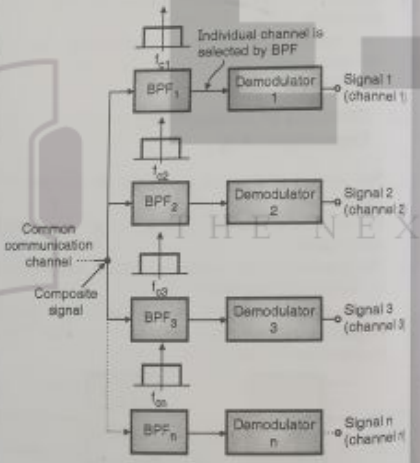
(L-108) **Fig. 5.3.1(b) : The FDM transmitter**

Operation of the FDM transmitter :

- Each signal modulates a separate carrier. The modulator outputs will contain the sidebands of the corresponding signals.
- The modulator outputs are added together in a linear mixer or adder. The linear mixer is different from the normal mixers. Here the sum and difference frequency components are not produced. But only the algebraic addition of the modulated outputs will take place.
- Different signals are thus added together in the time domain but they have their own separate identity in the frequency domain. This is as shown in the Fig. 5.3.1(a).
- The composite signal at the output of mixer is transmitted over the single communication channel as shown in Fig. 5.3.1(b). This signal can be used to modulate a radio transmitter if the FDM signal is to be transmitted through air.

5.3.2 FDM Receiver (Demultiplexing Process) :

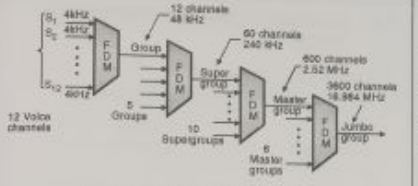
- The block diagram of an FDM receiver is as shown in Fig. 5.3.1(c). The composite signal is applied to a group of Band Pass Filters (BPF).
- Each BPF has a center frequency corresponding to one of the carriers used in the transmitter i.e. f_{c1}, f_{c2}, ... f_{cn} etc.
- The BPFs have an adequate bandwidth to pass all the channel information without any distortion.
- Each filter will pass through only its channel and reject all the other channels. Thus all the multiplexed channels are separated out.
- The channel demodulator then removes the carrier and recovers the original signal back.



(L-109) **Fig. 5.3.1(c) : FDM receiver**

5.3.3 The Analog Carrier System :

- To maximize the efficiency of their infrastructure, the telephone companies have used multiplexing techniques for lower bandwidth lines.
- In this way it is possible to combine many switched or leased lines into fewer but bigger channels.
- One of such hierarchical system is used by AT and T. It is as shown in Fig. 5.3.2 and is made up of groups, super groups, master groups and jumbo groups.



(L-110) **Fig. 5.3.2 : FDM hierarchy**

The levels of multiplexing is also called as multiplexing hierarchy.

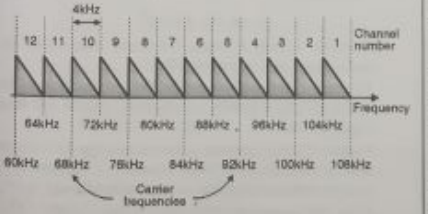
- The different levels of multiplexing which is also called multiplexing hierarchy is as follows :

Level (1) : Basic Group. [12 voice channels multiplexed together].
↓
Level (2) : Super Group. [Upto 5 basic groups multiplexed together
↓ i.e. upto 60 voice channels].
Level (3) : Master Group. [Upto 10 super groups multiplexed together
↓ i.e. upto 600 voice channels].
Level (4) : Jumbo Group. [Upto 6 master groups multiplexed together
i.e upto 3600 voice channels].

- This hierarchy is used by AT and T and shown in Fig. 5.3.2.

Basic Group [12 voice channels] :

- The frequency plan for the typical basic group is as shown in Fig. 5.3.3. Here the 12 voice channels such as telephone channels modulate the carrier frequencies in the range of 60 to 108 kHz range. The carrier frequencies are spaced at 4 kHz from each other.



(L-111) **Fig. 5.3.3 : Frequency plan for the basic group of FDM**

- SSB modulation technique is used to save the bandwidth. Each voice channel is applied to a balanced modulator along with a carrier. The output of a balanced modulator consists of the upper and lower sidebands.
- Frequency plans of groups of FDM are nothing but the frequency spectrums.
- The frequency plan for the basic group of FDM is shown in Fig. 5.3.3.

Super group :

The frequency plan for a super group is as shown in Fig. 5.3.4. A super group consists of at the most 60 voice channels.

(L-112) **Fig. 5.3.4 : Frequency plan for a super group of FDM**

5.4 Advantages, Disadvantages and Applications of FDM :

5.4.1 Advantages of FDM :

1. A large number of signals (channels) can be transmitted simultaneously.
2. FDM does not need synchronization between its transmitter and receiver for proper operation.
3. Demodulation of FDM is easy.
4. Due to slow narrow band fading only a single channel gets affected.

5.4.2 Disadvantages of FDM :

1. The communication channel must have a very large bandwidth.
2. Intermodulation distortion takes place.
3. Large number of modulators and filters are required.
4. FDM suffers from the problem of crosstalk.
5. All the FDM channels get affected due to wideband fading.

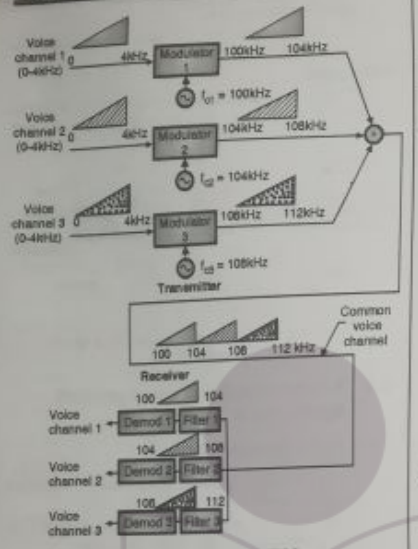
5.4.3 Applications of FDM :

Some of the important applications of FDM are :

1. Telephone systems.
2. AM (Amplitude Modulation) and FM (Frequency Modulation) radio broadcasting.
3. TV broadcasting.
4. First generation of cellular phones used FDM.

Ex. 5.4.1 : Draw the FDM system to combine three voice channels. Each voice channel occupies a bandwidth of 4 kHz. The common voice channel has a bandwidth of 12 kHz from 100 kHz to 112 kHz.

Soln. : Fig. P. 5.4.1 shows the required FDM system.

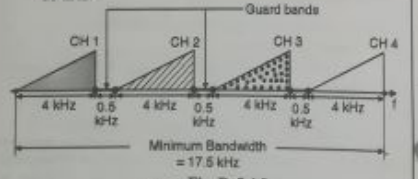


(L-117) **Fig. P. 5.4.1 : Required FDM system**

Ex. 5.4.2 : 4 voice channels each having a bandwidth of 4 kHz are to be multiplexed using FDM. A guardband of 500 Hz is to be inserted between the adjacent channels. Calculate the minimum bandwidth of the link.

Soln. :

- The frequency spectrum of the FDM signal is shown in Fig. P. 5.4.2.
- The minimum bandwidth is equal to 17.5 kHz as shown in Fig. P. 5.4.2.
- The bandwidth without guardbands would have been 16 kHz.



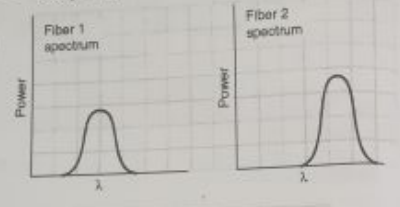
(L-118) **Fig. P. 5.4.2**

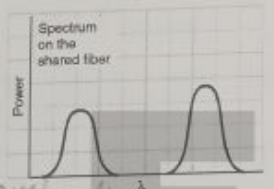
Conclusion :

Guardbands increase the bandwidth of FDM signal still they should be included in order to avoid interference between the adjacent channels.

5.5 Wavelength Division Multiplexing (WDM) :

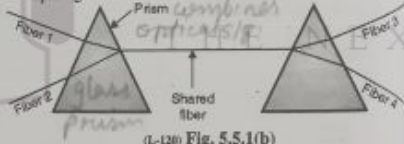
- WDM is the variation of FDM.
- It is specially used for fiber optic channels.





(L-119) **Fig. 5.5.1(a)**

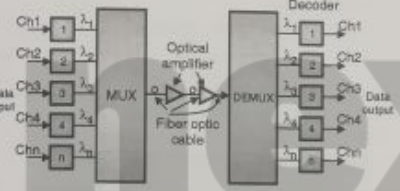
- As shown in Figs. 5.5.1(a) and (b), 2 fibres com together at a prism, each having energy in a differe frequency band. After passing through the prism beams are combined onto a single shared fiber, fo transmission to a distant destination, where they a split again.



(L-120) **Fig. 5.5.1(b)**

- Channels having different frequency ranges can b multiplexed on a single long fiber.
- The only difference between WDM and electrical FDM is that an optical system is completely passive and tha highly reliable.
- Reason WDM is popular, is that the energy on a singl fiber is a few gigahertz wide because it is impossible t convert between electrical and optical media any faster
- Since BW of a single fiber band is about 25,000 GHz there is great potential for multiplexing many optica channels together over long routes. Necessary condition is that incoming channels use differen frequency bands.
- Potential application of WDM is in the FTTC (Fiber T The Curb) systems or in SONET networks.

- In the Fig. 5.5.1(b) we have a fixed wavelength system bits from fiber 1 go to fiber 3 and bits from fiber 2 go to fiber 4.
- It is not possible to have bits going from fiber 1 to fiber 4. It is also possible to build WDM systems that are switched, which contain many input and output fibers, switching data among themselves.
- Although spreading energy over n outputs dilutes it by a factor n, such systems are practical for hundred of channels.
- If light from one of the incoming fibers have to go to any output fiber, all the output fibers need tunnable filters.
- Alternatively, input fibers could be tunnable and output ones fixed. Having both to be tunnable is unnecessary expense.
- A simple block diagram of WDM transmitter and receiver system with different channels is as shown in Fig. 5.5.1(c).



(L-121) **Fig. 5.5.1(c) : WDM system**

5.5.1 Application of WDM :

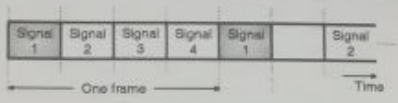
One important application of WDM is the SONET network in which a large number of optical fiber lines are multiplexed and demultiplexed.

5.5.2 DWDM (Dense WDM) :

- The long form of DWDM is dense WDM. It can multiplex a very large number of channels. The spacing between adjacent channels is small.
- Efficiency of DWDM is higher than that of WDM.

5.6 Synchronous Time Division Multiplexing :

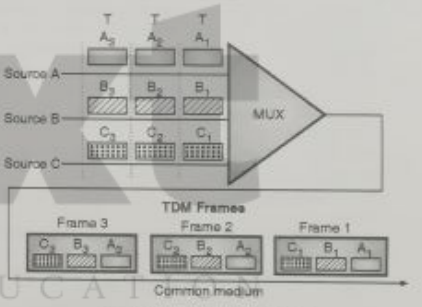
- The process called multiplexing is used in order to utilize common transmission channel or medium to transmit more than one signals simultaneously.
- TDM is a digital multiplexing process.
- In TDM all the signals to be transmitted are not transmitted simultaneously. Instead, they are transmitted one-by-one.
- Thus each signal will be transmitted for a very short time. One cycle or frame is said to be complete when all the signals are transmitted once on the transmission channel. The TDM principle is illustrated in Fig. 5.6.1.



(L-122) **Fig. 5.6.1 : Principle of TDM**

- As shown in the Fig. 5.6.1 one transmission of each channel completes one cycle of operation called as a "Frame".
- The TDM system can be used to multiplex analog or digital signals, however it is more suitable for the digital signal multiplexing.
- The concept of TDM will be more clear if you refer to Fig. 5.6.2.
- The data flow of each source (A, B or C) is divided into units (say A_1, A_2 or B_1, C_1 etc.)

  Then one unit from each source is taken and combined to form one frame. The size of each unit such as A_1, B_1 etc. can be 1 bit or several bits.



(L-123) **Fig. 5.6.2 : TDM system**

- Fig. 5.6.3 shows the frames of TDM signal. For 3 inputs being multiplexed, a frame of TDM will consist of 3 units i.e. one unit from each source.
- Similarly for n number of inputs, each TDM frame will consist of n units.

(L-124) **Fig. 5.6.3 : TDM frames**

- The TDM signal in the form of frames is transmitted on the common communication medium.

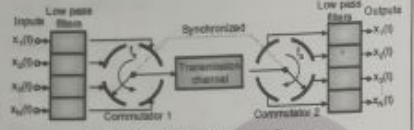
Data rate :

- For a TDM, the data rate of the multiplexed signal is always n times the data rate of individual sources, where n is the number of sources.

- So if three sources are being multiplexed, then the data rate of the TDM signal is three times higher than the individual data rate.
- Naturally the duration of every unit (A_1 or B_1 etc.) in TDM signal is n times shorter than the unit duration before multiplexing.

5.6.1 PAM - TDM System :

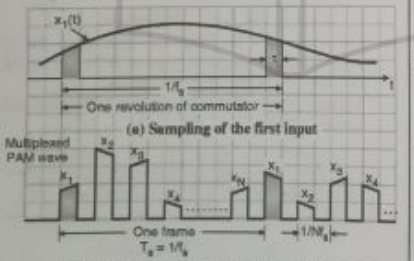
- The TDM system which is going to be discussed now, combines the concepts of PAM and TDM both. The TDM system is as shown in Fig. 5.6.4.



(L-125) **Fig. 5.6.4 : PAM/TDM system**

The operation of the system is as follows :

- The multiplexer here is a single pole rotating switch or commutator. It can be a mechanical switch or an electronic switch. It rotates at f_s rotations per second.
- As the switch arm rotates, it is going to make contact with the position 1, 2, 3 or N for a short time. To these contacts are connected the N analog signals which are to be multiplexed.
- Thus the switch arm will connect these N input signals one by one to the communication channel.
- The waveform of a TDM signal which is being transmitted is as shown in Fig. 5.6.5. It shows that the rotary switch samples each channel during each of its rotations. Each rotation corresponds to one frame. Hence 1 frame is completed in T_s seconds where $T_s = 1/f_s$.



(a) Sampling of the first input

(b) Multiplexed PAM signal, transmitted on the transmission channel

(L-126) **Fig. 5.6.5**

- At the receiver, there is one more rotating switch or commutator used for demultiplexing.
- It is important to note that this switch must rotate at the same speed as that of the commutator 1 at the transmitter and its position must be synchronized with commutator 1 in order to ensure proper demultiplexing.
- The same principle of multiplexing can be used for multiplexing more number of signals.

Interleaving :

- On the multiplexer side the commutator-1 opens in front of a connection, that connection has the opportunity to send its bit on to the channel.
- This process is called as interleaving.

Ex. 5.6.1 : 3 signals having a data rate of 2 kbps are grouped together by means of time division multiplexing. Each unit consists of 1 bit. Calculate :
1. The bit duration before multiplexing.
2. The transmission rate of TDM.
3. The duration of each time slot in TDM.
4. The duration of one TDM frame.

Soln. :

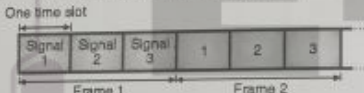
Step 1 : Duration of a bit before multiplexing :

- Each signal has a data rate of 2 kbps. That means 2000 bits per second.
- Hence the duration of each bit is,

$$T_b = \frac{1}{2000} = 0.5 \text{ mS} \quad \text{...Ans.}$$

Step 2 : Transmission rate of TDM :

- The TDM frame is shown in Fig. P. 5.6.1.



(L-127) **Fig. P. 5.6.1 : TDM frames**

- As discussed earlier the transmission rate of TDM is n times higher than the bit rate of each source.

Transmission rate of TDM = n × 2000
= 3 × 2000
= 6000 bps or 6 kbps ...Ans.

Step 3 : Duration of time slot in TDM :

Duration of each time slot in TDM = $\frac{1}{6000}$
= 166.67 ...Ans.

Step 4 : Frame duration :

Duration of 1 frame = n × duration of one slot
= 3 × 166.67 μS = 0.5 mS ...Ans.

Note : The duration of a TDM frame is always equal to the duration of one unit before multiplexing.

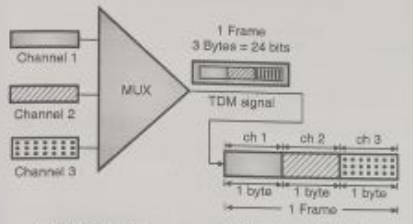
Ex. 5.6.2 : Three channels are to be multiplexed using TDM technique. The rate of each channel is 150 bytes per second. In TDM, one byte per channel is to be multiplexed.
Calculate :
1. Frame size
2. Frame duration
3. Frame rate and
4. Bit rate of the TDM signal.

Soln. :

Given : Bit rate of each channel 150 bytes per sec, n = 3. 1 byte per channel to be multiplexed.

Step 1 : Draw the system block diagram :

Fig. P. 5.6.2(a) shows the block diagram of the TDM system and Fig. P. 5.6.2(b) shows one TDM frame.



(a) TDM system (b) One frame of TDM

(L-128) **Fig. P. 5.6.2**

Step 2 : Frame size :

Each frame consists of one byte from each channel. So frame size is 3 bytes or 24 bits.

Step 3 : Frame duration and frame rate :

- The duration of a TDM frame is always equal to the duration of one unit before multiplexing.

Here one unit before multiplexing is 1 byte i.e. 8 bits.

∴ Frame duration = 1 byte duration

- But each channel transmits at 150 bytes/sec.

∴ Frame duration = $\frac{1}{150} = 6.666$ mS ...Ans.

Frame duration = $\frac{1}{\text{Frame duration}} = \frac{1}{6.666 \times 10^{-3}}$
= 150 frames/sec. ...Ans.

Step 4 : Bit rate of TDM signal :

Bit rate of TDM = Number of bits per frame × Number of frames per second.
= 24 × 150 = 3600 bps ...Ans.

5.6.2 Signaling Rate (r) :

The signaling rate of a TDM system is defined as the number of pulses transmitted per second. It is denoted by "r". Let us now derive the expression for the signaling rate of the PAM-TDM system.

- Let W = Maximum frequency of all the input signals x_1 to x_N.
- Therefore as per Nyquist criteria, the sampling frequency $f_s \geq 2$ W. Therefore the speed of rotation of the commutators is f_s rotations per second with $f_s \geq 2W$.
- As shown in Fig. 5.6.6, one revolution of commutators corresponding to one frame contains one sample from each input signal.

∴ 1 Revolution ⇒ 1 frame ⇒ N pulses ...(5.6.1)

- 1 frame period is $(1/f_s)$ i.e. T_s seconds. Therefore in "T_s" seconds "N" number of pulses are transmitted. Hence the pulse to pulse spacing within the frame is given by,

$$\text{Pulse to pulse spacing} = \frac{T_s}{N} = \frac{1}{Nf_s} \quad \text{...(5.6.2)}$$

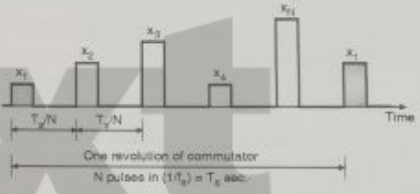
- As the period of one pulse (ON + OFF) is $(1/Nf_s)$ seconds, the number of pulses per second is given by,

Number of pulses per second = Nf_s

- This is nothing but the signaling rate.

∴ Signaling rate of a TDM system = r
= Nf_s pulses/second. But as $f_s \geq 2W$.
Signaling rate of a TDM system
= r ≥ 2NW pulses/second ...(5.6.3)

- A TDM system is supposed to have its signaling rate as high as possible. It is evident from the expressions above that the signaling rate can be increased by increasing the sampling rate f_s and/or the number of input signals N.



(L-129) **Fig. 5.6.6 : Calculation of number of pulses per second for PAM-TDM system**

5.6.3 Transmission Bandwidth of a TDM Channel :

- The minimum transmission bandwidth of a PAM-TDM channel is given by,

B_T = $\frac{1}{2}$ signaling rate

∴ Minimum transmission bandwidth

$$B_T \geq \frac{1}{2} \times 2NW$$

∴ Minimum transmission bandwidth

$$B_T = NW \quad \text{...(5.6.4)}$$

5.6.4 Empty Slots :

- The drawback of synchronous TDM is its low efficiency. This is because if a particular source does not have any data to send, then the time slot allotted to it in the output frame will be empty, as shown in Fig. 5.6.7.

(L-765) **Fig. 5.6.7 : Empty slot**

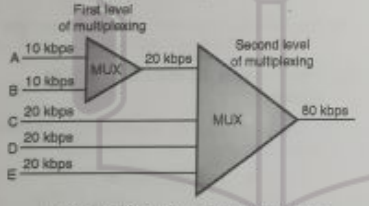
- In Fig. 5.6.7 the source B has one empty slot (due to discontinuous data).
- As shown in Fig. 5.6.7, the second and the third frame has three filled slots but the first frame has only two slots filled. Thus the first time slot is carrying less data than its maximum capacity.
- The statistical TDM system improves the efficiency by removing such empty slots.

5.6.5 Data Rate Management :

- The problem with synchronous TDM is it inability to handle the disparity in the input data rates. That means the data rates of different sources are not same but different.
- If the data rates of all the inputs are not same then three different strategies may be used. They are :
  1. Multilevel multiplexing.
  2. Multiple slot allocation.
  3. Pulse stuffing.
- We will discuss these strategies one by one.

1. Multilevel multiplexing :

- This technique is used when the data rate of an input line is exact multiple of data rates of other lines.
- For example in Fig. 5.6.8 sources A and B have data rates of 10 kbps while the remaining sources have a data rate of 20 kbps.

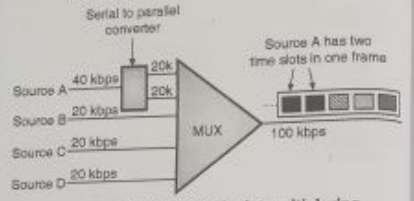


(L-766) **Fig. 5.6.8 : Multilevel multiplexing**

- As shown in Fig. 5.6.8, the input lines from sources A and B are multiplexed together to provide the data rate of 20 kbps (which is equal to the data rate of remaining three sources).
- These four inputs are given to second level of multiplexer which produces an output at 80 kbps.

2. Multiple slot allocation :

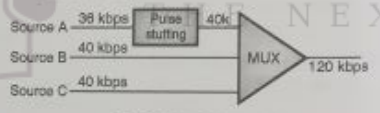
- Sometimes we can improve the efficiency by allotting more than one slot in a frame to a single source. This can be done for an input source which has a higher data rate as compared to the other input lines.
- For example refer Fig. 5.6.9 the source A has a data rate of 40 kbps and the data rate of the other sources is 20 kbps.
- So the source A is connected to a serial to parallel converter. This will give two slots for the source A.
- Thus there will be two slots for source A per frame and one slot each for sources B, C and D. The data rate of all these inputs will now be the same i.e. 20 kbps as shown in Fig. 5.6.9.



(L-767) **Fig. 5.6.9 : Multiple slot multiplexing**

3. Pulse stuffing :

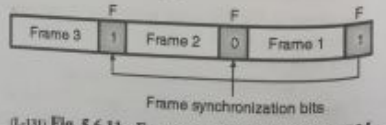
- If the bit rates of sources are not multiple integers of each other, then the two techniques discussed earlier are not useful.
- So another technique called pulse stuffing is used. Here the highest input data rate is made the dominant data rate and then dummy bits to the input lines having lower data rates.
- Pulse stuffing is also called as pulse stuffing, bit padding, or bit stuffing and it is demonstrated in Fig. 5.6.10.
- Note that the source A has a lower data rate of 36 kbps and so pulse stuffing is done to correct its rate to 40 kbps.



(L-768) **Fig. 5.6.10 : Pulse stuffing**

5.6.6 Frame Synchronization :

- The implementation of TDM is not as simple as that of FDM because in TDM, the synchronization of multiplxer and demultiplxer is essential.
- If the synchronization is not there then the bit that belongs to one channel may be received by some other channel.
- Therefore one or more synchronization bits are generally added to the beginning of each frame. They are known as the framing bits.



(L-131) **Fig. 5.6.11 : Frame synchronization in TDM**

- These bits are called frame synchronizing bits or simply framing bits.
- The framing bits will follow a pattern frame to frame. For example the pattern shown in Fig. 5.6.11 is 101.
- The framing bit pattern will allow the demux to synchronize itself to the mux.

5.6.7 Advantages of TDM :

1. Full available channel bandwidth can be utilized for each channel.
2. Intermodulation distortion is absent.
3. TDM circuitry is not very complex.
4. The problem of crosstalk is not severe.

5.6.8 Disadvantages of TDM :

1. Synchronization is essential for proper operation.
2. Due to slow narrowband fading, all the TDM channels may get wiped out.

5.6.9 Applications of TDM :

1. Multiplexing of digital signals.
2. Digital telephony.
3. Satellite communications.
4. Fiber optic communication.
5. Wireless communication applications.

5.6.10 Solved Examples on TDM / PAM :

Ex. 5.6.3 : Two analog signals m_1 (t) and m_2 (t) are to be transmitted over a common channel by means of time division multiplexing. The highest frequency of m_1 (t) is 4 kHz and that of m_2 (t) is 4.5 kHz. What is the minimum value of permissible sampling rate ?

Soln. :

The highest frequency component of the composite signal consisting of m_1 (t) and m_2 (t) is 4.5 kHz. Therefore the minimum value of permissible sampling rate is,

$$f_{s\,(min)} = 2 \times 4.5 \text{ kHz} = 9 \text{ kHz} \quad \text{...Ans.}$$

Ex. 5.6.4 : A signal x_1 (t) is bandlimited to 3 kHz. There are three more signals x_2 (t), x_3 (t) and x_4 (t) which are bandlimited to 1 kHz each. These signals are to be transmitted by a TDM system.

(a) Design a TDM scheme where each signal is sampled at its Nyquist rate.

(b) What must be the speed of the commutator ?

(c) Calculate the minimum transmission bandwidth of the channel.

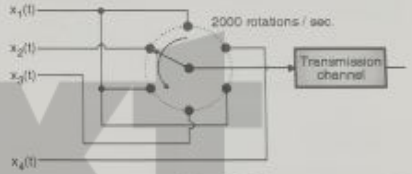
Soln. :

(a) Table P. 5.6.4 shows different message signal with corresponding Nyquist rates.

Table P. 5.6.4

Message signal	Bandwidth	Nyquist rate
x_1 (t)	3 kHz	6 kHz
x_2 (t)	1 kHz	2 kHz
x_3 (t)	1 kHz	2 kHz
x_4 (t)	1 kHz	2 kHz

- If the sampling commutator rotates at the rate of 2000 rotations per second then the signals x_2 (t), x_3 (t) and x_4 (t) will be sampled at their Nyquist rate. But we have to sample x_1 (t) also at its Nyquist rate which is three times higher than that of the other three.
- In order to achieve this we should sample x_1 (t) three times in one rotation of the commutator. Therefore the commutator must have atleast 6 poles connected to the signals as shown in Fig. P. 5.6.4.



(L-135) **Fig. P. 5.6.4**

(b) The speed of rotation of the commutator is 2000 rotations/sec.

(c) Number of samples produced per second is calculated as follows :

x_1 (t) produces $3 \times 2000 = 6000$ samples/sec.

x_2 (t), x_3 (t) and x_4 (t) produce 2000 samples/sec. each.

$\therefore$ Number of samples per second $= 6000 + (3 \times 2000)$

$= 12000$ samples/sec.

$\therefore$ Signaling rate $= 12000$ samples/sec.

(d) The minimum channel bandwidth is,

$$B_T = \frac{1}{2} \text{ signaling rate}$$

$$= 12000/2$$

$$\therefore \quad B_T = 6000 \text{ Hz} \quad \text{...Ans.}$$

Ex. 5.6.5 : Twenty four voice signals are sampled uniformly and then time division multiplexed. The sampling operation uses flat top samples with 1 μS duration. The multiplexing operation includes provision for synchronization by adding an extra pulse of appropriate amplitude and 1 μS duration. The highest frequency component of each voice signal is 3.4 kHz.

(a) Assuming a sampling rate of 8 kHz, calculate the spacing between successive pulses of the multiplexed signal.

(b) Repeat (a), assuming the use of Nyquist rate sampling.

Soln. :

(a) It has been given that :

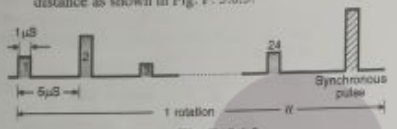
1. Sampling rate = 8 kHz = 8000 samples/sec.
2. There are 24 voice signals + 1 synchronizing pulse.
3. Pulse width of each voice channel and synchronizing pulse is 1 μS.

Time taken by the commutator for 1 rotation = $\frac{1}{8000}$

= 125 μsec.

Number of pulses produced in 1 rotation = 24 + 1 = 25

∴ The leading edges of the pulses are at $\frac{125}{25} = 5$ μsec. distance as shown in Fig. P. 5.6.5.



(L-134) **Fig. P. 5.6.5**

Therefore spacing between successive pulses = 5 – 1 = 4 μS ...Ans.

(b) Nyquist rate of sampling = 2 × 3.4 kHz = 6.8 kHz.

That means 6800 samples are produced per second.

One rotation of commutator takes

1/6800 = 147 μS time.

∴ 147 μsec corresponds to 25 pulses

∴ 1 pulse corresponds to 5.88 μsec.

As the pulse width of each pulse is 1 μ sec, the spacing between adjacent pulses will be 4.88 μsec. and if we assume τ = 0 then the spacing between the adjacent pulses will be 5.88 μ sec.

Ex. 5.6.6 : Six message signals each of bandwidth 5 kHz are time division multiplexed and transmitted. Calculate the signaling rate and the minimum channel bandwidth of the PAM/TDM channel.

Soln. :

The number of channels N = 6

Bandwidth of each channel, W = 5 kHz

Minimum sampling rate = 2 × 5 kHz = 10 kHz

Signaling rate = Number of bits per second

= 6 × 10 kHz

= 60 kbits/sec. ...Ans.

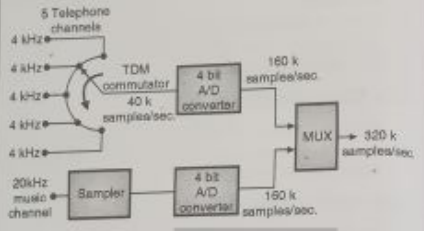
Minimum channel bandwidth to avoid cross talk in PAM/TDM is,

$B_T = NW = 6 \times 5$ kHz = 30 kHz ...Ans.

Ex. 5.6.7 : Sketch a channel interleaving scheme for time division multiplexing the following PAM signals : Five 4 kHz telephone channels and one 20 kHz music channel. Find the pulse repetition rate of the multiplexed signal and estimate the minimum system bandwidth required.

Soln. :

Each telephone channel of bandwidth 4 kHz must be sampled at Nyquist rate i.e. 2 × 4 kHz = 8 kHz using a TDM commutator. The 20 kHz music channel must be sampled at 40 kHz (Nyquist rate) hence a separate sampler is required. The sampled signals are applied to two 4-bit A-D converters to obtain the equivalent digital signals. These signals are finally multiplexed using a multiplexer as shown in Fig. P. 5.6.7.



(L-769) **Fig. P. 5.6.7 : PAM-TDM system for Ex. 5.6.7**

The TDM commutator output has a pulse repetition rate of 40k samples/sec as there are 5 channel and sampling rate is 8 kHz. Similarly the output of the separate sampler has a pulse repetition rate of 40 k samples/sec. The outputs of A-D converters have pulse repetition rates of 40 × 4 = 160 k samples/sec. Therefore pulse repetition rate at the output of a multiplexer is 160 + 160 = 320 k samples/sec.

∴ Pulse repetition rate of the system = 320 kHz

∴ Bandwidth required = Bit rate = 320 kHz ....Ans.

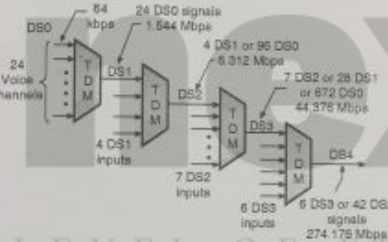
5.7 Comparison of FDM and TDM Systems :

Sr. No.	FDM	TDM
1.	The signals which are to be multiplexed are added in the time domain. But they occupy different slots in the frequency domain.	The signals which are to be multiplexed can occupy the entire bandwidth but they are isolated in the time domain.
2.	FDM is usually preferred for the analog signals.	TDM is preferred for the digital signals.
3.	Synchronization is not required.	Synchronization is required.
4.	The FDM requires a complex circuitry at the transmitter and receiver.	TDM circuitry is not very complex.
5.	FDM suffers from the problem of crosstalk due to imperfect band pass filters.	In TDM the problem of crosstalk is not severe.

Sr. No.	FDM	TDM
6.	Due to wideband fading in the transmission medium, all the FDM channels are affected.	Due to fading only a few TDM channels will be affected.
7.	Due to slow narrowband fading taking place in the transmission channel only a single channel may be affected in FDM.	Due to slow narrowband fading all the TDM channels may get wiped out.

5.8 Digital Signal (DS) Service :

- The telephone companies implement TDM (time division multiplexing) through the hierarchy of digital signals. This is called as digital signal (DS) service or digital hierarchy.
- Fig. 5.8.1 shows the DS hierarchy and the bit rates corresponding to various levels.



(G-1311) **Fig. 5.8.1 : D.S. hierarchy**

Explanation :

- A DS0 signal is the basic input signal which is a single digital channel (usually 64 kbps PCM channel).
- 24 DS0 signals are multiplexed using TDM to produce a DS1 signal. The bit rate of DS1 is 24 × 64 kbps = 1.544 Mbps plus 8 kbps of overhead.
- 4 such DS1 signals are multiplexed at the second level of multiplexing to obtain the DS2 signal.
- One DS2 signal is equivalent to 4 DS1 or 96 DS0 signals and it has a bit rate of 6.312 Mbps.
- 7 DS2 signals are multiplexed to produce a DS3 signal. Its bit rate is 44.376 Mbps and it is equivalent to 7 DS2 or 28 DS1 or 672 DS0 signals.
- Finally 6 DS3 lines are multiplexed to obtain a DS4 signal. Its bit rate is 274.176 Mbps. It is obtained as a result of 6 DS3 or 42 DS2 channels, 168 DS1 channels and 4032 DS0 channels.

5.8.1 T Lines :

- DS0, DS1, DS2, ... etc. are the names of the services. The telephone companies use the T lines (T0, T1, T2, ... etc.) to implement these services.
- The T lines have capacities which precisely match with the bit rates of the corresponding services as shown in Table 5.8.1.

Table 5.8.1 : Relation between DS and T lines

Service	Line	Rate (Mbps)	Number of voice channels
DS – 1	T – 1	1.544	24
DS – 2	T – 2	6.312	96
DS – 3	T – 3	44.736	672
DS – 4	T – 4	274.176	4032

- Thus T – 1 line implements DS – 1 service, T – 2 implements DS – 2 service and so on.
- DS0 is defined as the basic service.

Note : T lines are digital lines which are designed to carrying digital data, audio or video.

- But the T lines can also be used for analog communication. For example T1 line can be used for the telephone applications.

5.9 T Lines for Analog Transmission :

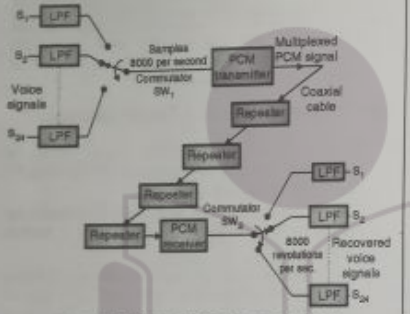
- When a large number of PCM signals are to be transmitted over a common channel, multiplexing of these PCM signals is required.
- Fig. 5.9.1 shows the basic time division multiplexing scheme for PCM voice channels called as the **T_1 digital system.**
- This system is used to convey a number of voice signals over telephone lines using wideband coaxial cable. Thus the communication medium used is a coaxial cable.

Operation of the T_1 system :

The operation of the PCM-TDM system shown in Fig. 5.9.1 is as follows :

- This system has been designed to multiplex 24 voice channels marked as S_1 to S_{24}. Each signal is bandlimited to 3.3 kHz, and the sampling is done at a standard rate of 8 kHz. This sampling rate is higher than the Nyquist rate. The sampling is done by the commutator switch SW_1.
- These voice signals are selected one by one and connected to a PCM transmitter by the commutator switch SW_1, as it completes its rotation. The commutator switch remains in contact with each voice channel for a short time. Thus it samples each of the 24 channels.
- Each sampled signal is then applied to the PCM transmitter which converts it into a digital signal by the process of A to D conversion and companding. Each sampled voice signal is converted into an 8-bit PCM word.
- The resulting digital waveform is transmitted over a co-axial cable. This waveform is called as the PCM-TDM signal.

- Periodically, after every 6000 ft., the PCM-TDM signal is regenerated by amplifiers called "Repeaters". They eliminate the distortion introduced by the channel and remove the superimposed noise and regenerate a clean noise free PCM-TDM signal at their output. This ensures that the received signal is free from the distortions and noise.
- At the destination the signal is companded, decoded and demultiplexed, using a PCM receiver. The PCM receiver output is connected to different low pass filters via the commutator switch SW_2. The LPF outputs are applied to the destination receivers (subscribers).
- Synchronization between the transmitter and receiver commutators SW_1 and SW_2 is essential in order to ensure proper communication.



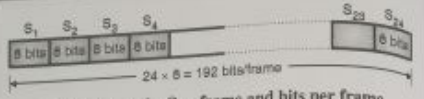
(G-1312) **Fig. 5.9.1 : Block diagram of a basic PCM-TDM system**

Bits/Frame :

- The commutators sweep continuously from S_1 to S_{24} and back to S_1 at the rate of 8000 revolutions per second (Sampling rate = 8000 samples/sec.).
- This will generate 8000 samples per second of each signal (S_1 to S_{24}). Each sample is then encoded (converted) into an eight bit digital word. One complete revolution of commutator switches corresponds to generation of one frame which consists of all 24 voice channels.
- Thus the digital signal generated during one complete sweep (revolution) of the commutator is given by :

$$\begin{aligned} 1 \text{ Frame} &= 1 \text{ revolution} \\ &= 24 \text{ channels} \\ &= 24 \times 8 \text{ bits} = 192 \text{ bits} \end{aligned}$$

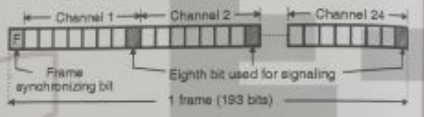
- One frame of PAM-TDM is shown in Fig. 5.9.2. Each voice signal from S_1 to S_{24} is encoded into eight bits.
- One frame corresponds to one revolution which is the time taken to transmit each signal once. Hence 1-frame corresponds to one-revolution of the commutator.



(G-1313) **Fig. 5.9.2 : One frame and bits per frame**

5.9.1 Frame Synchronization :

- As we have already seen, the synchronization between the transmitter and receiver commutators is essential.
- Without such synchronization the receiver cannot know which received bits were generated by whom at the transmitter and are meant for which subscriber on the receiving side.
- To provide such synchronization, an extra bit is transmitted preceding the 192 data bits carrying the information in each frame, as shown jn Fig. 5.9.3.
- This bit is called as the frame synchronizing bit "F". Thus one frame synchronizing bit is transmitted per frame.
- This makes the total number of bits per frame to be 193. The time slots for the 24 signals and the extra frame synchronizing bit is as shown in Fig. 5.9.3.



(G-1314) **Fig. 5.9.3 : The PCM T_1 frame using frame synchronization and channel associated signaling**

- Twelve successive F slots are used to transmit a 12 bit code. The code is 1101 1100 1000.
- This code is transmitted repeatedly once every 12 frames and it is used at the receiver to achieve synchronization between the transmitter and receiver commutators.

Bit rate :

- Bit rate means number of bits transmitted by a system per second. In the T1 system; as each signal is sampled 8000 times per second :

  1 frame (1 revolution of commutator) = 1/8000 = 125 μsec.
- But 1 frame consists of 193 bits.

  ∴ 193 bits are transmitted in 125μ sec.

$$\therefore \text{ Number of bits in 1 sec.} = \frac{193}{125 \times 10^{-6}} = 1.554 \times 10^6$$

$$\therefore \text{ Bit rate of } T_1 \text{ system} = 1.544 \text{ Mbits/sec.}$$

Bandwidth of T_1 system :

$$\text{Minimum bandwidth } B_T = \frac{1}{2} \text{ bit rate} = \frac{1}{2} \times 1.544 \times 10^6 = 772 \text{ kHz}$$

Duration of each bit :

$$\begin{aligned} 193 \text{ bits} &= 125 \ \mu S \\ 1 \text{ bit} &= (125 / 193) \ \mu S \\ &= 0.6476 \ \mu S \end{aligned}$$

5.9.2 Channel Associated Signaling :

- When the PCM-TDM system is being used for the telephony, it is expected to transmit certain control signals along with the voice information. The control information is of two types : signalling and supervisory.
- The signaling information consists of the signals such as a call is being initiated or a call is being terminated, or the address of calling party etc.
- In analog system such a signaling information is transmitted over a separate channel other than the voice channel. But in the T_1 system which is a digital system, a separate channel is not used.
- In T_1 system the signaling information is sent using the same data bit slots which are used to send the voice information. The technique used is "bit slot sharing".
- In the "bit slot sharing" method, for the first five frames, all the 24 channels are encoded into an 8 bit digital code. That means all the 8-bits in each PCM word will carry the voice information.
- However in the sixth frame, all the channels are coded into a 7 bit code and the LSB (least significant bit) of each channel is used to transmit the signaling information. This is as shown in Fig. 5.9.3. That means MSB 7-bits carry voice and the LSB bit carries the signalling information.
- This is called as "channel associated signalling". This pattern is repeated after every six frames.

5.9.3 E Lines :

- E-line is actually the European version of T lines. The T lines and E lines are conceptually identical but their capacities and number of voice channels which they can carry will be different.
- Table 5.9.1 shows the E lines, their capacities and the number of voice channels which they can carry.

Table 5.9.1 : E lines and their capacity

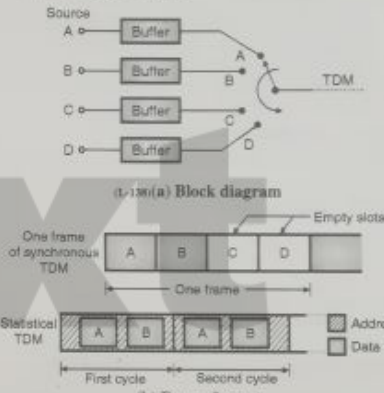
Line	Rate (Mbps)	Number of voice channels
E – 1	2.048	30
E – 2	8.448	120
E – 3	34.368	480
E – 4	139.264	1920

5.9.4 Applications of Synchronous TDM :

1. For analog telephone system (T – 1 system).
2. Some second generation cellular telephone companies use synchronous TDM.

5.10 Statistical TDM :

- The TDM system that we have discussed earlier is known as the synchronous TDM. This system has a major drawback. In synchronous TDM, many of the time slots in a frame are wasted due to absence of data on some of the time slots.
- Therefore an alternative system called as statistical TDM or asynchronous TDM or intelligent TDM is used.
- The block diagram of the statistical TDM system is as shown in Fig. 5.10.1(a) and its frame format is as shown in Fig. 5.10.1(b).



(L-139)(a) Block diagram

(b) Frame format

(L-139) **Fig. 5.10.1 : Statistical TDM**

Operating principle :

- In statistical TDM, the time slots are not permanently assigned to all the available users (like synchronous TDM). Instead, the time slots are allocated dynamically on demand only to those channels holding data for transfer.
- Each TDM channel is called as an I/O line. Thus the statistical TDM has many I/O lines and one high speed multiplexed line.
- Each I/O line has a buffer associated with it. As shown in Fig. 5.10.1, there are N number of I/O lines. Out of these only K channels are transmitted which hold data for transfer. The remaining (N – K) channels are not considered for transmission.
- In statistical TDM, the multiplexer will "scan" the input buffers of all the channels, sequentially. During the scan time, it collects the data until a frame is filled. As soon as a frame is filled, it is transmitted.
- The data is transferred on the transmission medium. The received frame is then distributed among the output buffers by the output multiplexer.

5.10.1 Data Rate of Statistical TDM :

- In statistical TDM system, all the channels are not transmitted in every frame. Hence the data rate on the multiplexed line will be less than the sum of the data rates of all the sources.
- Thus a statistical multiplexer can use a transmission medium of lower data rate to support the same number of sources as the synchronous multiplexer.
- That means if we have a synchronous and statistical TDM with equal data rates, then the statistical TDM will support more number of sources.

5.10.2 Slot Size :

- The slot carries both data and address, the ratio of the data size to address size should be reasonable to ensure high efficiency.
- In statistical TDM, the data block contains many bits while address bits are very few.

5.10.3 No Synchronization Bit :

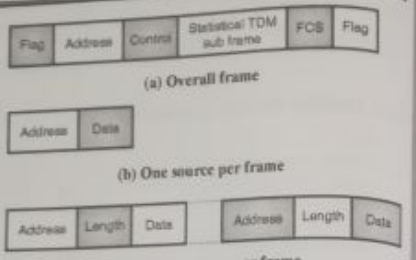
- The statistical TDM frames need not be synchronized. So it is not necessary to use the synchronizing bit.

5.10.4 Bandwidth :

- In statistical TDM, the capacity of multiplexed link is generally less than the sum of capacities of individual channels.
- Therefore the bandwidth requirement of the multiplexed link is less than that for the synchronous TDM.

5.10.5 Frame Format for Statistical TDM :

- Fig. 5.10.1(b) shows the comparison of synchronous and statistical TDM frames. It shows that there are four data sources and they are to be transmitted in four time slots t_0 to t_3.
- In the synchronous TDM, the data sources A and B only have data present whereas the data sources C and D do not have any data. Hence time slots ($t_1 - t_2$) and ($t_2 - t_3$) corresponding to these sources are empty.
- On the other hand, the statistical TDM system does not send the empty slots. Thus during the first cycle, data from the sources A and B only is transmitted.
- As the data from A and B only is being sent and the time slots of C and D are utilized for this purpose, the positional identity of slots is lost in statistical TDM. We cannot predict ahead of time about data from which source will be in which slot.
- Due to this unpredictability, the address information is required to be sent alongwith the data information. This address is then used for proper delivery of data.
- Thus each slot of statistical TDM carries an address as well as data. The detailed frame structure is as shown in Fig. 5.10.2.



(L-140) **Fig. 5.10.2 : Frame formats of statistical TDM**

Now refer to Fig. 5.10.2.

- The frame structure for the statistical TDM should be such that it should minimize the overhead bits. This is important for improving the throughout efficiency.
- The statistical TDM system uses a synchronous protocol such as HDLC. Therefore within the HDLC frame, the data frame contains the control bits for the sake of multiplexing operation.
- Fig. 5.10.2 shows two such frame formats. Fig. 5.10.2(b) shows only one source of data per frame. This source is identified by the associated address. The length of the data field is variable and its end is identified by the end of overall frame.

But the one source per frame scheme will work properly only for light loads. It is quite inefficient under the heavy load condition.

Improvement in efficiency :

1. The throughput efficiency can be improved by allowing the multiple data sources to be packaged in a single frame as shown in Fig. 5.10.2(c).
2. When many sources are packaged in a single frame, it is necessary to specify the length of data for each source. Therefore the statistical TDM subframe consists of a sequence of data fields. Each data field is labelled with an address and a length.

5.10.6 Comparison of FDM, Synchronous TDM and Statistical TDM :

Table 5.10.1 : Comparison of data multiplexer techniques

Sr. No.	Parameter	FDM	Synchronous TDM	Statistical TDM
1.	Line utilization efficiency	Poor	Good	Very good
2.	Flexibility	Poor	Good	Very good
3.	Channel capacity	Poor	Good	Excellent
4.	Error control	Not possible	Not possible	Possible
5.	Multidrop capacity	Very good	Difficult to achieve	Possible
6.	Transmission delay	Does not exist	Low	Random
7.	Cost	High	Low	Moderate

Review Questions

Q. 1 With the help of block schematic, explain the principle of FDM.

Q. 2 Compare FDM and TDM methods of multiplexing.

Q. 3 Illustrate working of FDM used for 96 channels of telephone.

Q. 4 With the help of block diagram explain the FDM system for telephone communication.

Q. 5 Explain the principles of Time Division Multiplexing.

Q. 6 What is statistical TDM ?

Q. 7 What is the difference between synchronous and statistical TDM ?

Q. 8 What are the advantages of statistical TDM ?

Q. 9 Why is it necessary to use time division multiplexing while transmitting PAM signals ?

Q. 10 Why is synchronization needed in TDM system ?

Q. 11 Describe how transmission distortion of a TDM signal can cause cross-talk between two adjacent channels ?

Q. 12 Describe the multiplexing hierarchy for an FDM system.

Q. 13 Describe the multiplexing hierarchy for digital multiplexing.

Q. 14 Explain the PCM-TDM system.

Q. 15 What do you understand by the channel associated signaling ?

Q. 16 State the applications of PCM-TDM.

Q. 17 How is synchronization achieved in PCM-TDM system ?

Q. 18 State advantages and disadvantages of TDM system.

❑❑❑

CHAPTER 6

Unit II

Transmission Media

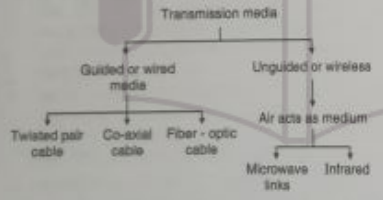
Syllabus :
Transmission media, Guided media, Twisted pair cable, Coaxial cable, Fiber optic cable.

6.1 Transmission Media :

- Media are what the message is transmitted over. In other words a communication channel is also called as a medium.
- Different media have different properties and used in different environments for different purposes.
- The purpose of the physical layer is to transport a raw bit stream from one computer to another.

6.1.1 Classification of Transmission Media :

- We can classify the transmission media as shown in Fig. 6.1.1 into two categories.

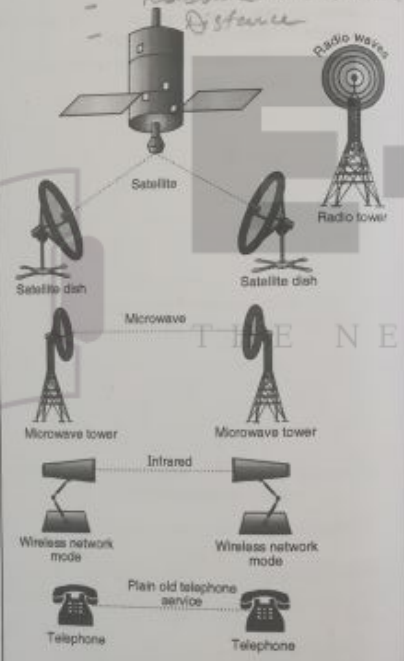


(L-571) **Fig. 6.1.1 : Classification of transmission media**

- Media are roughly grouped into two classes :
  1. Guided media 2. Unguided media

1. **Guided media** : Guided media is a communication medium which allows the data to get guided along it. For this the media need to have a point to point physical connection.
2. **Unguided media** : The wireless media is also called as an unguided media.

- The examples of guided media are copper wires and fiber-optics, whereas radio and lasers through the air are examples of unguided media as shown in Fig. 6.1.2.



(L-573) **Fig. 6.1.2 : Transmission medias**

6.1.2 Comparison of Wired and Wireless Media :

Comparison of wired and wireless media is given in Table 6.1.1.

Table 6.1.1 : Comparison of wired and wireless media

Sr. No.	Wired media	Wireless media
1.	The signal energy is contained and guided within a solid medium.	The signal energy propagates in the form of unguided electromagnetic waves.
2.	Twisted pair wires, coaxial cable, optical fiber cables are the examples of wired media	Radio and infrared light are the examples of wireless media.
3.	Used for point to point communication.	Used for radio broadcasting in all directions.
4.	Wired media lead to discrete network topologies.	Wireless media leads to continuous network topologies.
5.	Additional transmission capacity can be procured by adding more wires.	It is not possible procure additional capacity.
6.	Installation is costly, time consuming and complicated.	Installation needs less time and money.
7.	Attenuation depends exponentially on the distance.	Attenuation is proportional to square of the distance.

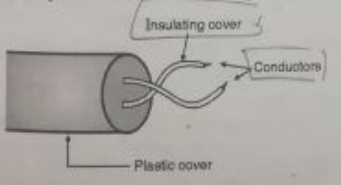
6.1.3 Types of Wired Media :

The most commonly used networking media are :
1. Co-axial cable 2. Twisted pair cable
3. Optical fiber cable.

The selection of networking media depends on various factors such as cost, connectivity, bandwidth, performance in presence of noise, geographical coverage etc.

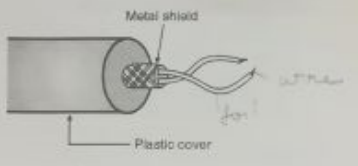
6.2 Twisted Pair Cables :

- The construction of twisted pair cable is as shown in Fig. 6.2.1. This is a very commonly used medium and it is cheaper than the co-axial cable or optical fiber cable.



(a) UTP

(L-574) **Fig. 6.2.1 (Contd..)**



(b) STP

(L-574) **Fig. 6.2.1 : Construction of twisted pair cables**

6.2.1 Types of Twisted Pair Cables :

- The two commonly used types of twisted pair cables are as follows :
  1. Unshielded Twisted Pair (UTP).
  2. Shielded Twisted Pair (STP).
- The construction of UTP and STP cables is shown in Fig. 6.2.1.

UTP :
- A twisted pair consists of two insulated conductors twisted together in the shape of a spiral as shown in Fig. 6.2.1. It can be shielded or unshielded.
- The unshielded twisted pair cables are very cheap and easy to install. But they are badly affected by the electromagnetic noise interference.

STP :
- STP cable as shown in Fig. 6.2.1(b) has a metal foil or braided mesh included in order to cover each pair of twisted insulating conductors.
- This is known as the metal shield which is normally connected to ground so as to reduce the interference of the noise. But this makes the cable bulky and expensive.
- So practically UTP is more used than STP. The STP was developed by IBM and is used primarily for the IBM company only.
- Applications of the twisted pair cables are in point to point and point to multipoint communications, telephone systems etc.
- Twisted pairs can be used for either analog or digital transmission. The bandwidth supported by the wire depends on the thickness of the wire and the distance to be travelled by a signal on it.
- Twisted pairs support several megabits/sec for a few kilometres and are less costly.

Why to twist the wires ?
- Twisting of wires will reduce the effect of noise or external interference. The induced emf into the two wires due to interference tends to cancel each other due to twisting.
- Number of twists per unit length will determine the quality of cable. More twists means better quality.

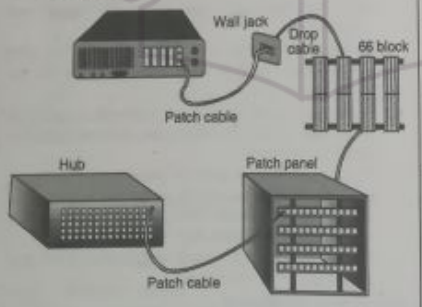
6.2.2 Categories (Cat) of UTP :

- Table 6.2.1 shows various categories of the unshielded twisted pair cables.
- These categories are decided by EIA i.e. electronic industries association. Different category cables are used for different applications.

Table 6.2.1 : Categories of UTP cables

Category	Data rate	Bandwidth	Application
1.	Extremely low upto 100 kbps	Low	Analog applications, telephony.
2.	Moderate upto 2 Mbps.	Moderate upto 2 MHz.	Analog and digital telephony
3.	Upto 10 Mbps	Upto 10 MHz	Local Area Networks (LANs)
4.	Upto 20 Mbps	Upto 20 MHz	Local Area Networks (LANs)
5.	Upto 100 Mbps	Upto 100 MHz	Local Area Networks (LANs)
6.	Upto 200 Mbps	Upto 200 MHz	Local Area Networks (LANs)
7.	Upto 600 Mbps	Upto 600 MHz	Local Area Networks (LANs)

- These cables ensure less crosstalk and a higher quality of signal over longer distances. Therefore these cables are popularly used for high speed computer communication.
- A connection diagram using the UTP is shown in Fig. 6.2.2.



(L-573) **Fig. 6.2.2 : A common UTP installation**

6.2.3 Category 3 and Category 5 (Cat 3 and Cat 5) UTP Cables :

- Most office buildings have been wired with twisted pair cable for telephones which is commonly called as voice grade UTP.
- Because these cables are already in place we can use them easily as LAN medium. The disadvantage of these voice grade twisted pair cables are low data rates and limited distances.
- Hence in 1991 the EIA published a new standard called EIA-568 in order to specify the use of voice grade unshielded twisted pair as well as shielded twisted pair for the in-building data applications.
- These standards were specified for the data rates upto 16 Mbps for LAN. But in the subsequent years, the LANS became faster with a data rate upto 100 Mbps.
- Hence a new standard EIA-568 A was published in 1995. EIA-568 A defined three categories of UTP cabling as follows :
  1. **Category 3 :** Characteristics of UTP cables and associated connecting hardware are specified upto 16 MHz.
  2. **Category 4 :** Under this category, the characteristics of UTP cables and associated connecting hardware have been specified for the data rates upto 20 MHz.
  3. **Category 5 :** Under this category, the characteristics of UTP cables and associated connecting hardware were specified for the data rates upto 100 MHz.
- The cat 3 and cat 5 cables were the most popular cables for LAN applications. Cat 3 cables are popularly used for the office building applications.
- The data rates upto 16 Mbps can be achieved by cat-3 cables provided that it is well designed and used over a limited distance.
- Cat-5 is a data-grade cable that can be used for data rates upto 100 Mbps if the distance is limited.

6.2.4 Category 6 (Cat 6) UTP :

Construction :

- Cat 6 UTP cable consists of 4-pairs of copper conductors. i.e. total 8-conductors. The jacket is made of thermoplastic polylefin or Fluorinated Ethylene Propylene (FEP).
- The material used for outside sheath of this cable can be of PVC, a fire retardant polyolefin or fluoropolymers.
- The design and manufacturing is done by taking a lot of care. Advanced connector design is essential.
- It is the best available UTP.
- Cat-6 that can operate upto 200 MHz and further increase is possible in the near future.

6.2.5 Category 7 (Cat 7) Shielded Screen Twisted Pair (SSTP) :

Construction :

- It is also called as PiMF (Pairs in Metal Foil) SSTP of 4 pairs of 22-23 copper conductors. The jacket is made of thermoplastic polyolefin or Fluorinated Ethylene Porpylene (FEP).
- A separate and improved shielding has been provided to each pair of conductors. Thus shielding has been improved.

Expected performance :

Cat-7 cable has a very large bandwidth between 6000 to 1200 MHz.

6.2.6 Applications of Twisted Pair Cables :

Some of the applications of twisted pair cables are as follows :

1. Local area networks for connecting computers to each other.
2. In the ISDN (Integrated Services Digital Network).
3. In the digital subscriber line (DSL).
4. In the analog telephony (conventional telephone line) to carry voice and data signals.
5. In digital telephony system (T_1 system)

Note :

- A modular RJ-45 telephone connector is used to connect a four-pair cable.
- A modular RJ-11 telephone connector is used to connect a two pair cable.
- Shielded twisted pair (STP) cables were introduced by IBM corporation.

6.2.7 Comparison of Twisted Pair Cables :

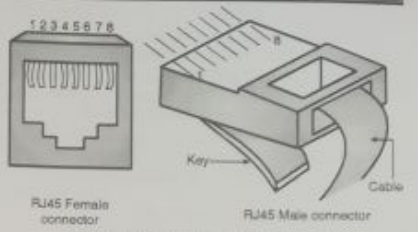
Sr. No.	Factors	UTP	STP
1.	Bandwidth	1 – 155 Mbps (typically 10 Mbps)	1 – 155 Mbps (typically 16 Mbps)
2.	Number of node connected per segment	2	2
3.	Attenuation	High	High
4.	Electromagnetic interference	Very high	High
5.	Ease of Installation	Easy	Fairly easy
6.	Cost	Lowest	Moderate

6.2.8 Connectors :

- For connecting one computer to the other, we need to use some transmission medium such as a cable.
- The cables are of different types such as twisted pair cables, coaxial cables or fiber optic cables.
- For connecting these cables between two computers we have to use connectors on both ends of a cable.
- Generally the connectors are male-female type to ensure reliable connection.

6.2.9 Connector for Twisted Pair Cable :

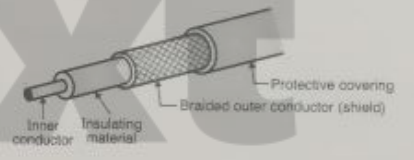
- The Unshielded Twisted Pair (UTP) cable is the most commonly used cable in computer communication.
- RJ45 is the most commonly used UTP where RJ is the short form of Registered Jack. It is a male-female type keyed connector as shown in Fig. 6.2.3.
- This connector can be inserted in only one way.



(G-348) **Fig. 6.2.3 : UTP RJ45 connector**

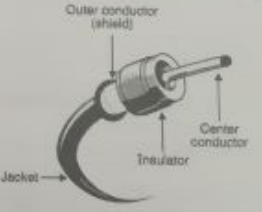
6.3 Co-axial Cables :

- The construction of co-axial cable is as shown in Fig. 6.3.1. It consists of two concentric conductors namely an inner conductor and a braided outer conductor separated by a dielectric material.
- The external conductor is in the form of metallic braid and used for the purpose of shielding. The co-axial cable may contain one or more co-axial pairs.



(L-577) **Fig. 6.3.1 : Construction of a co-axial cable**

- The construction of a co-axial cable with other accessories such as connector, jacket etc. is shown in Fig. 6.3.2.
- The wire mesh (braided conductor) protects the inner conductor from Electromagnetic Interference (EMI). It is often called a shield.
- A tough plastic jacket forms the cover of the cable as shown in Fig. 6.3.2 providing insulation and protection.



(L-577) **Fig. 6.3.2 : Co-axial cable**

- The co-axial cable was initially developed for analog telephone networks. A single co-axial cable would be used to carry more than 10,000 voice channels at a time.

- The digital transmission systems using the co-axial cable were developed in 1970s. These systems operated in the range of 8.5 Mb/s to 565 Mb/s.
- The most popular application of a co-axial cable is in the cable TV system. The existing co-axial cable system has a range from 54 MHz to 500 MHz.
- Other important application is cable modem, with the Cable Modem Termination System (CMTS).
- One more application is Ethernet LAN using the co-axial cable. The co-axial cable is used for its large bandwidth and high noise immunity.

6.3.1 Characteristics of a Co-axial Cable :

The important characteristics of a co-axial cable are as follows :

- Two types of cables having 75 Ω and 50 Ω impedance are available.
- Due to the shield provided, this cable has excellent noise immunity.
- It has a large bandwidth and low losses.
- This cable is suitable for point to point or point to multipoint applications. In fact this is the most widely used medium for local area networks.
- These cables are costlier than twisted pair cables but they are cheaper than the optical fiber cables.
- It has a data rate of 10 Mbps which can be increased with the increase in diameter of the inner conductor.
- The specified maximum number of nodes is upto 100.
- The attenuation is less as compared to the twisted pair cable.
- Co-axial cables are easy to install.
- Co-axial cables are relatively inexpensive (as compared to the optical fiber cable).

6.3.2 Co-axial Cable Standards :

Table 6.3.1 shows the co-axial cable standards. The co-axial cables are categorised by their RG ratings where RG stands for Radio Government.

Table 6.3.1 : Categories of co-axial cables

Category	Impedance	Application
RG-11	50 Ω	LAN
RG 58	50 Ω	LAN
RG 59	75 Ω	Cable TV.

6.3.3 Applications of Co-axial Cables :

1. Analog telephone networks.
2. Digital telephone network.
3. Cable TV.
4. Traditional Ethernet LANs.
5. Digital transmission.
6. Fast Ethernet.

6.3.4 Baseband Co-axial Cable :

The baseband co-axial cable is the one that makes use of digital signaling. The original Ethernet scheme makes use of baseband co-axial cable.

Characteristics of baseband co-axial cable :

- The baseband co-axial cables are used to allow digital signaling for the data.
- The digital signal used for data transfer on these cables is encoded using Manchester or Differential Manchester coding.
- The digital signals need larger bandwidth. Hence the entire frequency spectrum of the cable is consumed. So it is not possible to transmit multiple channel using FDM.
- The transmission of digital signal on the cable is bi-directional.
- The baseband co-axial cable was originally used for the Ethernet system that operates at 10 Mbps.
- These cables have a characteristic impedance of 50 Ω rather than 75 Ω of the cable TV co-axial cables.
- The maximum length of baseband co-axial cable between two repeaters is dependent on the data rates.
- Lower the data rate longer is the cable. The length has to be reduced with increased data rates so as to reduce the probability of errors getting introduced.
- There are two baseband coaxial cable used in bus LANs namely 10 BASE 5 and 10 BASE 2 which are compared based on various factors in Table 6.3.2.

Table 6.3.2 : IEEE 802.3 specifications for 10 Mbps baseband co-axial cable Bus LAN

Sr. No.	Parameter	10 BASE 5	10 BASE 2
1.	Data rate	10 Mbps	10 Mbps
2.	Maximum segment length	500 m	185 m
3.	Network span	2500 m	1000 m
4.	Nodes per segment	100	30
5.	Node spacing	2.5 m	0.5 m
6.	Cable diameter	1 cm	0.5 cm

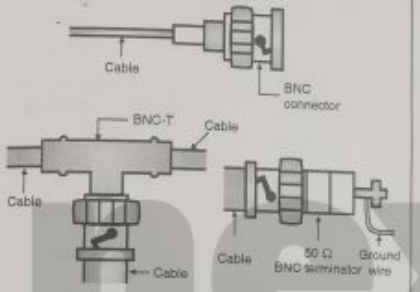
6.3.5 Broadband Co-axial Cable :

- This is the co-axial cable which is used in the cable TV system. It has higher bandwidth compared to the baseband cable.
- The type of signaling is analog at radio frequencies.
- This cable has certain disadvantages such as it is most expensive, more difficult to install and maintain as compared to the baseband co-axial cable.
- IEEE 802.3 standards have specified this as an option but practically the broadband co-axial cables are not popular.

6.3.6 Connector for Co-axial Cable :

- Coaxial cable is another important type of guided transmission media. It has higher bandwidth as compared to that of twisted pair cable.

- The coaxial cable connectors are required for connecting a coaxial cable to a computer or any other device.
- The most popular connector used for coaxial cables is the Bayone-Neill-Concelman or BNC connectors.
- Fig. 6.3.3 shows the various types of BNC connectors. The BNC connectors are available in three different types :
  1. BNC connector
  2. BNC–T connector
  3. BNC terminator.

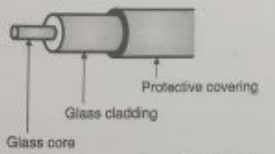


(G-341) **Fig. 6.3.3 : BNC connectors of different types**

6.4 Optical Fiber Cables :

Construction :

- The construction of an optical fiber cable is as shown in Fig. 6.4.1.
- It consists of an inner glass core surrounded by a glass cladding which has a lower refractive index and a protective covering.
- Digital signals are transmitted in the form of intensity - modulated light signal.

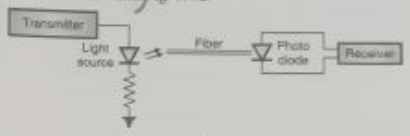


(G-103) **Fig. 6.4.1 : Construction of optical fiber cable**

- Light is launched into the fiber at one end using a light source such as a light emitting diode (LED) or laser.
- It is detected on the other side using a photo detector such as a phototransistor or photodiode.
- The optical fiber cables are costlier than the other two types but they have many advantages over the other two types.

6.4.1 Light Sources for Fiber :

- For data transmission to take place, the sending device that is the transmitter must be capable of inducing data bits 0 to 1 into the light source. At the receiver a photodiode is used to translate this light back into data bits as shown in Fig. 6.4.2.

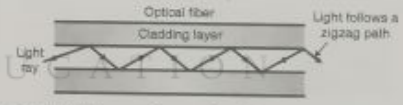


(G-104) **Fig. 6.4.2**

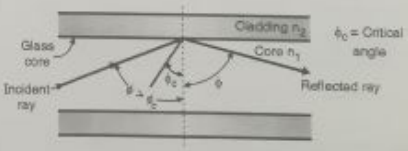
- The two light sources which are used popularly are :
  1. LED (Light Emitting Diode).
  2. Injection Laser Diode (ILD).
- The LED is cheaper but has a disadvantage that it provides an unfocussed light which hits the core boundaries and gets diffused.
- So LED is preferred only for short distances.
- The laser diode can provide a very focused beam which can be used for a long distance communication.

6.4.2 Principle of Light Propagation in a Fiber :

- The light enters into a glass fiber from one end, and gets reflected within the fiber. It follows a zigzag path along the length of the fiber as shown in Fig. 6.4.3(a).



(a) Light follows a zigzag path within the optical fiber



(b) Reflection at the interface of core and cladding

(G-145) **Fig. 6.4.3**

- Fig. 6.4.3(b) illustrates the principle of light travel through the optical fiber.
- When the light enters into a glass fiber from one end, most of it propagates along the length of the fiber and comes out from the far end.
- A small portion of the incident light escapes through the side walls of the fiber.

- The light which travels from one end to the other end of the glass fiber is said to have "guided" through the fiber.
- The light stays inside the fiber and does not escape through the walls because of the "total internal reflection" taking place inside the fiber.
- This total internal reflection can take place only if the following two conditions are satisfied :
  1. The glass fiber core must have a refractive index which is higher than the refractive index of the cladding around the core ($n_1 > n_2$).
  2. The angle of incidence of the light entering the fiber must be greater than the critical angle. "ϕ_c".

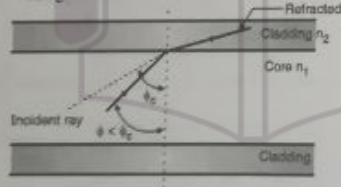
$$\sin \phi_c = \frac{n_2}{n_1}$$

This is as shown in Fig. 6.4.3.

Observations from Fig. 6.4.3(b) :

Some of the important observations from Fig. 6.4.3(b) are as follows :

1. The angle of incidence (angle made by the incident ray) i.e. ϕ is greater than the critical angle ϕ_c. Therefore the incident light ray will be reflected within the core totally. The reflected ray is at same angle as that of the incident ray.
2. If the incident light makes an angle which is less than the critical angle ϕ_c then it gets refracted as shown in Fig. 6.4.4. The refracted ray enters into the cladding and gets lost.

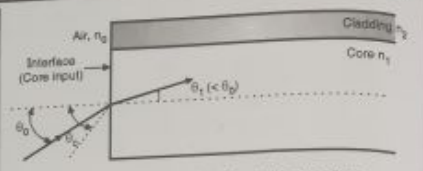


(G-186) **Fig. 6.4.4 : Refraction takes place at the core cladding interface if $\phi < \phi_c$**

6.4.3 Relation between Incident Angle and Emerging Angle :

Let us obtain the relation between the incident angle θ_0 and the emerging angle θ_1 by referring to Fig. 6.4.5.

- Assume that the refractive index of air is "n_0" and that of the fiber core is "n_1" such that $n_1 < n_0$.
- As shown in Fig. 6.4.5 the light ray enters the fiber core at an angle θ_0, through the air-core interface. The angle θ_0 is measured between the light ray and the dotted line which is normal to the air-core interface.



(G-187) **Fig. 6.4.5 : Refraction at the interface**

- When the incident light ray enters the core of refractive index, it undergoes refraction and makes an angle θ_1 with the dotted line normal to the air-core interface as shown in Fig. 6.4.5. This angle θ_1 is called as the emerging angle.
- The relation between the incident angle θ_0 and emerging angle θ_1 is given by "Snell's relationship" which states that,

$$n_0 \sin \theta_0 = n_1 \sin \theta_1 \quad ...(6.4.1)$$

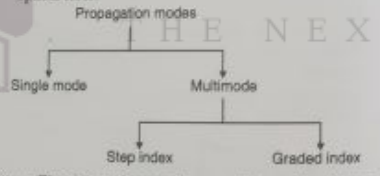
- Therefore the emerging angle θ_1 is given by,

$$\sin \theta_1 = \frac{n_0}{n_1} \sin \theta_0 \quad ...(6.4.2)$$

- As $n_0 < n_1$, $\frac{n_0}{n_1} < 1$ therefore the emerging angle will be less than the angle of incidence θ_0.

6.4.4 Modes of Propagation :

- The number of paths followed by light rays inside the optical cable is called as modes.
- Fig. 6.4.6 shows different modes of operation of an optical fiber.



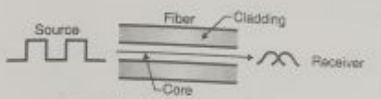
(G-188) **Fig. 6.4.6 : Propagation modes in optical fibers**

- There are two types namely single mode and multimodes fibers.
- In single mode light follows a single path through the core whereas in multimode, the light takes more than one paths through the core.

6.4.5 Single Mode Fibers :

- These are called as single mode fibers because they support on one mode of propagation (TE, TM or TEM).
- The optical signal travelling inside this fiber has only one group velocity.
- Due to single mode travelling, the amount of dispersion is less than that introduced in multimode fibers.

- These fibers can have either step index or graded index profile. They are high quality fibers used for wideband long haul communication and they are fabricated from doped silica to reduce internal attenuation.
- The light travel in a single mode fiber is shown in Fig. 6.4.7. This beam travel's almost horizontally and follows only one path from source to destination, as shown in Fig. 6.4.7. The critical angle of incident highly focused light beam is nearly equal to 90°.
- In the single mode fibers the delays are negligible and the signal reconstruction at the receiver is easier which results in almost no signal distortion.

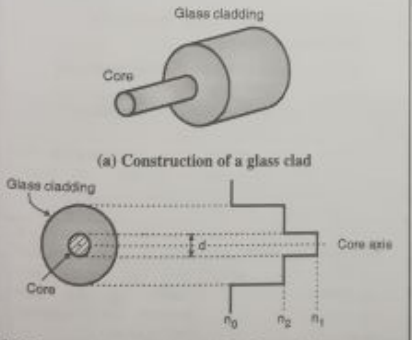


(G-189) **Fig. 6.4.7 : Single mode fiber**

6.4.6 Multimode Fibers :

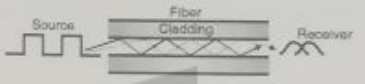
- These are called as multimode fibers because they support simultaneous propagation of many modes and the incident light follows different paths from the source to destination.
- Each mode has its own group velocity and each mode will follow its own path while travelling from the transmitter to receiver.
- Due to presence of more than one modes, the intermodal dispersion will exist.
- Multimode fibers can have the step index or graded index profile and they are fabricated using the multicomponent glasses or doped silica.

Step index fibers :



(G-110) **Fig. 6.4.8**

- The construction of an optical fiber with a core and glass cladding is as shown in Fig. 6.4.8(a).
- The refractive index of the core is n_1 and that of the glass cladding is n_2, with $n_1 > n_2$. Therefore the index profile of glass clad core fiber is as shown in Fig. 6.4.8(c).
- Due to the sudden change in refractive index at the boundary of core and cladding, this fiber is called **step index** fiber.
- Fig. 6.4.9 illustrates the propagation of light over a step index fiber.
- Multiple beams will follow different zigzag paths as shown in Fig. 6.4.9. The number of reflections that a beam undergoes, depends on the angle of incidence of that beam.

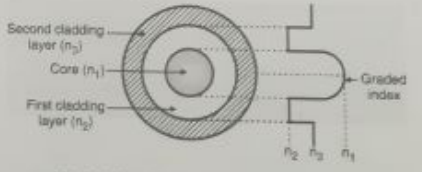


(G-111) **Fig. 6.4.9 : Multimode step index fiber**

- Hence, at the destination, all the beams do not reach simultaneously. This leads to diffusion of signal at the receiver.
- The step index multimode fibers are therefore not used for long distance communications.

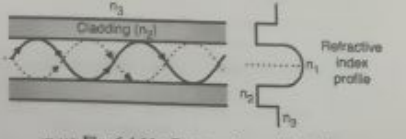
Graded index fibers :

- As shown in Fig. 6.4.10, the refractive index of the fiber core does not remain constant throughout its bulk.
- Instead it is maximum at the center of the core and reduces gradually towards the walls of the core. In order to get this type of index profile the material in the fiber core is modified.



(G-112) **Fig. 6.4.10 : Refractive index profile of a graded index fiber**

- Due to the modification in the index profile, the light gets refracted inside the fiber core and does not travel in straight line as shown in Fig. 6.4.11.



(G-113) **Fig. 6.4.11 : Propagation of light in a graded index fiber**

- Instead the light rays are curved towards the center of the core.
- These rays have been launched into the core within the acceptance cone. The acceptance cone of a graded index core is larger than that of the step index core.
- In graded index fibers as well different beams result in different curves or waveforms.

6.4.7 Characteristics of Optical Fiber Cables :

Fiber optic cables have the following characteristics :

1. Fiber optic cabling can provide extremely high bandwidths in the range from 100 Mbps to 2 Gbps because light has a much higher frequency than electricity.
2. The number of nodes which a fiber optic can support does not depend on its length but on the hub or hubs that connect cables together.
3. Fiber optic cable has much lower attenuation and can carry signal to longer distances without using amplifiers and repeaters in between.
4. Fiber optic cable is not affected by EMI effects and can be used in areas where high voltages are passing by.
5. The cost of fiber optic cable is more as compared to twisted pair and co-axial.
6. The installation of fiber optic cables is difficult and tedious.

Note :

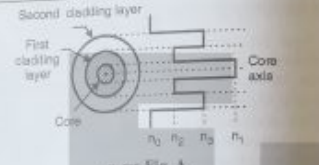
- Three wavelength bands are used for fiber optic communication respectively 850 nanometer, 1300 nanometer, 1550 nanometer.
- Single mode fiber devices are more expensive and more difficult to install than multi-mode devices.
- Fiber optic cable connectors and splice (joint) attenuate the signals.
- Fiber optic cable supports 75 nodes in an Ethernet network.
- Single mode fiber optic cable are used to provide network links of several hundred kilometres in length.
- Fiber optic cable does not leak signals so it is immune to eves dropping (tapping of signals).
- Fiber optic cable does not require a ground, hence it is not affected by potential shifts in the electrical ground, nor does it produce sparks.

6.5 Comparisons :

6.5.1 Comparison of Step Index and Graded Index Fibers :

Table 6.5.1 : Comparison of step index and graded index fibers

Sr. No.	Step index fibers	Graded index fibers
1.	The refractive index changes in steps or abruptly.	The refractive index changes gradually.
2.	The light rays travel in straight line through the step index fibers.	The light rays do not travel in straight line through the graded index fibers.
3.	Index profile : Refer Fig. A	Index profile : Refer Fig. B
4.	The light rays travel in a straight line due to constant refractive index of the fiber throughout the bulk of the core.	The light rays do not travel in straight line due to the continuous refraction. This is due to the continuously changing refractive index throughout the core bulk.
5.	Acceptance cone of these fibers is smaller than that of the graded index fiber.	Acceptance cone of these fibers is larger than that of the step index fiber.



(L-595) Fig. A

(L-595) Fig. B

6.5.2 Comparison of Single Mode and Multimode Fibers :

Table 6.5.2 : Comparison of single mode and multimode fibers

Sr. No.	Single mode fiber	Multimode fiber
1.	These fibers support only one mode of propagation (TE or TM or TEM).	These fibers support the propagation of many modes.
2.	The travelling signal inside the fiber has only one group velocity.	The different modes have different group velocities and each mode will follow its own path between the transmitter and receiver.
3.	The amount of dispersion introduced is less than that introduced in the multimode fibers.	The intermodal dispersion exists due to different group velocities of various modes.
4.	These fibers can have either a step index or graded index profile.	These fibers can have either step index or graded index profile.
5.	These are high quality fiber for wideband long haul transmission and are fabricated from doped silica for reducing the attenuation.	These are fabricated using the multicomponent glasses or doped silica.

6.5.3 Comparison of Optical Fiber with Coaxial and Twisted Pair Cables :

Sr. No.	Twisted pair cable	Co-axial cable	Optical fiber
1.	Transmission of signals takes place in the electrical form over the metallic conducting wires.	Transmission of signals takes place in the electrical form over the inner conductor of the cable.	Signal transmission takes place in an optical form over a glass fiber.
2.	Noise immunity is low. Therefore more distortion.	Higher noise immunity than the twisted pair cable due to the presence of shielding conductor.	Highest noise immunity as the light rays are unaffected by the electrical noise.
3.	Affected due to external magnetic field.	Less affected due to external magnetic field.	Not affected by the external magnetic field.
4.	Short circuit between the two conductors is possible.	Short circuit between the two conductors is possible.	Short circuit is not possible.
5.	Cheapest	Moderately expensive	Expensive
6.	Can support low data rates.	Moderately high data rates	Very high data rates.
7.	Power loss due to conduction and radiation.	Power loss due to conduction	Power loss due to absorption, scattering, dispersion and bending.
8.	Low bandwidth	Moderately high bandwidth	Very high bandwidth
9.	Node capacity per segment is 2	Node capacity per segment is 30 to 100	Node capacity per segment is 2.
10.	Attenuation is very high	Attenuation is low	Attenuation is very low.
11.	Installation is easy	Installation is fairly easy	Installation is difficult.
12.	Electromagnetic interference (EMI) can take place	EMI is reduced due to shielding	EMI is not present.

6.6 Advantages and Disadvantages of Fiber Optical Fibers :

6.6.1 Advantages of Optical Fibers :

Some of the advantages of fiber optic communication over the conventional means of communication are as follows :

1. **Small size and light weight :**
   The size (diameter) of the optical fibers is very small (it is comparable to the diameter of human hair). Therefore a large number of optical fibers can fit into a cable of small diameter.
2. **Easy availability and low cost :**
   The material used for the manufacturing of optical fibers is "silica glass". This material is easily available. So the optical fibers cost lower than the cables with metallic conductors.
3. **No electrical or electromagnetic interference :**
   Since the transmission takes place in the form of light rays the signal is not affected due to any electrical or electromagnetic interference.
4. **Large bandwidth :**
   As the light rays have a very high frequency in the GHz range, the bandwidth of the optical fiber is extremely large. This allows transmission of more number of channels. Therefore the information carrying capacity of an optical fiber is much higher than that of a co-axial cable.
5. **Other advantages :**
   In addition to the advantages discussed earlier, the optical fiber communication has the following other advantages :
   - No cross-talk inside the optical fiber cable.
   - Signals at higher data rates can be sent.
   - Intermediate amplifier are not required as the transmission losses in the fiber are low.
   - Ground loops are absent.
   - Installation is easy as the fiber optic cables are flexible.

- These cables are not affected by the drastic environmental conditions. Because of all these advantages the optical fiber cable is replacing the conventional metallic conductor cable rapidly in many areas.

6.6.2 Disadvantages of Optical Fibers :

Some of the disadvantages of optical communication system are :

1. Sophisticated plants are required for manufacturing optical fibers.
2. The initial cost incurred is high.
3. Joining the optical fibers is a difficult job.

6.6.3 Applications :

1. Optical fiber transmission systems are widely used in the backbone of networks.
2. Optical fibers are now used in the telephone systems.
3. In the Local Area Networks (LANs).

Review Questions

Q. 1 Name the layer which is associated with the transmission media.

Q. 2 Explain the classification of transmission media.

Q. 3 What is the difference between guided and unguided transmission media ?

Q. 4 State the types of guided media.

Q. 5 Explain the difference between UTP and STP.

Q. 6 What is the effect of twisting the wires in UTP cables ?

Q. 7 Give applications of co-axial cable.

Q. 8 What is the advantage of using shielding ?

Q. 9 Compare the guided transmission media.

Q. 10 State advantages of optical fiber cable.

Q. 11 State the three ways of wireless transmission.

Q. 12 Write a note on microwave communication.

Q. 13 State the applications of microwave communication.

Q. 14 Write a note on : Infrared transmission.

Q. 15 State applications of infrared transmission.

Q. 16 Compare twisted pair (UTP and STP).

Q. 17 Compare twisted pair, co-axial and fiber optic cable.

❑❑❑

CHAPTER 7

Unit II

Switching

Syllabus :

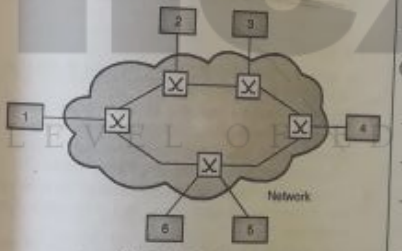
Three methods of switchings, Circuit switched networks, Packet switching.

7.1 Introduction :

A network consists of many switching devices. In order to connect multiple devices, one solution could be to have a point to point connection between each pair of devices. But this increases the number of connections.

The other solution could be to have a central device and connect every device to each other via the central device (Star topology).

Both these methods are wasteful and impractical for very large networks. The other topologies also cannot be used.



(L-616) **Fig. 7.1.1 : Switched network**

Hence a better solution is **switching**. A switched network is made of a series of interconnected nodes called switches.

Switch is a device that creates temporary connections between two or more devices. Fig. 7.1.1 shows a switched network.

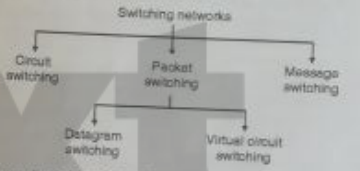
7.2 Switching Methods :

The three basic methods of switching are :

1. Circuit switching
2. Packet switching
3. Message switching

Out of these, the circuit and packet switching are commonly used today but the message switching has been phased out in general communication but is still used in the networking applications.

Fig. 7.2.1 shows the classification of switching methods.



(L-617) **Fig. 7.2.1 : Classification of switching methods**

7.3 Circuit Switching Networks :

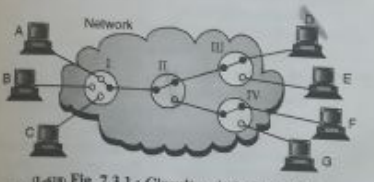
Circuit switching is used in public telephone networks. It was developed to handle voice traffic but it can also handle digital data.

However circuit switching cannot handle digital data efficiently.

Using the circuit switching, a dedicated path is established between two stations for communication.

The telephone network provide telephone service which involves the two way, real-time transmission of voice signals across a network.

The network connection allows electrical current and the associated voice signal to flow between the two users. The end to end connection is maintained for the duration of the call.



(L-618) **Fig. 7.3.1 : Circuit-switched network**

The telephone networks are connection oriented because they require the setting up of a connection before the actual transfer of information can take place.

- The transfer mode of a network that involves setting up a dedicated end to end connection is called circuit switching.
- In circuit switching the routing decision is made when the path is set up across the network. After the link has been set between the sender and receiver, the information is forwarded continuously over the link. After the link has been set up no additional address information about the receiver or destination machine is required.
- In circuit switching a dedicated path is established between the sender and the receiver which is maintained for the entire duration of conversation, as shown in Fig. 7.3.1.
- In telephone systems circuit switching is used. If circuit switching is used in computer networks the sending machine has to first establishes a link with the receiving machine.
- After the link is established the data is transmitted from the sender to the receiver. After the data flow stops, the link is released.
- In Fig. 7.3.1, I, II, III and IV are called as the switching nodes. They are used to connect one user to the other.
- The circuit switched networks operate in three phases :
  1. Set up phase
  2. Data transfer phase
  3. Tear down phase
- The circuit switching corresponds to the physical layer.
- Before starting communication in the setup phase the resources are reserved during communication. Some of these resources are channels, switch buffers, input/output ports etc.
- Data transferred between two stations is not in the packet form instead the data gets transferred continuously.
- No addressing is involved during the data transfer as the dedicated connection is established between the sender and receiver.
- The switches route the data on the basis of the allotted frequency band (FDM) or allotted time slot (TDM).

7.3.1 Three Phases :

- Communication via circuit switching takes place over three phases of operation as follows :
  1. Circuit establishment.
  2. Data transfer.
  3. Circuit disconnect. (tear down).

1. Circuit establishment :

- In a circuit switching network, before any signal is transmitted, it is necessary to establish an end-to-end (station to station) link.
- For example, in Fig. 7.3.1, if the communication is to be between A and D, then the path from A to node I to node II to node III and D has to be established first.
- The node to node links are usually multiplexed. They either use FDM or TDM.

2. Data transfer :

- The information can now be transferred from A to D through the network.
- The data can be analog or digital depending on the nature of network.
- Generally all the internal connections are duplex.

3. Circuit disconnect (tear down phase) :

- After some time the connection between two users is terminated usually by the action of one or two stations.
- Circuit switching is inefficient in most of the applications.
- The entire channel capacity is dedicated for the duration of connection, even if the data is not being transferred.
- Once the circuit is established, the network is effectively transparent to the users with no delays involved.

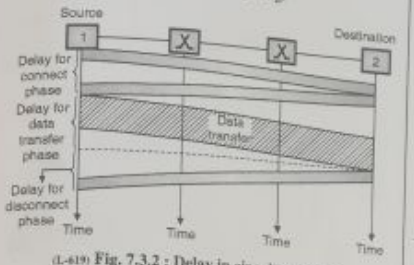
7.3.2 Efficiency :

- In circuit switching the resources remain dedicated as long as a connection is alive.
- Due to the allocation of resources during the entire duration of the connection, the efficiency of circuit switched networks is lower than the other two types of switching.

7.3.3 Delay :

- Eventhough the efficiency is low, the delay in this type of networks is very small.
- Fig. 7.3.2 explains the idea of delay in the circuit switched networks, when only two switches are used.
- During the data transfer the data is not delayed at any switch because there is no waiting time involved.
- The total delay is due to the time required for creating the connection, transfer data, and disconnect the connection.
- The delay at the time of set up is the sum of the following four parts :
  1. The propagation time related to the request message of the source computer (slope of the first gray box in Fig. 7.3.2).
  2. The time required for the transfer of request signal (height of the first gray box in Fig. 7.3.2).
  3. The time taken by the acknowledgement from the destination computer to propagate back to source (slope of the second gray box in Fig. 7.3.2).
  4. The propagation time required to transfer the acknowledgement from destination computer (height of second gray box.).
- The delay corresponding to the data transfer phase is equal to the sum of the following two components :
  1. The propagation delay (slope of hatched portion) for data transfer.

  2. Time required to transfer data (height of hatched portion) which can be very long.



(L-619) **Fig. 7.3.2 : Delay in circuit switching**

- The third component of delay is the delay corresponding to the disconnect or tear down phase. In Fig. 7.3.2 we have considered the situation in which the destination computer requests disconnection because this creates the maximum delay.,

Application :

The circuit switching is used in the telephone networks.

7.3.4 Advantages :

1. The major advantage of circuit switching is that the dedicated transmission channel the computers establish provides a guaranteed data rate.
2. In circuit switching because of the dedicated path there is no delay in data flow.

7.3.5 Disadvantages :

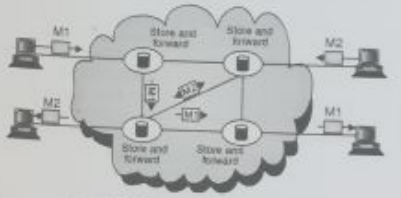
1. The disadvantage of circuit switching is that, since the connection is dedicated it cannot be used to transmit any other data even if the channel is free.
2. Dedicated channels require more bandwidth.
3. It takes long time to establish connection.

7.3.6 Circuit Switched Technology in Telephone Networks :

- The telephone companies previously used the circuit switching technology for switching and routing a call. This was a physical layer technology.
- However, today the tendency is to use other switching techniques. For example the telephone number is used as the global address and a signalling system (called SS7) is used for creating and disconnecting the connections.

7.4 Telegraph Networks and Message Switching :

- In telegraphy the text message is encoded using the Morse code into sequences of dots and dashes. Each dot or dash is communicated by transmitting short and long pulses of electrical current over a copper wire.



(L-620) **Fig. 7.4.1 : Message switching**

- In telegraph networks the text message is transmitted from the source telegraph office to the telegraph switching station. At this switching station an operator takes the decision of routing the message based on the destination address information. The operator will either forward the message if a communication line to the destination is free or store the message still the communication line becomes free.
- Message switching does not establish a dedicated path between two communicating devices. In message switching, each message is treated as an independent unit and includes its own destination and source address.
- Each complete message is then transmitted from device to device through the internetwork as shown in Fig. 7.4.1.
- Each intermediate device receives the message, stores it, until the next device is ready to receive it and then forwards it to the next device. For this reason, a message switching network is sometimes called as a store and forward network.
- Messages switches can be programmed with information about the most efficient routes as well as information regarding neighbouring switches that can be used to forward messages to their ultimate destination.

7.4.1 Advantages :

1. It provides efficient traffic management by assigning priorities to the messages to be switched.
2. It reduces network traffic congestion because it is able to store message until a communication channel becomes available.
3. With message switching, the network devices share the data channels.
4. It provides asynchronous communication across time zones.

7.4.2 Disadvantages :

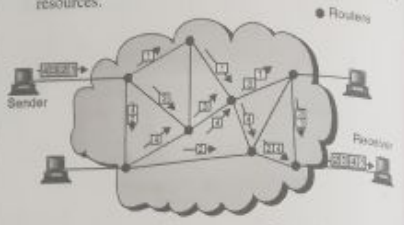
1. The storing and forwarding introduces delay hence cannot be used for real time applications like voice and video.
2. The intermediate devices require a large storing capacity since it has to store the message unless a free path is available.

7.5 Packet Switching :

- In packet switching, messages are broken up into packets. Each packets has a header with source, destination and intermediate node address information. The other part of the packet includes data load.
- Individual packets can take different routes to reach the destination. Independent routing of packets gives two advantages :
  1. Bandwidth is reduced due to splitting data onto different routes in a busy circuit.
  2. If a certain link in the network goes down during the transmission, the remaining packets can be sent through another route.
- The packets can arrive out of order at the receiver and have to be reassembled in proper sequence.
- In packet switchung, the packet length is restricted to a certain maximum length. This length is short enough to allow the switching devices to store the packet data in memory.
- There are two methods of packet switching :
  1. Datagram packet switching
  2. Virtual circuit packet switching.

7.5.1 Datagram Packet Switching :

- In this method a message is divided into a stream of packets. Each packet has its individually included address and treated as an independent unit with its own control instructors.
- The switching devices would route each packet independently through the network. Each intermediate node will determine the packet's next route segment.
- Before transmission starts, the sequence of packets and their destinations are communicated by exchanging control information between the sending terminal, the network and the receiving terminal.
- In packet switching, the resources are not allocated for any packet so there is no reserved bandwidth and no scheduled processing time allotted for each packet.
- No dedicated connection is established between the sender and receiver. The resource allocation is on demand and on the first come first serve basis.
- When a switch receives a packet, it has to wait if there are any other packets being processed. This will increase the delay.
- The datagram packet switching generally corresponds to the network layer. The packet are called as **datagrams.**
- Datagram packet switching is shown in Fig. 7.5.1.
- In this circuit, four packets are to be delivered from the sender to receiver. The switches in the datagram network are called as **routers.**
- All the four packets (datagrams) belong to the same message in this circuit however actually they can get originated from any computer.
- The four datagrams, as shown in Fig. 7.5.1 may travel different paths to reach the destination. Due to this the packets may arrive out of order at the destination.
- The delay associated with each packet will be different as a result of the different paths followed by them. The datagrams may get lost of dropped out due to lack of resources.



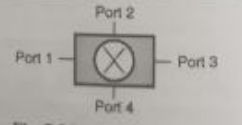
(L-623) **Fig. 7.5.1 : Datagram packet switching**

- The upper layer protocols are supposed to reorder the received datagrams or ask for the lost ones before passing them on the application.
- The datagram networks, are called as the **connectionless** networks. This is because the switch (packet switch) does not keep any information about the connection state. There are no connection set up or tear down processes in the packet switching networks.

7.5.2 Routing Table :

- In packet switched networks, each packet switch has a routing table. This table contains the destination address.
- The routing tables are dynamic and their information is updated on periodic basis. The routing table consists of destination address and the corresponding output port over which the packet is to be forwarded as shown in Fig. 7.5.2.

Destination address	Output port
1323	1
4360	2
9140	3
6436	4



(L-624) **Fig. 7.5.2 : Router and Routing table**

Destination Address :

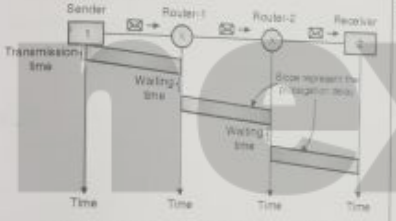
- Every packet in the datagram network consists of a header that contains the destination address where the packet is to be delivered and some additional information.
- When the router receives a packet, it examines the destination address of the packet and refers to its routing table to decide the port through which the packet is to be forwarded.
- For example in the routing table of Fig. 7.5.2, if the destination address on the received packet is 4360 then it will be forwarded through port 2.

7.5.3 Efficiency :

- As the resources are allocated only when the packets are to be transferred, the efficiency of datagram network is **higher** than that of the circuit switched network.

7.5.4 Delay :

- There are no set up or tear down phases in datagram circuit switching but each packet may have to wait at a switch before getting forwarded.
- All the packets in a message take different paths. Hence the delay associated with each packet is different.
- Fig. 7.5.3 illustrates the delays in a datagram network for one single packet.



(L-625) **Fig. 7.5.3 : Delay in datagram network**

- In Fig. 7.5.3, the packet travels through two switches while travelling from sender to receiver. The packet needs some transmission time (T) to travel from source to router-1. Then it has to wait for some time (w_1) before being forwarded.
- The total delay is made up of three transmission times (3T) and three propagation delays (3τ). The propagation delays correspond to the slopes of the lines as shown in Fig. 7.5.3 and the two waiting times w_1 and w_2.

$$\therefore \text{ Total delay} = 3T + 3\tau + w_1 + w_2 \quad \ldots(7.5.1)$$

- The datagram switching is used in Internet.

7.5.5 Advantages of Packet Switching :

1. Greater line utilization efficiency, as a single node-to-node link can be dynamically shared by many packets over time.
2. A packet switching network can perform data-rate conversion.
3. When traffic become heavy on circuit switching network, some calls are blocked. On a packet switching network, packets are still accepted, but delivery delay increases.
4. Priorities can be used.

5. Each terminal in group sharing the same physical circuit may be connected to a totally different destination. This versatility is one of the major strengths of packet switching.
6. No single user or large data block can tie up circuit or node resources indefinitely, making it well suited for interactive traffic.
7. Data protection against corruption or loss, errors are corrected by retransmission.
8. Users can select different destinations for each virtual call, overcoming the inflexibility of point to point dedicated networks.
9. Simultaneous calls allow PC users to access multiple windows to different remote applications.
10. Since many users can share transmission resources efficiently, the cost of intermittent data communication is reduced.
11. New calls can be added and old ones disconnected without affecting other users.

7.5.6 Disadvantages of Packet Switching :

1. Increased delay due to following reasons
   (a) Transmission delay = Length of packet divided by incoming channel rate.
   (b) Variable delay due to processing and queuing.
2. The amount of overall packet delay can vary substantially (jitter) due to the following reasons
   (a) Packets may vary in length.
   (b) Packets may take different routes.
   (c) Packets are subject to varying delays in switches.
   (d) This is not good for real time applications.
3. Header overhead reduces capacity to carry user data.
4. More processing required at node.

7.5.7 Datagram Networks in Internet :

The internet uses the datagram approach to switching at the Network layer. The routing of packets in Internet takes place on the basis of the universal addresses defined in the network layer.

7.6 Comparison of Message, Circuit and Packet Switching :

Parameter	Message switching	Circuit switching	Packet switching
Application	Telegraph network for transmission of telegrams.	Telephone network for bi-directional, real time transfer of voice signals.	Internet for datagram and reliable stream service between computers.
End terminal	Telegraph, teletype.	Telephone, modem.	Computer

Parameter	Message switching	Circuit switching	Packet switching
Information type	Data in the form of Morse, Baudot, ASCII codes.	Analog voice or PCM digital voice	Binary information
Transmission system	Digital data over different transmission media	Analog and digital data over different transmission media	Digital data over different transmission media.
Addressing scheme	Geographical addresses	Hierarchical numbering plan	Hierarchical address space
Routing scheme	Manual	Route selected during call setup.	Each packet is routed independently.
Multiplexing scheme	Character or message multiplexing	Circuit multiplexing.	Packet multiplexing shared media access networks.

Review Questions

- Q. 1 Explain the term circuit switching. How is it different from the packet switching ?
- Q. 2 Explain the three phases related to the communication via circuit switching.
- Q. 3 Write a short note on Space-Division switches.
- Q. 4 Explain the time-division switches.
- Q. 5 Write a short note on Time-space-Time switches.
- Q. 6 Explain the routing system in circuit switching networks.
- Q. 7 State the three switching methods.
- Q. 8 Name different types of switches used in circuit switching.
- Q. 9 How is space division switching better than time division switching ?
- Q. 10 Explain the concept of datagram packet switching.
- Q. 11 State the advantages and drawbacks of datagram packet switching.
- Q. 12 Explain the delays in datagram switching.

□□□

CHAPTER 8

Unit II

Data Link Layer

Syllabus :

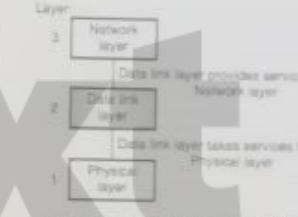
Introduction to data link layer, Nodes and links, Services, Two sub layers, Three types of addresses, Address Resolution Protocol (ARP), Error detection and correction, Introduction, Types of errors, Redundancy, Detection versus correction.

8.1 Introduction :

- The physical layer deals with the transmission of signals over different transmission medias.
- A reliable and efficient communication between two adjacent machines can be achieved via the data link layer.
- This layer basically deals with frame formation, flow control, error control, addressing and link management.
- While sending data from source to destination errors may get introduced. The data communication circuits have only a finite data rate and there is non-zero propagation delay between the instant a bit is sent and the instant at which it is received.
- These limitations affect the efficiency of data transfer. The data link layer protocols used for communication take care of all these problems.
- Data link layer is the second layer in OSI reference model. It is above the physical layer.
- It is interesting to know that the TCP/IP suite does not define any protocol corresponding to data link layer and physical layer. These two layers are known as territories of network.
- These territorial networks can provide services to all the upper layers of TCP/IP suite. They can be either wired or wireless networks.
- We know that various types of networks are connected to each other for the Internet. For interconnecting different networks, the connecting devices such as routers or switches are used.
- The packet sent by a sender host has to travel through all these networks to reach the destination host.

8.1.1 Position of Data Link Layer :

- Fig. 8.1.1 shows the position of data link layer in the five layer Internet model. It is the second layer.

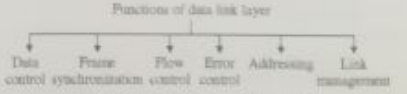


(1.4411) **Fig. 8.1.1 : Position of data link layer**

- It receives services from the physical layer and provides services to the network layer.

8.2 Data Link Layer Design Issues (Functions of Data Link Layer) :

- The data link layer is supposed to carry out many specified functions.
- For effective data communication between two directly (physically) connected transmitting and receiving stations the data link layer has to carry out a number of specific functions as follows :



(1.4412) **Fig. 8.2.1 : Functions of data link layer**

1. **Services provided to the network layer :**
The data link layer provides a well defined service interface to the network layer. The principle service is transferring data from the network layer on sending machine to the network layer on destination machine. This transfer always takes place via the DLL.

2. **Frame synchronisation :**
   The source machine sends data in the form of blocks called frames to the destination machine. The starting and ending of each frame should be identified so that the frames can be recognized by the destination machine.
3. **Flow control :**
   The source machine must not send data frames at a rate faster than the capacity of destination machine to accept them.
4. **Error control :**
   The errors introduced during transmission from source to destination machines must be detected and corrected at the destination machine.
5. **Addressing :**
   When many machines are connected together (LAN), the identity of the individual machines must be specified while transmitting the data frames. This is known as addressing.
6. **Control and data on same link :**
   The data and control information is combined in a frame and transmitted from the source to destination machine. The destination machine must be able to separate out the control information from the data being transmitted.
7. **Link management :**
   The communication link between the source and destination is required to be initiated, maintained and finally terminated for effective exchange of data. It requires co-ordination and co-operation among all the involved stations. Protocols or procedures are required to be designed for the link management.

8.3 Nodes and Links :

- The type of communication taking place at the data link layer level is called as the **node to node communication.**
- A packet sent by a computer in the Internet will have to travel through different types of networks (LANs and WANs) before reaching the destination.
- All these LANs and WANs are connected to each other using routers.

Node :

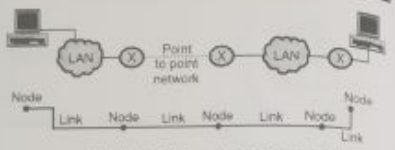
We can define a **node** as the two end hosts and the routers inbetween them.

Link :

The networks inbetween the two and hosts and the routers are called as **links.**

Source node and destination node :

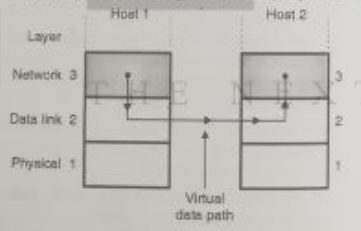
- The first node in the network is called as the source node while the last node is called as the destination node.
- Fig. 8.3.1, explains the concept of nodes and links.



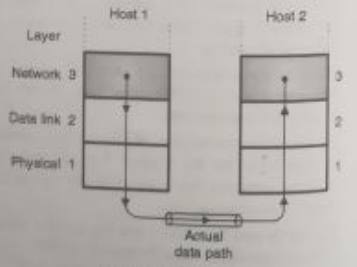
(G-2873) **Fig. 8.3.1 : Concept of nodes and links**

8.4 Services Provided to Network Layer :

- Network layer is the layer above the data link layer in the OSI model. So it is supposed to provide services to the network layer.
- The main service to be provided is to transfer data from the network layer on the sending machine to the network layer of the receiving machine.
- The virtual path followed for such a communication is shown in Fig. 8.4.1(a). It is not the actual path.
- The actual path followed by the data from sending machine to destination is shown in Fig. 8.4.1(b) which is via all the layers below the network layer, then the physical medium, then layers 1, 2, 3 of receiving machine.
- However it is always easier to think that the communication is taking place through the data link layers (Fig. 8.4.1(a)) using a data link layer protocol.



(a) Virtual communication



(b) Actual data path

(L-6657) **Fig. 8.4.1**

8.4.1 Types of Services Provided :

- Data link layer can be designed to offer different types of services. Some of them are as follows :
  1. Unacknowledged connectionless service.
  2. Acknowledged connectionless service.
  3. Acknowledged connection oriented service.

8.4.2 Unacknowledged Connectionless Service :

- In this type of service, the destination machine does not send back any acknowledgement after receiving frames.
- It is a connectionless service. So no connection is established before communication or released after it is over.
- If a frame is lost due to channel noise, then there are no attempts made to recover it.
- So this service is suitable only if the error rate is low. It is suitable for real time traffic such as speech. This type of service is highly unreliable.

8.4.3 Acknowledged Connectionless Service :

- This is the next step to improve reliability.
- In this service, there are no connections established for data transfer but for each frame received, the receiver sends an acknowledgement to the sender.
- If a frame is not received within some specified time it is assumed to be lost and the sender will retransmit it.
- This service is suitable for communication over unreliable channels such as wireless channels.

8.4.4 Acknowledged Connection Oriented Service :

- This is the most sophisticated one.
- The source and destination machines establish a connection before transferring the data.
- A specific number is given to each frame being sent and the data link layer guarantees that each transmitted frame is received.
- All the frames are guaranteed to be received in the same order as the order of transmission. Each received frame will be acknowledged individually by the destination machine.
- The data transfer takes place by following three distinct phases given below :
  1. Connection is established.
  2. The data frames are actually transmitted.
  3. The connection is released after completion of data transfer.

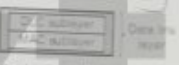
8.5 Two Sublayers :

8.5.1 Two Categories of Links :

- The medium which connects two nodes physically can be a cable or air. But the important point here is that the function of data link layer is to control how the medium is used.
- We can have a DLL which can utilize the capacity of the medium either fully or partially.
- A partially used medium is called as a **point to point link** whereas a fully used medium is called as the **broadcast link.**

8.5.2 Two Sublayers :

- We can divide the data link layer into two sublayers, in order to have a better understanding of its functionality and services provided by it.
- The two sublayers are as follows :
  1. Data link control sublayer (DLC).
  2. Media access control sublayer (MAC).
- The two sublayers are as shown in Fig. 8.5.1. The DLC sublayer is supposed to handle all the issues common to the point to point as well as broadcast links.
- However the MAC sublayer is supposed to handle the issues related only to broadcast links.



(G-2874) **Fig. 8.5.1 : Two sublayers in data link layer**

8.6 Three Types of Addresses :

- There are some data link layer protocols which define the following three types of addresses :
  1. Unicast address
  2. Multicast address
  3. Broadcast address

8.6.1 Unicast Address :

- The meaning of the word **unicast** is **one-to-one** communication. A unicast address is assigned to each host or each interface of a router.
- Therefore if a frame is having a unicast destination address, then it is destined to go to only one entity in the link.
- The example of a unicast address is the LAN address. Ethernet addresses are 48 bit in length (six bytes) which is written as 12 hexadecimal digits separated by colons.
- The example of a link layer unicast address of a computer is as shown in Fig. 8.6.1(a).

A4 : 36 : 41 : 12 : 94 : E1

Fig. 8.6.1(a) : A unicast address

8.6.2 Multicast Address :

- There are some protocols, which define multicast addresses.
- The meaning of the word **multicasting** is **one-to-many** communication. However the communication is local i.e. inside the link.
- The multicast link layer addresses are very commonly used in LANs. Ethernet. They are 48 bit

(6 bytes) long and are written as 12 hexadecimal digits separated by colons as shown in Fig. 8.6.1(b).

- Note that in the multicast address, the second digit should be an even number in hexadecimal.

A2 : 36 : 47 : 15 : 92 : E1

Fig. 8.6.1(b) : Multicast address

8.6.3 Broadcast Address :

- There are some protocols, which define the broadcast addresses.
- The meaning of the word **broadcasting** is **one-to-all** communication.
- If a frame has a destination broadcast address, then it will be sent to all the entities connected in the link.
- The broadcast address are very commonly used in LANs, Ethernets. They are 48 bit (6 bytes) long with all the bits equal to 1.
- They are written as 12 hexadecimal digits separated by colons as shown in Fig. 8.6.1(c).

FF : FF : FF : FF : FF : FF

Fig. 8.6.1(c) : Broadcast address

8.7 ARP (Address Resolution Protocol) :

- An internet consists of various types of networks and the connecting devices like routers.
- A packet starts from the source host, passes through many physical networks and finally reaches the destination host.
- At the network level, the hosts and routers are recognised by their IP addresses.

IP address :

- An IP address is an internetwork address. It is a universally unique address.
- Every protocol involved in internetworking requires IP addresses.

MAC address :

- The packets from source to destination hosts pass through physical networks. At the physical level the IP address is not useful but the hosts and routers are addressed by their MAC addresses.
- A MAC address is a local address. It is unique locally but it is not unique universally.
- The IP and MAC address are two different identifiers and both of them are needed, because a physical network can have two different protocols operating at the network layer at the same time.
- Similarly a packet may travel through different physical networks.
- So to deliver a packet to a host or a router, we require addressing to take place at two levels namely IP addressing and MAC addressing.
- Most importantly we should be able to map the IP address into a corresponding MAC address.

8.7.1 Mapping of IP Address into a MAC Address :

- We have seen the need of mapping an IP address into a MAC address.
- Such a mapping can be of two types :
  1. Static mapping and 2. Dynamic mapping.

1. **Static mapping :**
   - In static mapping a table is created and stored in each machine. This table associates an IP address with a MAC address.
   - If a machine knows the IP address of another machine then it can search for the corresponding MAC address in its table.
   - The limitation of static mapping is that the MAC addresses can change. These changed MAC addresses must be updated periodically in the static mapping table.
2. **Dynamic mapping :**
   - In dynamic mapping technique a protocol is used for finding the other address when one type of address is known.
   - There are two protocols used for carrying out the dynamic mapping. They are :
     1. Address Resolution Protocol (ARP).
     2. Reverse Address Resolution Protocol (RARP)
   - The ARP is used for mapping an IP address to a MAC address whereas the RARP is used for mapping a MAC address to an IP address.

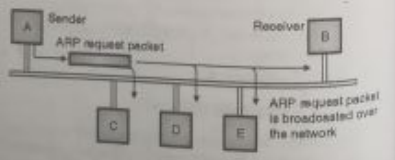
8.7.2 ARP Operation :

ARP is used for mapping an IP address to its MAC address. For a LAN, each device has its own physical or station address as its identification. This address is stored on the NIC (Network Interface Card) of that machine.

How to find the MAC address ?

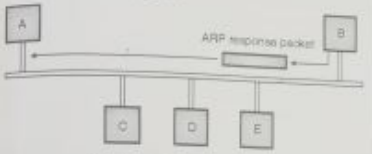
When a router or a host (A) needs to find the MAC address of another host (B) the sequence of events taking place is as follows :

1. The router or host A who wants to find the MAC address of some other router, sends an ARP request packet. This packet consists of IP and MAC addresses of the sender A and the IP address of the receiver (B).
2. This request packet is broadcasted over the network as shown in Fig. 8.7.1(a).



(G-876) **Fig. 8.7.1(a) : ARP request is broadcast**

3. Every host and router on the network will receive the ARP request packet and process it. But only the intended receiver (B) will recognize its IP address in the request packet and will send an ARP response packet back to A.
4. The ARP response packet has the IP and physical addresses of the receiver (B) in it. This packet is delivered only to A (unicast) using A's physical address in the ARP request packet. This is shown in Fig. 8.7.1(b). Thus host A has obtained the MAC address of B using ARP.



(G-876) **Fig. 8.7.1(b) : ARP response unicast**

8.7.3 ARP Packet Format :

The ARP message format is as shown in Fig. 8.7.2. The various fields in it are as follows :

1. **HTYPE (Hardware Type) :** This 16 bit field defines the type of network on which ARP is being run. ARP is capable of running on any physical network.
2. **PTYPE (Protocol Type) :** This 16 bit field is used to define the protocol using ARP. Note that we can use ARP with any higher-level protocol such as IPv4.
3. **HLEN (Hardware length) :** It is an 8 bit field which is used for defining the length of the physical address in bytes. For example, this value is 6 for Ethernet.

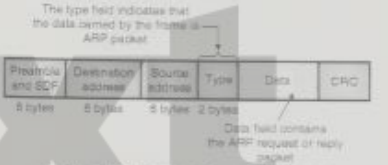
Hardware Type (16 bits)		Protocol type (16 bits)
Hardware length	Protocol length	Operation request 1, Reply 2
Sender hardware address		
Sender protocol address		
Target hardware address		
Target protocol address		

Fig. 8.7.2 : ARP message format

4. **PLEN (Protocol Length) :** This field is 8 bit long and it defines the length of the IP address in bytes. For IPv4 this value is 4.
5. **OPER (Operation) :** It is a 16 bit field which defines the type of packet. The two possible types of packets are : ARP request (1) and ARP reply (2).
6. **SHA (Sender Hardware Address) :** This field is used for defining the physical address of the sender. The length of this field is variable.
7. **SPA (Sender Protocol Address) :** This field defines the logical address of the sender. The length of this field is variable.
8. **THA (Target Hardware Address) :** It defines the physical address of the target. It is a variable length field. This field contains all zeros for the ARP request packet, because the receivers physical address is not known to the sender.
9. **TPA (Target Protocol Address) :** This field defines the logical address of the target. It is a variable length field.

8.7.4 Encapsulation :

- An ARP packet (request or reply) is inserted directly into the data link frame. Such an insertion is known as encapsulation.
- Fig. 8.7.3 shows an example of encapsulation in which an ARP packet being encapsulated in an Ethernet frame. The type field shows that the data carried by the frame is an ARP request or reply packet.



(G-877) **Fig. 8.7.3 : Encapsulation of ARP packet**

8.7.5 Operation of ARP on Internet :

The services of ARP can be used under the following working conditions when it is being operated on internet :

1. The sender is a host and wants to communicate with another host which is on the same network.
2. The sender is a host and wants to communicate with a host on another network.
3. The sender is a router. It has received a datagram with a destination address of a host on another network.
4. The sender is a router. It has received a datagram which is meant for a host in the same network.

- Now let us see how ARP works on the internet.

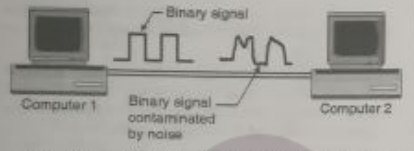
Operation :

1. The sender (host or router) knows the IP address of the target.
2. IP orders ARP to create an ARP request message. The request packet consists of senders physical and IP addresses plus the IP address of the target but the physical address of the target is not known.
3. This ARP request packet is sent to the data link layer. Here the ARP request packet is inserted in a frame.
4. Every router or host receives this frame because it is broadcast. All the machines except the target drop this packet as discussed earlier.
5. The target machine sends back a reply packet which contains the target's physical address. This reply is unicast and addressed only to the sender.
6. The sender receives the reply packet. Hence the physical address of the target has been obtained.

7. The IP datagram carrying data for the target machine is inserted in a frame and the frame is unicast to the target machine.

8.8 Introduction to Error Control Coding :

- When transmission of digital signals takes place between two systems such as computers as shown in Fig. 8.8.1, the signal get contaminated due to the addition of "Noise" to it.



(L-302)**Fig. 8.8.1 : Noise contaminates the binary signal**

- The noise can introduce an error in the binary bits travelling from one system to the other. That means a 0 may change to 1 or a 1 may change to 0.
- These error can become a serious threat to the accuracy of the digital system. Therefore it is necessary to detect and correct the errors.

8.8.1 Need of Error Control Coding :

- In data communication, errors are introduced during the transmission of data from the transmitter to receiver due to noise or some other reasons.
- The reliability of data transmission will be severely affected due to these errors.
- In order to improve the reliability of data transmission, the designer will have to increase the signal power or reduce the noise spectral density N_o so as to maximize the ratio E_b / N_o.
- But practically there is a limitation on the maximum value of the ratio E_b / N_o. We cannot increase the ratio beyond this limit. **Hence for a fixed value of E_b / N_o, we have to use some kind of "coding" in order to improve the quality of the transmitted signal.**
- Another advantage of using coding is that we can reduce the required value of E_b / N_o if the error rate is predecided and remains fixed at that value. This will inturn reduce the required transmitted power and the size of antenna.

How to detect and correct errors ?

- For the detection, and / or correction of these errors, one or more than one extra bits are added to the data bits at the time transmitting.
- These extra bits are called as parity bits. They allow the detection or sometimes correction of the errors.
- The data bits alongwith the parity bits form a code word.

Error control techniques :

- The error control techniques can be divided into two types :
  1. Error detection techniques.
  2. Error correction techniques.
- The error detecting techniques are capable of only detecting the errors. They cannot correct the errors.
- The error correcting techniques are capable of detecting as well as correcting the errors.

8.8.2 Types of Errors :

The errors introduced in the data bits during their transmission can be categorised as :

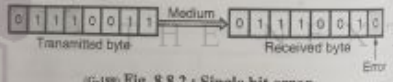
1. Content errors 2. Flow integrity errors.

- The content errors are nothing but errors in the content of a message e.g. a "0" may be received as "1" or vice versa. Such errors are introduced due to noise added into the data signal during its transmission.
- Flow integrity errors means missing blocks of data. It is possible that a data block may be lost in the network possibly because it has been delivered to a wrong destination.
- Depending on the number of bits in error we can classify the errors into two types as :

1. Single bit error 2. Burst errors.

1. Single bit error :

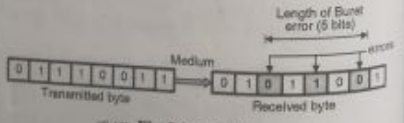
- The term single bit error suggests that only one bit in the given data unit such as byte is in error.
- That means only one bit in a transmitted byte will change from 1 to 0 or 0 to 1, as shown in Fig. 8.8.2.



(G-188) **Fig. 8.8.2 : Single bit error**

2. Burst errors :

- If two or more bits from a data unit such as a byte change from 1 to 0 or from 0 to 1 then burst errors are said to have occurred.
- Refer Fig. 8.8.3 in which the shaded bits in the received byte have been the erroneous bits. These are 3 bits but the length of the burst is shown to be of 5 bits.



(G-189) **Fig. 8.8.3 : Burst errors**

- The length of the burst error extends from the first erroneous bit to the last erroneous bit. Even though some of the bits in between have not been corrupted. The length of the burst error is shown to be 5 bits.
- Burst errors are illustrated in Fig. 8.8.3.

Disadvantages of coding :

Some of the disadvantages of the coding technique are :

1. An increased transmission bandwidth is required in order to transmit the encoded signal. This is due to the additional bits (redundancy) added by the encoder.
2. Use of coding make the system complex.

8.8.3 Disadvantages of Coding :

Some of the disadvantages of the coding technique are :

1. An increased transmission bandwidth is required in order to transmit the encoded signal. This is due to the additional bits (redundancy) added by the encoder.
2. Use of coding make the system complex.

8.8.4 Redundancy :

- Redundancy involves transmission of extra bits alongwith the data bits. These extra bits actually do not contain any data or information but they ensure the detection and correction of errors introduced during the data travel from sender to receiver.
- As these extra bits do not contain any information, they are known as redundant bits.
- The redundant bits are also known as **parity check** bits. They are produced from the data bits using some predecided rules.
- The data bits and redundant bits together form a code word as shown in Fig. 8.8.4.

(L-303)**Fig. 8.8.4 : Structure of a transmitted code word**

8.9 Detection Versus Correction :

- Detection and correction of errors are the two most important aspects of error control in data communication.
- The correction of errors is more difficult as compared to their detection. The process of error detection is much easier because we have to simply find if error is present or absent in the received code word.
- In **error detection** we are not interested even in the number of errors. The only question to be answered is whether an error has occurred or not.
- In **error correction,** multiple processes are involved such as detecting the errors, knowing their number, the location of errors and then correcting the erroneous bits.

Review Questions

Q. 1 State the various design issues for the data link layer.

Q. 2 State and explain the various services provided to the Network layer.

Q. 3 Define node and link.

Q. 4 Define node to node communication.

Q. 5 State three types of addresses.

Q. 6 Explain unicast and multicast addresses.

Q. 7 State the name of sublayers in data link layer.

Q. 8 Explain burst errors.

Q. 9 Explain the need, advantages and disadvantages of coding.

Q. 10 Explain the purpose of ARP.

Q. 11 Why is ARP request broadcast but ARP reply unicast ?

❑❑❑

CHAPTER 9

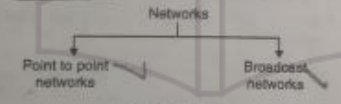
Unit III

Multiple Access

Syllabus :
Media access control (MAC), Random access, CSMA, CSMA/CD, CSMA/CA, Controlled access, Reservation, Polling, Token passing, Channelization, FDMA, TDMA, CDMA.

9.1 Introduction :

- We can classify the networks into two categories as shown in Fig. 9.1.1.
- In this chapter, we are going to discuss the broadcast networks and their protocols.
- The broadcast channels are also called as **multi-access channels** or **random access channels.**
- In the broadcast networks the most important point is the criteria by which we decide, who is allowed to use the common channel when more than one users want to use it.
- A protocol is used to make this decision.
- Such a protocol, belongs to a sublayer of data link layer called the MAC (Medium Access Control) sublayer.

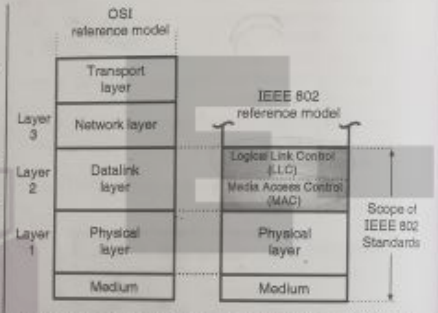


(G-264) **Fig. 9.1.1**

- The MAC sublayer is very important in LANs because it is a broadcast network.

9.1.1 MAC and LLC Sublayers :

- Fig. 9.1.2 shows the layered OSI model (partial) to show the position of MAC and LLC sublayers.
- We will discuss the broadcast protocols corresponding to the lower layers (1 and 2) of the OSI model as shown in Fig. 9.1.2.
- Fig. 9.1.2 relates the LAN protocols with the OSI architecture. This architecture was developed by IEEE 802 committee and it has been accepted as LAN standard.
- It is called as IEEE 802 reference model. Let discuss this model layer by layer.



(G-265) **Fig. 9.1.2 : IEEE 802 protocol layers compared to OSI model**

Functions of Media Access Control (MAC) sublayer :

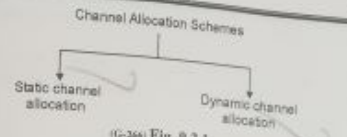
- To perform the control of access to media.
- It performs the unique addressing to stations directly connected to LAN.
- Detection of errors.

Functions of Logical Link Control (LLC) sublayer :

- Error recovery.
- It performs the flow control operation.
- User addressing.

9.2 The Channel Allocation Problem :

- In a broadcast network, the single communication channel is to be allocated to one transmitting user at a time. The other users connected to this medium should wait.
- This is called as channel allocation. There are two different schemes used for channel allocation as shown in Fig. 9.2.1.



(G-266) **Fig. 9.2.1**

9.2.1 Static Channel Allocation in LANs and MANs :

- The traditional way of allocating a single channel, among many users is by means of frequency division multiplexing (FDM).
- The Frequency Division Multiplexing (FDM) and Time Division Multiplexing (TDM) are the examples of static channel allocation.
- In these methods either a fixed frequency band or a fixed time slot is allotted to each user. Thus either the entire available bandwidth or entire time is shared.
- The problem in these methods is that if all the N number of users are not using the channel the channel bandwidth is wasted and if there are more than N users who want to use the channel they cannot do so for the lack of bandwidth.
- For a small number or users and fight traffic the static FDM is an efficient method of allocation but its performance is poor for large number of users, bursty and heavy traffic etc.
- The static channel allocation has a poor performance with bursty traffic and hence generally dynamic channel allocation is used, for computer networks where the traffic is of bursty nature.

9.2.2 Dynamic Channel Allocation in LAN and MAN :

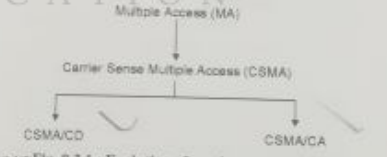
- In this method either a fixed frequency or fixed time slot is not allotted to the user. The user can use the single channel as per his requirement. Following assumptions are made for the implementation of this method :
  1. Station model – This model consists of N independent stations such as a PC, computer etc. which can generates frames for transmission.
  2. Single channel – A single channel is available for all communication.
  3. Collision – If frames are transmitted at the same time by two or more stations, there is an overlap in time and the resulting signal is garbled. This is called as collision.
  4. Continuous or slotted time – There is no master clock used to divide time into discrete time intervals. So frames can begin at any random instant. This is continuous time. For a slotted time, the time is divided into discrete time slots.
  5. Carrier or No carrier sense – Stations sense the channel before transmission or they directly transmit without sensing the channel.

9.3 Multiple Access :

- When a number of stations (users) use a common link of communication system we have to use a multiple access protocol in order to coordinate the access to the common link.
- The three techniques used to deal with the multiple access problem are as follows :
  1. Random Access
  2. Controlled Access
  3. Channelization.
- Let us discuss them one by one.

9.3.1 Random Access :

- In the random access technique there is no control station.
- Each station will have the right to use the common medium without any control over it.
- With increase in number of stations, there is an increased probability of **collision** or **access conflict.**
- The collisions will occur when more than one user tries to access the common medium simultaneously.
- As a result of such collisions some frames can be either modified (due to errors) or destroyed.
- In order to avoid collisions, we have to set up a procedure.
- The evolution of the random access methods is shown in Fig. 9.3.1.



(G-267) **Fig. 9.3.1 : Evolution of random access methods**

9.3.2 Evolution of Random Access Methods :

- The first method in the evolution ladder of Fig. 9.3.1, known as ALOHA used a simple procedure called Multiple Access (MA).
- It was improved to develop the Carrier Sense Multiple Access (CSMA).
- The CSMA further evolved into two methods namely CSMA/CD (CSMA with collision detection) and CSMA/CA (CSMA with collision avoidance) which avoids the collisions.

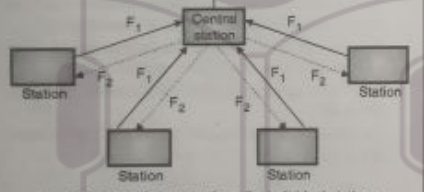
9.4 Multiple Access (ALOHA System) :

ALOHA System :

- Systems in which multiple users share a common channel in a way that can lead to conflicts are widely known as Contention systems.
- The ALOHA system is a contention protocol which was developed at the University of Hawaii in the early 1970's by Norman Abramson and his colleagues.
- The ALOHA system has two versions :
  1. Pure ALOHA – does not require global time synchronisation.
  2. Slotted ALOHA – requires time synchronisation.

9.4.1 Pure ALOHA :

- It works on a very simple principle. Essentially it allows for any station to broadcast at any time. If two signals collide, each station simply waits a random time and try again.
- Collisions are easily detected. As shown in the Fig. 9.4.1, when the central station receives a frame it sends an acknowledgement on a different frequency.



F_1 = Broadcast frequency from the individual stations.

F_2 = Broadcast frequency from the central station.

(G-268) **Fig. 9.4.1 : Pure ALOHA system**

- If a user station receives an acknowledgement it assumes that the transmitted frame was successfully received and if it does get an acknowledgement it assumes that collision had occurred and is ready to retransmit.
- The advantage of pure ALOHA is its simplicity in implementation but it's performance becomes worse as the data traffic on the channel increases.

9.4.2 Protocol Flow Chart for ALOHA :

Fig. 9.4.2 shows the protocol flow chart for ALOHA.

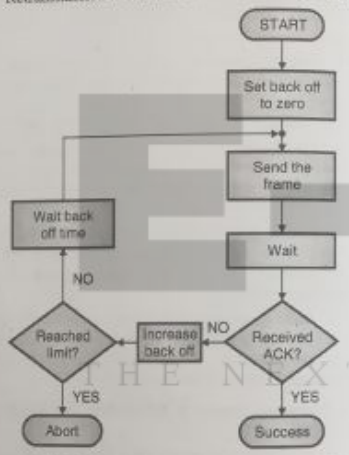
Explanation :

- A station which has a frame ready for transmission will send it.
- Then it waits for some time.
- If it receives the acknowledgement then the transmission is successful.
- Otherwise the station uses a backoff strategy, and will send the packet again.
- After sending the packet many times if there is no acknowledgement then the station aborts the idea of transmission.

Contention system :

Systems in which multiple users share a common channel in such a way that can lead to a conflict or collision are known as the contention systems.

- Whenever two frames try to occupy the channel at the same time, there is bound to be a collision and both will be garbled.
- Retransmission is essential for all the destroyed frames.



(G-269) **Fig. 9.4.2 : Protocol flow chart for ALOHA**

9.4.3 Efficiency of an ALOHA Channel :

- Efficiency of an ALOHA system is that fraction of all transmitted frames which escape collisions i.e. which do not get caught in collisions.
- Consider ∞ number of interactive users at their computers (stations). Each user is either typing or waiting. Initially all of them are in the typing state.
- When a user types a line, the user stops and waits. The station then transmits a frame containing this line and checks the channel to confirm the success. If it is successful then the user will start typing again, otherwise the user waits and its frame is retransmitted many time till it is sent successfully.

Frame time :

- Let the frame time be defined as the amount of time required to transmit the standard fixed length frame. Note that

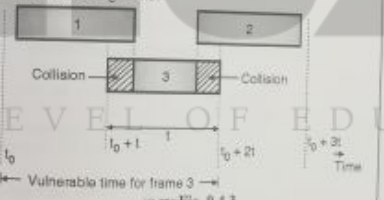
$$\text{Frame time} = \frac{\text{Frame length}}{\text{Bit rate}}$$

- We assume that ∞ number of users generate new frames according to the Poisson's distribution with an average N frames per frame time.
- The value of $N > 1$ indicates that the users are generating frames at a rate higher than that can be handled by the channel. So most of the frames will face collision. Hence $0 < N < 1$ in order to reduce number of collisions.
- Let there be k transmission attempts (including retransmissions) per frame time.
- The probability of k transmissions per frame time is also Poisson. Let the mean of number of transmissions be G per frame time. So $G \geq N$.
- At low load $N = 0$ there will be less number of collisions so less number of retransmissions and $G = N$.
- With increase in load there are many collisions so $G > N$. Combining all these we can say that for all the loads the throughput is given by,

$$S = GP_0$$

Where P_0 = Probability that a frame does not suffer a collision.

- Consider Fig. 9.4.3.



(G-270) **Fig. 9.4.3**

- What is the condition for frame 3 in Fig. 9.4.3 to arrive undamaged without collision ? Let t = Time required to send a frame. If frame 1 is generated at any instant between t_0 to $(t_0 + t)$ then it will collide with frame 3. Similarly any frame 2 generated between $(t_0 + t)$ and $(t_0 + 2t)$ also collides with frame 3.
- As per Poisson's distribution, the probability of generating k frames during a given frame time is given by,

$$P[k] = \frac{G^k e^{-G}}{k!}$$

- So the probability of generating zero frames i.e. $k = 0$ is

$$P_0 = \frac{G^0 e^{-G}}{0!} = e^{-G}$$

- If an interval is two frame time long, the mean number of frames generated during that interval is 2G.
- The probability that no other frame is transmitted during the Vulnerable period (time when collision can take place) is,

$$P_0 = e^{-2G}$$

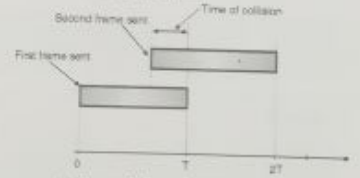
- But throughput $S = G P_0$

$$\therefore \quad S = Ge^{-2G}$$

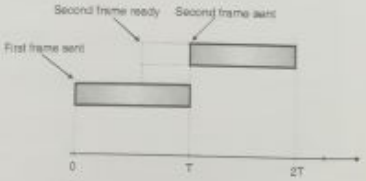
- Fig. 9.4.5 shows the relation between the offered traffic G and the throughput S. It shows that the maximum throughput occurs at $G = 0.5$ and $S_{max} = 0.184$. So the best possible channel utilization is on 18.4 percent.

9.4.4 Slotted ALOHA :

- To overcome the disadvantage of the pure ALOHA system (of low capacity) Robert published a method for doubling the capacity of traffic on the channel.
- In this method it was proposed that the time be divided up into discrete intervals and each interval correspond to one frame.
- This method requires that the users agree on the slot boundaries. In this method for achieving synchronisation one special station emits a pip at the start of each interval, like a clock. This method is known as the slotted ALOHA system.
- Collisions occur if any part of two transmission overlaps. Suppose that T is time required for one transmission and that two stations must transmit.
- The total time required for both stations to do so successfully is 2T as shown in Fig. 9.4.4. In case of pure ALOHA allowing a station to transmit at arbitrary times can waste time upto 2T.



(a) Transmission using pure ALOHA



(b) Transmission using slotted ALOHA

(G-271) **Fig. 9.4.4**

- As an alternative, in the slotted ALOHA method the time is divided into intervals (slots) of T units each and require each station to begin each transmission at the beginning of a slot.
- In other words, even if station is ready to send in the middle of a slot, it must wait until the beginning of the next one as shown in Fig. 9.4.4(b).
- In this method a collision occurs when both stations become ready in the same slot.
- Slotted ALOHA is thus a discrete time system whereas pure ALOHA is a continuous time system.
- The vulnerable period has been reduced to half that of pure ALOHA, the throughput for slotted ALOHA is given by,

$$S = Ge^{-G}$$

- The maximum throughput corresponds to G = 1 and it is given by $S_{max} = 1/e = 0.368$ as shown in Fig. 9.4.5. So for a slotted ALOHA with G = 1 the probability of success is 37%. The probability of empty slots is,

$$P(k) = \frac{G^k e^{-G}}{k!}$$

For G = 1 and k = 0 we get P (k = 0) = 0.368.

- And the probability of collisions is 26 %.
- The probability of transmission requiring exactly k attempts (i.e. k – 1 collisions followed by one success) is given by,

$$P_k = e^{-G}(1 - e^{-G})^{k-1}$$

- And the expected number of transmissions E per carriage return typed is

$$E = e^{G}$$

Conclusion : As E depends exponentially on G, with a small increase in G, there is a large increase in E and drastic fall in performance.

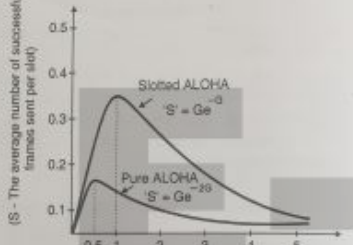
9.4.5 Comparison of Pure and Slotted ALOHA :

- A mathematical model can be created for the relationship between the number of frames transmitted and the number of frames transmitted successfully.
- Let G represent the traffic measured as the average number of frames generated per slot.
- Let S be the success rate measured as the average number of frames sent successfully per slot.
- The relationship between G and S for both pure and slotted ALOHA is given as follows :

  Pure ALOHA → $S = Ge^{-2G}$

  Slotted ALOHA → $S = Ge^{-G}$

  Where e is the mathematical constant = 2.718.
- From the above equation a success rate curve for pure and slotted ALOHA can be plotted as shown in Fig. 9.4.5.
- As seen in the Fig. 9.4.5 both graphs have the same shape. If G is small so is S, which means that if few frames are generated few frames will be transmitted successfully.
- As G increases so does S but upto a certain point. As G continues to increase S approaches to 0 which means that if more frames are generated there will be more collisions and the success rate will fall to 0.
- Similarly for pure ALOHA the maximum occurs at G = 0.5 for which S = 1/2e = 0.184 which means the rate of successful transmissions is approximately 18.4%.



(G-272)**Fig. 9.4.5 : Comparison of pure and slotted ALOHA**

As seen from the graph the maximum for slotted ALOHA occurs at G = 1 for which $S = 1/e = 0.368$. In other words the rate of successful transmissions is approximately 0.368 frames per slot time or 37% of the time will be spent on successful transmissions.

- Hence the slotted ALOHA has a double throughput efficiency than the pure ALOHA system.
- The maximum utilization achievable using CSMA can be increased much beyond that obtainable using ALOHA or slotted ALOHA.
- The maximum utilization is dependent on length of the frame and on the propagation time.
- With increase in the length of the frame or reduction in the propagation time the utilization gets improved.

9.5 Carrier Sense Multiple Access (CSMA) :

The CSMA protocol operates on the principle of carrier sensing. In this protocol, a station listens to see the presence of transmission (carrier) on the cable and decides to act accordingly.

Non-Persistent CSMA :

- In this scheme, if a station wants to transmit a frame and it finds that the channel is busy (some other station is transmitting) then it will wait for fixed interval of time.

- After this time, it again checks the status of the channel and if the channel is free it will transmit.

1-Persistent CSMA :

- In this scheme the station which wants to transmit, continuously monitors the channel until it is idle and then transmits immediately.
- The disadvantage of this strategy is that if two stations are waiting then they will transmit simultaneously and collision will take place. This will then require retransmission.

P-Persistent CSMA :

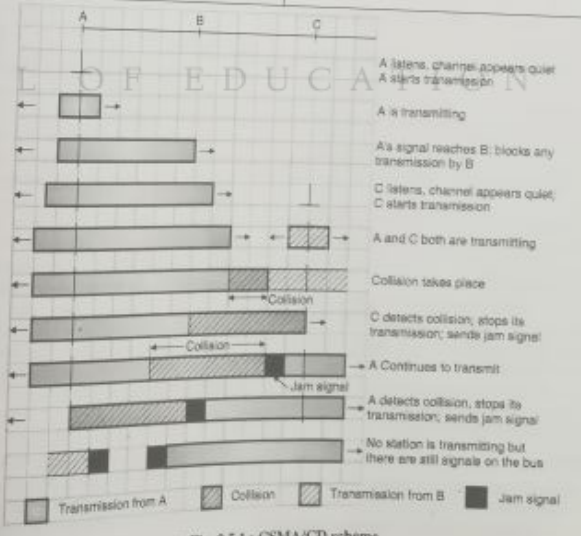
- The possibility of such collisions and retransmissions is reduced in the p-persistent CSMA. In this scheme all the waiting stations are not allowed to transmit simultaneously as soon as the channel becomes idle.
- A station is assumed to be transmitting with a probability "p". For example if p = 1/6 and if 6 stations are waiting then on an average only one station will transmit and others will wait.

9.5.1 Carrier Sense Multiple Access/Collision Detection (CSMA/CD) :

The CSMA/CD specifications have been standardized by IEEE 802.3 standard. It is a very widely used MAC protocol.

Media access control :

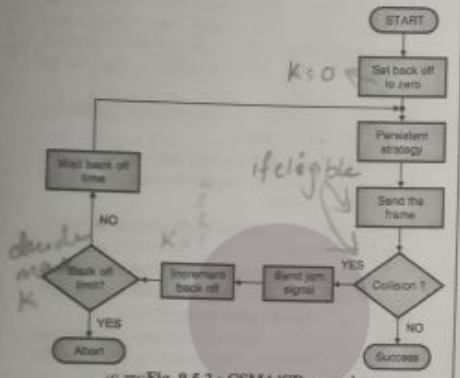
- The problem in CSMA explained earlier is that a transmitting station continues to transmit its frame even though a collision occurs.
- The channel time is unnecessarily wasted due to this. In CSMA/CD, if a station receives other transmissions when it is transmitting, then a collision can be detected as soon as it occurs and the transmission time can be saved.
- As soon as a collision is detected, the transmitting stations release a jam signal.
- The jam signal will alert the other stations. The stations then are not supposed to transmit immediately after the collision has occurred.
- Otherwise there is a possibility that the same frames would collide again.
- After some "back off" delay time the stations will retry the transmission. If again the collision takes place then the back off time is increased progressively.
- A careful design can achieve efficiencies of more than 90% using CSMA/CD. This scheme is as shown in Fig. 9.5.1.



(G-273)**Fig. 9.5.1 : CSMA/CD scheme**

9.5.2 CSMA/CD Procedure :

- Fig. 9.5.2 shows a flow chart for the CSMA/CD protocol.



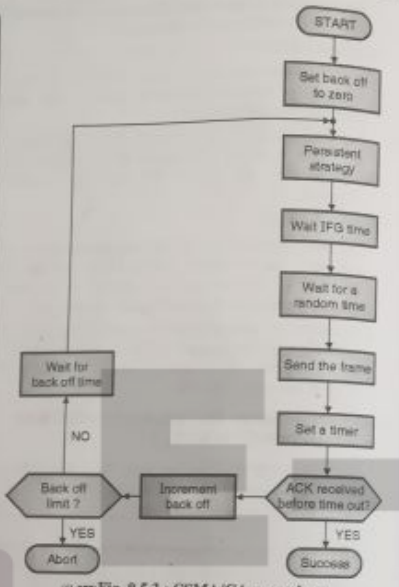
(G-276)**Fig. 9.5.2 : CSMA/CD procedure**

Explanation :

- The station that has a ready frame sets the back off parameter to zero.
- Then it senses the line using one of the persistent strategies.
- It then sends the frame, if there is no collision for a period corresponding to one complete frame, then the transmission is successful.
- Otherwise (in the event of collision) the station sends the jam signal to inform the other stations about the collision.
- The station then increments the back off time and waits for a random back off time and sends the frame again.
- If the back off has reached its limit then the station aborts the transmission.
- CSMA/CD is used for the traditional Ethernet.
- CSMA/CD is an important protocol. IEEE 802.3 (Ethernet) is an example of CSMA/CD. It is an international standard.
- The MAC sublayer protocol does not guarantee reliable delivery. Even in absence of collision the receiver may not have copied the frame correctly.

9.5.3 CSMA/CA :

- The long form CSMA/CA is CSMA protocol with collision avoidance.
- Fig. 9.5.3 shows the flow chart explaining the principle of CSMA/CA.



(G-277)**Fig. 9.5.3 : CSMA/CA procedure**

- The station ready to transmit, senses the line by using one of the persistent strategies.
- As soon as it find the line to be idle, the station waits for a time equal to an IFG (Interframe gap).
- It then waits for some more random time and sends the frame.
- After sending the frame, it sets a timer and waits for the acknowledgement from the receiver.
- If the acknowledgement is received before expiry of the timer, then the transmission is successful.
- But if the transmitting station does not receive the expected acknowledgement before the timer expiry then it increments the back off parameter, waits for the back off time and senses the line again. CSMA/CA completely avoids the collision.

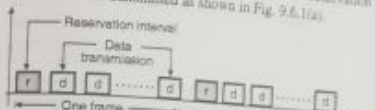
9.6 Controlled Access :

- In the previous section we have discussed the random access approach for sharing a transmission medium.
- The random access approach is simpler to implement and are useful in handling the light traffic.
- In this section we will discuss the scheduling approaches to the medium access control.

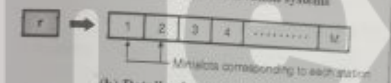
- There are three important approaches in the scheduling approach as follows :
  1. Reservation system
  2. Polling system
  3. Token passing ring networks.

9.6.1 Reservation Systems :

- The principle of reservation system can be understood from Fig. 9.6.1.
- In this system each station transmits a single packet at the full rate R bps. The transmissions from the stations can be organized into frames of variable length.
- Before each frame a reserved slot or reservation interval is transmitted as shown in Fig. 9.6.1(a).



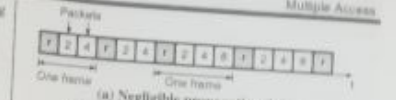
(a) Transmission in reservation systems



(b) Details of reservation interval

(L-733)**Fig. 9.6.1 : Basic reservation system**

- Fig. 9.6.1(b) shows the details of the reservation interval "r". The reservation interval consists of M minislots with one slot allotted to each station.
- These minislots are used by the stations to indicate that they have a packet to transmit in the corresponding frame.
- The station that wants to transmit packet by broadcasting their reservation bit during the appropriate minislot.
- All the stations will listen to the reservation interval, and then determine the order in which packet transmissions in the corresponding frame would take place.
- The frame length would correspond to the number of stations which have a packet to transmit.
- If the length of the packet is variable, then it can be handled if the reservation message includes packet length information.
- This reservation system that we discussed is called as the basic reservation system.
- The basic reservation system can be improved by using the time division multiplexing scheme. In the improved the time division multiplexing system the idle time slots are allotted to the reservation system the idle time slots are allotted to the other stations.
- The operation of the basic reservation system can be explained with the help of Fig. 9.6.2.



(a) Negligible propagation delay

(b) Non negligible propagation delay

(L-736)**Fig. 9.6.2 : Operation of reservation system with negligible and non-negligible delays**

- Refer Fig. 9.6.2(a) which shows a system with negligible propagation delay. In the first frame, only the stations 2 and 4 transmit their packets. But in the middle portion, station 8 also wants to transmit its packet. So the frame gets expanded from two slots to three slots.
- The maximum throughput from this system can be attained when all the stations transmit their packet in each frame.
- The corresponding maximum throughput is given by,

$$\rho_{max} = \frac{1}{1+v} \quad \text{... for one packet reservation/minislot}$$

- If $v << 1$ then the value of ρ_{max} can be very high.
- Now refer Fig. 9.6.2(b) which shows a reservation system with some finite non zero propagation delay which cannot be neglected. In this system the stations will transmit their reservations in the same way as they used to do before.
- It is possible to modify the basic reservation system so that stations can reserve more than one slot per packet transmission per minislot.
- Let us assume that a minislot can reserve say upto k packets.

Then the maximum achievable throughput is given by,

$$\rho_{max} = \frac{k}{1+(v/k)} \quad \text{... for k packet reservation/minislot}$$

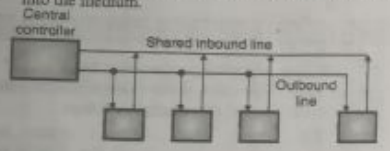
- Note that this value of ρ_{max} will be higher than that for the single packet reservation/minislot.

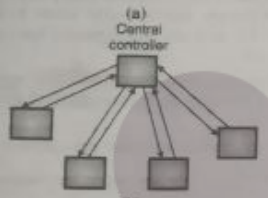
Effect of number of stations (M) :

- The reservation intervals introduce overhead which is proportional to M. That means the reservation interval becomes M * v.
- As the number of stations (M) become very large, this overhead will become significant. This then becomes a serious problem.
- This problem can be sorted out by not allocating a minislot to each station and then instead making the stations to compete for a reservation of minislot by using a random access technique such as ALOHA or slotted ALOHA.

9.6.2 Polling :

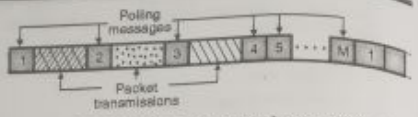
- Now consider polling system shown in Fig. 9.6.3. In this system the stations access the common medium one by one (by taking turns).
- At any given time only one of the stations will transmit into the medium.





(L-736)**Fig. 9.6.3 : Examples of polling systems**

- When a station finishes its transmitting, then some mechanism is used to pass the right of transmission to another station which wants to transmit next.
- There are different ways of passing the right of transmission from one station to the other station.
- Fig. 9.6.3(a) shows a scheme in which M stations communicate with a central controller. The outbound line is used for carrying the information from the central controller to the M users whereas the shared inbound line is required to carry the information from users to the central computer.
- Thus the inbound line acts as the shared medium that requires a medium access control (MAC).
- The host computer acts as a central controller. It sends control messages which co-ordinate the transmissions from the stations.
- The central controller sends a polling message to a particular station. That station sends its message on the shared inbound line. Once this process is over, the station gives a go-ahead message.
- It is possible that the central controller may poll the stations in a round robin (serial) fashion or it may do it according to some pre-determined rule.
- Fig. 9.6.3(b) shows another system where it is possible to use polling. The central controller of this system can make use of radio transmission.
- Fig. 9.6.4 shows the sequence of polling messages.

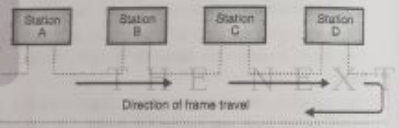


(L-736)**Fig. 9.6.4 : Polling messages and transmissions in a polling system**

- Station 1 gets the polling message first. The polling message will propagate. It is received by all stations but only station 1 begins transmission. All this process needs a time called **walk time.**
- The next period is occupied by the transmission from station 1.
- This period will then be followed by the walk time corresponding to station-2. This process will continue until all the M stations are polled. Thus in this system the stations are polled in the round robin manner.
- The walk time can be considered to be an overhead in the polling system because it is an unproductive time. The total walk time τ' is the sum of walk time corresponding to each station.

9.6.3 Token Passing :

- Token is a special frame which is used to authorize a particular station for transmission.
- In the token passing method, the token is given to that station, which is authorized to send its data. Thus the station that has the token with it can transmit others listen.



(L-737)**Fig. 9.6.5 : Token passing network**

- In a token passing network, each station has a predecessor and successor as shown in Fig. 9.6.5.
- The frames travel in one direction. They come from the predecessor and go to the successor as shown in Fig. 9.6.5.
- A token frame is circulated around the ring when no data is being transmitted and the line is idle.
- The stations which are ready to send data, will wait for the token. As the token circulates the first ready station in the ring will grab the circulating token and transmit one or more frames.
- This station will keep sending the frames as long as it has frames to send or the allotted time is not complete.
- It then passes this token on the ring from which the next ready to transmit station will grab it.

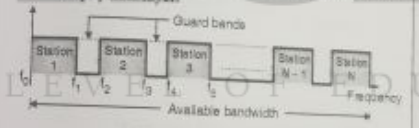
- This is the simplest possible token passing technique in which all the stations have equal priority or right to send.
- In the practical system, some other features such as priority and reservation are added.

9.7 Channelization :

- This is a multiple access method in which the total bandwidth of the common link is shared in the frequency domain, the time domain or through codes.
- Depending on the method of sharing there are three channelization techniques :
  1. FDMA : Frequency Division Multiple Access
  2. TDMA : Time Division Multiple Access
  3. CDMA : Code Division Multiple Access.

9.7.1 FDMA :

- In the frequency division multiple access (FDMA), the available channel (medium) bandwidth is shared by all the stations. That means each station will have its own specific slot reserved in the entire channel bandwidth.
- So each station uses its allocated frequency band to send its data. Each band is thus reserved for a specific station. e.g. the frequency band f_0 to f_1 is for station-1, then f_2 to f_3 is for station-2 and so on.
- The concept of FDMA is illustrated in Fig. 9.7.1.
- FDMA is a data link layer protocol which uses FDM at the physical layer.



(L-739)**Fig. 9.7.1 : Concept of FDMA**

- Guard bands are provided in between the adjacent frequency slots. e.g. $(f_1 - f_2)$ is a guard band between the bands allotted to stations 1 and 2. Guard bands avoid the adjacent channel interference.
- FDMA is used in cellular phones and satellite networks.

Advantages of FDMA :

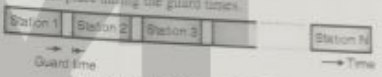
1. All the stations can operate continuously all 24 hours without having to wait for their turn to come.
2. The power required for transmission depends on the number of channels being transmitted.
3. The signal to noise ratio is improved due to the use of FM.
4. No synchronization is necessary.

Disadvantages of FDMA :

1. Each channel or earth station can use only a part of the total satellite bandwidth.
2. Inspite of guard bands being provided, there is some adjacent channel interference present.
3. As FM is used, it requires larger bandwidth, hence less number of channels will be accommodated in the bandwidth of a satellite.
4. Due to the nonlinearity of companders, the intermodulation products are generated.

9.7.2 TDMA :

- TDMA stands for Time Division Multiple Access.
- In TDMA, the entire bandwidth can be used by every user (station) but not simultaneously.
- A station can use the entire bandwidth only for the allocated time slot.
- Thus each channel is allocated a time slot only during which it can send its data. Thus the time is shared, frequency band is not shared.
- Fig. 9.7.2 illustrates the concept of TDMA. Guard times are inserted between the adjacent time slots in order to prevent any cross talk. Not data transmission takes place during the guard times.



(L-740)**Fig. 9.7.2 : Concept of TDMA**

- TDMA is a data link layer protocol which uses TDM at the physical layer.
- TDMA finds its application in cellular phones and satellite networks.

Advantage of TDMA :

Since only one station is present at any given time, the generation of intermodulation products will not take place.

Disadvantage of TDMA :

TDMA needs synchronization which makes it more complicated as compared to FDMA.

9.7.3 Code Division Multiple Access (CDMA) :

- An alternative to FDMA and TDMA is an another system called Code Division Multiple Access (CDMA). The most important feature of CDMA is as follows :

> In CDMA more than one user is allowed to share a channel or subchannel with the help of direct-sequence spread spectrum (DS-SS) signals.

- In CDMA each user is given a unique code sequence or signature sequence. This sequence allows the user to spread the information signal across the assigned frequency band.
- At the receiver the signal is recovered by using the same code sequence. At the receiver, the signals received from various users are separated by checking the cross-correlation of the received signal with each possible user signature sequence.

- In CDMA the users access the channel in a random manner. Hence the signals transmitted by multiple users will completely overlap both in time and in frequency.

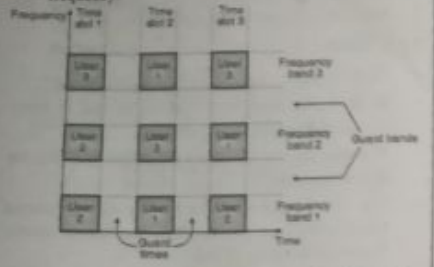


Fig. 9.7.3 : Structure of CDMA showing the guard bands and the guard times

- The CDMA signals are spread in frequency. Therefore the demodulation and separation of these signals at the receiver can be achieved by using the pseudorandom code sequence. CDMA is sometimes also called as Spread Spectrum Multiple Access (SSMA).
- In CDMA as the bandwidth as well as time of the channel is being shared by the users, it is necessary to introduce the guard times and guard bands as shown in Fig. 9.7.3.
- CDMA does not need any synchronization, but the code sequences or signature waveforms are required to be used.

9.7.4 Comparison of FDMA, TDMA and CDMA :

Sr. No.	FDMA	TDMA	CDMA
1.	Overall bandwidth is shared among many stations.	Time sharing takes place.	Sharing of bandwidth and time both takes place.
2.	Due to nonlinearity of devices inter modulation products are generated due to interference between adjacent channels.	Due to incorrect synchronization there can be an interference between the adjacent time slots.	Both type of interferences will be present.
3.	Synchronization is not necessary.	Synchronization is essential.	Synchronization is not necessary.
4.	Code word is not required.	Code word is not required.	Code words are required.
5.	Guard bands between adjacent channels are necessary.	Guard times between adjacent time slots are necessary.	Guard bands and Guard times both are necessary.

Review Questions

Q. 1 Explain the layered architecture of LAN explaining the function of the LLC and MAC sublayer.

Q. 2 What is static and dynamic channel allocation ?

Q. 3 Compare and explain the pure and slotted ALOHA system.

Q. 4 Explain the different CSMA protocols.

Q. 5 What is CSMA with collision detection ?

Q. 6 Explain the FDDI system.

Q. 7 What are the functions of a transceiver ?

Q. 8 Why there is no need of CSMA/CD for a full duplex Ethernet LAN ?

Q. 9 Explain CSMA/CD.

Q. 10 What is CSMA/CA ?

❑❑❑

CHAPTER 10

Unit III

Connecting Devices & Virtual LANs

Syllabus :
Connecting devices and virtual LANs, Connecting devices, Hubs, Link layer switches, Routers.

10.1 Network Connecting Devices :

- Different types of network connecting devices are as shown in Fig. 10.1.1.

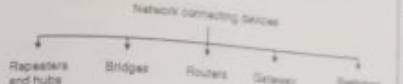


Fig. 10.1.1

- The relation between OSI reference model and various connecting devices is shown in Fig. 10.1.2.

Network connecting devices :

- Two or more devices are connected to each other for the purpose of sharing data or resources from a network.
- A LAN may be spread over a larger distance than its media can handle effectively. The number of stations also can be more than a number which can be handled and managed properly. Such networks should be subdivided into smaller networks and these smaller subnetworks should be connected to each other through connecting devices.
- A device called a repeater is inserted into the network to increase the coverable distance or a device called a bridge can be inserted for traffic management.
- When two or more separate networks are connected for exchanging data or resources it creates an internetwork. Routers and gateways are used for internetworking.
- Each of these device type interacts with protocols at different layers of the OSI model.
- Repeaters act only upon the electrical components of a signal and are therefore active only at the physical layer.
- Bridges utilize addressing protocols and can affect the flow control of a single LAN. Bridges are most active at the data link layer.
- Routers provide links between two separate but same type LANs and are active at the network layer.
- Finally gateways provide translation services between incompatible LANs or applications and are active in all of the layers. Connecting devices and the OSI model is shown in Fig. 10.1.2.

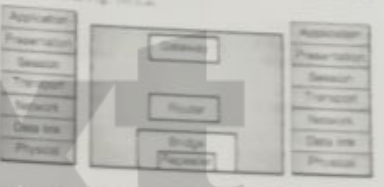


Fig. 10.1.2 : Connecting devices and OSI model

Categories of connecting devices

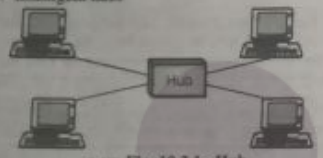
Fig. 10.1.2 shows the relationship between the connecting devices and various layers of the internet model.

Table 10.1.1 : Role of networking devices

Sr. No.	Name of the device	Role
1.	Passive hub	Operate below the physical layer.
2.	Repeater	Regenerates the original signal. Operates in the physical layer.
3.	Bridge	Bridges utilize the address protocol. They can carry out the traffic management. They are most active in the data link layer.
4.	Routers	Routers provide connections between two seperate but compatible networks. It works in the network layer.
5.	Gateways	Gateways provide translation services between incompatible networks and works in all the layers.

10.2 Hubs :

- The general meaning of the word hub is any connecting device. But its specific meaning is multiport repeater.
- It is normally used for connecting stations in a physical star topology.
- All networks require a central location to connect various segments of media coming from various nodes.
- Such a central location is called as a hub. A hub organises the cables and relays signals to the other media segments as shown in Fig. 10.2.1.
- There are three main types of hubs :
  1. Passive hubs 2. Active hubs
  3. Intelligent hubs



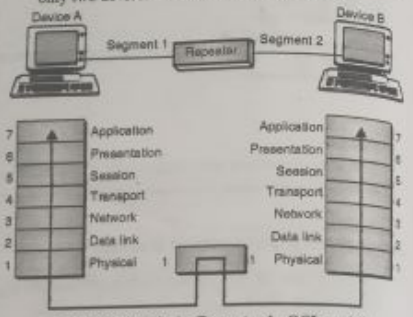
(G-350) **Fig. 10.2.1 : Hub**

10.2.1 Passive Hubs :

- A passive hub simply combines the signals of a network segments. There is no signal processing or regeneration. It merely acts as a connector.
- A passive hub reduces the cabling distance by half because it does not boost the signals and in fact absorbs some of the signal.
- With a passive hub, each computer receives the signals sent from all the other computers connected to the hub.
- This type of hub is a part of communication media. Hence its location is below the physical layer.

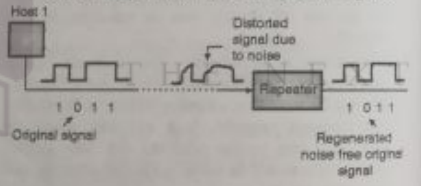
10.3 Repeaters :

- A repeater is a connecting device which can operate only in the physical layer.
- All transmission media weaken the electromagnetic waves that travel through them.
- Attenuation of signals limits the distance any medium can carry data. Devices that amplifies signals to ensure data transmission are called repeaters.
- A repeater receives a signal and before it gets attenuated or corrupted, regenerates the original signal.
- Thus we can use a repeater to extend the physical length of LAN as a shown in Fig. 10.3.1(a).
- Repeater is not an amplifier because amplifiers simply amplify the entire incoming signal along with noise.
- Signal – regenerating repeaters create an exact duplicate of incoming data by identifying it amidst the noise, reconstructing it and retransmitting only the desired information.
- The original signal is duplicated, boosted to its original strength and sent as shown in Figs. 10.3.1(a) and (b).
- A repeater does not connect two LANs. It connects only two devices connected in the same LAN.



(G-351) **Fig. 10.3.1(a) : Repeater in OSI model**

- It cannot connect two LANs of a different protocols.
- A repeater forwards every frame, it cannot filter out some frames and let the others pass through.
- A repeater should be placed at a precise point on the link. Such that the signal reaches it before the noise has induced an error in any of the transmitted bits.
- Fig. 10.3.1(b) illustrates the function of a repeater.
- Repeaters operate at the physical layer of the OSI model and they deal with the actual physical signals.



(G-352) **Fig. 10.3.1(b) : Function of a repeater**

Advantages of repeater :

1. Repeaters can regenerate the desired information.
2. They can reduce the effect of noise.
3. They can extend the network.
4. It reduces the number of errors introduced due to noise.

Disadvantages of repeater :

1. A repeater can not connect two LANs. It can only connect two devices connected in the same LAN.
2. It has no filtering capability.
3. Repeaters can operate only in the physical layer.
4. Repeaters must be placed at the precise point on the link so as to be effective.

10.3.1 Active Hubs :

- They are like passive hubs but have electronic components for regeneration and amplification of signals. By using active hubs the distance between devices can be increased. An active hub is equivalent to a multipoint repeater.
- The main drawback of active hubs is that they amplify noise as well along with the signals. They are more expensive than passive hubs as well.

10.3.2 Intelligent Hubs :

- In addition to signal regeneration, intelligent hubs perform some other intelligent functions such as network management and intelligent path selection.
- A switching hub chooses only the port of the device where the signal needs to go, rather than sending the signal along all paths.
- Hubs can also be used to create multiple levels of hierarchy as shown in Fig. 10.3.2.

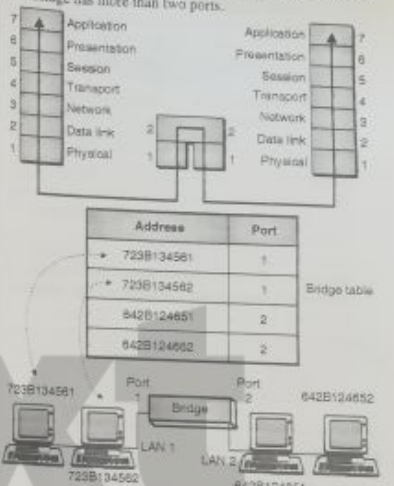
(L-446) **Fig. 10.3.2 : Hubs to create multiple levels of hierarchy**

10.4 Bridges :

- A bridge can operate in the physical layer as well as in the data link layer of the OSI model.
- It can regenerate the signal that it receives and it can check the physical (MAC) addresses of source and destination mentioned in the header of a frame.

Filtering :

- The major difference between the bridge and repeater is that the bridge has a filtering capability. That means a bridge will check the destination address of a frame and make a decision about whether the frame should be forwarded or dropped.
- If the frame is to be forwarded, then the bridge should specify the port over which it should be forwarded.
- In order to achieve this a bridge has a table relating the addresses and ports as shown in Fig. 10.4.1.
- If a frame for 723B134561 arrives at port 2 then the bridge goes through its table and understands that the frame is to be sent out on port 1 so it will do so.
- In Fig. 10.4.1 a two port bridge is shown but in reality a bridge has more than two ports.



(G-354) **Fig. 10.4.1 : Bridge and bridge table**

- It is important to note that the bridges do not change the physical address contained in the frame.

Types of bridges :

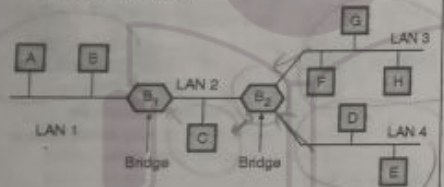
The bridges are of two types :

1. Transparent bridges and
2. Routing bridges.

- Transparent bridge is a bridge in which the stations are not at all aware of the existence of the bridge.
- Transparent bridges keep a table of addresses in memory to determine where to send data.
- The duties of a transparent bridge are as follows :
  1. Filtering frames
  2. Forwarding and
  3. Blocking.
- In source routing a sending station defines the bridges that should be visited by the frames.
- The addresses of these bridges are included in the frame. So a frame contains not only the source and destination address but also the bridge addresses.
- Source routing bridges are used to avoid a problem called looping. These bridges were designed for the token ring LANs. But these LANs are not very common now a days.

10.4.1 Transparent Bridge :

- A transparent bridge builds its table of station addresses on its own as it performs its bridge function. When this bridge is first installed, its table is empty.
- As it comes across each packet it looks at both the destination and source addresses.
- It checks the destination to decide where to send the packet. If it does not yet recognise the destination address it relays the packet to all of the stations on both segments.
- It uses the source address to build its table. As it reads the source address it notes which side the packet came from and associates that address with the segment to which it belongs.
- As an example, consider the configuration of Fig. 10.4.2. As shown in the Fig. 10.4.2 bridge B_1 is connected to LANs 1 and 2 and bridge B_2 is connected to LANs 2, 3 and 4.
- A frame arriving at bridge B_1 on LAN 1 destined for A can be discarded immediately because it is already on the right LAN, but a frame arriving on LAN 1 for C or F must be forwarded.



(L-648) **Fig. 10.4.2 : Configuration of bridge and LAN**

- When a frame arrives, a bridge must decide whether to discard or forward it, and if the latter is true, then decide on which LAN to put the frame.

Bridge learning :

- When a frame arrives at one of the ports of a bridge, it has to make a decision about forwarding the frame to another port. This decision is made based on the destination address of the frame.
- In order to make such decisions every bridge needs a table called **forwarding table** or **forwarding database**.
- This table indicates which side of the port the destination station is attached to, directly or indirectly. The format of a forwarding table is shown in Table 10.4.1.

Table 10.4.1 : Format of a forwarding table

MAC address	Port

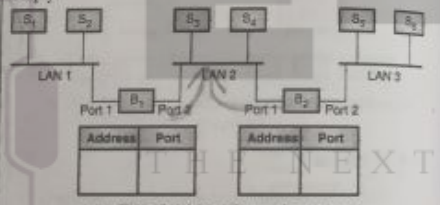
- Note that in practice there are a few thousand entries in a forwarding table.
- Let us see how to fill up these forwarding tables. It is filled up by a process called as "bridge learning".
- The basic bridge learning process is as follows :

Bridge learning procedure :

1. When a bridge receives a frame, it first compares the source address of the frame with each entry in the forwarding table. If no match is found, then the bridge will add this source address alongwith the port number on which the frame was received, to the forwarding table.
2. The bridge compares the destination address of the received frame with each entry in the forwarding table. If a match is found, then the bridge forwards the frame to the port indicated in the entry. But if this port is same as the one on which the frame was received, then the frame is discarded. Finally if a match is not found, then the bridge will send that frame on all its ports except the one on which the frame was received.

Example on bridge learning :

Consider the network shown in Fig. 10.4.2(a). Assume that forwarding tables of both the bridges are initially empty.



(L-649) **Fig. 10.4.2(a) : Example network**

1. **S_2 sends a frame to S_1 :**
   - If S_2 sends a frame to S_1, then B_1 compares the source address of the received frame with the existing entries. So here S_2 is the sender and S_1 is destination.
   - But there are no entries in B_1 table. So it adds the address of S_2 in its forwarding table as shown in Fig. 10.4.2(b).
   - Then B_1 compares the destination address of the received frame with the existing entries. But the table is empty. So the bridge B_1 thinks of flooding the frames. But then it understands that the **destination S_1 is connected on the same port (Port 1) on which the frame has been received.**
   - So B_1 will note down the address of S_2 in its table and **discard the frame.** This is because bridge B_1 is not required to be used when a communication between S_1 and S_2 is to be made.

- The traffic is now completely isolated in LAN 1, and the updated bridge tables are shown in Fig. 10.4.2(b).

B_1

Address	Port
S_2	1
S_1	1

B_2

Address	Port

Fig. 10.4.2(b) : Forwarding tables after $S_2 \rightarrow S_1$

2. **S_5 transmits to S_4 :**
   - The two stations correspond to two different LANs. S_5 is the sender and S_4 is the destination.
   - First B_2 records the address of S_5 and port number (Port 2) because the address of S_5 is not found in its forwarding table.
   - Then B_2 checks the destination address. Since there are no entries, it will add S_4 and port 1 in its table as shown in Fig. 10.4.2(c). Bridge B_2 will forward the frame to port 2 of B_1 as well as to LAN 2 where S_4 will receive it.
   - When this frame arrives at port 2 of B_1 it also adds the source address i.e. S_5 and port 2 in its table as shown in Fig. 10.4.2(c).
   - However the destination address (S_4) is on the same port (2) of B_1 on which it has received the frame. So it will note down S_4 and port 2 in its table but **discard** the frame.

B_1

Address	Port
S_2	1
S_1	1
S_5	2
S_4	2

B_2

Address	Port
S_5	2
S_4	1

$S_2 \rightarrow S_1$, $S_5 \rightarrow S_4$, $S_5 \rightarrow S_4$

Fig. 10.4.2(c) : Forwarding tables after $S_5 \rightarrow S_4$

- The table entries for the remaining transmissions are given in Figs. 10.4.2(d) and (e).

3. **S_3 transmits to S_5 :**

B_1

Address	Port
S_2	1
S_1	1
S_5	2
S_4	2
S_3	2

B_2

Address	Port
S_5	2
S_4	1
S_3	2

$S_2 \rightarrow S_1$, $S_5 \rightarrow S_4$, $S_7 \rightarrow S_5$, $S_3 \rightarrow S_4$, $S_3 \rightarrow S_5$

Fig. 10.4.2(d) : Tables after $S_3 \rightarrow S_5$

4. **S_1 transmits to S_2 :**
   No change in the tables.
5. **S_6 transmits to S_5 :**

B_1

Address	Port
S_2	1
S_1	1
S_5	2
S_4	2
S_3	2
S_7	2

B_2

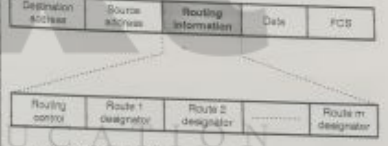
Address	Port
S_5	2
S_4	1
S_3	2
S_6	2

$S_2 \rightarrow S_1$, $S_5 \rightarrow S_4$, $S_3 \rightarrow S_4$, $S_7 \rightarrow S_5$, $S_3 \rightarrow S_5$, $S_6 \rightarrow S_5$

Fig. 10.4.2(e) : Table after $S_6 \rightarrow S_5$

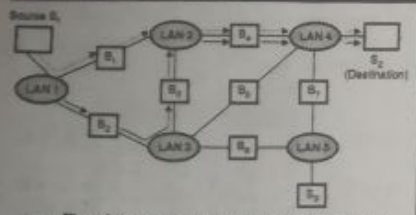
10.4.2 Source Routing Bridges :

- The source routing bridges were developed by the IEEE 802.5 committee and they are used basically to interconnect token ring networks.
- The main idea of source routing is that each station should determine the route to the destination when it wants to send a frame and therefore include the route information in the header of the frame.



(L-664) **Fig. 10.4.3 : Frame format for source routing**

- The frame format for source routing is shown in Fig. 10.4.3.
- Note that the routing information field is inserted only if the two communicating stations are on different LANs.
- Fig. 10.4.4 shows the LAN interconnection with source routing bridges. If station-1 wants to send a frame to station-2 then a possible route can be LAN-1 $\rightarrow B_1 \rightarrow$ LAN 2 $\rightarrow B_4 \rightarrow$ LAN 4.
- Many more routes are available for the same source destination pair.
- In general when a station wants to transmit a frame to another station on a different LAN, the station consults its routing table.
- If the route to the destination is found, then the station simply inserts the routing information into the frame.



(L-425) **Fig. 10.4.4 : LANs interconnected with source routing bridges**

How to discover a route ?

To discover a route the basic idea is as follows :

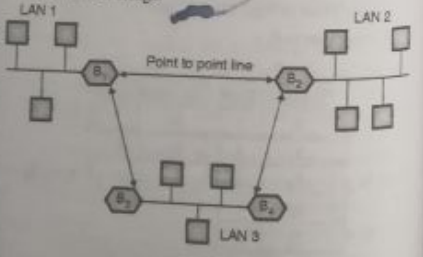
1. The station who wants to discover a route first broadcasts a special frame called single route broadcast frame.
2. This frame will visit every LAN exactly once and eventually reaches the destination.
3. Then the destination station responds with another special frame called the all routes special frame which generates all possible routes back to the source station.
4. After collecting all routes the source chooses the best possible route and saves it.

10.4.3 Comparison of Transparent and Source Routing Bridge :

Sr. No.	Parameters	Transparent bridge	Source routing bridge
1.	Ability to reconfigure	High. Bridges keep information on location of stations.	High. Each station must learn the route to its destination before sending.
2.	Stations responsibilities	None. They just send the frames and let the bridges do the work.	They determine and maintain addresses.
3.	Bridges requirements	Routing tables and the ability to both update them and execute a spanning tree algorithm.	Ability to broadcast or forward, depending on routing designators and ability to execute a spanning tree algorithm.
4.	Routes used	Always along the spanning tree, but not necessarily the cheapest.	Stations can choose the cheapest routes to one another.
5.	Dependence on topology	None. Bridges learn where stations are relative to their ports dynamically and stations have no need to know.	Some Bridges respond to routing information and spanning tree algorithms, but stations must determine a route to a destination.
6.	Orientation	Connectionless	Connection-oriented
7.	Configuration	Automatic	Manual
8.	Failures	Handled by the bridge	Handled by the host
9.	Complexity	In the bridge	In the hosts.

10.4.4 Remote Bridges :

- If bridges are used to connect LANs, having large distance between them they are called remote bridges. Many point to point links can be used to connect these bridges as shown in the Fig. 10.4.5.
- Various protocols can be used on these point to point lines. One of them is to use a point to point data link protocol (PPP), putting complete MAC frames in the payload field.
- Another option is to strip off the MAC header and trailer at the source bridge and put what is left in the payload of the point to point protocol. A new MAC header and trailer can then be generated at the destination bridge.



(L-426) **Fig. 10.4.5 : Configuration of remote bridges**

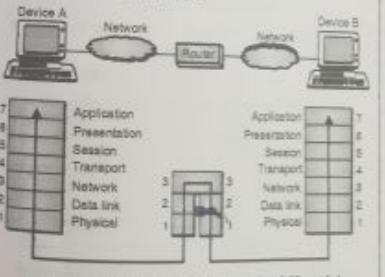
10.4.5 Bridges Connecting Different LANs :

Ideally a bridge should be able to connect LANs that use different protocols at the data link layer. For example, wired LAN and wireless LAN. But in practice the following issues are needed to be considered.

1. Frame format
2. Maximum data size
3. Bit order
4. Data rate
5. Security issues
6. Multimedia support

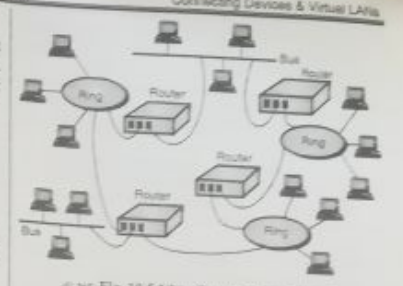
10.5 Routers :

- Routers are devices that connect two or more networks as shown in Figs. 10.5.1(a) and (b). They consist of a combination of hardware and software.
- The hardware can be in the form of a network server, a separate computer or a special device, as well as the physical interfaces to the various networks in the internetwork.
- Various types of networks can be interconnected through routers as shown in Fig. 10.5.1(b).
- The software in a router are the operating systems and the routing protocol. Management software can also be used.
- Routers use logical and physical addressing to connect two or more logically separate networks.
- The large network is organized into small network segments called as subnets and these subnets are interconnected via routers.



(G-364) **Fig. 10.5.1 (a) : A router in the OSI model**

- Each of the subnet is given a logical address. This allows the networks to be separate but still access each other and exchange data.
- Data is grouped into packets, or blocks of data. Each packet has a physical device address as well as logical network address.
- The network address allows routers to calculate the optimal path to a workstation or computer.

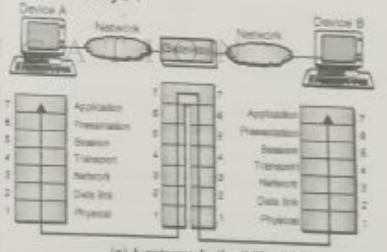


(G-365) **Fig. 10.5.1(b) : Routers in an internet**

- Route discovery is the process of finding the possible routes through the internetwork and then building routing tables to store that information. The two methods of route discovery are :
  1. Distance vector routing
  2. Link state routing

Note :
- Routers work at the network layer of the OSI model.
- With static route selection, packets always follow a pre-determined path.

10.6 Gateways :



(a) A gateway in the OSI model

(b) A gateway

(G-366) **Fig. 10.6.1**

- When the networks that must be connected are using completely different protocols from each other, a powerful and intelligent device called a gateway is used.

- A gateway is a device that can interpret and translate the different protocols that are used on two distinct networks as shown in Figs. 10.6.1(a) and (b).
- Gateways comprise of software, dedicated hardware or a combination of both. Gateway operate through all the seven layers of the OSI model and all five layers of the internet model.
- A gateway can actually convert data so that it works with an application on a computer on the other side of the gateway. For e.g. a gateway can receive e-mail message in one format and convert them into another format.
- Gateways can connect systems with different communication protocols, languages and architecture. For e.g. IBM networks using Systems Network Architecture (SNA) can be connected to LANs using a gateway.

Note : Gateways are slow because they need to perform intensive conversions.

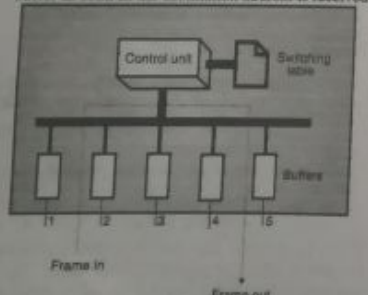
10.7 Switches :

- A switch is a device which provides bridging functionality with greater efficiency. A switch acts as a multiport bridge to connect devices or segments in a LAN.
- The switch has a buffer for each link to which it is connected. When it receives a packet, it stores the packet in the buffer of the receiving link and checks the address to find the outgoing link.
- If the outgoing link is free, the switch sends the frame to that particular link.

Switches are of two types :

1. Store - and - forward switch
2. Cut - through switch.

- A store - and - forward switch stores the frame in the input buffer until the whole packet has arrived.
- A cut-through switch, forwards the packet to the output buffer as soon as the destination address is received.



(G-367) **Fig. 10.7.1 : Switch**

- Concept of a switch is shown in Fig. 10.7.1. As shown in the Fig. 10.7.1 a frame arrives at port 2 and is stored in the buffer.
- The CPU and the control unit, using the information in the frame consult the switching table to find the output port. The frame is then sent to port 5 for transmission.

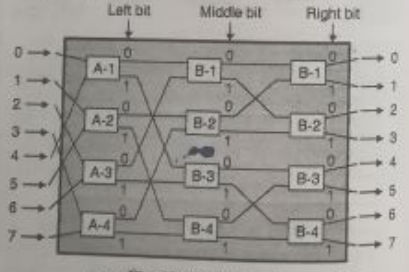
Note : Routing switches use the network layer destination address to find the output link to which the packet should be forwarded.

10.7.1 Two Layer Switch :

- The switches can be of two types namely the two layer switches and the three layer switches.
- A two layer switch operates at the physical as well as data link layer.
- The two layer switch is basically a bridge. It has many ports and it is designed to allow better performance.
- A bridge with few ports is used for connecting a few LANs together. But a bridge with many can allocate a unique port to each station. Thus each station will have its own separate identity.
- Therefore there is no competing traffic and so there are no collisions.

10.7.2 Three Layer Switch :

- A three layer switch is used at the network layer and it is a kind of router.
- A three layer switch is shown in Fig. 10.7.2.
- It has n = 8 inputs and same number of outputs. A three bit number is used to decide the internal path over which the input is passed to output.
- The number of microswitches at each stage is n/2 i.e. 4 switches.



(G-368) **Fig. 10.7.2 : A three layer switch**

- The first stage routes the cell based on the high order bit in the binary bit string.
- The second stage routes the cell based on the middle bit and last stage routes it based on the low order bit.
- Note that number of stages = $\log_2 (n) = \log_2 8 = 3$.

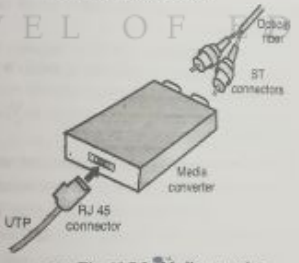
10.7.3 Comparison of Hub and Switch :

Sr. No.	Hub	Switch
1.	It is a broadcast device.	It is a point to point device.
2.	It operates at physical layer.	It operates at datalink layer.
3.	It is not an intelligent device.	It is an intelligent device.
4.	It simply broadcasts the incoming packet.	It uses switching table to find the correct destination.
5.	It can not be used as a repeater.	It can be used as a repeater.
6.	Not a sophisticated device.	It is a sophisticated device.
7.	Not very costly.	Costly.

10.7.4 Media Converters :

- It is a device which is used to connect two different networking media. For example connection between a shielded cable and twisted pair can be achieved through the media converter.
- Fig. 10.7.3 shows the media converter.
- Various functions performed by a media converter are as follows :

1. Connect two different types of wiring systems without an additional repeater.



(G-369) **Fig. 10.7.3 : Media converter**

2. Connect two 10 Base T or 100 Base TX networks in different buildings using fiber-optic cable.
3. To allow the use of different types of cables such as UTP cabling, thin net fiber optic cable, thick net etc. in a single network.

10.7.5 Comparison of Router and Bridge :

Sr. No.	Parameter	Router	Bridge
1.	Layer in OSI model.	Network layer	Physical or data link.
2.	Operation.	Connect two or more network.	Regeneration, check MAC address.
3.	Types.	Distance vector, Link state	Transparent, Routing.
4.	Principle of working.	Uses hardware and software.	Uses tables relating the addresses and ports.
5.	Used for	Connecting networks	Connecting computers.

10.7.6 Comparison of Bridge, Switch and Hub :

Sr. No.	Parameter	Hub	Switch	Bridge
1.	Type of device	Broadcast	Point to point	Both
2.	Layer of operation	Physical	Data link	Physical and data link
3.	Intelligence	Not intelligent	Intelligent	Highly intelligent
4.	Duties	Simply broadcast the incoming packet	Uses switching table to find correct destination	Filtering, forwarding and blocking of frames
5.	Sophistication	Low	High	Very high
6.	Cost	Low cost	Expensive	Very expensive

10.7.7 Comparison of Bridges, Routers and Switches :

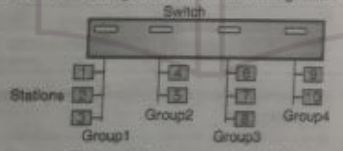
Table 10.7.1

Sr. No.	Parameter	Router	Bridge	Switch
1.	Layer in OSI model	Network layer	Physical or data link	Data link and network layer
2.	Type of device	Point to point	Point to point or broad cast	Point to point
3.	Operation	It connects two or more networks	It regenerates, checks MAC address	It provides bridging operation with greater accuracy

Sr. No.	Parameter	Router	Bridge	Switch
4.	Types	Distance vector, link state	Transparent, Routing	Two layer, three layer.
5.	Intelligence	Highly intelligent	Highly intelligent	Highly intelligent
6.	Used for	Connecting networks	Filtering forwarding and blocking frames.	Uses switching table to find correct destination.

10.8 Virtual LANs :

- The virtual local area network (VLAN) is a LAN configured not by physical wiring like conventional LAN but it is configured by software.
- It is developed in order to establish a connection between two stations belonging to two different physical LANs.
- The concept of VLAN technology is based on dividing a LAN into logical instead of physical segments. A LAN can be divided into several logical LANs called VLANs.
- The concept of VLAN will be clear after referring to Fig. 10.8.1 which shows the conventional switched LAN.
- The total number of stations are grouped into 4 groups and the groups are connected by a switch.
- The LAN is configured to allow this arrangement.

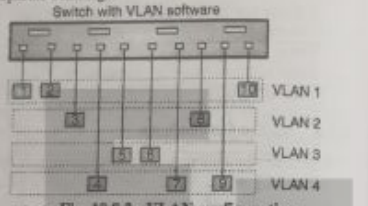


(L-746) **Fig. 10.8.1 : A conventional LAN**

- In a conventional LAN if we wish to transfer stations 1 and 2 from group 1 to any other group, then the LAN configuration needs to be changed. Rewiring would be required.
- This is the biggest problem in the conventional LAN. But the problem is solved by using a VLAN.

10.8.1 VLAN Configuration :

- Fig. 10.8.2 shows the configuration of VLAN connecting the same ten stations in four VLAN segments.
- The total stations are divided in four logical instead of physical segments. A LAN can be divided into several logic LANs called VLANs.
- It is possible to move one station from one group to any other group without changing the physical configuration because the membership of a group is defined by software and not by hardware.
- Any station can be moved logically to another VLAN.
- All the members corresponding to a VLAN can receive the broadcast message broadcast sent to that VLAN.
- Using the VLAN technology it is even possible to group stations connected to different switches.
- VLAN is very useful for a company having two separate buildings.



(L-746) **Fig. 10.8.2 : VLAN configuration**

- Thus VLANs group stations belonging to one or more physical LANs into broadcast domain. The stations in a VLAN communicate with each other as if they belong to the physical segment.

10.8.2 Membership :

- There are various characteristics used to group various stations in a VLAN.
- Some of such characteristics are port number, MAC addresses, IP addresses, IP multicast address. We can use them singly or a combination of two or more characteristics for grouping various stations in VLAN.

1. Port number :

- An administrator may define that the stations connected to ports 1, 3, 5, 7 of a switch correspond to VLAN-1, the stations connecting to ports 2, 4, 6, 8 correspond to VLAN-2 etc. So in this case the port number of the switch is being used as a membership characteristic.

2. MAC addresses :

- The 48 bit MAC address is used as a membership characteristics by some vendors. The stations having particular MAC addresses are grouped together to form VLAN-1. The stations having some other MAC addresses form VLAN-2 etc.

3. IP addresses :

- The 32-bit IP addresses are used as membership characteristics by some vendors. The principle is same as that for the MAC addresses.

4. Multicast IP addresses and combinations :

- Some vendors use the multicast IP addresses as a membership characteristics and some of them use the combination of all the characteristics mentioned earlier to form the VLANs.

10.8.3 Configurations :

The grouping of stations in a VLAN can be carried out by using one of the following configurations :

1. Manual configuration
2. Automatic configuration
3. Semiautomatic configuration.

1. Manual configuration :

- In the manual configuration, the network administrator assigns the stations to different VLANs using VLAN software but manually.
- If required, the migration of a station from one VLAN to the other also takes place manually.
- Since VLAN is a logical configuration, all the assignments and migrations take place via VLAN software.

2. Automatic configuration :

- In an automatic configuration, the stations are brought into or taken out of the VLANs automatically.
- The criteria for connection or otherwise is defined by the administrator.

3. Semi automatic configuration :

- This configuration is in between the manual and automatic configurations. The initialization process is done manually whereas the migration is an automatic process.

10.8.4 Communication between Switches :

- In the backbone using multiple switches each switch is supposed to know the stations and their VLANs and the membership of stations connected to each switch.
- Methods devised for the purpose of communication between switches are as follows :

1. Table maintenance
2. Frame tagging.
3. Time division multiplexing.

1. Table maintenance :

- This method works in the following manner. When a station sends its broadcast frame to all the other members of its own group, the switch will create an entry in a table and note down the membership number of the broadcasting station.
- The modified tables are sent by the switches to each other periodically for updating.

2. Frame tagging :

- In this method, when a frame is moving from one switch to the other, an extra header is added to its MAC frame.
- This extra header defines the destination VLAN. The receiving switches then use this header to determine the destination VLANs. The extra header is known as frame tag.

3. Time division multiplexing (TDM) :

- In this method, the connection between switches is done on the basis of time sharing of channels which is the principle of TDM.
- If there are six VLANs connected in a backbone network, then each trunk (connection) is divided into six channels.
- Channel 1 is designated to VLAN-1 channel 2 to VLAN-2 and so on.

IEEE standard :

- In late 1998 the IEEE 802.1 committee passed a new standard called 802.1 Q to define the format for frame tagging.
- This standard defines the format used in multiswitched backbones as well.
- 802.1 Q is the first step towards the future standardization and most vendors have already accepted this standard.

10.8.5 Advantages of VLANs :

Some of the advantages are

1. Reduction in cost.
2. Saving of time required for rewiring.
3. No need to change the physical configuration.
4. VLANs provide additional security. The message broadcast in one group cannot be listened by members of other groups.

Review Questions

- Q. 1 State and discuss various types of connectors.
- Q. 2 What is NIC ?
- Q. 3 Write a note on Transceivers.
- Q. 4 Explain the function of repeaters. Is it as amplifier ?
- Q. 5 Compare repeater and hub.
- Q. 6 Write a note on : Bridges.
- Q. 7 What is the uses of bridge table ?
- Q. 8 With the help of suitable explanatory diagram, explain the routers and gateways.
- Q. 9 Explain different types of switches.
- Q. 10 Compare switches and hub.
- Q. 11 What is backbone network ? What are its types ?
- Q. 12 Explain the following : 1. Bus backbone 2. Star backbone.

□□□

CHAPTER 11

Unit III

Network Layer

Syllabus :

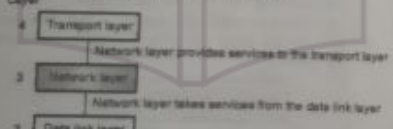
Introduction to network layer, Network layer services, Packetizing, Routing and forwarding, Other services, IPv4 addressing, Address space, Classful addressing, Unicast routing, General idea, Least cost routing, Routing algorithms, Distance vector routing, Link state routing, Path vector routing.

11.1 Network Layer :

- The network layer is responsible for carrying the packet from the source all the way to destination. In short it is responsible for host-to-host delivery.
- The network layer has a higher responsibility than the data link layer, because the data link layer is only supposed to move the frames from one end of the wire to the other end.
- Thus network layer is the lowest layer that deals with the end-to-end transmission.

Position of network layer :

- Fig. 11.1.1 shows the position of network layer in the 5 layer internet model. It is the third layer.



(G-433) **Fig. 11.1.1 : Position of network layer**

- It receives services from the data link layer and provides services to the transport layer.
- The network layer was designed to solve the problem of delivery through several links. The network layer is also called as the **Internetwork** layer.
- In addition to the host-to-host delivery the network layer is also responsible for **routing** the packets through the router.
- As a pure concept we can imagine that the Internet is a black box which connects a very large number of computers in the entire world together.
- But the Internet also is not a single network. It is infact the network of many networks or links.
- That means the Internet is an **internetwork** which is actually a combination of LANs and WANs.
- All these LANs and WANs are connected to each other via a connecting device such as a **router** which acts as a switch.

Routers :

Routers have many ports or interfaces. When it receives a packet at one of its ports, it forwards the packet through another port to the next switch or the final destination.

11.2 Network Layer Services :

- The duty of the network layer in TCP/IP is to provide the host-to-host delivery of datagrams.
- In this section we are going to discuss the services that are expected from the network layer.
- At the sending end, the network layer will accept a packet from its transport layer, encapsulate the packet into datagram and will deliver the packet to the data link layer.
- At the destination, exactly opposite process takes place. That means, at the destination the received datagram is decapsulated to extract the packet from it and the packet is delivered to the transport layer.

11.2.1 Packetizing :

- **Packetizing** is the first duty of the network layer in which it encapsulates the payload (data received from the transport layer) in a packet at network layer at the source. Then at the destination the decapsulation process takes place.
- In the way the network layer is doing the job of a postal service in delivering the packages from source to destination.

At the source :

At the sending end the events take place in the following sequence :

1. The payload (data) from the upper layer is received.
2. A header containing the source and destination address and some other information is added to the payload.
3. This packet is then delivered to the data link layer.
4. If the payload is too large, then the host carries out **fragmentation** on it. Otherwise the host is not allowed to modify the contents of the payload.

At the destination :

The sequence of events taking place at the destination is as follows :

1. The network layer packet is received from the data link layer.
2. The received packet is decapsulated and the payload is delivered to the upper layer protocol.
3. If a large packet is fragmented by either the source host or a router, then the responsibility of the network layer at the destination is to wait until all fragments are received, reassemble them and deliver them to the upper layer protocol.

11.2.2 Router's Role :

- The router's present in between the source and destination are supposed to check the source and destination addresses in the packet in order to forward it to the next network on the path.
- The router is not allowed to decapsulate the received packet unless it is too big and fragmentation needs to be carried out on it.
- The routers are not supposed to change the source and destination addresses.
- In the event of fragmentation, a router has to copy the header in all the fragments.

11.2.3 Routing and Forwarding :

The other two important duties of the network layer, which are related to each other are routing and forwarding.

Routing :

- The responsibility of the network layer is to route the packets from its source to destination. The physical network through which the packets travel consists of LANs, WANs and routers.
- Due to this the source and destination are connected to each other via more than one routes.
- It is the responsibility of the network layer to find the best route out of all the possible routes.
- In order to achieve this goal, the network layer must have some concrete strategy for defining the best route.
- In the modern days, this is done by running an appropriate routing protocol which helps the routers to co-ordinate their knowledge about the neighbouring routers and prepare routing tables which can be used on the arrival of a packet.
- These routing protocols should be run before commencement of any communication.

Forwarding :

- We can define the process of forwarding as the action taken by a router when it receives a packet at one of its interfaces.
- A router takes such an action with the help of the decision making tables called as **forwarding table** or **routing table.**
- When a packet arrives at one of the interfaces of a router from one of the attached network, the router has to forward it to another attached network.
- The router has to make this decision with the help of piece of information present in the packet header.
- This piece of information can be the **destination address** or a **label.** The router can use this information to find the corresponding output interface number in the **forwarding table.**

11.3 Other Services :

- The other services expected from the network layer are as follows :
  1. Error control
  2. Flow control
  3. Congestion control
  4. Quality of service (QoS)
  5. Security

Let's discuss these one by one.

11.3.1 Error Control :

- Eventhough it is possible to implement the error control at the network layer level, the design engineers have neglected this issue.
- One possible reason for this is that the packets may get fragmented at every router due to which the error checking becomes inefficient.
- However a **checksum** field has been added to the datagram in order to control any corruption in the header only. The error control is not applicable to the whole datagram.
- Thus there is no direct error control provided by the network layer in the Internet. But an auxiliary protocol ICMP is used by the Internet for providing some error control to the datagram.

11.3.2 Flow Control :

- The purpose of providing the flow control is to regulate the data rate of the source so as to avoid the receiver getting overwhelmed.
- The receiver will be overwhelmed if the upper layers at the sending end are producing data at a rate which is higher than the rate at which the upper layers at the destination can consume it (data).

- So as to control the sender's data rate, some kind of a feedback mechanism should be setup so that the receiver can tell the source that it (receiver) has overwhelmed with excess data.
- It is important to remember that the network layer does not directly provide any flow control
- The flow control is not provided at the network layer level because it is provided for most of the upper layer protocols and there is no need to provide flow control again which makes the design of network layer complex.

11.3.3 Congestion Control :

- This is another important issue to be handled at the network layer. **Congestion** will take place if the source computer sends more datagrams than the capacity of the network or routers.
- In this situation, the routers will drop some of the received packets.
- But this will make the congestion worse because the error control mechanism present at the upper layers will retransmit the packets dropped by the routers.
- Sometimes the congestion becomes so bad that the system collapses and no datagrams are delivered at the destination.
- The congestion control at the network layer is never implemented in the Internet.

11.3.4 Quality of Service (QoS) :

- The quality of service in the Internet has become more important since new applications like multimedia communication have been introduced.
- The Internet has grown as it successfully provides the quality of service to support all the modern day applications.
- However the QoS provisions are not implemented in the network layer. They are mostly implemented in the upper layers.

11.3.5 Security :

- During the early days of the Internet, security was not a major design concern due to limited (small) number of users. Hence the network layer was designed without any security provisions.
- But security has become a big concern now. But network layer is connectionless. Hence to provide security at the network layer we need to have another virtual level in order to change the connectionless service to connection oriented one.
- The virtual layer is known as IPsec.

11.4 IPv4 Addresses :

- Each computer connected to the Internet should be identified uniquely. The identifier used for this purpose is called as the **Internet address** or IP address.
- The hosts and routers on the Internet have unique IP addresses.
- The current version of IP (Internet Protocol) is IPv4 whereas the advanced version is IPv6.
- The IPv4 address is a 32-bit address and it is used for defining the connection of a host or router to the Internet. **Thus an IP address is an address of the interface.**

11.4.1 Uniqueness of IP Addresses :

- The IP address is **unique** and **universal.** That means each IP address defines only **one connection** to the Internet.
- At any given time, no two devices connected to the Internet can have the same IP address.
- But if a device is connected to the Internet via two connections through two different networks, then it can have two different IP addresses.
- All the IPv4 addresses are 32 bit long and they are used in the source address and destination address fields of the IP header.
- The IP addresses for hosts are assigned by the network administrator. For Internet it has to be obtained from the network information center.

11.4.2 Address Space :

- The IPv4 protocol has an address space. It is defined as the total number of addresses used by the protocol.
- If N number of bits are used for defining an address then the address space will be 2^N addresses.
- For IPv4, N is 32 bits. Hence its address space is 2^{32} or 4, 294, 967, 296 (more than 4 billion). So theoretically more than 4 billion devices could be connected to the Internet.
- Thus **the address space** of IPv4 is 2^{32}.

11.4.3 Notation :

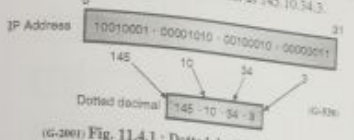
- The IPv4 addresses can be shown, use three different notations as follows :
  1. Binary notations (base 2).
  2. Dotted decimal notation (base 256).
  3. Hexadecimal notation (base 16).
- Out of these the **dotted decimal** notation is most commonly used.

Dotted decimal notation :

- This notation has become popular because of the two advantages it offers. This notation makes the IPv4 address more compact and easy to read.

- The 32 bit IPv4 address is grouped into groups of 8-bits each separated by decimal points (dots).
- Each 8-bit group is then converted into an equivalent decimal number as shown in Fig. 11.4.1.
- Each octet (byte) can take a value between 0 and 255. Therefore the IPv4 address in the dotted decimal notation has a range from 0.0.0.0 to 255.255.255.255.
- For example the IPv4 address of

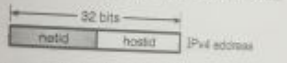
  1001 0001.00001010 00100010 00000011 is denoted in the dotted decimal form as 145.10.34.3.



(G-2001) **Fig. 11.4.1 : Dotted decimal notation**

11.4.4 IPv4 Address Format :

- A 32 bit IPv4 address consists of two parts. The first part is called as **net id** i.e. network identification which identifies a network on the Internet and the second part is called as the **host id** which identifies a host on that network.
- Fig. 11.4.2 shows the IPv4 address format. Note that the **net id** and **host id** are of variable lengths depending on the class of address.
- Note that class D and E addresses are not divided into net id and host id for the reasons discussed later on.



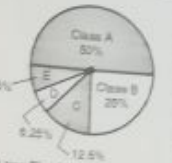
(G-2002) **Fig. 11.4.2 : IPv4 address format**

11.5 Classful Addressing :

- The concept of IP addresses is few decades old. It uses the concept of **classes.** This architecture is called as the **classful addressing.**
- Later on in mid 1990s a new architecture of addressing was introduced which was known as **classless addressing.** This new architecture has superseded the original architecture.
- In this section we are going to discuss the classful addressing.

11.5.1 IPv4 Address Classes :

- In the classful addressing architecture, the IP address space has been divided into five classes : A, B, C, D and E.
- Fig. 11.5.1 shows the percentage of occupation of the address space by each class.
- The number of class A addresses is the highest i.e. 50% and those of classes D and E is the lowest i.e. 6.25%.



Class	No. of addresses	
A	2^{31}	50%
B	2^{30}	25%
C	2^{29}	12.5%
D	2^{28}	6.25%
E	2^{28}	6.25%

(G-2003) **Fig. 11.5.1 : Classful addressing occupation of address space**

11.5.2 Formats of Various Classes :

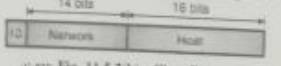
Class A format :

- The formats used for IPv4 addresses are as shown in Fig. 11.5.2. The IPv4 address for class A networks is shown in Fig. 11.5.2(a).
- The network field is 7 bit long as shown in Fig. 11.5.2(a) and the host field is of 24 bit length. So the network field can have numbers between 1 to 126.
- But the host numbers will range from 0.0.0.0 to 127.255.255.255.
- Thus in class A, there can be 126 types of networks and 17 million hosts.
- The "0" in the first field identifies that it is a class A network address.

(G-530) **Fig. 11.5.2(a) : Class A IPv4 address formats**

Class B format :

- The class B address format is shown in Fig. 11.5.2(b).
- The first two fields identify the network, and the number in the first field must be in the range 128 - 191.



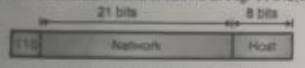
(G-531) **Fig. 11.5.2(b) : Class B format**

- Class B networks are large. Host numbers 0.0 and 255.255 are reserved, so there can be upto 65,534 (216-2) hosts in a class B network. Most of the 16,382 class B addresses have been allocated. The first block covers address from 128.0.0.0 to 128.255.255.255 and the last block covers from 191.255.0.0 to 191.255.255.255.

Example : 128.89.0.26, for host 0.26 on net 128.89.

Class C format :

- The class C address format is shown in Fig. 11.5.2(c).



(G-4355) **Fig. 11.5.2(c) : Class C format**

- The first block in class C covers addresses from 192.0.0.0 to 192.0.0.255 and the last block covers addresses from 223.255.255.0 to 223.255.255.255.

Class D format :

- The class D address format is shown in Fig. 11.5.2(d).

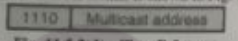


Fig. 11.5.2(d) : Class D format

- The class format allows for upto 2 million networks with upto 254 hosts each and class D format allows the multicast in which a datagram is directed to multiple hosts.

Class E address format :

- Fig. 11.5.2(e) shows the address format for a class E address. This address begins with 11110 which shows that it is reserved for the future use.

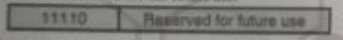
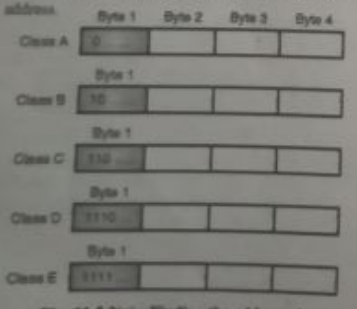


Fig. 11.5.2(e) : IPv4 address for class E network

- The 32 bit (4 byte) network addresses are usually written in dotted decimal notation. In this notation each of the 4-bytes is written in decimal from 0 to 255.
- So the lowest IP address is 0.0.0.0 i.e. all the 32 bits are zero and the highest IPv4 address is 255.255.255.255.

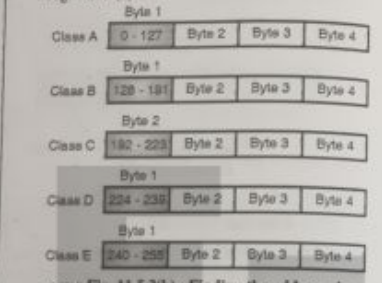
11.5.3 How to Recognize Classes :

- When an IPv4 address is given to us either in the binary or dotted decimal notation, we can find the class of the address.



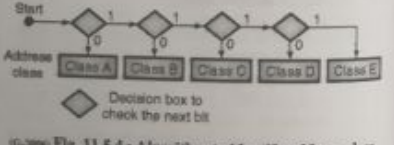
(G-2004) **Fig. 11.5.3(a) : Finding the address class**

- If the given address is in the binary notation then we can identify its class by inspecting the first few bits of the address. This is as shown in Fig. 11.5.3(a).
- If the given address is in the dotted decimal notation then we can identify the address class by inspecting the first byte of the address. This is as shown in Fig. 11.5.3(b).



(G-2005) **Fig. 11.5.3(b) : Finding the address class**

- It is important to note here that there are some special addresses which fall in class A or E. These special addresses are to be treated as the exceptions to the classful addressing. We have discussed them later in the chapter.
- In computers, the IPv4 addresses are generally stored in the binary notation format. Therefore it is possible to write an algorithm which can identify the address class by using the continuous checking process.
- The principle of such an algorithm has been shown in Fig. 11.5.4.

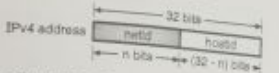


(G-2006) **Fig. 11.5.4 : Algorithm to identify address class**

11.5.4 Two Level Addressing :

- The IPv4 addressing is used for defining a destination for an Internet packet at the network layer.
- At the time when classful addresses were designed, the Internet was considered as the network of networks. In other words the whole Internet was divided into a number of smaller networks with many hosts connected to each network.

- Normally an organization which wants to connect to the Internet creates a network and the Internet authorities allocate a block of address to the organization. These addresses can be in class A, B or C.
- All the addresses allotted to an organization belong to a single block. Therefore each IPv4 address in classful addressing system is made up of two parts namely **net id** and **host id** as shown in Fig. 11.5.5.



(G-2007) **Fig. 11.5.5 : Two level addressing in classful addressing**

- The job of the **net id** is to define a network and that of the **host id** is to define a particular host in that network.
- As shown in Fig. 11.5.5 if n bits define **net id** then the remaining (32-n) bits define **host id**.
- The value of 'n" is not same for all the classes. Infact it is depend on the class as shown in Table 11.5.1.

Table 11.5.1

Class	Value of n
A	n = 8
B	n = 16
C	n = 24

11.5.5 Extracting Information in a Block :

- A block is nothing but a range of addresses. For any given block we would be interested to extract the following three pieces of information :
  1. The total number of addresses in the block.
  2. The first address of the block.
  3. The last address in the block.
- Before extracting all this information, we have to identify the class of the address as discussed earlier.
- Once we find the class of the block, we will have the values of "n" (the length of **net id** in bits) and (32 – n) i.e. the length of the **host id** in bits.
- It is now possible to obtain the three pieces of information mentioned above as shown in Fig. 11.5.6.

1. Total number of addresses in the block :

The total number of IPv4 addresses in the given block will be equal to,

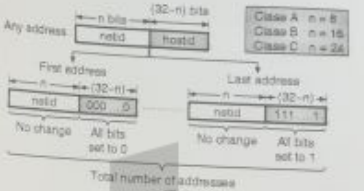
$$N = 2^{(32-n)} \quad \ldots(11.5.1)$$

2. First address in the block :

The first address in the given block can be obtained by keeping the leftmost "n" bits in the address as it is and setting all the (32 – n) rightmost bits to 0 as shown in Fig. 11.5.6.

3. Last address in the block :

The last address in the given block can be obtained by keeping the leftmost "n" bits in the address as it is and then setting all the (32 – n) rightmost bits to 1 as shown in Fig. 11.5.6.



(G-2008) **Fig. 11.5.6 : Information extraction in classful addressing**

11.5.6 Network Address :

- The network address is an address that defines the network itself. It cannot be assigned to a host. Fig. 11.5.7 shows the examples of network addresses for different classes.

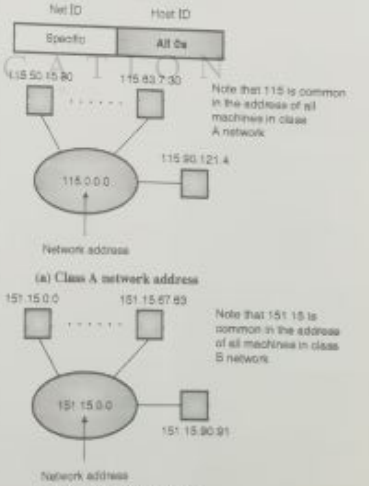
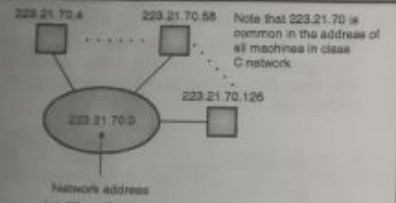


Fig. 11.5.7 (Contd...)



(c) Class C network address

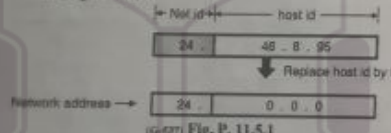
(G-536) Fig. 11.5.7

- The following examples will enable you to find the network address.

Ex. 11.5.1 : For the address 24.46.8.95 identify the type of network and find the network address.

Soln. :

- Examine the first byte. Its value is 24 i.e. it is between 0 and 127. So it is a class A network.
- So only the first byte defines the Net id. So we can find the network address by replacing the host id with 0s.
- The process of obtaining the network address is shown in Fig. P. 11.5.1.



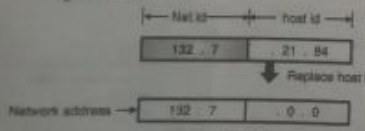
(G-537) Fig. P. 11.5.1

So the network address is 24.0.0.0.

Ex. 11.5.2 : For the address 132.7.21.84 find the type of network and the network address.

Soln. :

- Examine the first byte. It is 132 i.e. between 128 and 192. So it is a class B network.
- So the first two bytes define the net id. Replace the host id with 0's to get the network address as shown in Fig. P. 11.5.2.

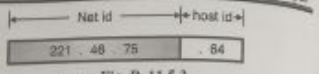


(G-538) Fig. P. 11.5.2

So the network address is 132.7.0.0.

Ex. 11.5.3 : Find the class of the network if the address is 221.46.75.84.

Soln. : The first byte is 221 i.e. between 192 and 255. So this is a class C network. The net id and host id are as shown in Fig. P. 11.5.3.



(G-539) Fig. P. 11.5.3

What is the difference between net id and network address ?

The network address is different from a net id. A network address has both net id and host id, with 0s for the host id.

Where to use the network address ?

The network address is used to route the packets to the desired location.

11.5.7 Network Mask or Default Mask :

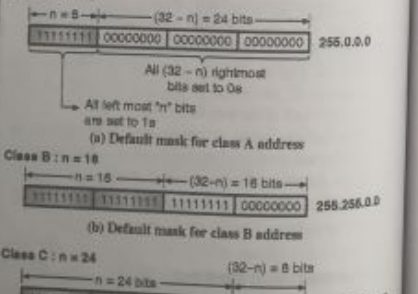
- Earlier we have discussed the methods for extracting different pieces of information. But all these methods are theoretical methods which are useful in explaining the concept.
- But practically these methods are not used. When a packet arrives at the input of the router in the Internet, it uses an algorithm to extract the **network address** from the destination address in the received packet.
- This can be achieved by using a **network mask.**

Definition of default mask :

A **network mask** or **default mask** in classful addressing is defined as a 32-bit number obtained by setting all the "n" leftmost bits to 1s and all the (32 – n) rightmost bits to 0.

11.5.8 Default Masks for Different Classes :

- We know that the value of n is different for different classes. Therefore their default masks also will be different.
- The default masks for class A, B and C addresses are as shown in Fig. 11.5.8.



(a) Default mask for class A address

(b) Default mask for class B address

(c) Default mask for class C address

(G-3009) Fig. 11.5.8

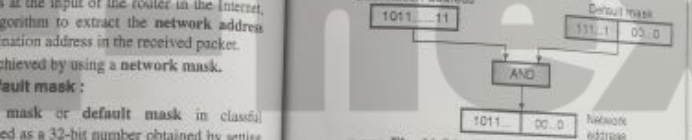
- Table 11.5.2 enlists the default masks of the three classes of IPv4 addresses.

Table 11.5.2 : Default masks

Address class	Default mask
A	255.0.0.0
B	255.255.0.0
C	255.255.255.0

11.5.9 Finding Network Address using Default Mask :

- The router uses the AND operation for extracting the network address from the destination address of the received packet.
- The router ANDs the destination address with the default mask to extract the network address as shown in Fig. 11.5.9.
- It is possible to use the default mask to find the number of addresses and the last address in the block.



(G-2010) Fig. 11.5.9 : Finding a network address using the default mask

11.5.10 Three Level Addressing : Subnetting :

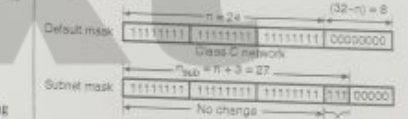
- As discussed earlier, the originally designed IP addresses were with two level addressing with **net id** and **host id.**
- The two level addressing is based on the principle that in order to reach a host on the Internet, we have to reach the network first and then the host.
- But very soon it became evident that the two level addressing would not be sufficient for the following two reasons :
  1. First it was needed to divide a large network of an organization (to which a block in class A or B is allotted) into many smaller **subnets** (subnetworks) for improved management and security.
  2. Second reason is more important. The blocks in class A and B were almost depleted and the blocks in class C were smaller than the needs of most organization. Therefore the organizations had to divide their allotted class A or B block into smaller subnetworks and share them.

Definition of subnetting :

- We can define the **subnetting** as the principle of splitting a block of addresses into smaller blocks of addresses.
- In the process of **subnetting** we divide a big network into smaller subnetworks or **subnets.**
- Each such subnet has its own **subnet address.**

Subnet mask :

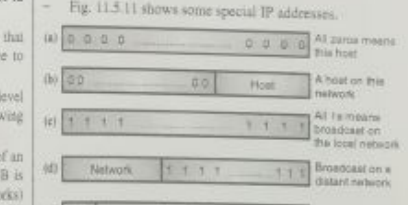
- The **network mask** or **default** mask that we discussed earlier is used when the given network is **not** to be divided into smaller subnetworks i.e. when **subnetting** is **not** to be done.
- But when the given network is to be divided into smaller subnets i.e. when subnetting is to be done, we need to create a **subnet mask** for each subnet.
- Fig. 11.5.10 shows the format of a subnet mask. Each subnet has its own **net id** and **host id.**
- If we want to divide a network into 8 subnets then the corresponding subnet mask will have three extra 1's because $2^3 = 8$, as compared to the default mask, as shown in Fig. 11.5.10.
- In Fig. 11.5.10, we have shown the default mask and subnet mask when a class C network is to be divided into 8 subnets.



(G-2011) Fig. 11.5.10 : Default and subnet masks

11.5.11 Special IP Addresses :

- Fig. 11.5.11 shows some special IP addresses.



(G-540) Fig. 11.5.11 : Special IP addresses

- All zeros means this host or this network and all 1s means broadcast address to all hosts on the indicated network.
- The IP address 0.0.0.0 is used by the hosts when they are being booted but not used afterward.
- The IP addresses with 0 as the network number refer to their own network without knowing its number as shown in Fig. 11.5.11(b).

- The address having all ones is used for broadcasting on the local network such as a LAN as shown in Fig. 11.5.11(c).
- Refer Fig. 11.5.11(d). This is an address with proper network number and all 1s in the host field. This address allow machines to send broadcast packets to distant LANs anywhere in the Internet.
- If the address is "127. Anything" as shown in Fig. 11.5.11(e) then it is a reserved address **loopback testing.** This feature is also used for debugging network software.

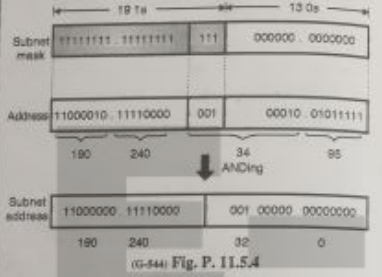
11.5.12 Limitations of IPv4 :

- The most obvious limitation of IPv4 is its address field. IP relies on network layer addresses to identify end-points on networks, and each networked device has a unique IP address.
- IPv4 uses a 32-bit addressing scheme, which gives it 4 billion possible addresses. With the proliferation of networked devices including PCs, cell phones, wireless devices, etc., unique IP addresses are becoming scarce, and the world could theoretically run out of IP addresses.
- If a network has slightly more number of hosts than a particular class, then it needs either two IP addresses of that class or the next class of IP address. For example, let use say a network has 300 hosts, this network needs either a single class B IP address or two class C IP addresses. If class B address is allocated to this network, as the number of hosts that can be defined in a class B network is $(2 \wedge 16 - 2)$, a large number of host IP addresses are wasted.
- If two class C IP addresses are allocated, as the number of networks that can be defined using a class C address is only $(2 \wedge 21)$, the number of available class C networks will quickly exhaust. Because of the above two reasons, a lot of IP addresses are wasted and also the available IP address space is rapidly reduced.
- Other identified limitations of the IPv4 protocol are : Complex host and router configuration, non-hierarchical addressing, difficulty in re-numbering addresses, large routing tables, non-trivial implementations in providing security, QoS (Quality of Service), mobility and multi-homing, multicasting etc.
- To overcome these problems the internet protocol version 6 (IPv6) which is also known as internet protocol, next generation (IPng) was proposed.
- In IPv6 the internet protocol was extensively modified for accommodating the unforeseen growth of the internet.
- The format and length of the IP addresses has been changed and the packet format also is changed.

Ex. 11.5.4 : A router inside an organization receives the same packet with a destination address 190.240.34.95. If the subnet mask is /19 (first 19-bits are 1s and following bits are 0s). Find the subnet address.

Soln. :

- To find the subnet address, AND the destination address with the subnet mask as shown in Fig. P. 11.5.4.



(G-544) **Fig. P. 11.5.4**

Thus the subnet address is 190.240.32.0

11.5.13 Classless Addressing :

- Even though the number of actual devices connected to Internet is much less than 4 billion, the address depletion has taken place due to flaws in the classful addressing scheme.
- We have run out of class A and B addresses. To overcome these problems, the classless addressing is now being tried out.
- In the classless addressing, there are no classes but the address generation take place in blocks.

Address blocks :

- Address block is defined as the range of addresses.
- In the classless addressing, when an entity wants to get connected to the internet, a block (range) of addresses is granted to it.
- The size of this block i.e. number of addresses depends on the size of the entity as well as its nature.
- That means for a small entity such as a household only one or two addresses will be given whereas for a larger entity like an organization, thousands of addresses can be allotted.

Restrictions :

Some of the restriction on classless address blocks have been imposed by the internet authorities in order to simplify the process of address handling.

1. The addresses in a block should be continuous, i.e. serial in manner.
2. The total number of addresses in a block has to be equal to some power of 2 i.e. $2^1, 2^2, 2^3 \ldots$etc.

3. The first address should be evenly divisible by the number of addresses.

11.5.14 Supernetting :

- The class A and class B addresses are almost depleted. But class C addresses are still available.
- But the size of class C address with a maximum number of 256 addresses does not satisfy the needs of an organization. More addresses will be required.
- The solution to this problem is **supernetting.**
- In supernetting an organization combines several class C blocks to create a large range of addresses i.e. several networks are combined to create a supernetwork.
- By doing this the organization can apply for a set of class C blocks instead of just one.

Example of supernetting :

- If an organization needs 1000 addresses, they can be obtained by using four C blocks (one C block corresponds to 256 addresses).
- The organization can then use these addresses as one supernetwork as a whole.

Note : The classful addressing is almost obsolete now and it is being replaced with classless addressing.

11.5.15 Who Decides the IP Addresses ?

- No two IP addresses should be same. This is ensured by a central authority that issues the prefix for the network number portion of the IP address.
- Locally an ISP is to be contacted in order to get a unique IP address prefix.
- At the global level the Internet Assigned Number Authority (IANA) allots an IP address prefix to the ISP. Thus it is ensured that the IP addresses are not duplicated.
- Conceptually IANA is a wholesales and ISP is a retailer of the IP addresses because ISP purchases IP addresses from IANA and sells them to the customers.

11.5.16 Registered and Unregistered Addresses :

- Registered IP addresses are required for computers which are accessible from the Internet but not every computer that is connected to the Internet.
- For security reasons, networks use firewalls or some other technologies for protecting the computers.
- The firewalls will enable the workstations to access the Internet but do not allow the other systems on the Internet to access them.
- These workstations are given the unregistered private IP addresses. These addresses are assigned by the network administrator without obtaining them from an ISP (Internet Service Provider) or IANA.
- These are special network addresses in each class as shown in Table 11.5.3. These addresses are to be used for private networks and are called **unregistered addresses.**
- We can choose any of these unregistered address while building our own private network.

Table 11.5.3 : IP addresses for private networks

Class	Network address
A	10.0.0.0 through 10.255.255.255
B	172.16.0.0 through 172.31.255.255
C	192.168.0.0 through 192.168.255.255

11.5.17 Solved Examples :

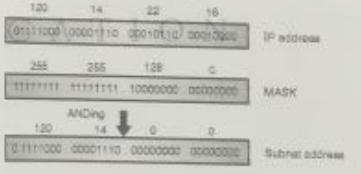
Ex. 11.5.5 : Find the sub-network address and the host id for the following :

Qr. No.	IP Address	MASK
(a)	120.14.22.16	255.255.128.0
(b)	140.11.36.22	255.255.255.0
(c)	141.181.14.16	255.255.224.0
(d)	200.34.22.156	255.255.255.240

Soln. :

Step 1 : To find the subnet address :

In order to find the subnet address we have to AND the IP address and the mask as follows :



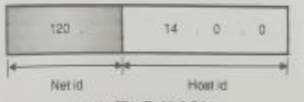
(G-553) **Fig. P. 11.5.5(a)**

So the subnet address is 120.14.0.0.

Similarly we can find the other subnet addresses.

Step 2 : Host id :

- Examine the first byte of the subnet address. It is 120 which is between 0 and 127. Hence this is a class A network.
- So only the first byte corresponds to the net id and the remaining three bytes correspond to the host id as shown in Fig. P. 11.5.5(b).



(G-554) **Fig. P. 11.5.5(b)**

So the host id is 14.0.0.

Similarly we can find the other host id.

Ex. 11.5.6 : The IP address of a host on class C network is 198.123.46.237. Four networks are allowed for this network. What is subnet mask ?

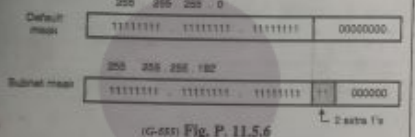
Soln. :

The default mask for a class C network is,

255.255.255.0

In order to have four networks, we must have two extra 1s.

Hence the default mask and subnet mask are shown in Fig. P. 11.5.6.



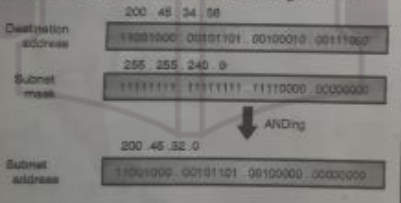
(G-555) **Fig. P. 11.5.6**

Thus the required subnet mask is 255.255.255.192.

Ex. 11.5.7 : What is the subnet address if the destination address is 200.45.34.56 and subnet mask is 255.255.240.0 ?

Soln. :

To find the subnet address we have to AND the IP address and the subnet mask as shown in Fig. P. 11.5.7.



(G-556) **Fig. P. 11.5.7**

Thus the required subnet address is 200.45.32.0.

Ex. 11.5.8 : A company is granted a site address 201.70.64.0. The company needs six subnets. Design the subnets.

Soln. :

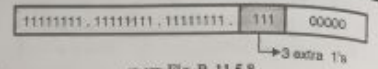
- This is a class C network. So the default mask is,

  255.255.255.0

- As we need 6 subnets, we need three extra 1s. So the subnet mask is,

  255.255.255.224

- In the binary form the subnet mask is as shown in Fig. P. 11.5.8.



(G-557) **Fig. P. 11.5.8**

- In order to have six subnets, we can have 6 different combinations of the 3-extra 1s as shown in Table P. 11.5.8(a).

Table P. 11.5.8(a)

Combination	Subnet number
000	Subnet 1
001	Subnet 2
010	Subnet 3
011	Subnet 4
100	Subnet 5
101	Subnet 6

- So the various addresses of 6 subnets are as shown in Table P. 11.5.8(b).

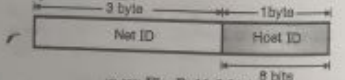
Table P. 11.5.8(b)

Subnet Number	Addresses
1	201.70.64.0 to 201.70.64.31
2	201.70.64.32 to 201.70.64.63
3	201.70.64.64 to 201.70.64.95
4	201.70.64.96 to 201.70.64.127
5	201.70.64.128 to 201.70.64.159
6	201.70.64.160 to 201.70.64.191

Ex. 11.5.9 : For a given class C network 195.188.65.0 design equal subnets in such a way that each subnet has atleast 60 nodes.

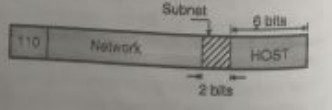
Soln. :

- Fig. P. 11.5.9(a) shows the structure of a class C address in which 3-bytes are reserved for net ID and 1-byte for host ID.



(G-558) **Fig. P. 11.5.9(a)**

- We are expected to design equal subnets such that each subnet has atleast 60 nodes (i.e. 60 users).
- In order to identify at least 60 users we need 6-bits in the host ID.
- The remaining 2-bits are assigned for subnetting as shown in Fig. P. 11.5.9(b).



(G-559) **Fig. P. 11.5.9(b)**

- This shows that there will be four equal subnets each one having at least 60 nodes.

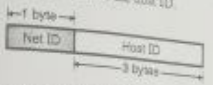
Ex. 11.5.10 : Show by calculations how many network each IP address class can have with one example ?

Soln. :

Number of networks in different IP address :

Class A address :

- The format of class A address is shown in Fig. P. 11.5.10(a). Here one byte defines the network ID and three bytes define the host ID.

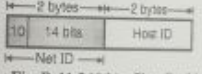


(G-560) **Fig. P. 11.5.10(a) : Class A address**

- The MSB in the network field is reserved. So actually there are only 7-bits in the network field.
- So the number of networks in class A address will be 128.

Class B address :

- The format of class B address is shown in Fig. P. 11.5.10(b). Here 2-bytes are reserved for network field and remaining two bytes are for the host field.
- Out of 16-bits in the network field the first two bits (MSBs) are reserved. So actually 14 bits are available in the network field.

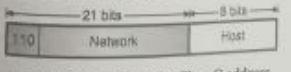


(G-561) **Fig. P. 11.5.10(b) : Class B address**

- So the number of networks in class B address is $2^{14} = 16,368$.

Class C address :

- The format of class C is shown in Fig. P. 11.5.10(c). Here 3-bytes are reserved for network field and only one byte for the host field.
- Out of 24-bits in the network field 3-bits are again reserved. So actually only 21-bits are available.



(G-562) **Fig. P. 11.5.10(c) : Class C address**

- So the number of networks in class C addresses is 2,097,152.

Ex. 11.5.11 : How many host per network in each IP address class can exist, show with example ?

Soln. :

Number of hosts in different IP addresses :

Class A :

There are 3-bytes (24-bits) in the host field. Hence the number of hosts in class A address will be $2^{24} = 16,7772,16$.

Class B :

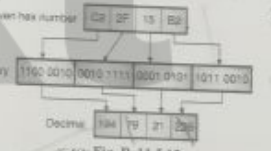
There are 2-bytes (16-bits) in the host field. So the number of hosts in class B address will be 65536 i.e. 2^{16} per network.

Class C :

There is 1-byte (8-bits) in the host field. So number of hosts in class C address will be $2^8 = 256$ per network.

Ex. 11.5.12 : Convert the IP address whose hexadecimal representation is C22F1582 to dotted decimal notation.

Soln. :



(G-563) **Fig. P. 11.5.12**

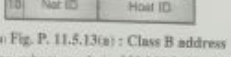
The IP address in the dotted decimal notation is as follows :

194.47.21.226

Ex. 11.5.13 : A class B network on internet has a subnet mask of 255.255.240.0. What is the maximum number of hosts per subnet ?

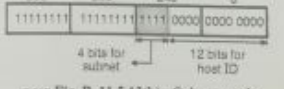
Soln. :

The structure of class B address is as shown in Fig. P. 11.5.13(a).



(G-564) **Fig. P. 11.5.13(a) : Class B address**

The given subnet mask is 255.255.2400. So it is as shown in Fig. P. 11.5.13(b).



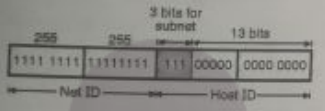
(G-565) **Fig. P. 11.5.13(b) : Subnet mask**

Thus there are 4 extra 1s as shown in Fig. P. 11.5.13(b). So there will be 16 subnets and each subnet can have $2^{12} = 4096$ hosts.

Ex. 11.5.14 : Perform the subnetting of the following IP address 160.111. X.X
Original subnet mask 255.255.0.0
Number of subnets 6 (six)

Soln. :

- The original subnet mask indicates that we are dealing with a class B address.
- In order to have six subnets we need to use 3 extra bits from the bits that are reserved for host ID. So the subnet mask is as shown in Fig. P. 11.5.14.



(G-566) **Fig. P. 11.5.14**

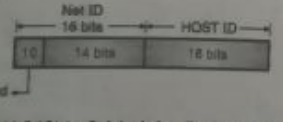
- The 3-bits reserved for subnetting will have 8 combinations from 000 to 111 out of which any six combinations can be used for 6 subnets.
- Let us decide that the combinations 000 to 001 are not to be used. Then the subnet masks for the 6 possible subnets will have the following addresses.

Subnet 1	255.255.64.0
Subnet 2	255.255.96.0
Subnet 3	255.255.128.0
Subnet 4	255.255.160.0
Subnet 5	255.255.192.0
Subnet 6	255.255.224.0

Ex. 11.5.15 : Suppose that instead of using 16-bits for the part of class B address originally, 20-bits had been used. How many class B network addresses would there have been? Give the range of IP addresses in decimal dotted form.

Soln. :

- Fig. P. 11.5.15(a) shows the original class B address format :



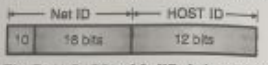
(G-567) **Fig. P. 11.5.15(a) : Original class B address format**

- The first two MSB bits of Net ID part are reserved. Hence, the number of bits actually available for network ID is 14.
- Hence the number of class B networks = 2^{14} = 16382.

Modification :

Now with 20 bits instead of 16 being available for the Net ID part the actually available number of bits for Network part becomes 18. This is shown in Fig. P. 11.5.15(b).

∴ Number of class B networks = 2^{18} = 2, 61, 888



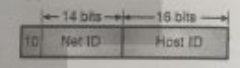
(G-568) **Fig. P. 11.5.15(b) : Modified class B address format**

The range of IP addresses in the decimal dotted form would be 128.0.0.0 to 191.255.255.255.

Ex. 11.5.16 : A network on the Internet has a subnet mask of 255.255.240.0. What is the maximum number of host it can handle? Give the range of IP addresses in decimal dotted form.

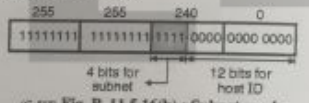
Soln. :

The structure of class B address is as shown in Fig. P. 11.5.16(a).



(G-564) **Fig. P. 11.5.16(a) : Class B address**

The given subnet mask is 255.255.240.0. So it is as shown in Fig. P. 11.5.16(b).



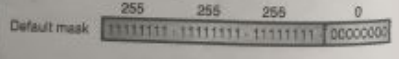
(G-565) **Fig. P. 11.5.16(b) : Subnet mask**

Thus there are 4 extra 1s as shown in Fig. P. 11.5.16(b). So there will be 16 subnets and each subnet can have 2^{12} = 4096 hosts.

Ex. 11.5.17 : For a given class-C network, design 4 equal subnets having minimum 50 nodes in each subnetwork.

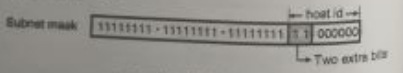
Soln. :

The default mask for a class C network is
255.255.255.0
This is as shown in Fig. P. 11.5.17(a).



(G-571) **Fig. P. 11.5.17(a)**

In order to design 4 equal subnets having a minimum 50 nodes in each subnetwork, we have to use two extra bits from the host id field. So the subnet mask is as shown in Fig. P. 11.5.17(b).



(G-572) **Fig. P. 11.5.17(b)**

In order to have four subnets, we can have four different combinations of the two extra bits as shown in Table P. 11.5.17(a).

Table P. 11.5.17(a)

Combination	Subnet
00	subnet 1
01	subnet 2
10	subnet 3
11	subnet 4

Let the class C address be 201.70.64.0. Then the addresses of the four subnets are as shown in Table P. 11.5.17(b).

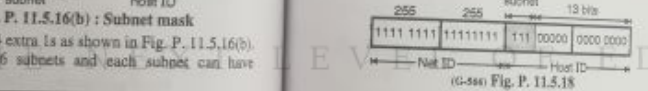
Table P. 11.5.17(b)

Subnet number	Addresses
1	201.70.64.0 to 201.70.64.63
2	201.70.64.64 to 201.70.64.127
3	201.70.64.128 to 201.70.64.191
4	201.70.64.192 to 201.70.64.255

Ex. 11.5.18 : For a given class B network 144.155.0.0 with default subnet mask, how can you divide it into 8 equal subnets ? How many hosts can be accommodated in each sub-network ?

Soln. :

Given class B network : 144.155.0.0. The default subnet mask is 255.255.0.0. In order to have 8 subnets we need to use 3 extra bits from the host id field as shown in Fig. P. 11.5.18.



(G-586) **Fig. P. 11.5.18**

- The 3-bits reserved for subnetting will have 8 combinations from 000 to 111 which can be used for 8 subnets.
- The subnet masks for the 8 possible subnets will have the following subnet masks :

Subnet	Mask
1	255.255.0.0
2	255.255.32.0
3	255.255.64.0
4	255.255.96.0
5	255.255.128.0
6	255.255.160.0
7	255.255.192.0
8	255.255.224.0

Number of hosts in each subnet :

Due to use of extra 3-bits for subnetting, now we have only 13-bits left in the host id field.

∴ No. of hosts in each subnet = 2^{13} = 8192 **...Ans.**

Ex. 11.5.19 : Consider any class – C network with default subnet mask. How many actual hosts can be connected in that network ? Divide that network into 4 equal subnets. What is the new subnet mask ? How many hosts can be connected in each subnet ?

Soln. :

- For a class C network, the default mask is
255.255.255.0
- For a class C network we can connect 2^8 = 256 total hosts.
- As we need 4 subnets, we need two extra 1s. So the subnet mask is
255.255.255.192
- In the binary from the subnet mask is as shown in Fig. P. 11.5.19.

(G-574) **Fig. P. 11.5.19**

- In order to have four subnets we can have the 4 combinations of the two extra 1s as shown in Table P. 11.5.19.

Table P. 11.5.19

Combination	Subnet number
0 0	Subnet 1
0 1	Subnet 2
1 0	Subnet 3
1 1	Subnet 4

- As we have used the 2 MSB bits of host ID field for subnet mask, we have only 6 bits remaining in the host id field.

∴ No. of hosts/subnet = 2^6 = 64.

Ex. 11.5.20 : Consider any class – C network with default subnet mask. Design the subnet in such a way that each has 62 nodes. Write the range of IP addresses for all subnets.

Soln. : Refer Ex. 11.5.19.

- But we want only 62 nodes on each subnet. So 2 nodes on each subnet will be inactive.
- Let the class C address be 201.70.64.0. Then the addresses of the four subnets are as shown in Table P. 11.5.20.

Table P. 11.5.20

Subnet number	Addresses
1	201.70.64.0 to 201.70.64.61
2	201.70.64.64 to 201.70.64.125
3	201.70.64.128 to 201.70.64.189
4	201.70.64.192 to 201.70.64.253

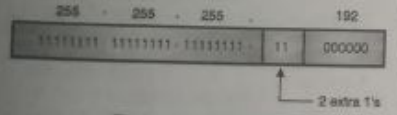
Ex. 11.5.21 : For a given C class network 210.50.60.0, how will you divide it into 4 equal subnets ? What will be the new subnet mask ? Give the network and broadcast address of each subnetwork.

Soln. :

Given : IP address : 210.50.60.0 (class C)

Step 1 : Subnet mask :

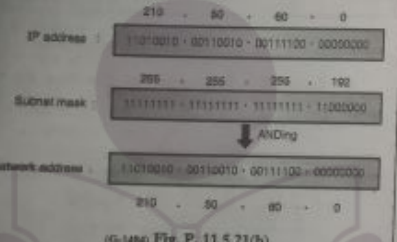
This is class C network. So default mask is given by 255.255.255.0.



(G-1483) **Fig. P. 11.5.21(a) : Subnet mask**

The new subnet mask is 255.255.255.192. ...Ans.

Step 2 : Find network address :

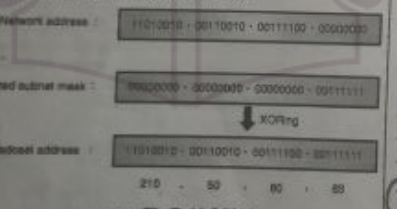


(G-1484) **Fig. P. 11.5.21(b)**

Network address is 210.50.60.0 ...Ans.

Step 3 : Find broadcast address :

To find broadcast address, take inverted subnet mask and perform XOR with network address.



(G-1485) **Fig P. 11.5.21(c)**

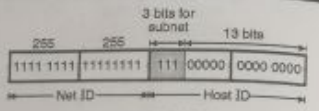
The broadcast address is 210.50.60.63 ...Ans.

Ex. 11.5.22 : For a given class B network 144.155.0.0 with default subnet mask, how can divide it into 8 subnets ? Write the :

1. Range of each subnet.
2. Network IP for 7th subnet.
3. Broadcast IP for the 7th subnet.
4. Subnet mask in subnets.

Soln. :

Given class B network : 144.155.0.0. The default subnet mask is 255.255.0.0. In order to have 8 subnets we need to use 3 extra bits from the host id field as shown in Fig. P. 11.5.22.



(G-566) **Fig. P. 11.5.22**

- The new subnet mask is 255.255.224.0.
- The 3-bits reserved for subnetting will have 8 combinations from 000 to 111 which can be used for 8 subnets.
- The subnet masks for the 8 possible subnets will have the following subnet masks :

Subnet	Mask
1	255.255.0.0
2	255.255.32.0
3	255.255.64.0
4	255.255.96.0
5	255.255.128.0
6	255.255.160.0
7	255.255.192.0
8	255.255.224.0

- The following is the range of subnets :

Subnet	Subnet range
1	144.155.0.0 to 144.155.31.255
2	144.155.32.0 to 144.155.63.255
3	144.155.64.0 to 144.155.95.255
4	144.155.96.0 to 144.155.127.255
5	144.155.128.0 to 144.155.159.255
6	144.155.160.0 to 144.155.191.255
7	144.155.192.0 to 144.155.223.255
8	144.155.224.0 to 144.155.255.255

- Network IP for 7th subnet is 144.155.192.0.
- Broadcast IP for 7th subnet is 144.155.223.255.
- Subnet mask is 255.255.224.0.

11.6 Routing :

- Routing is a very important issue in the network layer. A router creates its routing table so as to help forwarding a datagram in the connectionless services. It also helps in creating a virtual circuit in the connection oriented service.
- In the following sections we are going to discuss about the types of routing and different routing algorithms such as distance vector routing, link state routing and hierarchical routing.

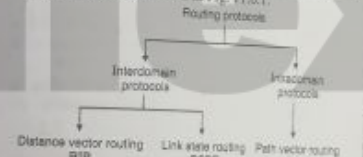
11.6.1 Types of Routing :

- Routing can be broadly classified into three types :
  1. Unicast routing.
  2. Broadcast routing
  3. Multicast routing.

- We can also classify the routing into two types as follows :
  1. Intradomain routing.
  2. Interdomain routing.

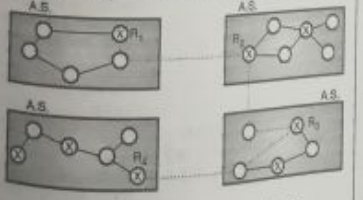
11.6.2 Intra and Interdomain Routing :

- Today the size of the internet is so big that one routing protocol cannot handle the task of updating the routing tables of all the routers.
- Hence an internet is divided into **Autonomous Systems** (AS). An Autonomous System (AS) is a group of networks and routers which is controlled by a single administrator. An AS is shown in Fig. 11.6.1.
- The **intradomain routing** is defined as the routing inside an autonomous system whereas the routing between autonomous system is known as the **interdomain routing.**
- Several intradomain and interdomain protocols are used. They are as shown in Fig. 11.6.1.



(G-1291) **Fig. 11.6.1 : Classification of routing protocols**

- The examples of interdomain routing protocols are :
  1. Distance vector routing
  2. Link state routing.
- An example of intradomain routing protocol is path vector routing.
- Each A.S. is allowed to choose one or more intradomain routing protocols in order to handle the routing inside the A.S. But only one interdomain routing protocol will handle routing between autonomous systems.
- The Routing Information Protocol (RIP) is an implementation of distance vector routing. Whereas the OSPF is an implementation of link state protocol. The BGP is an implementation of the path vector protocol.



(G-1292) **Fig. 11.6.2 : Autonomous systems**

11.6.3 Unicast Routing :

- In unicast routing there is a one to one relation between the source and the destination. That means only one source sends packets to only one destination.
- The type of source and destination addresses included in the IP datagram are unicast addresses assigned to the hosts.
- The concept of unicast routing is illustrated in Fig. 11.6.3.

(G-448) **Fig. 11.6.3 : Unicast routing**

- In unicast routing when a router receives a packet, it forwards that packet through only one of its ports which corresponds to the optimum path.
- The router can discard the packet if it can not find the destination address.

Metric :

- A metric is defined as the cost assigned for passing through a network.
- The metric assigned to each network depends on the type of protocol.

Interior and exterior routing :

- An Internet is so large that for one routing protocol it is impossible to handle the task of updating the routing tables of all the routers.
- So an Internet is divided into a number of Autonomous Systems (AS). An AS is group of networks and routers.

Interior routing :

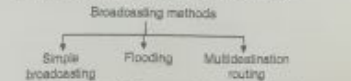
The routing that takes place inside an AS is called as interior routing.

Exterior routing :

The routing that takes place among various autonomous systems is called as exterior routing.

11.6.4 Broadcast Routing :

- In certain applications, the host has to send packets to many or all other hosts.
- If the sender send a packet to all destinations simultaneously then it is called as **broadcasting.**
- Various methods of broadcasting are as follows :



(G-449) **Fig. 11.6.4 : Various methods of broadcasting**

1. **Simple broadcasting :**
   - In this method the source will simply send a distinct (a separate) packet to each destination.
   - This method has two drawbacks :
     1. A lot of bandwidth is wasted.
     2. The source has to have a complete list of all destinations.
2. **Flooding :**
   - Flooding is another method used for broadcasting. The problem with flooding is that it has a point to point routing algorithm.
   - So it consumes a lot of bandwidth and generates too many packets.
3. **Multidestination routing :**
   - This is the third algorithm used for broadcasting.
   - In this algorithm each packet will contain a list of destinations or a bit map which indicates the desired destination.
   - When such a packet arrives at a router, the router first checks all the destinations. Then it decides the set of output lines that will be required based on the destination addresses.
   - The router then generates a new copy of the received packet for each output line to be used. It includes a list of only those destinations that are to use the line in each packet going out on that line. This will save bandwidth to a great extent. Also generation of too many packets right from the sending end will also be avoided.

11.6.5 Multicast Routing :

- In multicasting a message from a sender is to be sent to a group of destinations but not all the destinations in a network.
- A process has to send a message to all other processes in the group. For a small group it is possible to send a point-to-point message.
- But this is expensive if the group is large. So we have to send messages to a well defined groups which are small compared to the network size.
- Sending message to such a group is called **multicasting** and the routing algorithm used for multicasting is **multicast routing.**
- Multicast routing is a special class of broadcast routing.

11.7 Routing Algorithms :

- One of the important functions of the network layer is to route the packets from the source machine to the destination machine.
- The major area of network layer design includes the algorithms which choose the routes and the data structures which are used.
- **Routing algorithm** is a part of network layer software. It is responsible for deciding the output line over which a packet is to be sent.
- Such a decision is dependent on whether the subnet is a virtual circuit or it is datagram switching.

11.7.1 Desired Properties of a Routing Algorithm :

- There are certain desirable properties of a routing algorithm as follows :
  1. Correctness
  2. Robustness
  3. Stability
  4. Fairness and
  5. Optimality.

11.7.2 Types of Routing Algorithms :

Routing algorithms can be divided into two groups :
1. Non-adaptive algorithms.
2. Adaptive algorithms.

1. **Non-adaptive algorithms :**
   - For this type of algorithms, the routing decision is not based on the measurement or estimation of current traffic and topology.
   - However the choice of the route is done in advance, off-line and it is downloaded to the routers.
   - This is called as static routing.
2. **Adaptive algorithms :**
   - For these algorithms the routing decision can be changed if there are any changes in topology or traffic etc.
   - This is called as dynamic routing.
   - In the following sections we are going to discuss various static and dynamic algorithms.

11.7.3 Optimality Principle :

- A general statement about optimality is called as optimality principle.
- It states that if router J is on the optimal path from router I to router K, then the optimal path from J to K will also be along the same route.

Sink tree :
- A set of optimal routes from all the sources to a given destination form a tree called sink tree and it is shown in Fig. 11.7.1. The root of the sink tree is at the destination.
- Note that a sink tree need not be unique. Other trees with the same path lengths may also exist.
- All the routing algorithms are supposed to discover and use the sink trees for all routers.
- In the sink tree of Fig. 11.7.1, the distance metric is the number of hops. In Fig. 11.7.1(b) a sink tree for router B has been shown. The paths from B to every router with minimum number of hops.

(a) A subnet

(b) A sink tree for router B

(G-4W) Fig. 11.7.1

11.8 Static Algorithms :

The examples of static algorithms are :
1. Shortest path routing.
2. Flooding.
3. Flow based routing.

11.8.1 Shortest Path Routing :

- This algorithm is based on the simplest and most widely used principle. Here a graph of subnet is prepared in which each node represents either a host or a router and each arc represents a communication link.
- So as to choose a path between any two routers, this algorithm simply finds the shortest path between them.

How to decide the shortest path ?
- One way of measuring the path length is the number of hops. Another way (metric) is the geographical distance in kilometres.
- Some other metrics are also possible. For example we can label each arc (link) with the mean queuing and transmission delay and obtain the shortest path as the fastest path.

Labels on the arcs :
- The labels on the arcs can be computed as a function of distance bandwidth, average traffic, mean queue length, cost of communication, measured delay etc.
- The algorithm compares various parameters and calculates the shortest path, on the basis of any one or combination of criterions stated above.

Various shortest path algorithms :
- There are many algorithms for computing the shortest path between two nodes.
- One of them is Dijkstra algorithm. The other one is Bellman-Ford algorithm.

11.9 Dynamic Routing Algorithms :

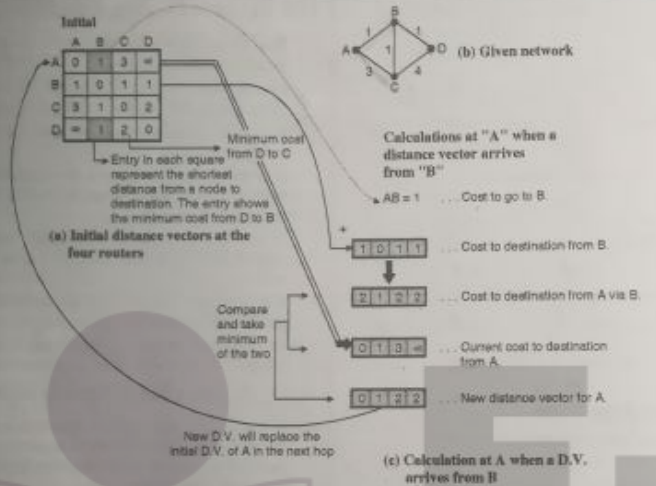
- The modern computer networks normally use the dynamic routing algorithms.
- Two dynamic routing algorithms namely distance vector routing and link state routing are used popularly.
- Both these algorithms are suitable for the packet switched networks.
- Both these algorithms assume that a router knows the address of each neighbouring router and the cost of reaching each neighbour.
- In the distance vector routing, each node tells its neighbours about its distance to every other node in the network.
- In the link state routing, a node tells every other node in the network the distance to its neighbours.
- So both these routing algorithms are distributed type and so they are suitable for large internetworks.

11.9.1 Distance Vector Routing Algorithm :

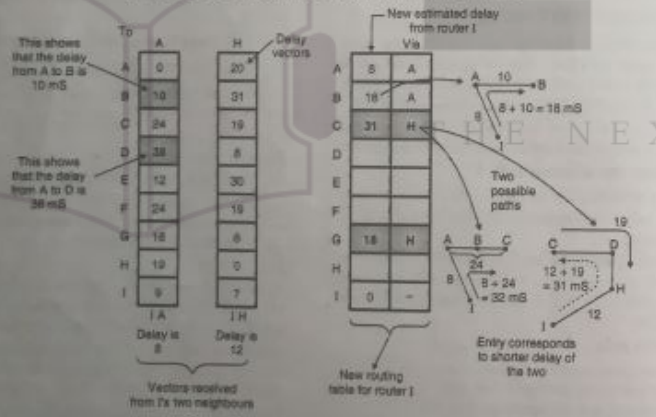
- In this algorithm, each router maintains a table called vector, such a table gives the best known distance to each destination and the information about which line is be used to reach there.
- This algorithm is sometimes called by other names such as :
  1. Distributed Bellman-Ford routing algorithm.
  2. Ford-Fulkerson algorithm
- In distance vector routing, each router maintains a routing table. It contains one entry for each router in the subnet.
- This entry has two parts :
  1. The first part shows the preferred outgoing line to be used to reach the specific destination.
  2. Second part gives an estimate of the time or distance to that destination.

Distance vector :
- In distance vector routing, we assume that each router knows the identity of every other router in the network, but the shortest part to each router is not known.
- A **distance vector** is defined as the list of <destination, cost> tuples, one tuple per destination. Each router maintains a distance vector.
- The cost in each tuple is equal the sum of costs on the shortest path to the destination.

Updation of router tables :
- A router periodically sends a copy of its distance vector to all its neighbours.
- When a router receives a distance vector from its neighbour, it tries to find out whether its cost to reach any destination would decrease if it routed packets to that destination through that particular neighbouring router. This is illustrated in Fig. 11.9.1.



(G-465) **Fig. 11.9.1 : Distance vector algorithm at router A**



(G-466) **Fig. 11.9.2(b) : Routing tables**

- Fig. 11.9.1 shows how the D.V. at A is automatically modified when a D.V. is received from B.
- A similar calculation takes place at the other routers as well. So the entries at every router can change. In Fig. 11.9.1(a) the initial distance vector is shown. The entries indicate to the costs corresponding to the shortest distance between the routers indicate to that square.
- For example, AC = 3 indicates the cost corresponding to the shortest path in terms of number of hops from A to C.
- Even if nodes asynchronously update their distance vectors the routing tables eventually converge.
- The well known example of distance vector routing is the Bellman-Ford algorithm.

Routing procedure in distance vector routing :

- The example of a subnet is shown in Fig. 11.9.2(a) and the routing tables are shown in Fig. 11.9.2(b).

(G-464) **Fig. 11.9.2(a) : A subnet**

- The entries in router tables of Fig. 11.9.2(b) are the delay vectors. For example consider the shaded boxes of Fig. 11.9.2(b).
- The entry in the first shaded box shows that the delay from A to B is 10 msec, whereas the entry in the other shaded box indicates that the delay from A to D is 38 msec.
- Consider how router I computes its new route to router G. Fig. 11.9.2(c) shows the two possible routes between I and G.

(G-466) **Fig. 11.9.2(c)**

I knows that the reach G via A, the delay required is :

1 to A Delay = 8 mS
A to G Delay = 16 mS
$\therefore$ I to G Delay = 8 + 16 = 24 msec

Whereas the delay between I and G via H (route IHG) is :

I to H Delay = 12 mS
H to G Delay = 6 mS
$\therefore$ I to G Delay = 12 + 6 = 18 msec

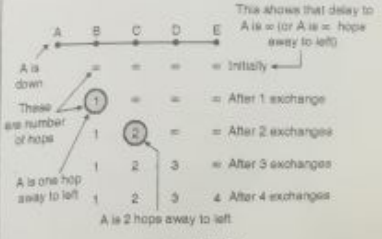
- The best of these values is 18 msec corresponding to the path IHG. Hence it makes an entry in its routing table (I's table) that the delay to G is 18 msec and that the route to use it is via H.
- The new routing table for router I is shown in Fig. 11.9.2(b).
- Similarly we can calculate the delays, from I to different destinations from A to I and enter the minimum possible delay into the I's router table.

Disadvantages :

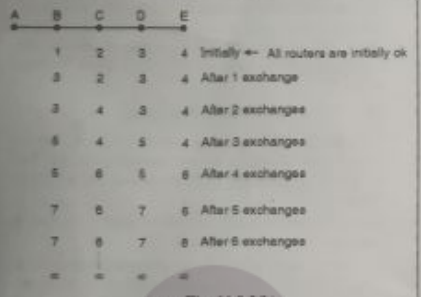
1. The distance vector routing takes a long time in converging to the correct answer. This is due to a problem called count-to-infinity problem. This problem can be solved by using the split horizon algorithm.
2. Another problem is that this algorithm does not take the line bandwidth into consideration when choosing a route. This is a serious problem due to which this algorithm was replaced by the Link State Routing algorithm.

11.9.2 Count to Infinity Problem :

- Theoretically the distance vector routing works properly but practically it has a serious problem. The problem is that we get a correct answer but we get it slowly.
- In other words it reacts quickly to good news but it reacts too slowly to bad news.
- Consider a router whose best route to destination X is large. If on the next exchange neighbour A suddenly reports a short delay to X, the router will switch over and start using the line to A for sending the traffic to destination X.
- Thus in one vector exchange, the good news is processed.
- Let us see how fast does a good news propagate. Consider a linear subnet of Fig. 11.9.3 which has five nodes. The delay metric used is the number of hops.
- Assume that A is initially down and that all the other routers know this. So all the routers have recorded that the delay to A is infinity.
- When A becomes OK, the other routers come to know about it via the vector exchanges. Then suddenly a vector exchange at all the routers will take place simultaneously.
- At the time of first vector exchange, B comes to know that its left neighbour has a zero delay to A. So as shown in Fig. 11.9.3(a), B makes an entry in its routing table that A is one hop away to the left.
- All the other routers still think that A is down. So in the second row of Fig. 11.9.3(a), the entries below C D E are ∞.
- On the second vector exchange, C comes to know that B has a path of 1 hop length to A, so C updates its routing table and indicates a path of 2 hop length. But D and E do not change their table entries.

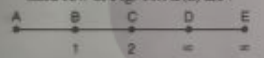


(G-467) **Fig. 11.9.3(a)**



(G-469) **Fig. 11.9.3(b)**

- So after the second vector exchange the entries in the third row of Fig. 11.9.3(a) are :

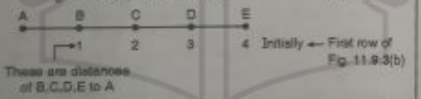


(G-468(a))

- Similarly D and E will update their routing tables after 3 and 4 exchanges respectively.
- So we conclude that the good news of A has recovered has spread at a rate of one hop per exchange.

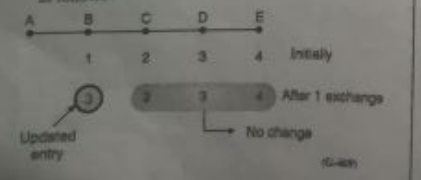
Explanation of Fig. 11.9.3(b) :

- Now refer Fig. 11.9.3(b). Here initially all routers are OK. The routers B, C, D and E have distances of 1, 2, 3 and 4 respectively to A. So the first row of Fig. 11.9.3(b) is as follows :



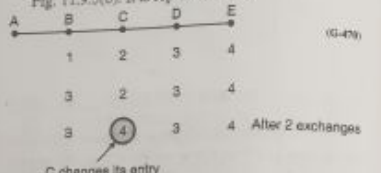
(G-468(b))

- Now imagine that suddenly A goes down or line between A and B is cut.
- At the first packet exchange B does not hear anything from A (because A is down). But C says " I have a path of length 2 to A". But poor B does not understand that this path is through B itself.
- So B thinks that it can reach A via C with a path length 3. (B to C 1 hop and C to A 2 hops) so it accordingly updates its routing table. But D and E do not update their entries. So the second row of Fig. 11.9.3(b) looks as follows :



(G-469)

- On the second exchange C realizes that both its neighbours (B and D) claim to have a path of length 3 to A. So it picks one of them at random and makes its new distance to A as 4. This is shown in row 3 of Fig. 11.9.3(b). It is repeated below.



(G-470)

- Similarly the other routers keep updating their tables after every exchange.
- It is expected that finally we should get ∞ in the router tables of B, C, D and E indicating that A is down. We do reach this state at the end in Fig. 11.9.3(b) but after a very long time.
- The conclusion is bad news propagates slowly. This problem is called as **count-to-infinity** problem.
- The solution to this problem is to use the split horizon algorithm.

Split horizon algorithm :

- To avoid the count to infinity problem, several changes in the algorithm have been suggested. But none of them work satisfactorily in all situations.
- One particular method which is widely implemented is called as the **split horizon algorithm.**
- In this algorithm, the minimum cost to a given destination is not sent to a neighbour if the neighbour is the next node along the shortest path.
- For example if node A thinks that the best route to node B is via node C, then node A should not send the corresponding minimum cost to node C.

11.9.3 Link State Routing :

- Distance vector routing was used in ARPANET upto 1979. After that it was replaced by the link state routing.
- Variants of this algorithm are now widely used.
- The link state routing is simple and each router has to perform the following five operations.

Router operations :

1. Each router should discover its neighbours and obtain their network addresses.
2. Then it should measure the delay or cost to each of these neighbours.
3. It should construct a packet containing the network addresses and the delays of all the neighbours.
4. Send this packet to all other routers.

5. Compute the shortest path to every other router.

- The complete topology and all the delays are experimentally measured and this information is conveyed to each and every router.
- Then a shortest path algorithm such as Dijkshtra's algorithm can be used to find the shortest path to every other router.

Protocols :

Link state routing is popularly used in practice.

- The OSPF protocol which is used in the Internet uses the link state algorithm.
- IS-IS i.e. Intermediate system – Intermediate system is the other protocol which uses the link state algorithm.
- IS-IS is used in Internet backbones and in some digital cellular systems such as CDPD.

11.9.4 Comparison of Link State Routing and Distance Vector Routing :

Sr. No.	Distance vector routing	Link state routing
1.	Each router maintains routing table indexed by and containing one entry for each router in the subnet.	It is the advanced version of distance vector routing.
2.	Algorithm took too long to converge.	Algorithm is faster.
3.	**Bandwidth is less.**	Wide bandwidth is available.
4.	Router measure delay directly with special ECHO packets.	All delays measured and distributed to every router.
5.	**It doesn't take line bandwidth into account when choosing the routes.**	It considers the line bandwidth into account when choosing the routes.

11.10 Path Vector Routing :

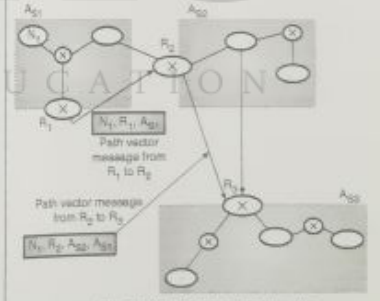
- It is different from both distance vector routing and link state routing.
- Table 11.10.1 shows the example of a path routing table. Each entry in the routing table will have the information about the destination network, the next router and the path to reach the destination.

Table 11.10.1 : Path vector routing table

Network	Next router	Path
N01	R01	AS 12, AS 21, AS 56
N02	R08	AS 20, AS 57, AS 06
.	.	.
.	.	.
.	.	.

11.10.1 Path Vector Messages :

- The autonomous boundary routers participate in path vector routing. Their job is to advertise the reachability of networks present in their A.S. to the neighbour autonomous boundary router.
- Each router that receives a path vector message verifies whether or not the advertised path is according to its policy. Such a policy is made up of rules that are imposed by the router controlling administrator.
- If yes then the router will update its routing table and will modify the message before it is sent to the next neighbour.
- In the modified message it sends its own AS number and replaces the next router entry with its own identification. This process is demonstrated in Fig. 11.10.1.
- Fig. 11.10.1 shows an internet containing three autonomous systems A_{S1} through A_{S3}.
- Router R_1 sends a path vector message to advertise that it is reachable to network N_1. Router R_2 on receiving this message will update its routing table. It then adds its own autonomous system (A_{S2}) to the path, inserts itself as the next router and sends this message to router R_3 as shown in Fig. 11.10.1.



(G-1799) **Fig. 11.10.1 : Path vector messages**

11.10.2 Loop Prevention :

- When a message is received, a router checks it to see if its autonomous system is in the path list to the destination. If it is present it indicates looping is involved which is undesirable and the message is ignored.
- In this way the looping problem and the associated instability which is present in distance vector routing is avoided in path vector routing.

11.10.3 Path Attributes :

- The path is specified in terms of attributes. Each attribute gives some information about the path. Hence the list of attributes helps the receiving router to make a better decision about when to apply its policy.
- Attributes are of two types :
  1. A well known attribute
  2. An optional attribute
- An attribute is called as a well known attribute if it is recognised by every BGP router.
- An optional attribute is the one that need not be recognised by every BGP router.
- The well known attributes are further classified into two categories :
  1. Well known mandatory attributes
  2. Well known discretionary attributes.
- The optional attributes also are classified into two types :
  1. An optional transitive attribute
  2. An optional nontransitive attribute.

Review Questions

Q. 1 State and explain the various services provided by network layer.

Q. 2 What is packetizing ?

Q. 3 Write short note on : routing and forwarding.

Q. 4 Explain error control and flow control.

Q. 5 Write short note on : IPv4 addresses.

Q. 6 What do you mean by uniqueness of IP addresses.

Q. 7 Draw IPv4 address format.

Q. 8 Define classful addressing.

Q. 9 Draw class B IPv4 address format.

Q. 10 How to recognize IPv4 classe.

Q. 11 Write short note on : Two level addressing in classful addressing.

Q. 12 How information is extracted in classful addressing ?

Q. 13 Define default mask.

Q. 14 Write default masks for different classes.

Q. 15 Define subnetting.

Q. 16 Write down limitations of IPv4.

Q. 17 Who decides the IP addresses ?

Q. 18 State the types of routing.

Q. 19 Explain unicast and broadcast routing.

Q. 20 Write down desired properties of a routing algorithm.

Q. 21 Write short note on : optimality principle.

Q. 22 Explain shortest path routing.

Q. 23 Explain distance vector routing algorithms.

Q. 24 Write short note on : Link state routing.

Q. 25 Compare link state routing and distance vector routing.

Q. 26 Write short note on : path vector routing.

□□□

CHAPTER 12

Unit III

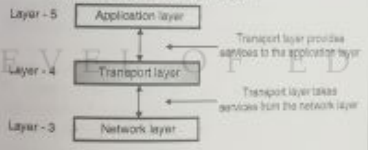
Introduction to Transport Layer

Syllabus :

Introduction to transport layer, Transport layer services, Connectionless and connection oriented protocols, Transport layer protocols, Services, Port number, User datagram protocol, User datagram, UDP services, UDP applications, Transmission control protocol, TCP services, TCP features, Segment.

12.1 Introduction :

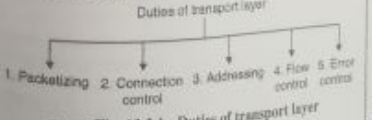
- The transport layer is the core of the Internet model. The application layer programs interact with each other using the services of the transport layer.
- Transport layer provides services to the application layer and takes services from the network layer.
- Fig. 12.1.1 shows the position of the transport layer in the 5-layer internet model. The transport layer is fourth layer in this model. It connects the lower three layers to upper three layers of an OSI layer.



(G-892) **Fig. 12.1.1 : Position of transport layer**

12.2 Transport Layer Duties and Functionalities :

- Transport layer is meant for the process to process delivery and it is achieved by performing a number of functions.
- Fig. 12.2.1 lists the functions of a transport layer.



(G-1407) **Fig. 12.2.1 : Duties of transport layer**

1. Packetizing :

- The transport layer creates packets with the help of encapsulation on the messages received from the application layer. Packetizing is a process of dividing a long message into smaller ones.
- These packets are then encapsulated into the data field of the transport layer packet. The headers containing source and destination address are then added.
- The length of the message which is to be divided can vary from several lines (e-mail) to several pages.
- But the size of the message can become a problem. The message size can be larger than the maximum size that can be handled by the lower layer protocols.
- Hence the messages must be divided into smaller sections. Each small section is then encapsulated into a separate packet.
- Then a header is added in each packet to allow the transport layer to perform its other functions.

2. Connection control :

- Transport layer protocols are divided into two categories :
  1. Connection oriented.
  2. Connectionless.

Connection oriented delivery :

- A connection oriented transport layer protocol establishes a connection i.e. virtual path between sender and receiver.
- This is a virtual connection. The packet may travel out of order. The packets are numbered consecutively and communication is bi directional.

Connectionless delivery :

A connectionless transport protocol will treat each packet independently. There is no connection between them. Each packet can take its own different route.

3. Addressing :

The client needs the address of the remote computer it wants to communicate with. Such a remote computer has a unique address so that it can be distinguished from all the other computers.

4. Flow and error control :

For high reliability the flow control and error control should be incorporated.

- **Flow control :** We know that data link layer can provide the flow control. Similarly transport layer also can provide flow control. But this flow control is performed end to end and not across a single link.
- **Error control :** The transport layer can provide error control as well. But error control at transport layer is performed end to end and not across a single link. Error correction is generally achieved by retransmission of the packets discarded due to errors.

Congestion control and QoS :

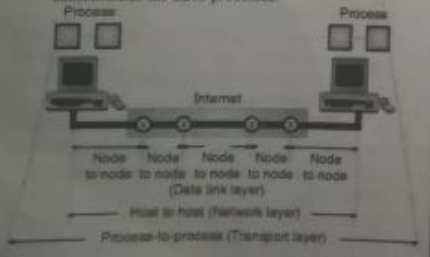
- The congestion can take place in the data link, network or transport layer. But the effect of congestion is generally evident in the transport layer.
- Quality of Service (QoS) can be implemented in other layers but its actual effect is felt in the transport layer.
- The transport layer enhances the QoS provided by the network layer.

12.3 Transport Layer Services :

In this section we are going to discuss the services provided by the transport layer.

12.3.1 Process-to-Process Communication :

- The data link layer performs a node to node delivery. The network layer carries out the datagram delivery between two hosts (host to host delivery).
- But the real communication takes place between two processes or application programs for which we need the **process-to-process delivery.**
- The transport layer takes care of the **process-to-process delivery.** In this a packet from one process is delivered to the other process.
- The relationship between the communicating processes is the client-server relationship. Fig. 12.3.1 demonstrates the three processes.



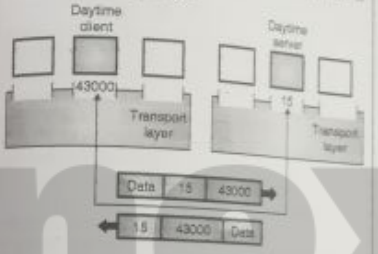
(G-894) **Fig. 12.3.1 : Types of data deliveries**

- There is a difference between host-to-host communication and process to process communication that we need to understand clearly.
- The host to host (computer to computer) communication is handled by the network layer. But this communication only ensures that the message is delivered to the destination computer. But this is not enough.
- It is necessary to handover this message to the correct process. The transport layer will take care of this.

12.3.2 Addressing : Port Number :

- There are several ways of achieving the process-to-process communication, but the most common method is using the client-server paradigm.
- **Client** is defined as the process on the local host. It needs services from another process called **server** which is on the other (remote) host.
- Both client and server have the same name. Some of the important terms related to the client-server paradigm are :
  1. Local host
  2. Remote host
  3. Local process
  4. Remote process
- We can use the IP addresses to define the local host and remote host. But this is not enough to define a process.
- In order to define a process, we have to use one more identifier called **Port Numbers.** In TCP/protocol suite, the port numbers are integers and they are numbered between 0 and 65,535.
- At the data link layer we need a MAC address, at the network layer we need to use an IP address. A datagram uses the destination IP address to deliver the datagram and uses the source IP address for the destination's reply.
- At the transport layer a transport layer address called a **port number** is required to be used to choose among multiple processes running on the destination host.
- The destination port number is required to make the packet delivery and the source port number is needed to return back the reply.
- In the Internet model, the port numbers are 16 bit integers. Hence the number of possible port numbers will be $2^{16} = 65,535$ and the port numbers range from 0 to 65,535.
- The client program identifies itself with a port number which is chosen randomly. This number is called as **ephemeral port number.** Ephemeral means short lived. It is used because life of a client is generally short.
- The server process should also identify itself with a port number but this port number can not be chosen randomly.

- The Internet uses universal port numbers for servers and these numbers are called as **well known port numbers.**
- Every client process knows the well known port numbers of the pre identified server process.
- For example, a Day time client process can use an ephemeral (temporary) port number 43000 for identifying itself, the Day time server process must use the well known (permanent) port number 15. This is illustrated in Fig. 12.3.2.



(G-595) **Fig. 12.3.2 : Concept of port numbers**

What is difference between IP Addresses and Port Numbers ?

- The IP addresses and port numbers have altogether different roles in selecting the final destination of data.
- The destination IP address is used for defining a particular host among the millions of hosts in the world.
- After a particular host is selected, the port number is used for identifying one of the processes on this selected host.

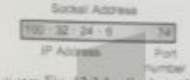
IANA Ranges :

- The port numbers are divided into three ranges by IANA (International Assigned Number Authority).
- The ranges are as follows :
  1. Well known ports
  2. Registered ports
  3. Dynamic or private ports.

1. **Well known ports :** The ports from 0 to 1023 are known as well known ports. They are assigned as well as controlled by IANA.
2. **Registered ports :** The ports from 1024 to 49,151 are neither controlled nor assigned by IANA. We can only register them with IANA to avoid duplication.
3. **Dynamic or private ports :** The ports from 49,152 to 63,535 are known as dynamic ports and they are neither controlled nor registered. They can be used by any process. Dynamic ports are also known as private ports and dynamic port are called as ephemeral ports.

Socket Address :

- Process to process delivery (transport layer communication) has to use two addresses, one is IP address and the other is port number at each end to make a connection. Hence a process to process delivery uses the combination of these two.
- The combination of IP address and port number is as shown in Fig. 12.3.3 and it is known as the socket address.
- The client socket address defines the client process uniquely whereas the server socket address defines the server process uniquely.



(G-596) **Fig. 12.3.3 : Socket address**

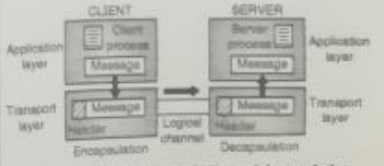
- A transport layer protocol requires the client socket address as well as the server socket address. These two addresses contain four pieces.
- These four pieces go into the IP header and the transport layer protocol header.
- The IP header contains the IP addresses while the UDP and TCP headers contain the port numbers.
- If we want to use the transport layer services in the Internet, then we have to use a pair of socket addresses namely the clients socket address and the server's socket address.

12.3.3 Encapsulation and Decapsulation :

- The transport layer carries out the **Encapsulation** of the message at the sending end and then **Decapsulation** at the receiving end when two computers communicate. This process has been illustrated in Fig. 12.3.4.

Encapsulation :

- At the sending end the process that has a message to send, will pass it to the transport layer alongwith a pair of socket addresses and some additional information.
- The transport layer adds its own header to this data. This packet at the transport layer in the Internet is known by different names such as **user datagram, segment or packet.**



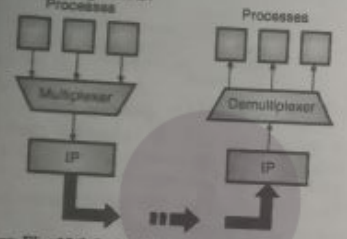
(G-3912) **Fig. 12.3.4 : Encapsulation and decapsulation**

Decapsulation :

- When the segment or datagram arrives at the receiving end, the header is isolated and destroyed, and the message is delivered to the process running at the application layer as shown in Fig. 12.3.4.
- The socket address of the sender process is then handed over to the destination process.

12.3.4 Multiplexing and Demultiplexing :

- The addressing mechanism allows multiplexing and demultiplexing taking place at the transport layer as shown in Fig. 12.3.5.



(G-2017) **Fig. 12.3.5 : Multiplexing and demultiplexing**

Multiplexing :

- At the sending end, there are several processes that are interested in sending packets. But there is only one transport layer protocol (UDP or TCP). Thus it is a many processes-one transport layer protocol situation.
- Such a many-to-one relationship requires multiplexing.
- The protocol first accepts messages from different processes. These messages are separated from each other by their port numbers. Each process has a unique port number assigned to it.
- Then the transport layer adds header and passes the packet to the network layer as shown in Fig. 12.3.5.

Demultiplexing :

- At the receiving end, the relationship is one as to many. So we need a demultiplexer.
- First the transport layer receives datagrams from the network layer.
- The transport layer then checks for errors and drops the header to obtain the messages and delivers them to appropriate process based on the port number.

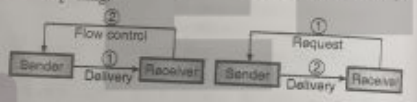
12.3.5 Flow Control :

- If the packets produced by the sender are at a rate X and the receiver is receiving them at a rate Y, then for $X = Y$, there will be a perfect balance observed in the system.
- But if X is higher than Y (source is producing packets at a rate which is higher than the rate at which the receiver is accepting them), then the receiver can be overwhelmed and has to **discard** some packets.
- And if X is less than Y (i.e. source is producing packets at slower rate than the rate of acceptance at the receiver) then system becomes **less efficient.**
- Flow control is related to the situation in which $X > Y$ because it is very important to prevent data loss (due to discarding of packets) at the receiver site.

Pushing and pulling for flow control :

- There are two different ways of delivering the packets produced by the sender to the receiver. They are pushing or pulling.

1. **Pushing :**
   If the sender is sending the packets soon as they are produced, without receiving any prior request from the receiver then this type of deliver is called as **pushing.** Fig. 12.3.6(a) illustrates this concept.
2. **Pulling :**
   If the sender sends the produced packets only when they are requested by the receiver then the delivery is called as **pulling.** Fig. 12.3.6(b) illustrates the principle of pulling.



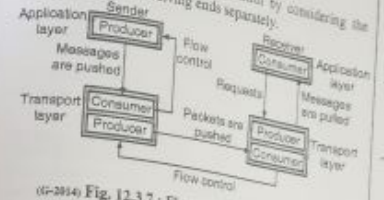
(a) Concept of pushing (b) Concept of pulling

(G-2013) **Fig. 12.3.6**

- In case of **pushing** type delivery, if the packets are being sent at a higher rate than that of receiving, then the receiver will be **overwhelmed,** and some received packets will have to be discarded.
- In order to avoid discarding of packets, the **flow control** will have to be exercised. For this the receiver has to warn the sender to stop the delivery when it is overwhelmed and it has to inform the sender again to start delivery when it (receiver) is ready, to receive the packets.
- In case of **pulling type delivery,** the receiver is actually pulling the packets from the sender. It requests for the packets when it is ready. Therefore the flow control is not required in this case.

12.3.6 Flow Control at Transport Layer :

- The concept of flow control at transport layer has been illustrated in Fig. 12.3.7. It shows the communication taking place between a sender and a receiver.
- As shown in Fig. 12.3.7, there are four entities involved in this communication. They are as follows :
  1. Sender process.
  2. Sender transport layer.
  3. Receiver process.
  4. Receiver transport layer.
- We will discuss the flow control by considering the sending and receiving ends separately.



(G-2014) **Fig. 12.3.7 : Flow control at transport layer**

Sending end :

- The first entity on the sending end is the **sender process,** at the application layer. It works only as a **producer** which produces chunk of messages and pushes them to the transport layer on the sending end, as shown in Fig. 12.3.7.
- The second entity on the sending end is the **sender transport layer.** It has two different roles to play.
- First it acts as a **customer** and consumes all the messages produced and pushed by the producer. Then it encapsulates those messages into packets and pushes them to the receiver transport layer as shown in Fig. 12.3.7. Here it acts as a **producer.**

Receiving end :

- The first entity on the receiving end is the receiver **transport layer.** It also has two different roles to play.
- It acts as a **consumer** for the packets pushed by the senders transport layer and it also acts as the **producer.** It has decapsulate the messages and deliver them to the application layer as shown in Fig. 12.3.7.
- However the delivery of decapsulated messages to the application layer is a **pulling** type delivery. That means the transport layer waits till the application layer process requests for the decapsulated messages.

Flow control :

- As shown in Fig. 12.3.7, the flow control is needed for atleast two cases. First is from transport layer of sender to the application layer of sender.
- And secondly form the transport layer of receiver to the transport layer of sender.

Buffers :

- It is possible to implement the flow control in many different ways. One of the ways of implementation is to use two **buffers** one each at the sending and receiving transport layers.
- A **buffer** is nothing but a set of memory locations which can temporarily hold (store) packets.
- It is possible to exercise flow control communication by sending signals from the consumer to producer.
- The **flow control** at the **sending end** takes place as follows : As soon as the buffer at the transport layer becomes full it sends the stop message to its application layer in order to stop the chunk of messages that are being pushed into the buffer.
- The second flow control takes place at the receiver transport layer as follows : As soon as the buffer at receiver transport layer becomes full, it will inform the sender transport layer to stop pushing the packets.
- Whenever the buffer becomes partially empty, it again informs the sender transport layer to start sending the packets again.

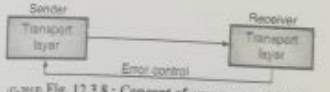
12.3.7 Error Control :

Need of error control :

- In the Internet, the network layer protocol IP has the responsibility to carry the packets from the transport layer at the sending end to the transport layer at the receiving end.
- But IP is unreliable. Therefore transport layer should be made reliable, in order to ensure reliability at the application layer.
- We can make the transport layer reliable by adding the **error control service** to the transport layer.

Duties of error control mechanism :

- Following are the important responsibilities of the error control mechanism introduced at the transport layer :
  1. To find and discard the corrupted packets.
  2. To keep the track of lost and discarded packets and to resend them.
  3. Identify the duplicate packets and discard them.
  4. To buffer out of order packets until the missing packets arrive.
- In the error control process, only the sending and receiving transport layers are involved. That means it is assumed that the chunk of messages exchanged between the application layers and transport layers are error free.
- The concept of error control at the transport layer level is demonstrated in Fig. 12.3.8.
- The receiving transport layer manages the error control by communicating with the sending transport layer about the problem.



(G-2015) **Fig. 12.3.8 : Concept of error control at the transport layer**

Sequence numbers :

- In order to exercise the error control at the transport layer following two requirements should be satisfied :

1. The sending transport layer should know about the packet which is to be resent.
2. The receiving transport layer should know about the packets which are duplicate or the ones that have arrived out of order.

- The requirements can be satisfied only if each packet has a unique **sequence number.**
- If a packet is either corrupted or lost the receiving transport layer will somehow inform the sending transport layer about the sequence number of those packets and request it to resend those packets.
- Due to the unique sequence number assigned to each packet it is possible for the receiving transport layer to identify the duplicate packets received. The out of order packets can also be recognized by observing gaps in the sequence numbers of the received packets.
- Packet numbers are given sequentially. But the length of the sequence number cannot be too long because the sequence number is to be included in the header of the packets.
- If the header of a packet allows "m" bits per sequence number, then the range of sequence number will be from 0 to 2^m-1. For example if m = 3 then the range of sequence numbers will be from 0 to 7.
- Thus sequence numbers are modulo 2^m.

Acknowledgement :

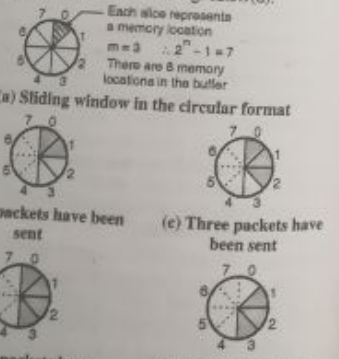
- The receiver side can send an acknowledgement (ACK) signal corresponding to each packet or each group of packets which arrived safe and sound.
- The question is what happens if a received packet is corrupted ? The answer is that the receiver simply discards the corrupted packet and does not send any ACK signal for it.
- The sender can detect a lost packet with the help of a timer. A timer is started at the sending end as soon as a packet is sent. If the ACK does not arrive before the expiry of the timer, then the sender treats the packet to be either lost or corrupted and resends it.
- The receiver silently discards the duplicate packets. It will either discard the out of order packets or stored until the missing packet is received.
- Note that every discarded packet is treated as a lost packet by the sender.

12.3.8 Combination of Flow and Error Control :

- Till now we have discussed the following important concepts :
  1. We need to use buffers at the sending and receiving ends for exercising the flow control.
  2. Also we have to use the sequence numbers and acknowledgements for exercising the error control.
- We can combine these two concepts together by using two numbered buffers one at the sender and the other at the receiver, in order to exercise a combination of flow and error control.
- At the sending end, when a packet, is prepared to be sent, the number of the next free location (x) in the buffer is used as the sequence number of that packet.
- As soon as the packet is sent, its copy is stored at location (x) in the sending end buffer and the sender waits for the acknowledgement from the receiver.
- On reception of the acknowledgement of the sent packet, the copy of that packet is purged to make the memory location (x) free again.
- At the receiver, when a packet having a sequence number "y" arrives, it is stored at the memory location "y" in the receiver buffer until the receiver application layer is ready to receive it. The receiver will send the ACK message back to sender to inform it that packet "y" has arrived.

Sliding window :

- As the sequence numbers are modulo 2^m, we can use a circle as shown in Fig. 12.3.9 to represent the sequence number from 0 to 2^m-1.
- We can represent the buffer as a set of slices, called as the **sliding window** which will occupy a part of the circle at any time.
- In Fig. 12.3.9, we have assumed that m = 3. Therefore $2^m - 1 = 7$ and the sequence numbers are from 0 to 7. Hence the number of memory locations in a buffer will also be 8 i.e. 0 to 7.
- The sliding windows will correspond to the sender as well as receiver.
- On the sending side, when a packet is sent we will mark the corresponding slice. Therefore when marking of all the slices is done, it means the **sending buffer is full,** and it cannot accept any further messages from the application layer as shown in Fig. 12.3.9(d).



(a) Sliding window in the circular format

(b) Two packets have been sent

(c) Three packets have been sent

(d) Four packets have been sent. The window is full

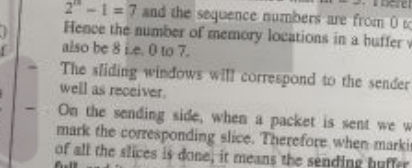
(e) Packet 0 has been acknowledged and the window slides

(G-3017) **Fig. 12.3.9**

- When the acknowledgement for segment "0" arrives at the sending end, the corresponding segment (segment 0) is unmarked and window slides ahead by one slice as shown in Fig. 12.3.9(e). The size of the **sending window** is 4.
- Note that the sliding window is just an abstraction. In actual practice, computer variables are used to hold the sequence number of the next packet to be sent and the last packet sent.

Sliding window in the linear format :

- This is another way to diagrammatically represent a sliding window. It is as shown in Fig. 12.3.10.
- The principle of this type of sliding window is same as that of the circular representation. The linear format is the most preferred format. It needs less space on paper.
- Fig. 12.3.10(a), (b), (c) and (d) are the sliding windows presented in the linear format corresponding to Figs. 12.3.9(b), (c), (d) and (e) respectively in the circular presentation.



(a) Two packets have been sent

(b) Three packets have been sent

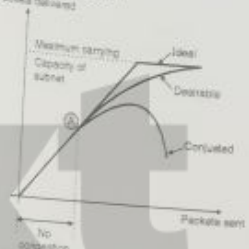
(c) Four packets have been sent. The window is full.

(d) Packet 0 has been acknowledged and the window slides

(G-3018) **Fig. 12.3.10 : Sliding windows presented in the linear format**

12.3.9 Congestion Control :

- An important issue in a packet switching network is congestion.
- If an extremely large number of packets are present in a part of a subnet, the performance degrades. This situation is called as congestion.
- Congestion in a network may occur when the load on the network i.e. the number of packets sent to the network is greater than the capacity of the network (i.e. the number of packets a network can handle).
- Fig. 12.3.11 explains the concept of congestion graphically.
- Upto point A in Fig. 12.3.11, the number of packets sent into the subnet by the host is within the capacity of the network. So all these packets are delivered. In short the number of packets delivered is proportional to number of packets sent and no congestion takes place.
- But after point A, the traffic increases too far. The routers cannot cope with the increased traffic and they begin to lose packets. The congestion begins here.
- As the traffic increases further, the performance degrades more and more packets are lost and congestion worsens.
- At very high traffic, the performance collapses completely and almost all packets are lost. This is the worst possible congestion.



(G-4153) **Fig. 12.3.11 : Concept of congestion**

Need of congestion control :

- We may define the **congestion control** as the mechanisms and techniques to control the congestion and keep the load below the capacity.
- It is not possible to completely avoid the congestion but it is necessary to avoid it otherwise control it.
- Congestion still result in long queues, which results in buffer overflow and loss of packets.
- So congestion control is necessary to ensure that the user gets the negotiated QoS (Quality of Service).

Causes of congestion :

- Congestion happens in any network due to waiting, and due to the abnormality in the flow.
- It also occurs due to the fact that routers and switches have queues at the buffers which store packet before and after their processing.

12.3.10 Connectionless and Connection Oriented Services :

- A transport layer protocol is capable of providing two types of services :
  1. Connectionless services.
  2. Connection oriented services.

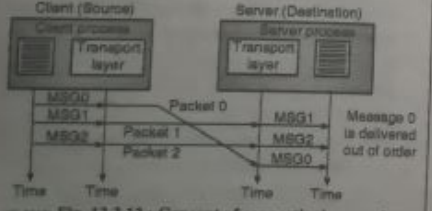
- The meaning of the words connectionless and connection oriented is different at the transport layer than that at the network layer.
- A connectionless service at the network layer means different datagrams of the same message following different paths.
- However at the transport layer, the meaning of connectionless service is independency between different packets.
- On the other hand a connection oriented service means the packets are interdependent.

Connectionless service :

- Refer Fig. 12.3.12 to understand the concept of connectionless service.
- The source process at the application layer first divides its message in chunks of data the size of which is acceptable to the transport layer.
- These data chunks are then delivered to the transport layer one by one. These chunks are treated as independent units by the transport layer.
- Every data chunk arriving from the application layer is encapsulated in a packet by the transport layer and sent to the destination transport layer as shown in Fig. 12.3.12.

Out of Order Delivery :

- In Fig. 12.3.12 we have considered three chunks of independent messages 0, 1 and 2. As the corresponding packets also are independent of each other and as they are free to follow their own path, these packets can arrive out of order at the destination as shown in Fig. 12.3.12.
- Naturally they are delivered to server process in an out of order manner.



(G-2016) **Fig. 12.3.12 : Concept of connectionless service**

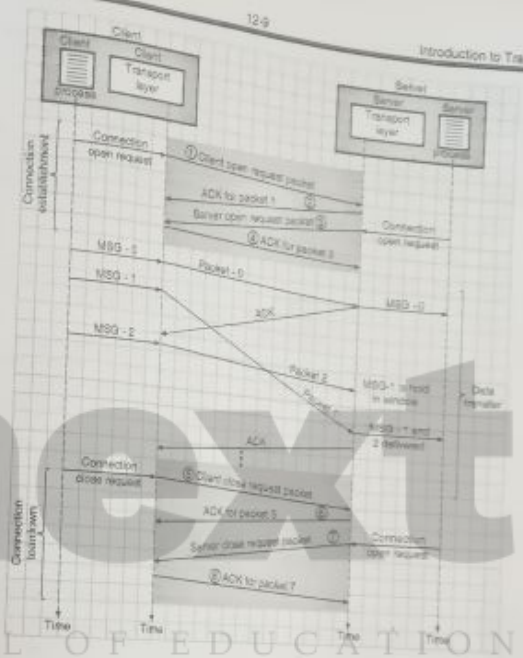
- As seen in Fig. 12.3.12, at the sending end (client) the three chunks of messages 0, 1 and 2 are delivered to the transport layer in the order 0, 1, 2.
- But packet 0 travels a longer path and undergoes an extra delay. Therefore the packets are not delivered in order at the destination (server) transport layer.
- Therefore the message chunks delivered to the server process will also be out of order (1, 2, 0).
- If these chunks are of the same message then due to their out of order delivery the server will receive a strange message.

One packet is lost :

- The UDP packets are not numbered. So if one of the packets is lost, then the receiving transport layer will not have any idea about the lost packet. It will simply deliver the received chunks of messages to the server process.
- The above problems arise due to **lack of coordination** between the two transport layers. Due to this lack of co-ordination it is not possible to implement flow control, error control or congestion control in the connectionless service.

Connection oriented service :

- As we know, there are three stages involved in the connection oriented service. They are :
  1. Connection establishment.
  2. Exchange of data.
  3. Connection teardown.
- The connection oriented service is present at the network layer as well, but it is different from that at the transport layer.
- At the network layer, the meaning of connection oriented service involves the co-ordination between the hosts on either sides and all the routers between them.
- But at the transport layer, the meaning of connection oriented service is the end to end service that involves only the two hosts.
- Refer Fig. 12.3.13 to understand the concept of connection oriented service at the transport layer.
- In Fig. 12.3.13, all the three stages namely connection establishment, data exchange and connection teardown have been shown.
- It is important to note that it is possible to implement the flow control, error control and congestion control in the connection oriented service.



(G-2016) **Fig. 12.3.13 : Concept of connection oriented service**

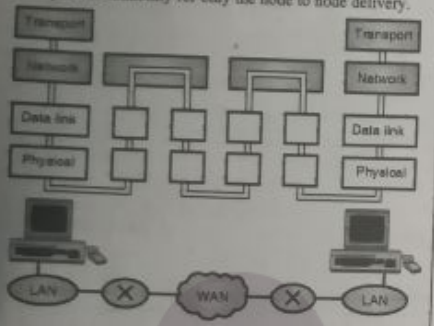
Comparison of Connection Oriented and Connectionless Services :

Sr. No.	Parameter	Connection oriented	Connectionless
1.	Reservation of resources	Necessary	Not necessary
2.	Utilization of resources	Less	Good
3.	State information	Lot of information required	Not much information is required to be stored
4.	Guarantee of service	Guaranteed	No guarantee
5.	Connection	Connection needs to be established	Connection need not be established
6.	Delays	More	Less
7.	Overheads	Less	More
8.	Packets travel	Sequentially	Randomly
9.	Congestion due to overloading	Not possible	Very much possible

12.3.11 Reliability at Transport Layer Versus Reliability at DLL :

- The transport layer services can be of two types :
  1. Reliable services 2. Unreliable services.
- If the application layer program needs reliability then the reliable transport layer protocol is used which implements the flow and error control at the transport layer. But this service will be slow and more complex.
- But some application layer programs do not need reliability because they have their own flow and error control mechanisms. Such programs use an unreliable service.
- UDP is connectionless and unreliable, but TCP is connection oriented and reliable protocol. Both these are the transport layer protocols.

- We need reliability at the transport layer eventhough data link layer is reliable because the data link can provide reliability for only the node to node delivery.



(G-898) **Fig. 12.3.14 : Error control**

- The error control at the data link layer does not guarantee error control at the transport layer. The network layer service in the Internet is unreliable. Hence reliability at the transport layer must be ensured independently.
- Therefore flow and error controls are implemented in TCP using the sliding window protocols. This is reliability assurance at the transport layer.

Note that the error is checked only upto the data link layer by the data link error control system.

12.3.12 Quality of Service (QoS) :

- As mentioned earlier, the QoS parameters are as follows :

1. **Connection establishment delay :**
   - The time difference between the instant at which a request for transport connection is made and the instant at which it is confirmed is called as **connection establishment delay.**
   - This delay should be as short as possible to ensure better service.
2. **Connection establishment failure probability :**
   - Sometimes the connection may not get established even after the maximum connection establishment delay.
   - This can be due to network congestion, lack of table space or some other problems.
3. **Throughput :**
   - It is defined as the number of bytes of user data transferred per second, measured over some time interval.
   - Throughput is measured separately for each direction.
4. **Transit delay :**
   It is the time duration between a message being sent by the transport user from the source machine and its being received by the transport user at the destination machine.
5. **Residual error ratio :**
   - It measures the number of lost or garbled messages as a percentage of the total messages sent.
   - Ideally the value of this ratio should be zero and practically it should be as small as possible.
6. **Protection :**
   This parameter provides a way to protect the transmitted data against reading or modifying it by some unauthorised parties.
7. **Priority :**
   - Using this parameter the user can show that some of its connections are more important (have higher priority) than the other ones.
   - This is important when congestions take place. Because the higher priority connections should get service before the low priority connections.
8. **Resilience :**
   Due to internal problem or congestion the transport layer spontaneously terminates a connection. The resilience parameter gives the probability of such a termination.

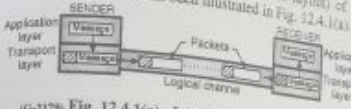
12.4 Transport Layer Protocols :

- We have discussed a few transport layer services in the previous section. By combining a set of these services as per requirement, we can create a transport layer protocol.
- It is important to understand the behavior of these general protocols, before we discuss the transport layer protocols such as UDP and TCP.
- In this section we will discuss the following protocols :
  1. Simple protocol.
  2. Stop and wait protocol.
  3. Go back N (GBN) protocol.
  4. Selective repeat protocol.
  5. Bidirectional protocol. (Piggybacking).
- Initially we will discuss all these protocols as **simplex** i.e. **unidirectional** protocols and then we will see how to make them the **full duplex** i.e. **bidirectional** protocols.

12.4.1 Simplex Protocol :

- This is the simplest type of connectionless protocol which has the following characteristics :
  1. No flow control.
  2. No error control.
  3. The receiver does not get overwhelmed.

- Because the receiver does not get overwhelmed due to the incoming packets even at very high rate, the receiver can handle any packet immediately as soon as it is received.
- The principle of operation (or protocol layout) of the simple protocol has been illustrated in Fig. 12.4.1(a).



(G-2179) **Fig. 12.4.1(a) : Layout of the simple protocol**

Operation :

At the sender :

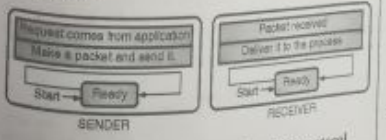
- The application layer at the sender, sends its message to the transport layer.
- The sender transport layer receives the message and makes a packet out of it.
- This packet is then sent over the logical channel between the transport layers on the two ends.

At the receiver :

- The network layer at the receiver (not shown in Fig. 12.4.1(a)) delivers the received packet to the transport layer.
- The receiver transport layer extracts the message from the packet (decapsulation) and sends the message to the application layer.

FSM :

- In this protocol, the sender should not send a packet as long as its application layer does not have a message to send.
- Whereas the receiving transport layer should not deliver a message to its application layer unless it receives a packet from the sender.
- These two requirements suggest, the sender and the receiver have only one state : **Ready state.**
- The sending machine remains in the **ready state** until a process in its application layer sends a request to send its message.
- As soon as the request comes, the sending machine will encapsulate the message and send it to the receiver.
- The receiving machine also remains in **ready state** until it receives a packet from the sender.
- On arrival of a packet, the receiver decapsulates it and delivers the extracted message to the application layer process.



(G-2180) **Fig. 12.4.1(b) : FSM for the simple protocol**

- Note that the UDP protocol is a slight modification of this protocol. The FSM (Finite State Machine) for this protocol has been shown in Fig. 12.4.1(b) and its flow diagram is as shown in Fig. 12.4.1(c).

Flow Diagram :

- The communication between the sender and receiver using the simple protocol has been shown in Fig. 12.4.1(c).
- The sender keeps sending the packets, without taking the receiver into consideration at all.

(G-2181) **Fig. 12.4.1(c) : Flow diagram for the simple protocol**

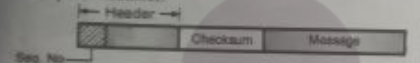
12.4.2 Stop and Wait Protocol :

- The second transport layer protocol that we will discuss now is a **connection oriented** protocol called as **stop and wait protocol.**
- The operation of this protocol are as follows :
  1. It is a connection oriented protocol.
  2. It provides both flow and error control.
  3. Sender sends one packet at a time and waits for its acknowledgement from receiver before sending the next packet.
  4. A checksum is added to each data packet so as to detect a corrupted packet.
  5. At the receiver, the checksum in each packet is checked. If found incorrect, the receiver considers it as the corrupted packet and discards it silently. Such a packet is not acknowledged by the receiver.
  6. If the sender does not receive an acknowledgement for a packet within a predecoded time, it understands that the packet is either corrupted or lost.
  7. The sender starts a timer everytime it sends out a packet. If it receives the acknowledgement for the packet before the expiry of the timer, it stops the timer, and sends the next packet. But if the timer expires before the arrival of acknowledgement, the sender resends the previous packet which was either corrupted or lost.
- Fig. 12.4.2(a) shows the principle of the stop and wait protocol. Note that at any given time there can be only one packet and one acknowledgement in the channel.

(G-2182) **Fig. 12.4.2(a) : Principle of stop and wait protocol**

Sequence number :

- In this protocol, sequence numbers and acknowledgement numbers are used for preventing duplicate packets.
- As shown in Fig. 12.4.2(b), an additional field is created in the packet header of each packet to hold its sequence number.



(G-2183) **Fig. 12.4.2(b) : Packet**

- A very important consideration about the sequence number is the **range of sequence numbers.**
- In order to provide an unambiguous communication with the minimum packet size, we look for the smallest range of sequence numbers.
- Let x be the sequence number of a packet, then the next sequence number should be (x + 1). There is no need for (x + 2). We can show it using the following discussion.
- Suppose that a packet with the sequence number x has been sent by the sender. Then the following three things can possibly happen.

1. Everything is normal :

- The first possibility is that the packet reaches its destination safe and sound without getting corrupted or lost. The receiver sends the acknowledgement for it.
- The acknowledgement reaches the sender safe and sound.
- The sender sends the next packet having a sequence number of (x + 1).

2. Packet corrupted or lost :

- The second possibility is that the sent packet either gets corrupted or gets lost and does not reach the receiving end at all.
- The receiver discards the corrupted packet silently. In either case (corrupted or lost packet), the acknowledgement is not sent back.
- The sender waits for the timer to expire and resends the packet numbered x. The receiver sends back the acknowledgement for it.

3. The acknowledgement is corrupted or lost :

- The packet (numbered x) arrives safe and sound at the receiving end for which it sends an acknowledgement back to the sender.
- However the acknowledgement either get corrupted or gets lost on its way back. Therefore the sender resends the packet (numbered x) again after the expiry of the timer.
- Thus packet x has a duplicate now. The receiver will understand this fact because it was expecting packet numbered (x + 1) to arrive but instead it received the packet numbered x again.

Conclusions :

- From the above discussion we can conclude that sequence numbers x and x + 1 are required so that the receiver can distinguish between cases 1 and 3 discussed above. But it is not necessary to number the packet as (x + 2).
- In case 1, we can number the packet as x again because both the packets (x and x + 1) are acknowledged by the receiver and neither the sender nor the receiver has any ambiguity about it.
- Finally in the case 2 and 3, the new packet is (x + 1) and not (x + 2). Therefore we conclude that only two sequence numbers x and x + 1 are needed and x +2 is not needed.
- So let x = 0 then (x + 1) = 1. Thus there will be only two sequence numbers 0 and 1 and the packet sequence would be 0, 1, 0, 1, 0,.. and so on. Due to the presence of only two distinct sequence numbers, this is called as modulo-2 arithmetic.

Acknowledgement numbers :

- For both types of packets i.e. data packets and acknowledgements, the same sequence numbers should be suitable.
- For this to happen successfully the following convention is used.
- The acknowledgement number always indicates the sequence number of the **next packet** that the receiver is expecting to receive.
- For example, the packet with a sequence number 0 arrives at the receiver safe and sound. Then the corresponding ACK sent by the receiver will have a number 1 on it which means that the next expected packet to be received is packet 1.
- Similarly if packet 1 arrives safe and sound then ACK with acknowledgement 0 is sent back which means that packet – 0 is the next expected packet at the receiver.
- The **control variable** at the sender is called as the **sender (s)** and it points to the only slot present in the send window as shown in Fig. 12.4.2(a).
- Similarly the **control variable** at the receiving end called as the **Receiver (R)** and it points to the only slot present in the receive window as shown in Fig. 12.4.2(a).

FSMs of stop and wait protocol :

- This protocol is a connection oriented protocol. Therefore a connection between the two ends should be established before transferring the data.

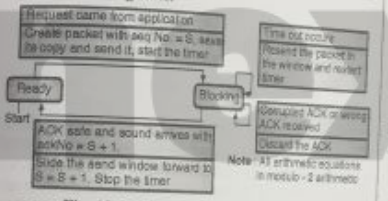
- In other words both sender and receiver must be in the **established state** before the beginning of data exchange.

1. Sender FSM :

- The sender's FSM is shown in Fig. 12.4.2(c). Initially it is in the **ready** state. However it can move between the ready and blocking state.
- The initial value of variable "s" is set to 0.

1. Ready state :

- The sender, when in the ready state waits only for one event to happen, that is the request coming from application layer.
- As soon as such a request comes from the application layer, the sender makes a packet with the sequence number same as "s."
- It stores a copy of this packet and sends the packet. The sender starts the timer and moves into its **blocking state.**



(G-2184) **Fig. 12.4.2(c) : Sender's FSM for stop and wait protocol**

2. Blocking state :

- When the sender is in the blocking state as shown in Fig. 12.4.2(c), the following three possible events can happen :

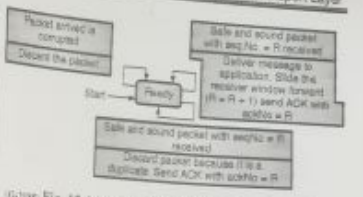
a. An error free ACK is received by the sender. It's ackNo is also correct i.e. (S + 1). The sender then stops the timer, slides the sending window to S = (S + 1) modulo - 2 and moves to the ready state.

b. The ACK received by the sender is either corrupted or a wrong ACK i.e. the one having the ackNo other than (S + 1). The sender discards the ACK.

c. In case if the timer expires (time out condition), the sender resends the only outstanding packet with it. It then restarts the timer as shown in Fig. 12.4.2(c).

2. Receiver FSM :

- The receiver's FSM is shown in Fig. 12.4.2(d). Note that there is no blocking state in the receiver's FSM. There is only the ready state.
- At the receiver also there is a possibility of following three events happening after the arrival of a packet.



(G-2185) **Fig. 12.4.2(d) : FSM of receiver for stop and wait protocol**

a. A safe and sound packet (without corruption) is received with seq.No = R. Then the messages is extracted (decapsulation) and delivered to the application layer. The receive window slides forward to (R = R + 1) modulo - 2 and the receiver sends an ACK with ackNo = R.

b. A safe and sound packet (with any error) arrives, but its seq.No ≠ R. This shows that it is a duplicate packet. The receiver will discard this packet but sends an ACK with ackNo = R.

c. The received packet is corrupted. The receiver silently discards it. No ACK is sent back.

Efficiency of stop and wait protocol :

- The efficiency of the stop and wait protocol is very very low. This is because it sends a packet and simply waits for its ACK before sending the next packet.
- This is a gross underutilization of the communication channel especially if the channel is **thick and long.** A channel is thick if it has a large bandwidth and it is **long** if it has a long round trip time.
- The product of these two parameters is called as **bandwidth delay product.**
- A channel is equivalent to a pipe. If it is underutilized, then it will be called inefficient.
- The number of bits a sender can transmit through the channel can be measured from the value of bandwidth delay product.
- On all these accounts the stop and wait protocol proves to be extremely inefficient.

Pipelining :

- In networking and even other areas, a task is started before the ending of previous task. This is known as **pipelining.**
- In the stop and wait protocol, the senders sends a packet and waits for its acknowledgement before sending the next packet.
- This shows that there is no pipelining in the stop and wait protocol.
- But in the other protocols that we are going to discuss after the concept of pipelining will be used.

- Therefore it is possible for the sender to send several packets before it receives only acknowledgements for the previously sent packets.
- The process of pipelining improves the efficiency of the protocol.

12.4.3 Go Back-N Protocol (GBN) :

- The efficiency of transmission can be improved by transmitting multiple packets while the sender is waiting for acknowledgment.
- That means we should allow more than one outstanding packets even when the sender is waiting for acknowledgment because this will keep the channel busy.
- A protocol which can achieve this goal is our next protocol called Go Back-N (GBN) protocol.
- The most important part in the operation of GBN protocol is that we can send several packets before receiving acknowledgement. But the receiver can buffer only one packet.
- A copy of every sent packet is kept by the sender until it receives the acknowledgement of that packet.
- Fig. 12.4.3(a) shows the outline of GBN protocol which explains its principle of operation. Note the simultaneous presence of multiple packets and multiple acknowledgements in the channel at any given time.

(G-2186) **Fig. 12.4.3(a) : Principle of Go Back-N (GBN) protocol**

Sequence numbers :

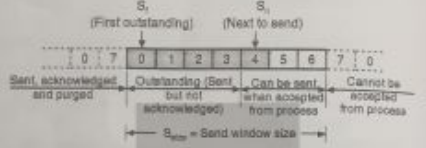
In GBN protocol, the sequence numbers are modulo 2^m, where m denotes the size of sequence number field in bits.

Acknowledge numbers :

- In the GBN protocol, the acknowledgement number is cumulative and it carries the sequence number of the next packet that is expected to be received at the receiver.
- If the ackNo = 6, its an indication that the receiver has received all the packets having sequence number upto 5 safe and sound. Hence the receiver is expecting the packet with seq.No = 6 to arrive next.

Send window :

- We can define the send window as an imaginary box, which covers the sequence numbers of the data packets that can be sent.
- The **maximum size** of the send window is $(2^m - 1)$ for the reasons discussed later on in the chapter.
- In each send window position (it can slide), some sequence numbers indicate the packets that have been already sent whereas the other sequence numbers indicate the data packet that are to be sent.
- In this chapter we assume that the send window size is fixed and has been set to its maximum possible value. But in some protocols the send window size is variable.
- The structure of a send window for the GBN protocol with m = 3 has been shown in Fig. 12.4.3(b). Note that the window size = $2^m - 1 = 2^3 - 1 = 7$.



(G-2187) **Fig. 12.4.3(b) : Format of the send window of GBN**

- At any given time, the send window divides the possible sequence numbers into four regions.
- As shown in Fig. 12.4.3(b), the first region corresponds to the portion to the left of the send window. It consists of the sequence numbers which belong to the packet which are already acknowledged. The sender does not keep any copy of these packets.
- The second region which is shaded in Fig. 12.4.3(b) contains the sequence numbers belonging to the packets that are already sent but not acknowledged by the receiver. That means the exact status of these packets is not known.
- These packets are called as **outstanding packets.**
- The third range, which is not shaded in Fig. 12.4.3(b), contains the sequence numbers belonging to the packets which the sender can send. But the corresponding data is yet to be received from the application layer.
- And finally the fourth range, which is at the right of the send window in Fig. 12.4.3(b), consists of the sequence numbers that cannot be used by the sender until the send window slides to the right hand side.

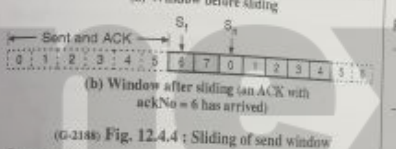
Size and location of send window :

- There are three variables that define the size and location of the send window at any given time. They are :
  1. S_f : Send window, the first outstanding packet.
  2. S_n : Send window ; the next packet to be sent.
  3. S_{size} : Send window, size.
- The sequence number of the first (oldest) outstanding packet is defined by the variable S_f.

- The sequence number, that will be assigned to the next packet to be sent is defined by the variable S_n.
- And finally the size of the send window which is fixed in GBN protocol is defined by the variable S_{size}.

Sliding of send window :

- A send window will slide right on the arrival of acknowledgements.
- Fig. 12.4.4 shows the send window before sliding and after the arrival of an acknowledgement with ackNo = 6. This means that all packets upto seq.No = 5 have reached safe and sound and the receiver is expecting the packet with seq.No = 6 to arrive.



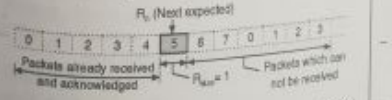
(G-2188) **Fig. 12.4.4 : Sliding of send window**

Conclusion :

From all this discussion we conclude that the send window will slide by one or more slots when the sender receives an errorfree ACK whose ackNo is greater than or equal to S_f and less than S_n.

Receive window :

- The receive window has two tasks : First it has to ensure that correct data packets are received and second is to make sure that correct acknowledgements are sent.
- The size of receive window in the GBN protocol is always 1. Therefore, the receiver is always expecting a specific packet to arrive.
- That means the receiver will discard any packet which arrives out of order and the sender has to resend the discarded packet.
- The receive window for the GBN protocol is shown in Fig. 12.4.5. It has only one variable R_n i.e. receive window, next packet expected.



(G-2189) **Fig. 12.4.5 : Structure of receive window of GBN**

- The sequence numbers to the left of the receive window correspond to the already received and acknowledged packets.
- The sequence numbers to the right of receive window correspond to the packets which cannot be received.
- The receiver discards any packet that belongs to these two ranges. It will only accept that packet whose sequence number exactly matches with the value of R_n.
- Like the sliding window, the receive window also slides but only by one slot at a time. On reception of a correct packet, the receive window slides to $R_n = (R_n + 1) \text{ modulo } 2^m$.
- If a corrupted packet is received, the receive window does not slide at all.

Timers :

- Ideally there should be one timer per packet, which is sent. In GBN protocol only one timer is used.
- The reason for this is that the timer for the first outgoing packet will always expire first. If so, then all the outstanding packets will be resent by the sender.

Resending the packets :

- As stated earlier, on the expiry of the only timer (also called as time out), all the outstanding packets will be resent.
- As an example, let us assume that the sender has already sent the packet having seq.No 6 (S_n = 7) but the time out takes place (that means the only timer in GBN has expired).
- If S_f = 3, then it is an indication that the packets 3, 4, 5 and 6 are all outstanding packets i.e. they are sent but not acknowledged.
- Hence, as soon as the timer expires, the sender will go back and resend all the outstanding packets i.e. packets 3, 4, 5 and 6.
- This is the reason behind the name of this protocol which Go Back N. The sender goes back by N slots and resends all the packets from there as soon as the timer expires.

Send window size :

- Now we are going to discuss, why in GBN protocol the size of send window should be less than 2^m.
- Let m = 2. Therefore the size of the send window will be $2^m - 1 = 3$. With this send window size if all the acknowledgements are lost and the timer expires, then the sender resends all packets.
- As the receiver is expecting packet 3 and not 0, it will successfully identify the resent packet 0 as the duplicate and discard it.
- But if the send window size is $2^m = 4$, and if all the acknowledgement are lost and the timer expires, then the sender will retransmit packet 0.
- But this time, the receiver also is expecting packet 0 to arrive (next cycle). Hence it won't treat the resent packet 0 as the duplicate packet and won't discard it.
- In fact the duplicate packet 0 is accepted as the legitimate packet 0 of the next cycle. This is an error.

- From this example we conclude that the size of send window in GBN protocol should be **less than 2^m.**

Comparison of GBN with stop and wait :

- The GBN and stop and wait protocols are somewhat similar to each other.
- The stop and wait protocol is actually a GBN protocol with only two sequence numbers (0 and 1) and send window size of 1.
- In stop and wait protocol, the modulo 2 arithmetic is used whereas in GBN protocol, modulo 2^m arithmetic is said to have been used.
- Thus stop and wait protocol is a GBN protocol with m = 1.

12.4.4 Selective Repeat Protocol :

- The process at the receiving end is simplified in the GBN protocol to a great extent. This is because R_n is the only variable which is to be tracked by the receiver and the out of order received packets need not be buffered. They are to be simply discarded.
- But the problem with this protocol is its **inefficiency** if the underlying protocol tends to loose a lot of packets.
- This is because everytime with the loss of a packet the sender has to send all the outstanding packets.
- It is possible that some of these packets may have been received without any error but out of order.
- If the network congestion is already existing, then it will become worse due to these frequently resent packets. The worsened network congestion will result in the loss of more packets which leads to retransmission on of more packets and so on.
- This is called as an **avalanche effect** which may eventually cause total collapse of the network.
- In order to overcome these problems of the GBN protocol, a new protocol has been deviced which is called as the **Selective Repeat Protocol.**
- This new protocol, as the name suggests, resends only **selected packets**, that are actually corrupted or lost. It does not resend all the outstanding packets like the GBN protocol.
- Thus will reduce the number of resent packets and therefore reduces the possibility of network congestion.

(G-2190) **Fig. 12.4.6 : Outline of selective repeat protocol**

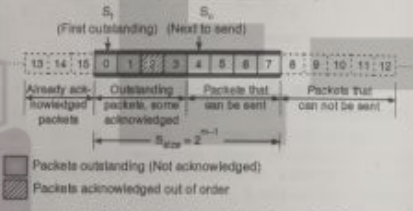
- The principle of selective repeat protocol has been illustrated in Fig. 12.4.6.

Windows :

- In the selective request protocol also there are two windows used : a send window and a receive window.
- However these windows are different from those in the GBN protocol. In this protocol the maximum size of send window is (2^{m-1}). This size is much smaller than that in the GBN protocol. Also the size of receive window is same as that of the send window.

Send and receive windows :

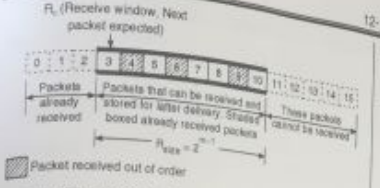
- If m = 4, then the maximum size of the send window is $2^{m-1} = 2^3 = 8$ (It is 15 in the GBN protocol). Fig. 12.4.6(a) shows the structure of the send window.
- Fig. 12.4.6(b) shows the structure of receive window in the selective repeat protocol. Note that it is totally different from that in the GBN protocol.
- The receive window here has the same size as that of the send window (Maximum size = 2^{m-1}).



(G-2191) **Fig. 12.4.6(a) : Send window for selective repeat protocol**

Principle :

- In the selective repeat protocol, the packets equal to the size of the receive window are allowed to arrive out of order.
- The receiver is allowed to keep them until it has a set of consecutive packets which can be delivered to the application layer.
- As the send and receive windows are of the same size, all the packets in the send window can arrive out of order at the receiver and the receiver is allowed to store them until it can deliver them to the application layer.
- However the selective repeat is a reliable protocol. Therefore the receiver is not expected to deliver packets out of order to the application layer.
- The structure of the receiver window for selective repeat protocol is as shown in Fig. 12.4.6(b). It shows that there are packets received out of order. These packets have to wait for the earlier transmitted packets to arrive before all of them are finally delivered to the application layer.



(G-2192) **Fig. 12.4.6(b) : Receive window for selective repeat protocol**

Timer :

- Theoretically in SR protocol a timer is assigned to each outstanding packet in the send window. When a timer expires, only the corresponding packet is resent.
- This is totally different from the GBN protocol which has only one timer for a group of outstanding packets.
- But practically, almost all the transport layer protocols which are based on selective repeat principle use only one timer.

Acknowledgements :

- In GBN protocol, the ackNo is cumulative. It carries the number of the next expected packet to be received. It also confirms that all the previous packets have been received safe and sound.
- But in the SR protocol it is totally different. In SR the ackNo defines the sequence number of only one packet which is received safe and sound. It does not give any feedback about the other packets.

Window sizes :

- The maximum size of send and receive windows in the SR protocol is 2^{m-1} that means $2^m/2$ i.e. half of 2^m.
- If m = 2, all the acknowledgements are lost and if the time out takes place (i.e. timer expires) then sender retransmits packet 0.
- But the receiver window is expecting packet 2 and not packet 0. Hence the receiver will identify packet 0 as the duplicate packet and will discard it. (The sequence number 0 is not in the window).
- Now imagine that the window size is 3, all acknowledgements lost and the timer expires. Now the sender will resend packet 0.
- At this time the receiver is also expecting packet 0 of the next cycle to arrive (0 is the part of the window). Therefore the receiver cannot recognize that packet 0 is a duplicate packet. This is an error.
- That is why in S.R. protocol, the maximum size of the send and receive windows is 2^{m-1} or half of 2^m.

12.4.5 Bidirectional Protocols : Piggybacking :

- Note that in all the protocols discussed so far the data packets flow in only one direction and acknowledgements travel in the opposite direction. Therefore all these four protocols are said to be **unidirectional protocols.**
- However in reality the data packets are travelling in both the directions, client to server and vice versa. The acknowledgements also are travelling in both the directions.
- Thus all the transport layer protocols in real life are bidirectional. We can improve the efficiency of these bidirectional protocols with a technique called **piggybacking.**
- In piggybacking, the data packet going from A to B can also carry acknowledgement for the data packet arrived from B to A.
- Similarly a data packet sent by B to A can carry acknowledgement for the data packet arrived from A to B.

12.5 The Internet Transport Protocols (TCP and UDP) :

- The Internet has two main protocols in the transport layer. One of them is connection oriented and the other one supports the connectionless service.
- TCP (Transmission Control Protocol) is a connection oriented protocol and UDP (User's Data Protocol) is the connectionless protocol.
- UDP is basically just IP with an additional short header.

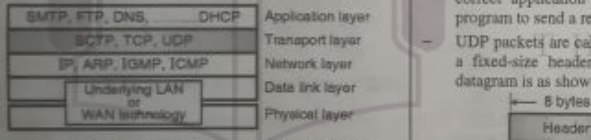
12.6 User Datagram Protocol (UDP) :

- The User Datagram Protocol is a very simple protocol. It adds little to the basic functionality of IP. Like IP, it is an unreliable, connectionless protocol.
- You do not need to establish a connection with a host before exchanging data with it using UDP, and there is no mechanism for ensuring that data sent is received.
- A unit of data sent using UDP is called a Datagram. UDP adds four 16-bit header fields (8 bytes) to whatever data is sent.
- These fields are : a length field, a checksum field, and source and destination port numbers. "Port number", in this context, represents a software port, not a hardware port.
- The concept of port numbers is common to both UDP and TCP. The port numbers identify which protocol module sent (or is to receive) the data.

- Most protocols have standard ports that are generally used for this. For example, the Telnet protocol generally uses port 23. The Simple Mail Transfer Protocol (SMTP) uses port 25. The use of standard port numbers makes it possible for clients to communicate with a server without first having to establish which port to use.
- The port number and the protocol field in the IP header duplicate each other to some extent, though the protocol field is not available to the higher-level protocols. IP uses the protocol field to determine whether data should be passed to the UDP or TCP module.
- UDP or TCP use the port number to determine which application-layer protocol should receive the data.
- Although UDP isn't reliable, it is still a preferred choice for many applications. It is used in real-time applications like Net audio and video where, if data is lost, it's better to do without it than send it again out of sequence. It is also used by protocols like the Simple Network Management Protocol (SNMP).

Relationship with other protocols :

- The relationship of UDP with the other protocols and layers of TCP/IP suite is as shown in Fig. 12.6.1. As shown, UDP is located between IP and application layer. It therefore works as an intermediary between application program and the network layer.



(G-3819) **Fig. 12.6.1 : Relation between UDP and other protocols**

12.6.1 Responsibilities of UDP :

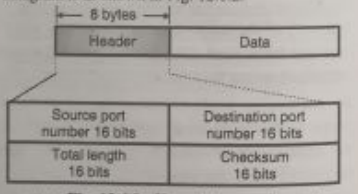
- Being a transport layer protocol, the UDP has the following responsibilities :
  1. To create a process to process communication, UDP uses port numbers to accomplish this.
  2. To provide control mechanisms at the transport layer, UDP does not provide flow control or acknowledgements. It provides error detection. The erroneous packet is discarded.
  3. UDP does not add anything to the services of IP except for providing process to process communication.

12.6.2 Advantages of UDP :

- UDP, despite all its simplicity and powerlessness is still used because it offers the following advantages :
  1. UDP has minimum overheads.
  2. UDP can be easily used if the sending process is not too bothered about reliability.
  3. UDP reduces interaction between sender and receiver.

12.6.3 User Datagram :

- User Datagram Protocol (UDP) provides a connectionless packet service that offers unreliable "best effort" delivery. This means that the arrival of packets is not guaranteed, nor is the correct sequencing of delivered packets.
- Applications that do not require an acknowledgement of receipt of data, for example, audio or video broadcasting uses UDP.
- UDP is also used by applications that typically transmit small amounts of data at one time, for example, the Simple Network Management Protocol (SNMP).
- UDP provides a mechanism that application programs use to send data to other application programs. UDP provides protocol port numbers used to distinguish between multiple programs executing on a single device.
- That is, in addition to the data sent, each UDP message contains both a destination port number and a source port number. This makes it possible for the UDP software at the destination to deliver the message to the correct application program, and for the application program to send a reply.
- UDP packets are called as **user datagrams.** They have a fixed-size header of 8-bytes. The format of user datagram is as shown in Fig. 12.6.2.



(G-4240) **Fig. 12.6.2 : User datagram format**

- The UDP header is divided into the following four 16-bit fields :
  1. Source port
  2. Destination port
  3. Total length
  4. Checksum.

Source Port Number :

- Source port is an optional field, when meaningful, it indicates the port of the sending process, and may be assumed to be the port to which a reply should be addressed in the absence of any other information. If not used, a value of zero is inserted.

- This is a 16 bit field. That means the port numbers can range from 0 to 65,535.
- If the source host is a client, means if a client is sending a request using UDP, then generally a **ephemeral (temporary)** port number is requested by the process and chosen by the UDP.
- If the source host is a server that means if a server is sending a response message, mostly the **well known port** number is used.

Destination Port Number :

- The destination port number also is a 16 bit number and this port number is used by the process running on the destination host.
- If the destination host is a server that means if a client is sending a request to it, then a **well known port** number is used in most cases.
- However if the destination host is a client than means if a server is sending its response to it, then the chosen port number is generally an **ephemeral port** number.

Length :

- It is also a 16 bit field which is used for defining the total length of the UDP datagram including header as well as data. Due to 16 bit length it can define a total length of the datagram upto 65,535 bytes.
- However practically the total length of a UDP datagram is much smaller than 65,535 bytes. This is because the UDP datagram is to be stored in an IP datagram which itself has a length of 65,535 bytes.
- The **length** field in the UDP datagram is actually not necessary, because this UDP datagram is actually encapsulated in an IP datagram and the IP datagram has its own length field.
- So without using the length field in UDP datagram, we can obtain the length of the UDP datagram as follows :

  UDP length = IP length – IP header length

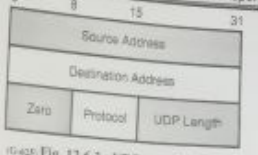
- Note that while delivering the UDP datagram to UDP layer, the IP software drops the IP header.

UDP Checksum :

This is used to verify the integrity (i.e. to detect errors) of the UDP header. The checksum is performed on a "pseudo header" consisting of information obtained from the IP header (source and destination address) as well as the UDP header.

12.6.4 UDP Pseudo Header :

- The purpose of using a pseudo-header is to verify that the UDP packet has reached its correct destination.
- The correct destination consists of a specific machine and a specific protocol port number within that machine.



(G-4235) **Fig. 12.6.3 : UDP pseudo header**

- The UDP header itself specifies only the protocol port number. Thus, to verify the destination, UDP on the sending machine computes a checksum that covers the destination IP address as well as the UDP packet.
- At the ultimate destination, UDP software verifies the checksum using the destination IP address obtained from the header of the IP packet that carried the UDP message.
- If the checksum agrees, then it must be true that the packet has reached the intended destination host as well as the correct protocol port within that host.

User Interface :

A user interface should allow the creation of new receive ports, receive operations on the receive ports that return the data octets and an indication of source port and source address, and an operation that allows a datagram to be sent, specifying the data, source and destination ports and addresses to be sent.

IP Interface :

- The UDP module must be able to determine the source and destination Internet addresses and the protocol field from the Internet header.
- One possible UDP/IP interface would return the whole Internet datagram including the entire Internet header in response to a receive operation. Such an interface would also allow the UDP to pass a full Internet datagram complete with header to the IP to send.
- The IP would verify certain fields for consistency and compute the Internet header checksum.

Protocol Application :

The major uses of this protocol are the Internet Name Server, and the Trivial File Transfer.

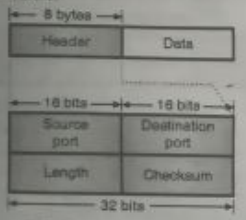
Protocol Number :

This is protocol 17 (21 octal) when used in the Internet Protocol.

Ex. 12.6.1 : The dump of a UDP header in hexadecimal format is as follows :
B C 9 2 0 0 0 D 0 0 2 B 0 0 1 D Obtain the following from it :
1. Source port number
2. Destination port number
3. Total length
4. Length of the data.
5. Packet direction.
6. Name of client process.

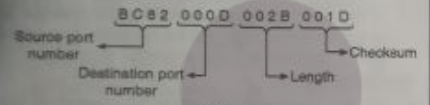
Soln. :

- The standard format of UDP header has been shown in Fig. P. 12.6.1.



(G-2020) **Fig. P. 12.6.1 : UDP header format**

- Therefore we can split the given UDP header in 4 equal parts as follows :



(G-2021)

1. Source port number = $(BC82)_{16}$ ...Ans.
2. Destination port number = $(000D)_{16}$ ...Ans.
3. Total length of UDP packet = $(002B)_{16}$ = $(43)_{10}$ bytes ...Ans.
4. Length of data = Total length – Length of the header = 43 – 8 = 35 bytes ...Ans.
5. Destination port number is $(000D)_{16} = (13)_{10}$

It is a well known port. Hence the direction of UDP packet travel is from client to server.

6. The client process can be obtained from Table 3.6.1 which shows that for well known port number 13, the corresponding client process is "Daytime".

12.7 UDP Services :

In this section we are going to discuss the following important services provided by the UDP :

1. Process to process communication.
2. Connectionless services.
3. Flow control.
4. Error control.
5. Checksum.
6. Congestion control.
7. Encapsulation and decapsulation.
8. Queuing.
9. Multiplexing and demultiplexing.

12.7.1 Process to Process Communication :

- We have already discussed the process to process communication in a general sense, earlier in this chapter.
- UDP also does it with the help of sockets which is a combination of IP address and port numbers. Table 12.7.1 shows different port numbers used by UDP.

Some of these ports can be used by UDP as well as TCP.

Table 12.7.1 : Well known ports used with UDP

Port	Protocol	Description
7	Echo	The received datagram is echoed back to sender.
9	Discard	Any received datagram is discarded.
11	Users	Active users.
13	Daytime	Return the day and the current time.
17	Quote	Return the quote of the day.
19	Chargen	To return a string of characters.
53	Nameserver	Domain Name Service (DNS).
67	BOOT PS	This is the server port to download the bootstrap information.
68	BOOT PC	This is the client port to download bootstrap information.
69	TFTP	Trivial File Transport Protocol.
111	RPC	Remote Procedure Call.
123	NTP	Network Time Protocol.
161	SNMP	Simple Network Management Protocol.
162	SNMP	Simple Network Management Protocol (Trap).

12.7.2 Connectionless Services :

- As UDP is a connectionless, unreliable protocol, each user datagram sent using UDP is an independent datagram.
- Different user datagrams sent by the UDP have absolutely no relationship between them. This is true even for those datagrams which are originating from the same process and being sent to the same destination. The user datagrams **do not have any number.**
- Also the connection establishment and release are not at all required. So each datagram is free to travel any path.
- Only those processes which are sending very short messages can successfully use the UDP.

12.7.3 Flow and Error Control :

- Being a connectionless protocol, UDP is a simple, unreliable protocol. It does not provide any flow control, hence the receiver can overflow with incoming messages.

- UDP does not support any other error control mechanism, except for the checksum.
- There are no acknowledgements sent from destination to sender. Hence the sender does not know if the message has reached, lost or duplicated. If the receiver detects any error using the checksum, then that particular datagram is discarded.

12.7.4 Checksum :

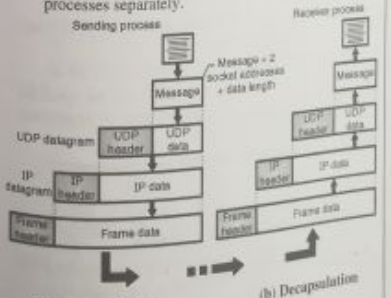
- The calculation of checksum for UDP is different than that for IP. In UDP the checksum is calculated by considering the following three sections :
  1. A pseudoheader
  2. The UDP header.
  3. The data coming from the application layer.
- The checksum in UDP is **optional.** That means the sender can make a decision of not calculating the checksum. If so, then the checksum field is filled with all zeros before sending the UDP packet.
- In case if the calculated checksum is all zeros (when the sender decides to send checksum) then an all 1 checksum is sent.
- This solution works without any problem because, a checksum will never have an all 1 value.

12.7.5 Congestion Control :

- UDP does not provide any congestion control. It assumes that the UDP packets being small, will not create any congestion.
- But this assumption may not always be correct.

12.7.6 Encapsulation and Decapsulation :

- The UDP encapsulates and decapsulates messages in an IP datagram in order to exchange the message between two communicating processes.
- This is as shown in Fig. 12.7.1. We will discuss the two processes separately.



(a) Encapsulation (b) Decapsulation

(G-2022) **Fig. 12.7.1**

Encapsulation :

- Refer Fig. 12.7.1(a). The message produced by a process is to be sent with the help of UDP. The process passes the message and two socket addresses alongwith the length of data to UDP.
- UDP receives this data and adds the UDP header to it as shown. This is called as UDP datagram which is passed to IP with the socket address.
- IP adds its own header to UDP datagram as shown. It enters value 17 into the protocol field. This is an indication that UDP is being used. The IP datagram is then passed on to the data link layer.
- The DLL adds its own header and possibly a trailer to create a frame and sends it to the physical layer.
- Finally the physical layer converts these bits into electrical or optical signals and sends them to the destination machine.

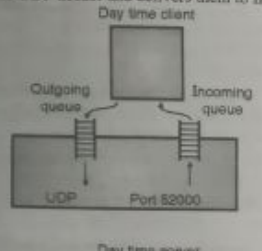
Decapsulation :

- Refer Fig. 12.7.1(b) for understanding of the decapsulation process. The encoded message arrives at the destination physical layer where it decodes the electrical/optical signals into bits and passes them to the DLL.
- The DLL checks the data using header and trailer. The header and trailer are discarded if no errors are found, and the datagram is passed to IP.
- The IP carries out its checking to find the errors and if none are found, the datagram is passed on to UDP, after dropping the IP header.
- The datagram from IP to UDP also contains the sender and receiver IP addresses. This entire user datagram is checked by the UDP with the help of checksum.
- If there is no error detected, then the UDP header is dropped and the application data plus senders socket address are handed over to the process.
- The process can use this senders socket address if it wants to respond to the message received.

12.7.7 Queuing :

- The queues in UDP are related with ports as shown in Fig. 12.7.2.
- A process starts at the client site by requesting a port number from the operating system. In some implementations both incoming and outgoing queues are created in association with each process.
- Every process gets only one port number and hence it can create one outgoing and another incoming queue. The queues function only when the process is running. They are destroyed as soon as the process is terminated.
- The client process uses the source port number mentioned in the request to send message to its outgoing queue.

- UDP removes the queue messages one by one by adding the UDP header and delivers them to IP.

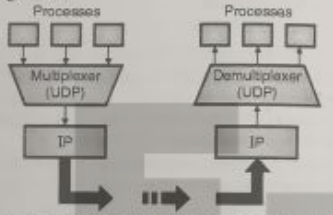


(G-426) **Fig. 12.7.2 : Queues in UDP**

- If the outgoing queue overflows, then operating system tells that client process to wait before sending the next message.
- When the client receives a message, UDP checks if the incoming queue has been created or not. If the queue has been created, then the UDP sends the received datagram to the end of the queue.
- If the queue is not present then UDP will simply discard the user datagram. If the incoming queue overflows, then UDP discards the user datagram and arranges to send the port unavailable message to the server.
- The mechanism to create the server queue is different. The server creates the incoming and outgoing queues using its well known port as soon as it starts running. The queues exist as long as the server is running.
- When a message is received at the server, the UDP checks if the incoming queue has been created or not.
- If the queue is not present, the UDP discards the user datagram. If the queue is present then UDP sends the datagram at the end of the queue.
- If the incoming queue overflows, then UDP drops the user datagram and arranges to send the port unavailable message to the client.
- When the server wants to send a message to client it sends that message to the outgoing queue. These messages are then removed one by one after adding the UDP header. They are delivered to IP.
- If the outgoing queue overflows then the operating system will ask the server to wait before it sends the next message.

12.7.8 Multiplexing and Demultiplexing :

- We have discussed the general principle of multiplexing and demultiplexing in the transport layer.
- Now let us see how to apply the same principle to UDP. Imagine that a host is running a TCP/IP protocol suite and that there is only one UDP and a number of processes which would like to use the services of UDP.
- UDP handles such a situation by using the principle of multiplexing and demultiplexing as shown in Fig. 12.7.3.



(G-2023) **Fig. 12.7.3 : Multiplexing and demultiplexing**

Multiplexing :

- At the sending end, there are several processes that are interested in sending packets. But there is only one transport layer protocol (UDP or TCP). Thus it is a many processes-one transport layer protocol situation.
- Such a many-to-one relationship requires multiplexing.
- The UDP first accepts messages from different processes. These messages are separated from each other by their port numbers. Each process has a unique port number assigned to it.
- Then the UDP adds header and passes the packet to IP as shown in Fig. 12.7.3.

Demultiplexing :

- At the receiving end, the relationship is one as to many. So we need a demultiplexer.
- First the UDP layer receives datagrams from the IP.
- The UDP then checks for errors and drops the header to obtain the messages and delivers them to appropriate process based on the port number.

12.7.9 Comparison of UDP and Generic Simple Protocol :

- In this section we will compare UDP with a simple connectionless transport layer protocol.
- The only difference between the two is that the UDP provides an **optional checksum.**

- If the checksum is added to the UDP packet then at the destination, the receiving UDP can check the packet for any error with the help of the checksum.
- If any error is detected, the receiving UDP will discard that packet, without sending any feedback to the sender.

12.8 UDP Applications :

- Despite being connectionless, unreliable, no flow control, no error control, UDP is still preferred for some applications.
- This is because UDP has some advantages too. An application designer has to sometimes compromise between advantages and drawbacks to get the optimum.
- Here we will discuss some important features of UDP that are useful in designing an application program.

12.9 UDP Features :

12.9.1 Connectionless Service :

- The feature of UDP is that it is a connectionless protocol and that each UDP packet is independent from the other packets, can be considered as an advantage or a disadvantage depending on the requirements of an application.
- In an application, if we want to send only short messages to server and receive short messages from the server. Then the above mentioned feature becomes an advantage.
- The feature of being connectionless is an advantage if request and respond each can fit in one single user datagram.
- The overhead (number packets to be exchanged) required to establish and close a connection is zero in case of UDP. This can be a very important advantage for some applications.
- Similarly the delay involved with the connectionless delivery is very short as compared to that with the connection oriented delivery. Hence the connectionless service provided by UDP is preferred for the applications in which delay is important.

12.9.2 Lack of Error Control :

- UDP is an unreliable protocol which does not provide any error control. Now this is actually a disadvantage but it becomes an advantage for some applications as explained below.
- If TCP is used for reliable service and if a packet is lost, then TCP will resend it. So the receiver transport layer is unable to deliver that part of the message to the application immediately. Due to this an uneven delay is introduced between different parts of the messages which is undesirable for some delay sensitive applications.
- This delay is actually a side effect of the reliable operation of TCP.
- Some applications are not affected by this delay but for some others it is very crucial.

12.9.3 Lack of Congestion Control :

- We know that there is no provision for congestion control in UDP. But this disadvantage can become an advantage for some applications.
- A good side effect of lack of congestion control is that UDP does not create any additional traffic that is created by TCP for congestion control.
- Hence the UDP is preferred fro some congestion prone networks.

12.9.4 Typical Applications of UDP :

1. UDP is suitable for the applications (processes) that have the following requirements :
   (a) A simple response to request is to be made.
   (b) Flow and error controls not essential.
   (c) Bulk data is not to be sent (like FTP).
2. UDP is used for RIP (Routing Information Protocol).
3. UDP is used for management processes such as SNMP.
4. UDP is suitable for the processes having inbuilt flow and error control mechanisms, such as TFTP.
5. UDP is suitable for the multicasting applications.
6. UDP is also used in the real time applications which do not tolerate the uneven delays.

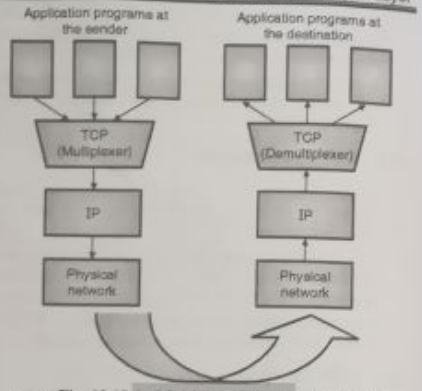
12.10 Transmission Control Protocol (TCP) :

- The TCP provides reliable transmission of data in an IP environment. TCP corresponds to the transport layer (Layer 4) of the OSI reference model.
- Among the services TCP provides are stream data transfer, reliability, efficient flow control, full-duplex operation, and multiplexing.
- TCP is the layer 4 protocol in the TCP/IP suite and it is a very important and complicated protocol. TCP has been revised multiple times in last few decades.
- With stream data transfer, TCP delivers an unstructured stream of bytes identified by sequence numbers.
- This service benefits applications because they do not have to chop data into blocks before handing it off to TCP. Instead, TCP groups bytes into segments and passes them to IP for delivery.
- TCP offers reliability by providing connection-oriented, end-to-end reliable packet delivery through an internetwork.

- It does this by sequencing bytes with a forwarding acknowledgment number that indicates to the destination the next byte the source expects to receive.
- Bytes not acknowledged within a specified time period are retransmitted.
- The reliability mechanism of TCP allows devices to deal with lost, delayed, duplicate, or misread packets. A time-out mechanism allows devices to detect lost packets and request retransmission.
- TCP offers efficient flow control, which means that, when sending acknowledgments back to the source, the receiving TCP process indicates the highest sequence number that it can receive without overflowing its internal buffers.
- TCP supports a full-duplex operation means that TCP processes can both send and receive at the same time.
- Finally, TCP's multiplexing means that numerous simultaneous upper-layer conversations can be multiplexed over a single connection.

12.10.1 Relationship Between TCP and IP :

- The relationship between TCP and IP is very interesting. Each TCP message gets encapsulated or inserted in an IP datagram and then this datagram is sent over the Internet to the destination.
- IP transports this datagram from sender to destination, without bothering about the contents of the TCP message.
- At the final destination the IP hands over the message to the TCP software running on the destination computer.
- IP acts like a postal service and transfers the datagrams from one computer to the other.
- Thus TCP deals with the actual data to be transferred and IP takes care of transfer of that data.
- Many applications such as FTP, Remote login TELNET etc. keep sending data to TCP software on the sending computer.
- The TCP software acts as a multiplexer at the sending computer. It receives data from various applications, multiplexes the data and hands it over to the IP software at the sending end as shown in Fig. 12.10.1.
- IP adds its own header to this TCP packet and creates an IP packet out of it. Then this packet is sent to its destination.
- At the destination exactly opposite process will take place. The IP software hands over the multiplexed data to the TCP software.
- The TCP software at the destination computer then demultiplexes the multiplexed data and gives it to the corresponding applications as shown in Fig. 12.10.1.

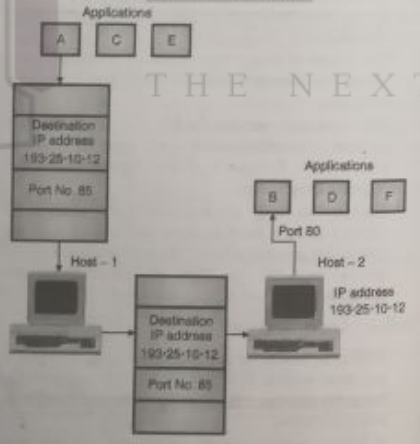


(G-1440) **Fig. 12.10.1 : Multiplexing and demultiplexing using TCP**

12.10.2 Ports and Sockets :

1. Ports :

- Applications running on different hosts communicate with TCP with the help of ports. Every application has been allotted a unique 16 bit number which is known as a **port.**



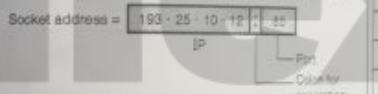
(G-1437) **Fig. 12.10.2 : Use of port numbers**

- When an application on one computer wants to communicate using a TCP connection to another application on some other computers these ports prove to be very helpful.

- Let an application A on host 1 wants to communicate with an application B on host 2. So the process takes place as shown in Fig. 12.10.2 and explained below.
- Application A running on computer 1 provides the IP address of computer 2 and the port number corresponding to application B as shown in Fig. 12.10.2.
- Computer 1 communicates with computer 2 using the IP address and computer 2 uses the port number to direct the message to application B.

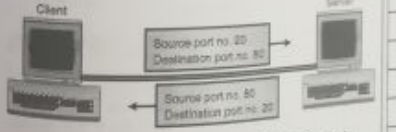
2. Sockets :

- A port is a 16 bit unique number used for identification of a single application.
- But socket address or simply socket would identify the combination of the IP address and the port number concatenated together as shown in Fig. 12.10.3.
- For example if the IP address = 193.25.10.12 and the port number is 85. Then this port of this computer will have the following socket address.



(G-1438) **Fig. 12.10.3**

- So a pair of sockets is required to identify a TCP connection between two applications on two different hosts. These two socket addresses specify the end points of the connection as shown in Fig. 12.10.4.



(G-1436) **Fig. 12.10.4 : Source and destination port numbers**

- Generally the server port numbers are known as the **well known ports.** Some of the well known port numbers have already been mentioned for UDP and TCP earlier in this chapter.
- Multiple TCP connections between different applications or same applications on two hosts exist in practice. Here the IP addresses of the two hosts are same but the port numbers are different.
- The communication using port numbers is illustrated in Fig. 12.10.4.

12.11 TCP Services :

Following are some of the services offered by TCP to the processes at the application layer :

1. Stream delivery service
2. Sending and receiving buffers
3. Bytes and segments
4. Full duplex service
5. Connection oriented service
6. Reliable service.
7. Process to process communication.

12.11.1 Process to Process Communication :

- The TCP uses port numbers a transport layer addresses. Table 12.11.1 shows some well known port numbers used by TCP.
- Note that if an application can use both UDP and TCP, the same port number is assigned to this application.

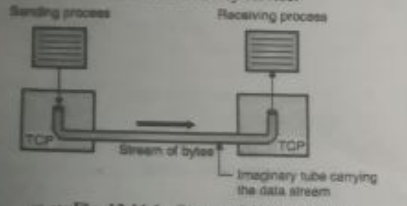
Table 12.11.1 : Well known ports used by TCP

Port	Protocol	Description
7	Echo	Sends received datagram back to sender
9	Discard	Discards any received packet
11	Users	Active users
13	Daytime	Sends the date and the time
17	Quote	Sends a quote of the day
19	Chargen	Sends a string character
20	FTP, Data	File Transfer protocol for data
21	FTP, Control	File Transfer protocol for control
23	TELNET	Terminal network
25	SMTP	Simple Mail Transfer Protocol
53	DNS	Domain Name server
67	BOOTP	Bootstrap Protocol
79	Finger	Finger
80	HTTP	Hypertext Transfer Protocol
111	RPC	Remote Procedure Call

12.11.2 Stream Delivery Service :

- TCP is a stream oriented protocol. The sending process delivers data in the form of a stream of bytes and the receiving process receives it in the same manner.
- TCP creates a working environment in such a way that the sending and receiving processes seem to be connected by an imaginary "tube" as shown in Fig. 12.11.1.

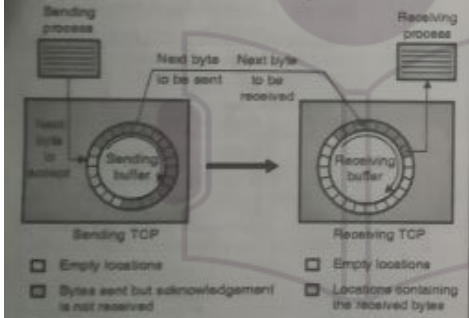
- This is called as stream delivery service.



(G-421)**Fig. 12.11.1 : Stream delivery service**

12.11.3 Sending and Receiving Buffers :

- The sending and receiving processes may not produce and receive data at the same speed.
- Hence TCP needs buffers for storage of data at both the ends. There are two types of buffers used in each direction :
  1. Sending buffer 2. Receiving buffer.
- A buffer can be implemented by using a circular array of 1 byte locations as shown in Fig. 12.11.2.

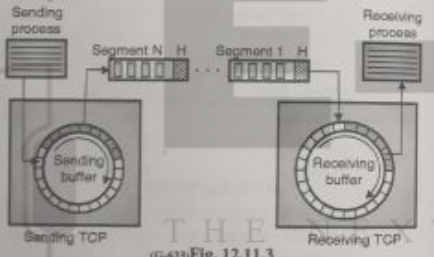


(G-422)**Fig. 12.11.2 : Sending and receiving buffers**

- Fig. 12.11.2 shows the direction of movement of data. The sending buffer has three types of locations :
  1. Empty locations.
  2. Locations containing the bytes which have been sent but not acknowledged. These bytes are kept in the buffer till an acknowledgement is received.
  3. The locations containing the bytes to be sent by the sending TCP.
- In practice, the TCP may be able to send only a part of data which is to be sent, due to slowness of the receiving process or congestion in the network.
- The buffer at the receiver is divided into two parts :
  1. The part containing empty locations.
  2. The part containing the received bytes which can be consumed by the sending process.

12.11.4 Bytes and Segments :

- Buffering is used to handle the difference between the speed of data transmission and data consumption.
- But only buffering is not enough. We need one more step before sending the data.
- The IP layer, which provides service to TCP, has to send data in the form of packets instead of stream of bytes.
- At the transport layer, TCP groups a number of bytes to form a packet called a segment.
- A header is added to each segment for the purpose of exercising control.
- The segments are then inserted in an IP datagram and transmitted. The entire operation is transparent to the receiving process.
- The segments may be received out of order, lost or corrupted when it reaches the receiving end.
- Fig. 12.11.3 shows the creation of segments from the bytes in the buffers.



(G-423)**Fig. 12.11.3**

- The segments are not of the same size. Each segment can carry hundreds of bytes.

12.11.5 Full Duplex Service :

- TCP offers full duplex service where the data can flow in both the directions simultaneously.
- Each TCP will then have a sending buffer and receiving buffer. The TCP segments can travel in both the directions, therefore TCP provides a full duplex service.

12.11.6 Connection Oriented Service :

- TCP is a connection oriented protocol. When process – 1 wants to communicate (send and receive) with another process (process - 2), the sequence of operations is as follows :
  1. TCP of process – 1 informs TCP of process – 2 and create a connection between them.
  2. TCP of process – 1 and TCP of process – 2 exchange data in both the directions.
  3. After completing the data exchange, when buffers on both sides are empty, the two TCPs destroy their buffers to terminate the connection.

- The type of connection in TCP is not physical, it is virtual. The TCP segment is encapsulated in an IP datagram and these packets can be transmitted without following the sequence :
- These segments can get lost or corrupted and may have to be resent.
- Each segment may take a different path to reach the destination.

12.11.7 Reliable Service :

TCP is a reliable transport protocol and not unreliable like UDP. Different acknowledgements are used by the receiver to convey sender the status of data.

12.12 Features of TCP :

In order to provide the services mentioned in the previous section, TCP has a number of features as follows :

12.12.1 Numbering System :

- The TCP software keeps track of the segments being transmitted or received. However in the segment header there is no field for a segment number value.
- But there are fields called **sequence number** and the **acknowledgement number.**
- Note that these fields correspond to the byte number and not the segment number.

Byte numbers :

- TCP give numbers to all the data bytes which are transmitted. The numbering is independent of the direction of data travel.
- The numbering does not always start from 0, but it can start with a randomly generated number between 0 and $2^{32} - 1$.

Sequence number :

- After numbering the bytes, the TCP assigns a sequence number to each segment that is being transmitted.
- The sequence number for each segment is same as the number assigned to the first byte present in that segment.

Acknowledgement number :

- The TCP communication is duplex. So both the communicating processes can send and receive data at the same time.
- Each process will give numbers to the bytes with a different starting byte number.
- Each party also uses an ackNo to confirm the reception of bytes.

The acknowledgement number is cumulative i.e. the receiver takes the number of the last byte received, adds 1 to it and uses this sum as the acknowledgement number.

12.12.2 Flow Control :

- TCP provides flow control (UDP does not). The receiver will control the amount of data to be sent by the sender.
- This will avoid data overflow at the receiver. The TCP uses byte oriented flow control.

12.12.3 Error Control :

- The error control mechanism is inbuilt for TCP. This allows TCP to provide a reliable service.
- The error control mechanism considers a segment as the unit of data for error correction however the byte oriented error control is provided.

12.12.4 Congestion Control :

- TCP takes the congestion in network into account. UDP does not do this.
- The amount of data sent by the sender depends on the following factors :
  1. The receiver's decision (flow control).
  2. The network congestion.

Summary of TCP features :

1. TCP is a process-to-process protocol.
2. TCP uses port numbers.
3. It is a connection oriented protocol (creates a virtual connection).
4. It uses flow and error control mechanisms.
5. TCP is a reliable protocol.

12.13 The TCP Protocol :

- Let us take a general overview of the TCP protocol.
- Every byte on a TCP connection has its own 32-bit sequence number. These numbers are used for both acknowledgement and for window mechanism.

Segments :

The sending and receiving TCP entities exchange data in the form of segments. A segment consists of a fixed 20 byte header (plus and optional part) followed by zero or more data bytes.

Segment size :

The segment size is decided by the TCP software. Two limits restrict the segment size as follows :

1. Each segment including the TCP header, must fit in the 65535 byte IP payload.
2. Each segment must fit in the **MTU (Maximum Transfer Unit).** Each network has a maximum transfer unit. Practically an MTU which is a few thousand bytes defines the upper limit on the segment size.

Fragmentation :

- If a segment is too large, then it should be broken into small segments. Using fragmentation by a router.

- Each new segment gets a new IP header. So the fragmentation by router will increase the overhead.

Timer :

- The basic protocol used by TCP entities is the sliding window protocol. A sender starts a timer as soon as a sender transmits a segment.
- When the segment is received by the destination, it sends back acknowledgement alongwith data if any. The acknowledgement number is equal to the next sequence number it expects to receive.
- If the timer at the sender goes out before the acknowledgement reaches back, it will **retransmit** that segment again.

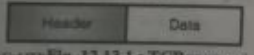
Possible problems :

- As the segments can be fragmented, a part of the transmitted segment only may reach the destination with the remaining part lost.
- Segments can arrive out of order.
- Segments can get delayed so much that timer is out and unnecessary retransmission will take place.
- If a retransmitted segment takes a different route than the original segment is fragmented then the fragments of original and retransmitted segments can reach the destination in a sporadic way. So a careful administration is required to achieve reliable byte stream.
- There is a possibility of congestion or broken network along the path.
- TCP should be able to solve these problems in an efficient manner.

12.13.1 TCP Segment :

The TCP segment as shown in Fig. 12.13.1 consists of two parts :

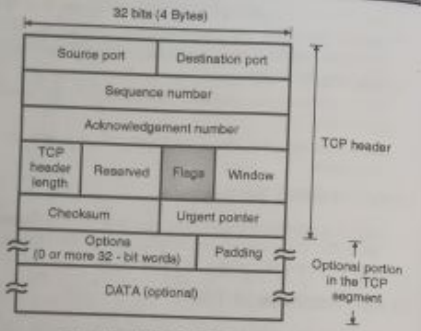
1. Header 2. Data



(G-1423) **Fig. 12.13.1 : TCP segment**

12.13.2 The TCP Segment Header :

- Fig. 12.13.2 shows the layout of a TCP segment. Every segment begins with a 20 byte fixed format header.
- The fixed header may be followed by header options.
- After the options, if any, upto 65535 – 20 – 20 = 65495 data bytes may follow. Note that the first 20 bytes correspond to the IP header and the next 20 correspond to the TCP header.
- The TCP segment without data are used for sending the acknowledgements and control messages.



(G-611) **Fig. 12.13.2 : TCP header format**

Source port : A 16-bit number identifying the application the TCP segment originated from within the sending host. The port numbers are divided into three ranges, well-known ports (0 through 1023), registered ports (1024 through 49,151) and private ports (49,152 through 65,535). Port assignments are used by TCP as an interface to the application layer.

Destination port : A 16-bit number identifying the application the TCP segment is destined for on a receiving host. Destination ports use the same port number assignments as those set aside for source ports.

Sequence number : A 32-bit number identifying the current position of the first data byte in the segment within the entire byte stream for the TCP connection. After reaching $2^{32} - 1$, this number will wrap around to 0.

Acknowledgement number :

A 32-bit number identifying the next data byte the sender expects from the receiver. Therefore, the number will be one greater than the most recently received data byte. This field is only used when the ACK control bit is turned on.

Header length or offset :

A 4-bit field that specifies the total TCP header length in 32-bit words (or in ultiples of 4 bytes if you prefer). Without options, a TCP header is always 20 bytes in length. The largest a TCP header may be is 60 bytes. This field is required because the size of the options field(s) cannot be determined in advance. Note that this field is called "data offset" in the official TCP standard, but header length is more commonly used.

Reserved :

A 6-bit field currently unused and reserved for future use.

Control bits or flags :

1. **Urgent pointer (URG) :** If this bit field is set, the receiving TCP should interpret the urgent pointer field.
2. **Acknowledgement (ACK) :** If this bit field is set, the acknowledgement field described earlier is valid.

3. **Push function (PSH) :** If this bit field is set, the receiver should deliver this segment to the receiving application as soon as possible. An example of its use may be to send a Control-BREAK request to an application, which can jump ahead of queued data.
4. **Reset the connection (RST) :** If this bit is present, it signals the receiver that the sender is aborting the connection and all queued data and allocated buffers for the connection can be freely relinquished.
5. **Synchronize (SYN) :** When present, this bit field signifies that sender is attempting to "synchronize" sequence numbers. This bit is used during the initial stages of connection establishment between a sender and receiver.
6. **No more data from sender (FIN) :** If set, this bit field tells the receiver that the sender has reached the end of its byte stream for the current TCP connection.

Window :

A 16-bit integer used by TCP for flow control in the form of a data transmission window size. This number tells the sender how much data the receiver is willing to accept. The maximum value for this field would limit the window size to 65,535 bytes, however a "window scale" option can be used to make use of even larger windows.

Checksum : A TCP sender computes a value based on the contents of the TCP header and data fields. This 16-bit value will be compared with the value the receiver generates using the same computation. If the values match, the receiver can be very confident that the segment arrived intact.

Urgent pointer :

In certain circumstances, it may be necessary for a TCP sender to notify the receiver of urgent data that should be processed by the receiving application as soon as possible. This 16-bit field tells the receiver when the last byte of urgent data in the segment ends.

Options :

In order to provide additional functionality, several optional parameters may be used between a TCP sender and receiver. Depending on the option(s) used, the length of this field will vary in size, but it cannot be larger than 40 bytes due to the size of the header length field (4 bits). The most common option is the Maximum Segment Size (MSS) option. A TCP receiver tells the TCP sender the maximum segment size it is willing to accept through the use of this option. Other options are often used for various flow control and congestion control techniques.

Padding :

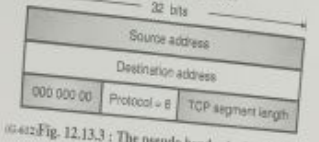
Because options may vary in size, it may be necessary to "pad" the TCP header with zeros so that the segment ends on a 32-bit word boundary as defined by the standard.

Data :

Although not used in some circumstances (e.g. acknowledgement segments with no data in the reverse direction), this variable length field carries the application data from TCP sender to receiver. This field coupled with the TCP header fields constitutes a TCP segment.

12.13.3 Checksum :

- A checksum is provided to ensure extreme reliability. It checksums the header, the data and the conceptual pseudo header shown in Fig. 12.13.3.



(G-612) **Fig. 12.13.3 : The pseudo header included in the TCP checksum**

- When the checksum is being computed, the TCP checksum field is set to zero, and the data field is padded out with an additional zero byte if its length is an odd number.
- Then all the 16 bit words are added in 1's complement and then 1's complement of the sum is taken to get the checksum.
- When a receiver performs the calculation on the entire segment including the checksum field, the result has to be zero.
- The pseudo header contains the 32 bit IP address of the source and destination machines, the protocol number for TCP i.e. 6 and the TCP segment length as shown in Fig. 12.13.3.

12.13.4 Encapsulation :

- The data coming from the application layer is encapsulated in a TCP segment. This TCP segment is then encapsulated in an IP datagram.
- The IP datagram is encapsulated in a frame at the data link layer. The process of encapsulation is shown in Fig. 12.13.4.

(G-3972) **Fig. 12.13.4 : Encapsulation**

Review Questions

Q. 1 Write down duties of transport layer.

Q. 2 What are the services provided by transport layer ?

Q. 3 Write short note on port numbers.

Q. 4 Explain the concept of socket address.

Q. 5 State the difference between IP addresses and port numbers.

Q. 6 Write short note no encapsulation and decapsulation.

Q. 7 Define multiplexing and demultiplexing.

Q. 8 Explain flow control at transport layer.