9

Telephone Networks

Public switched telephone network (PSTN) or the plain old telephone system (POTS) is perhaps the most stupendous telecommunication network in existence today. There are over 400 million telephone connections and over 60,000 telephone exchanges the worldover. The length of telephone wire-pairs buried underground exceeds a billion kilometers. A unique feature of the telephone network is that every piece of equipment, technique or procedure which has evolved in the last 100 years from a number of different giant corporations, is capable of working with each other. Compare this with the fact that it is almost impossible to interface the first IBM computer with its own latest system. The enormous complexity of the telephone network is managed by using a hierarchical structure, worldwide standardisation, and decentralisation of administration, operation and maintenance.

Any telecommunication network may be viewed as consisting of the following major systems:

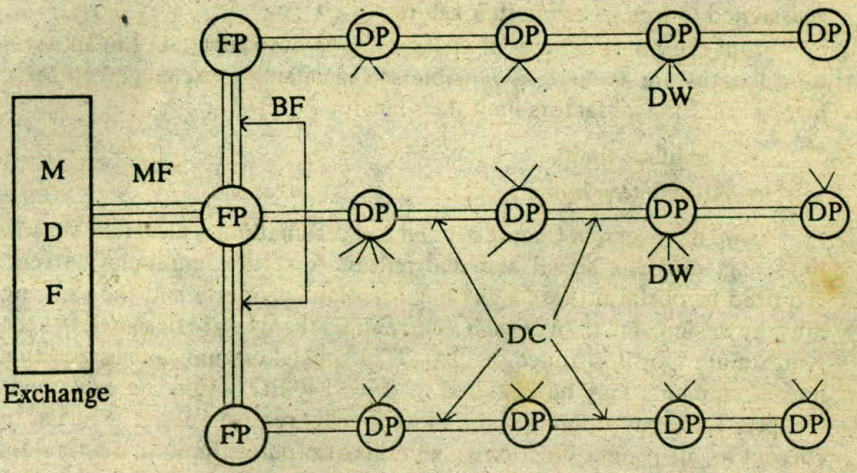
1. Subscriber end instruments or equipments
2. Subscriber loop systems
3. Switching systems
4. Transmission systems
5. Signalling systems.

Telephone instruments and switching systems have been dealt with in Chapters 2–6. Fibre optic transmission system has been discussed in Chapter 7. In this chapter, we discuss other transmission systems, subscriber loop systems, signalling systems and network management related aspects.

9.1 Subscriber Loop Systems

Every subscriber in a telephone network is connected generally to the nearest switching office by means of a dedicated pair of wires. Subscriber loop refers to this pair of wires. It is unwieldy to run physically independent pairs from every subscriber premises to the exchange. It is far easier to lay cables containing a number of pairs of wires for different geographical locations and run individual pairs as required by the subscriber premises.

Generally, four levels of cabling are used as shown in Fig. 9.1. At the subscriber end, the drop wires are taken to a distribution point. The drop



MDF = main distribution frame MF = main feeder
FP = feeder point BF = branch feeder DP = distribution point
DC = distribution cable DW = drop wires

Fig. 9.1 Cable hierarchy for subscriber loops.

wires are the individual pairs that run into the subscriber premises. At the distribution point, the drop wires are connected to wire pairs in the distribution cables. Many distribution cables from nearby geographical locations are terminated on a feeder point where they are connected to branch feeder cables which, in turn, are connected to the main feeder cable. The main feeder cables carry a larger number of wire pairs, typically 100–2000, than the distribution cables which carry typically 10–500 pairs. The feeder cables are terminated on a main distribution frame (MDF) at the exchange. The subscriber cable pairs emanating from the exchange are also terminated on the MDF.

Subscriber pairs and exchange pairs are interconnected at the MDF by means of jumpers. The MDF thus provides a flexible interconnection mechanism which is very useful in reallocating cable pairs and subscriber numbers. For example, if a subscriber moves his house to a nearby area served by the same exchange but a different distribution point, he can be permitted to retain the same telephone number by a suitable jumper connection at the MDF. Similarly, the wire pair released by him can be given to a new number and assigned to another subscriber. In newer installations, distribution and feeder points also have flexible crosspoint connection capability. This enables a subscriber drop wire to be easily reconnected to any pair in the distribution cables, and similarly any pair from the distribution cable can be connected to any other pair in the feeder cable at

the feeder point. This arrangement permits efficient utilisation of the cable pairs as well as helps in cable management during faults. For example, if a particular cable is faulty, important subscribers assigned to this cable may be reassigned to free pairs in other cables.

From the point of view of economy, it is desirable that the subscriber loop lengths are as large as possible so that a single exchange can serve a large area. But two factors limit their length:

- Signalling limits
- Attenuation limits.

As we have seen in Chapters 1 and 2, d.c. signalling is used for subscriber lines, e.g. off-hook signal and dial pulses. A certain minimum current is required to perform these signalling functions properly and the exchanges must be designed to deliver such a current. Exchanges are designed to accept a maximum loop resistance of 1300 Ω. In special circumstances, additional line equipments may be installed to drive 2400 Ω loops. As explained in Chapter 1, the microphone in the subscriber set requires about 25 mA as bias current for its proper functioning and this also puts a limit on the total loop resistance. A bound on the loop resistance, in turn, limits the loop length for a given gauge of wire. The d.c. loop resistance R_{dc} for copper conductors can be calculated from the following formula:

$$R_{dc} = \frac{21.96}{d^2} \text{ ohms/km} \tag{9.1}$$

where d is the diameter of the conductor in mm. Since the conductor resistance is a function of temperature, the equation holds good for resistance values at 20°C. Subscriber instruments are usually connected to the exchanges using copper conductors of sizes AWG 19 to AWG 26 American wire gauge. Smaller gauge wires use thicker conductors and offer less d.c. resistance per unit length. They are used to connect subscribers located at far away distances and are more expensive for obvious reasons. Tabic 9.1 gives the technical specifications for the commonly used wire gauges.

Table 9.1 Technical Specifications for Subscriber Lines

Gauge No. (AWG)	Diameter (mm)	D.C. resistance (Ω/km)	Attenuation (dB/km)
19	0.91	26.40	0.71
22	0.64	52.95	1.01
24	0.51	84.22	1.27
26	0.41	133.89	1.61

EXAMPLE 9.1 An exchange uses a −40 V battery to drive subscriber lines. A resistance of 250 Ω is placed in series with the battery to protect it

from short circuits. The subscribers are required to use a standard telephone set which offers a d.c. resistance of 50 Ω. The microphone requires 23 mA for proper functioning. Determine the farthest distance from the exchange at which a subscriber can be located if 26 AWG conductor is used.

Solution Let R_L be the line loop resistance. Then, we have

$$23 \times 10^{-3} = 40/(250 + 50 + R_L)$$

Hence, $R_L = 1439$.

From Table 9.1, for 26 AWG wire, we have

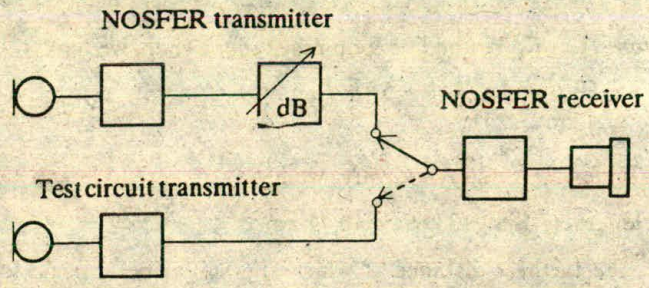
Loop length = 1439/133.89 = 10.74 km

Therefore, the farthest distance at which the subscriber can be located is 10.74/2 = 5.37 km.

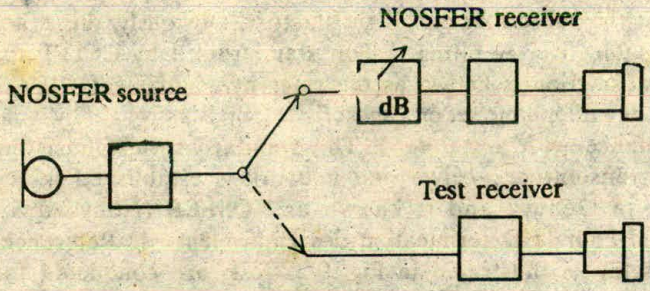
Attenuation limits arise from the a.c. response of the loop and refers to loop loss in decibels. The criterion here is to ensure that the quality of reception at the subscriber end is satisfactory. Obviously, this criterion is somewhat subjective. A rating system standardised by CCITT to grade customer satisfaction is known as the **reference equivalent (RE)**. In this system, RE of a telephone set or a subscriber loop is arrived at by comparing its performance with a standard. The standard setup consisting of a telephone transmitter, receiver and network is established in the ITU laboratory in Geneva and is known as **NOSFER** (Nouveau Systame Fondamental Pour la Determination des Equivalents de Reference). The NOSFER setup is illustrated in Fig. 9.2. Tests are conducted for both transmit and receive qualities and the corresponding REs are known as transmit reference equivalent (TRE) and receive reference equivalent (RRE) respectively. Since the transmitter and receiver technologies have evolved differently, the TRE and RRE are usually unequal. The REs for cables are measured at a standard frequency of 800 Hz as specified by CCITT. The American standard uses 1000 Hz. Trained listeners and talkers are engaged to judge the quality of reception and to transmit speech signal using a standard test language. Calibrated attenuations are inserted into the system until the speech quality of the test circuit is judged to be equal to that of the NOSFER. The value of the balancing attenuation is the RE. If 5 dB loss is added to the NOSFER system for balancing, the test connection is said to have an RE of 5 dB and if 5 dB is taken out of NOSFER, the test connection RE is –5 dB. Negative value for RE indicates that the test circuit performs better than the laboratory standard. The standard language is made up of logatoms. A logatom is a 1-syllable word comprising a constant, a vowel and another consonant in that sequence.

EXAMPLE 9.2 Determine the reference equivalent of a local call circuit, given the following REs:

For the telephone set, TRE = 3 dB, RRE = –3 dB
For the subscriber line loop, TRE = 10 dB, RRE = 8 dB
For the exchange switch, RE = 1 dB.



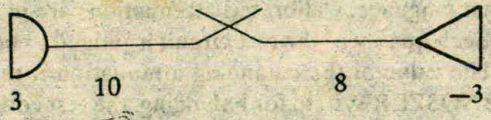
(a) Transmit reference equivalent



(b) Receive reference equivalent

Fig. 9.2 NOSFER reference equivalent systems.

Solution Overall reference equivalent (ORE) is illustrated as in the following figure:



Therefore, ORE = 3 + 10 + 1 + 8 + (– 3) = 19 dB.

Sometimes it is convenient to specify symmetric values for reference equivalents. This is done by taking the average of TRE and RRE as

$$RE = \frac{TRE + RRE}{2} \tag{9.2}$$

A need often arises to connect to an existing exchange, subscribers who are located beyond the maximum prescribed distance. For example, it is

uneconomical to install a new exchange for a lone or a few remote subscribers. In such cases special techniques are used to meet the resistance and attenuation constraints. The d.c. resistance constraint is met by

1. use of higher diameter (lower gauge) wire,
2. use of equalised telephone sets, and
3. unigauge design or use of higher supply voltage.

The attenuation constraint is usually overcome by the use of loading coils. A significant portion (30–40%) of the cost of a telecommunication network comes from the cost of copper in the subscriber lines. Hence the first technique of using higher diameter copper wire is rarely resorted to. Equalised telephone sets require about 8–12 mA of d.c. bias current as against 22–30 mA required by a normal telephone. As a result, a larger value of loop resistance is acceptable.

A typical unigauge design is shown in Fig. 9.3. This design attempts to use a wire with as small a diameter as possible while retaining the resistance

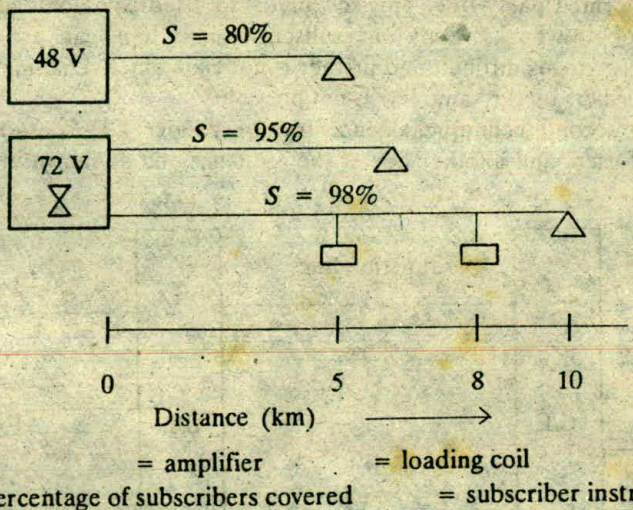


Fig. 9.3 Unigauge design of subscriber loops.

and attenuation limits. For both long and short distances, the same gauge wire is used and hence the name unigauge design. For up to 5 km, a 48 V battery is used with 26 gauge wire and for longer distances a combination of a higher battery voltage (72 or 96 V) and an amplifier that gives a midband gain of 5 dB is used. Beyond 8 km, loading coils are required to obtain better frequency vs. attenuation characteristics. The role of loading coils in improving frequency response has been explained in Section 5.7. Loading coils are identified by using a standard convention as 19–D–44, 24–A–88 etc.

The first number indicates the wire gauge with which the coil is used, the letter specifies the spacing between the coils and the number at the end gives the inductance value in mH. Table 9.2 gives the standard letters and their associated spacing. The most commonly used spacings are B, D and H.

Table 9.2 Loading Coil Spacing

Letter code	A	B	C	D	E	F	H	X	Y
Spacing (km)	0.21	0.92	0.28	1.37	1.7	0.85	1.83	0.2	0.65

In rural areas, subscribers are generally dispersed. It is both unnecessary and expensive to provide a dedicated pair for every subscriber. Three techniques are used to gain on the number pairs:

1. Party lines
2. Concentrators
3. Carrier systems.

In the first technique, two or more subscribers are connected to one line which is termed **party line**. This scheme is not used commonly as it has a number of drawbacks. Only one subscriber at a time can use the line. Selective ringing is difficult and privacy is not maintained. Dialling between two subscribers on the same line is not possible.

In the second technique, a concentrator expander (CE) is used near the cluster of users and another one at the exchange end as shown in Fig. 9.4.

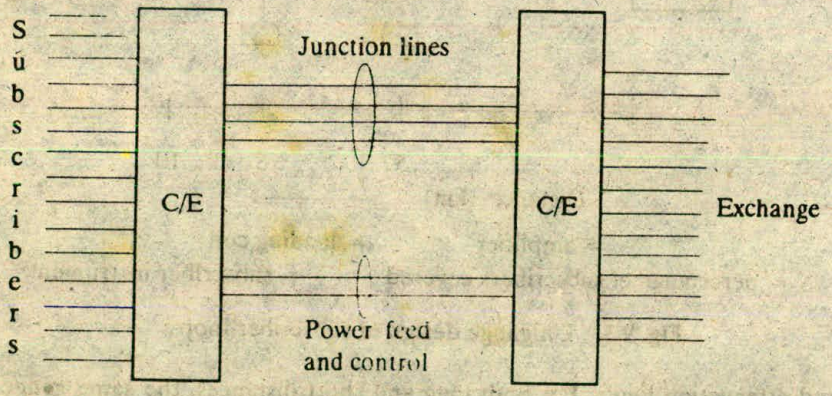


Fig. 9.4 Concentrator-expander connection for dispersed subscribers.

Only a few junction lines are run between the CEs which have switching capability. Typically, a ratio of 1:10 between the junction lines and the subscriber lines is used. The CE at the exchange end remotely powers and controls the CE at the subscriber end.

Carrier systems employ multiplexing techniques and enable all the users to access the exchange over a single line. Analog FDM or digital TDM systems are used.

Signalling and voice transmission on the subscriber lines requires that the exchange performs a set of functions. These functions are performed by an interface at the exchange end known as **subscriber loop interface**. Some functions are required in analog networks, some in digital networks and others in both. The complete set of functions are known by an acronym **BORSCHT** which stands for:

B = battery feed
O = overvoltage protection
R = ringing
S = supervision
C = coding
H = hybrid
T = test

Functions B and R are well known. Overvoltage protection deals with equipment and personnel protection from lightning strikes and power line surges. Detection of off-hook condition is a supervisory function. Functions C and H are exclusive to digital switch interfaces. As we know, digital switching demands that analog-to-digital (A–D) and digital-to-analog (D–A) conversions and some form of coding/decoding be done. This aspect has been discussed in Chapter 5. Subscribers are connected to the exchange via 2-wire circuits. These circuits use a balanced connection as shown in Fig. 9.5. Balanced connections overcome many drawbacks of unbalanced

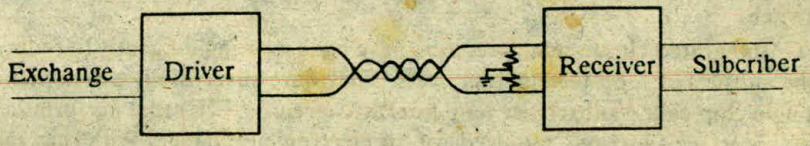


Fig. 9.5 Balanced circuit connection.

circuits. The transmission lines have equal impedances to ground and hence do not act as an antenna to pick up signals. Since the ground is not part of the signal path, hum is eliminated. Differential inputs improve noise immunity as any interference affects both lines equally and does not introduce differential currents.

Digital exchanges require receive and transmit signals on separate two-wire circuits. This calls for two-wire to four-wire conversion. Such a conversion is normally required for trunk transmissions in analog exchanges. The circuit that performs the 2-wire to 4-wire conversion is called **hybrid**. A transformer based hybrid circuit is shown in Fig. 9.6. The main function of a hybrid is to ensure that there is no coupling of signal from the input to the

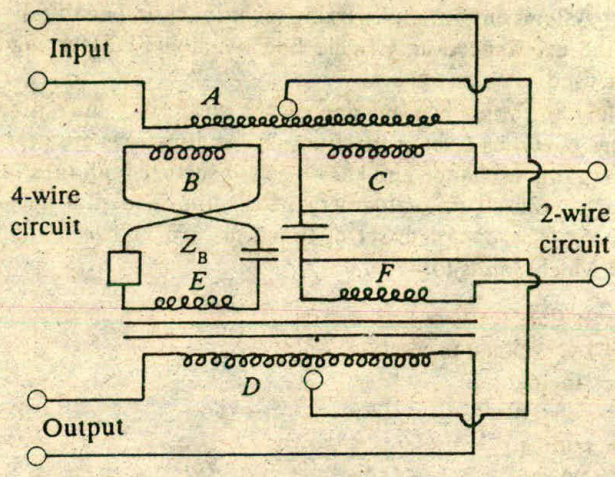


Fig. 9.6 Two-wire to 4-wire transformer hybrid.

output in the 4-wire circuit. The operation of the circuit is as follows: The input signal is coupled to the *B* and *F* windings equally. Through the *C* winding, the input is coupled to the 2-wire circuit. The same signal when it flows through the balanced 2-wire couples the signal to winding *D* through winding *C*. The signal induced in *B* flows through *E* and induces a current in *D* that opposes the current induced by *F*. If the impedance Z_B exactly matches that of the 2-wire circuit, the effect of input signal on the output winding *D* is completely nullified. In a similar way, the input signal from the subscriber end is completely nullified from coupling into the winding *A*.

Test function enables forward test of the loop or inward test of the switch.

Integrated circuit manufacturers have been working diligently to reduce the subscriber loop termination costs. They have successfully developed a single chip called **subscriber loop interface circuit (SLIC)** that performs all the interface functions required for an electronic exchange. With this, the current trend is to use one SLIC per subscriber unlike the earlier systems where one A/D and D/A converter was being shared by a number of the lines.

Another interesting development on account of VLSI technology is the all-electronic telephone. Single chip telephones that can be powered completely from the telephone loop and having microprocessor interface is a reality now. The use of a microprocessor in the telephone instrument enables features such as automatic redialling.

9.2 Switching Hierarchy and Routing

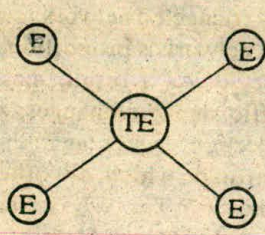
Telephone networks require some form of interconnection of switching exchanges to route traffic effectively and economically. Exchanges are

interconnected by groups of trunk lines, usually known as **trunk groups** that carry traffic in one direction. Two trunk groups are required between any two exchanges. Three basic topologies are adopted for interconnecting exchanges: mesh, star, and hierarchy. Mesh is a fully connected network (see Section 1.1). The number of trunk groups in a mesh network is proportional to the square of the exchanges being interconnected. As a result, mesh connections are used only when there is heavy traffic among exchanges, as may happen in a metropolitan area. A star connection utilises an intermediate exchange called a **tandem exchange** through which all other exchanges communicate. A star configuration is shown in Fig. 9.7(a). Star networks are used when the traffic levels are comparatively low. Many star networks may be interconnected by using an additional tandem exchange, leading to a two-level star network as shown in Fig. 9.7(b). An orderly construction of multilevel star networks leads to hierarchical networks.

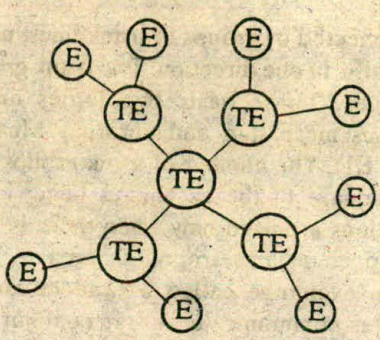
Hierarchical networks are capable of handling heavy traffic where required, and at the same time use minimal number of trunk groups. A 5-level switching hierarchy is recommended by CCITT as shown in Fig. 9.7(c). In a strictly hierarchical network, traffic from subscriber A to subscriber B and vice versa flows through the highest level of hierarchy, viz. quaternary centres in Fig. 9.7(c). A traffic route via the highest level of hierarchy is known as the **final route**. However, if there is a high traffic intensity between any pair of exchanges, direct trunk groups may be established between them as shown by dashed lines in Fig. 9.7(c). These direct routes are known as **high usage routes**. Wherever high usage routes exist, the traffic is primarily routed through them. Overflow traffic, if any, is routed along the hierarchical path. No overflow is permitted from the final route. In Fig. 9.7(c), the first choice routing for traffic between subscribers A and B is via the high usage route across the primary centres. The second and the third choice routes and the final route are also indicated in Fig. 9.7(c). A hierarchical system of routing leads to simplified switch design. Three methods are commonly used for deciding on the route for a particular connection:

1. Right-through routing
2. Own-exchange routing
3. Computer-controlled routing.

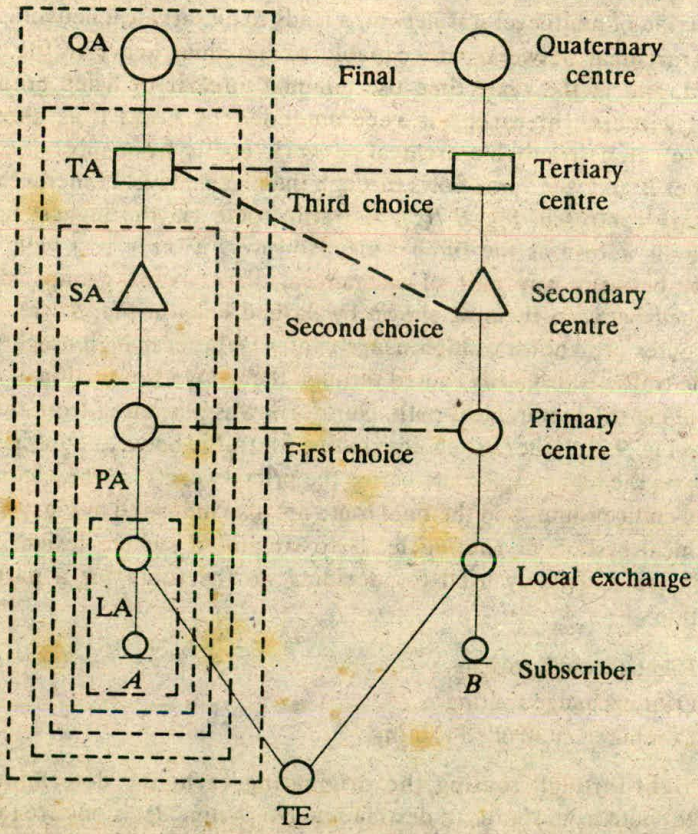
In right-through routing the originating exchange determines the complete route from source to destination. No routing decisions are taken at the intermediate routes. In the absence of a computer, only a predetermined route can be chosen by the originating exchange. However, there may be more than one predetermined route and the originating node may select one out of these, based on certain criteria like time of the day, even distribution of traffic etc.



(a) Star topology



(b) Two-level star



(c) CCITT hierarchical structure

LA, PA, SA, TA, QA = local, primary, secondary, tertiary and quarternary areas respectively E = exchange TE = tandem exchange

Fig. 9.7 Telecommunication network topologies.

Own-exchange routing or distributed routing allows alternative routes to be chosen at the intermediate nodes. Thus the strategy is capable of responding to changes in traffic loads and network configurations. Another advantage of distributed routing is that when new exchanges are added, modifications required in the switch are minimal.

Computers are used in networks with common channel signalling (CCS) features. In CCS, there is a separate computer-controlled signalling network. With computers in position, a number of sophisticated route selection methods can be implemented. Computer based routing is a standard feature in data networks. A detailed discussion of computer-controlled routing technique is presented in Section 10.4 while dealing with data networks.

A strictly hierarchical network suffers from one serious drawback, i.e. its poor fault tolerance feature. A good network design should maintain communication, though may be with reduced capability and increased blockage, even in the event of a failure of one or several links due to causes such as fire, explosion, sabotage and natural disaster. Total breakdown of the network should never occur unless under calamity. In a hierarchical network, as we go higher in the hierarchy, the nodes of each rank become fewer and fewer. A failure of a node or communication links at higher levels might seriously jeopardise communications. Alternative routing paths and redundant nodes have to be provided for in higher levels. The current tendency is to reduce the number of levels in the hierarchy, and fully interconnect the high level nodes to provide a large number of alternative routes. It is expected that the future national networks may be built with only three levels of hierarchy.

9.3 Transmission Plan

For reasons of transmission quality and efficiency of operation of signalling, it is desirable to limit the number of circuits connected in tandem. In a tandem chain, the apportionment of links between national and international circuits is necessary to ensure 'quality' telecommunications. CCITT lays down certain guidelines in this regard in its recommendations Q.40:

1. The maximum number of circuits to be used in an international call is 12.
2. No more than four international circuits be used in tandem between the originating and the terminating international switching centres.
3. In exceptional cases and for a low number of calls, the total number of circuits may be 14, but even in this case, the international circuits are limited to a maximum of four.

Taking the guidelines 1 and 2 above, we have eight links available for national circuits, which implies a limit of four for each national circuit. National network designs should take into account this limit.

The transmission loss is defined in terms of reference equivalents TRE, RRE and ORE as discussed in Section 9.1. CCITT recommends that for 97% of the connections the maximum TRE be limited to 20.8 dB and RRE to 12.2 dB between the subscriber and the international interface in the national network. This gives an overall reference equivalent ORE of 33 dB. Telephone administrations and companies attempt to design networks in such a way as to reduce as much as possible the ORE to improve subscriber satisfaction. From country to country OREs range from 6 dB to 26 dB.

Transmission loss budget should provide for two factors other than the line and switch losses:

- Keeping echo levels within limits
- Control singing.

In Section 9.1, we saw that a hybrid is required to convert a 2-wire circuit into a 4-wire circuit between the subscriber and a digital exchange. In analog exchanges, local calls are established on 2-wire circuits. But, long distance calls require 2-wire to 4-wire conversion at the subscriber line-trunk interfaces. Interexchange or intercity trunk lines carry a number of conversations on a single bearer circuit which may be derived from a coaxial cable, microwave or satellite system. Due to the long distances involved, the bearer circuits need amplifiers or repeaters at appropriate intervals to boost the signals. The amplifiers are almost invariably one-way devices and cannot handle bidirectional signals. Since the telephone conversation calls for signal transmission both ways, long distance trunks require separate circuits for each direction, leading to 4-wire circuits. Hence, the need for 2-wire to 4-wire conversion in long distance connections. The conversion is done by the hybrids.

As discussed in Section 9.1, an important function of the hybrid is to ensure that the received signal is not coupled. The coupling is zero only when the 2-wire circuit and the 4-wire circuit impedances are perfectly matched. While it is relatively easy to control the impedances of the trunk circuits, the subscriber loop impedances vary from subscriber to subscriber, depending on the distance at which a subscriber is located from the exchange. As a result, an impedance mismatch occurs for most of the connections at the subscriber line-trunk interface. The effect of such a mismatch is to reflect a part of the incoming speech signal on to the outgoing circuit, which returns to the speaker as echo. The echo may be loud enough to annoy the speaker as it is amplified like other signals in the return path. However, it may not be as strong as the speech signal from the other party, since it experiences attenuation on two lengths of the transmission channel before reaching the originator as shown in Fig. 9.8. If the distances are short, the round trip delay experienced by the echo is small such that the echo superimposes on the speaker's own voice and becomes unnoticeable. In fact, the sidetone, discussed in Section 1.2, may be considered as echo with zero time delay. As

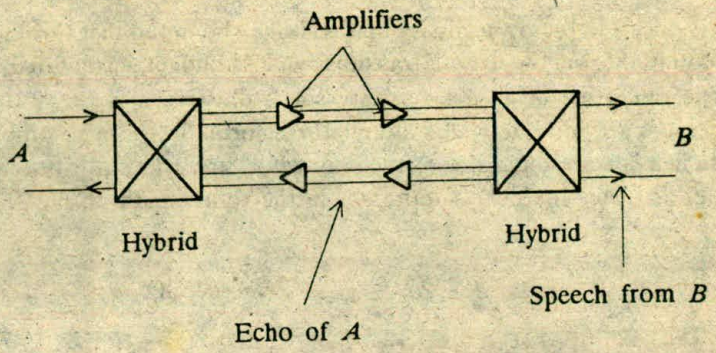


Fig. 9.8 Echo as reflected signal.

the time delay increases, the echo becomes noticeable and annoying to the speaker.

Short delay echos are controlled by using attenuators and the long delay ones by **echo suppressors** or **echo cancellers**. CCITT recommends the use of echo suppressors or cancellors if the round trip delay exceeds 50 ms. Use of echo suppressors is mandatory in satellite circuits as the round trip delay involved is several hundred milliseconds. For delays up to 50 ms, simple attenuators in the transmission path limit the loudness of echo to a tolerable level. The attenuation required increases as the delay increases as shown in Fig. 9.9. For instance, if the echo path delay is 20 ms, 11-dB attenuators must

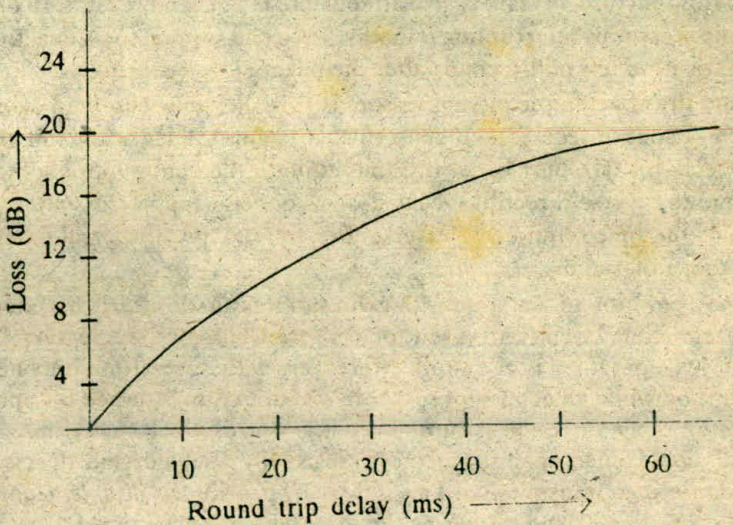


Fig. 9.9 Attenuation vs. echo delay.

be introduced in the transmission paths. It may be noted that this loss must be accounted for in the overall transmission loss budget, i.e. in ORE.

The operation of an echo suppressor is illustrated in Fig. 9.10. Echo suppressors are voice activated attenuators. Normally, the echo suppressors remain in a deactivated state, i.e. the attenuators are bypassed. Speech in one channel activates the echo suppressor in the return path. In Fig. 9.10, *A*'s

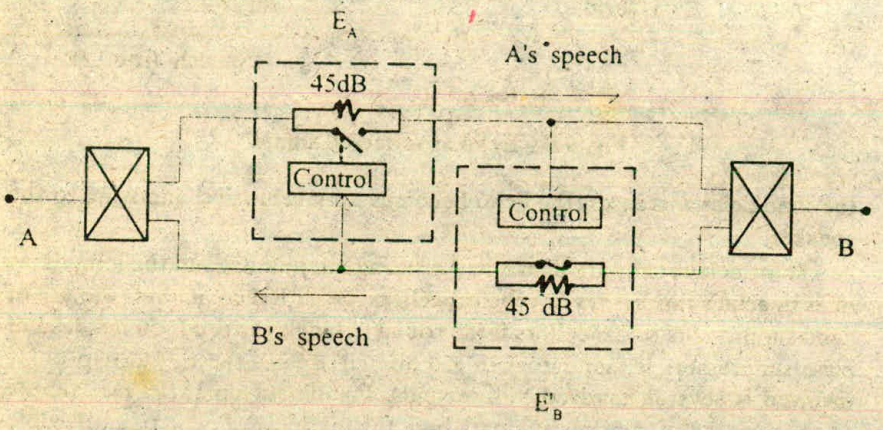


Fig. 9.10 Echo suppressor operation

speech activates the echo suppressor E_B and *B*'s speech E_A. Figure 9.10 depicts the situation when *B* is talking and *A* is silent. Should *A* attempt to talk at this juncture, his talk is also attenuated. He can, however, turn off the echo suppressor by interrupting *B* loudly. The echo suppressor is deactivated automatically a few milliseconds after the talker stops speaking.

One drawback of echo suppressors is that they may clip the beginning portion of speech segments. If subscriber *A* begins talking at the tail end of *B*'s speech, his talkspurt is not transmitted until the echo suppressors have had time to reverse directions. New designs of echo suppressors attempt to minimise the time required to reverse directions. Typical reversal times are in the range of 2–5 ms.

The operation of a system with echo suppressors is clearly half duplex. When telephone circuits are used for data transmission (see Section 10.1), full duplex operation is required. Moreover, with several milliseconds of interruption while an echo suppressor in one direction is turned off and the one in the other direction is turned on, it is very difficult to organise data transmission. Hence, echo suppressors are usually disabled while the circuits are used for data transmission. This is done by providing a *disabler feature* in the echo suppressor and triggering the same with a special signal. Usually, a 2025-Hz or 2100-Hz tone, transmitted for at least a duration of 300 ms with a signal level not less than –5 dBm, is used to trigger the disabler. Once

disabled, an echo suppressor remains so, as long as signals in the frequency range of 300 to 3000 Hz are being transmitted in either direction. A no-signal interval of 100 ms or more switches the echo suppressor back into the circuit.

Recent developments in electronics technology have paved the way for a new form of echo control by echo cancellation. Echo cancellors do not physically insert attenuators or bypass them, instead they process the incoming signal to eliminate the reflected component from it. Transmitted speech is stored for a period of time equal to the round trip delay of the circuit. The stored signal is attenuated by a quantity equal to the loop loss and then subtracted from the incoming signal. An integrated circuit echo cancellor incorporating the functions just described is now available for use on satellite circuits. It may be noted that echo cancellors eliminate speech clipping and permit full duplex operation.

In the discussions so far, we considered only one reflection of the signal at the listener's end being echoed at the talker's end. This is referred to as 'talker echo'. If a second reflection takes place at the talker's end, 'listener echo' occurs. If repeated multiple reflections occur, oscillations are produced and a condition known as 'singing' is said to exist. Under this condition, the circuit is said to experience instability. Singing occurs if the loop gain at some frequency is greater than unity. If the loop gain is close to but less than unity, damped oscillations or near singing conditions result. Singing or near singing conditions have a disturbing effect on both the talker and the listener. In general, the procedures used to control echos also control singing. But in some short connections where no control is necessary for echos, singing may become a problem. Singing can occur in idle circuits whereas echos cannot. To control singing under such circumstances, the 4-wire circuits must have a certain minimum loss. CCITT recommends a minimum loss of 10 dB on national networks to avoid singing.

The amount by which the reflected signal is attenuated is known as **return loss**. This is given by

$$RL = 20 \log \frac{Z_4 + Z_2}{Z_4 - Z_2} \text{ dB} \tag{9.3}$$

where

RL = return loss
Z_4 = impedance of the 4-wire circuit
Z_2 = impedance of the 2-wire circuit

or, in terms of power,

$$RL = 10 \log \frac{P_4}{P_4 - P_2} \text{ dB} \tag{9.4}$$

where

P_4 = incoming power on the 4-wire circuit
P_2 = power reaching the 2-wire circuit
$P_4 - P_2$ = power reflected on to the return path

or, in terms of signal voltages,

$$RL = 20 \log \frac{V_4}{V_4 - V_2} \text{ dB} \tag{9.5}$$

$$= 20 \log \frac{1}{r_c} \tag{9.6}$$

where r_c is the reflection coefficient defined as

$$r_c = \frac{\text{reflected signal}}{\text{incident signal}} \tag{9.7}$$

If the two networks are perfectly balanced, then $Z_4 = Z_2$. Therefore, from Eq. (9.3) we have

$$RL \text{ (balanced)} = 20 \log \frac{2Z_2}{0} = \infty$$

The return loss is infinite, i.e. the return signal experiences an infinite attenuation and hence there is no reflected signal.

EXAMPLE 9.3 In a national transmission system the characteristic impedances of the 4-wire circuit and the 2-wire circuit are 1000 Ω and 1200 Ω respectively. The average phase velocity of the signal in the circuit is 3×10^7 m/s. If the largest distance of a connection is 300 km, determine the attenuation to be inserted in the circuit.

Solution From equation (9.3) we have

$$RL = 20 \log \frac{2200}{200} = 20.8 \text{ dB}$$

$$\text{Round trip delay for echo} = \frac{300 \times 10^3}{3 \times 10^7} = 10 \text{ ms}$$

No echo suppressor is required in this case. From Fig. 9.9, for a round trip delay of 10 ms, we need about 7 dB loss to contain the echo. As per CCITT recommendations, the circuits would have been provided with a loss of 10 dB to control singing. Hence, no additional attenuator needs to be inserted in the circuit to control echos.

9.4 Transmission Systems

Modern long distance transmission systems can be placed under three broad categories:

1. Radio systems
2. Coaxial cable systems
3. Optical fibre systems.

Fibre optic transmission systems have been discussed in detail in Chapter 7. In this section, we concentrate on radio and coaxial cable systems.

Radio communication deals with electronic radiation of electromagnetic energy from one point to another through the atmosphere or free space. It is possible only in a certain portion of the electromagnetic frequency spectrum. Presently, this portion includes frequencies from 9 kHz to 400 GHz. While there are international allocations for the radio spectrum up to 275 GHz, most of the commercial uses take place between 100 kHz and 20 GHz. Very few experimental systems have been operated beyond 100 GHz.

Different layers of the atmosphere play a role in propagating radio waves. The atmosphere consists of four layers as shown in Fig. 9.11. Of the

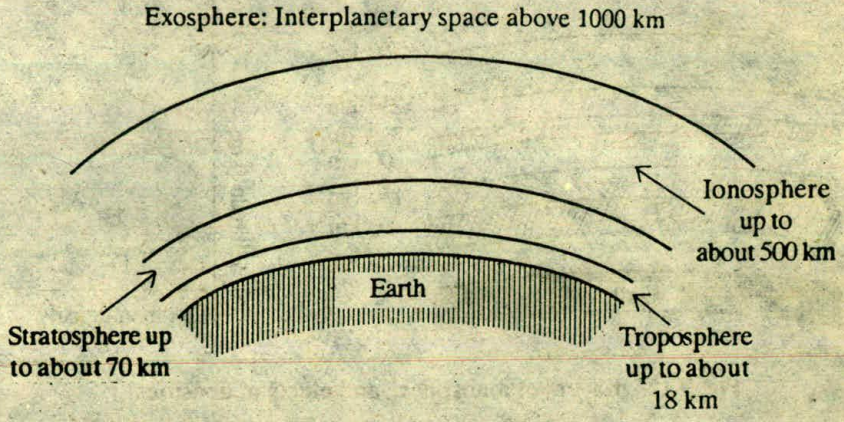


Fig. 9.11 Layers of atmosphere.

four layers, the ionosphere and troposphere are useful for radio communication in certain frequency ranges. Certain other radio frequencies pass straight through the atmosphere and can be beamed towards satellites placed in the interplanetary space. Depending on the mechanism of propagation, radio communication can be placed under four categories:

1. Skywave or ionospheric communication
2. Line-of-sight (LOS) microwave communication limited by horizon
3. Tropospheric scatter communication
4. Satellite communication.

9.4.1 Ionospheric Communication

Skywave or ionospheric communication deals with the frequency range 3–30 MHz, usually classified as high frequency (HF) band. Ultraviolet radiation from the sun ionises the gas molecules of the atmosphere producing positively charged ions and free electrons. A free electron may recombine with an ion when it comes near it and the two form a neutral atom which in turn may be ionised once again. Thus, a continuous process of ionisation and recombination takes place in the atmosphere. The degree of ionisation is dependent upon the intensity of the ultraviolet radiation and the density of the atmosphere. At high altitudes the intensity of the ultraviolet radiation is low. The recombination rate is lower in the rarer atmosphere than in the denser atmosphere. As a result, the ionospheric region is formed between 50–500 kilometres of altitude. Within the ionosphere, there are regions of concentration of electrons, and these may be considered as sublayers within the ionosphere. Four such layers are present in the day time as shown in Fig. 9.12. At night time, the absence of ultraviolet radiation and

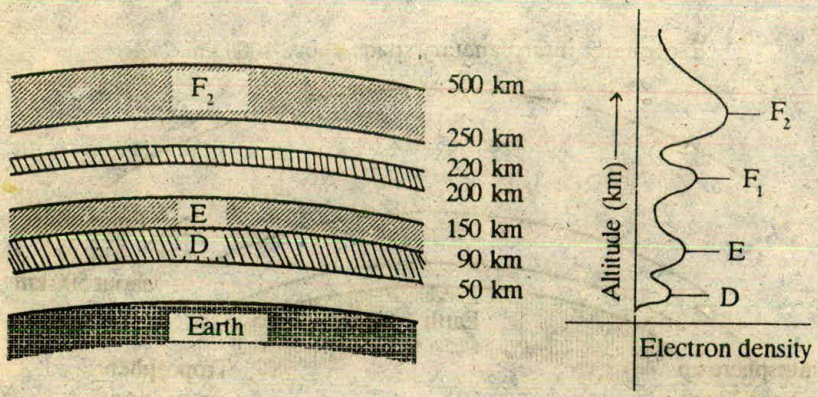


Fig. 9.12 Layers of ionosphere and electron densities.

the high recombination rate in the dense lower atmosphere make the D layer disappear. The heights of the ionospheric layers are not constant but vary both daily and seasonally as the intensity of sun's radiation fluctuates. At sunset, the F_1 layer rises, merging with the next higher F_2 layer and forming a common F layer. Thus, there are only two layers present in the night.

The layers of the ionosphere may be considered as mirrors or partial mirrors depending on the amount of ionisation present. Thus, the transmitted waves striking an ionospheric layer may be refracted back to the earth and received at a distance after one refraction. If this is the case, the transmission is known as **single hop transmission**. Alternatively, the waves may be reflected back from the earth to the ionosphere and the process of refraction and reflection repeated several times before reaching the distant receiver. In this case, the transmission is known as **multihop transmission**.

The mode of propagation is known as **skywave communication.** Single and multihop skywave transmissions are illustrated in Fig. 9.13. As the frequency

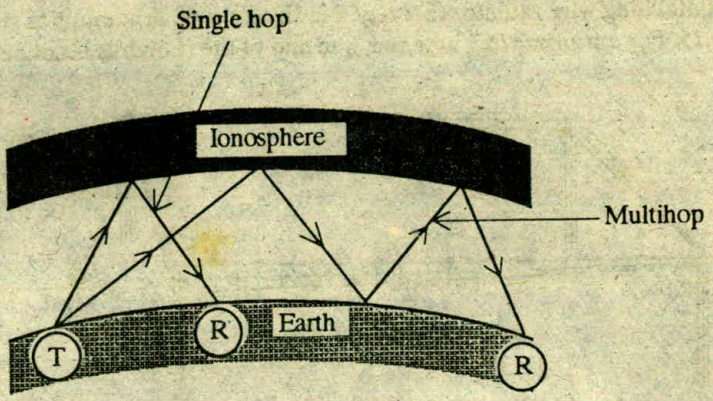


Fig. 9.13 Skywave communication.

of transmission increases, higher and more strongly ionised layers are required for refraction. Very low frequencies (VLF) and low frequencies (LF) are refracted by the less dense D layer whenever it is present. MF and HF are refracted by the F layer. As the frequency increases, at some point, the wave is not refracted by the ionosphere and pierces through it. This frequency is known as the **critical frequency** which varies with the time of day, season, geographical position, etc. Skywave communication suffers from a blackout of signals to some part of the earth known as **skip zones**. The skip zone extends from the outer limit of groundwave communication to the inner limit of skywave communication as illustrated in Fig. 9.14.

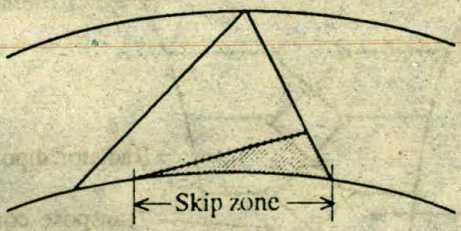


Fig. 9.14 Skip zone in skywave transmission.

Perhaps the most important element in an ionospheric (HF) telecommunication system is the antenna. For point-to-point HF operation, two basic antenna types are available:

- Rhombic
- Log periodic.

For transmission distances greater than 4500 km, precision-designed rhombic antennas are preferred. For shorter distances, both rhombic and log periodic antennas may be considered. A rhombic antenna comprises several horizontal long wire radiators arranged in the form of a rhombus as shown in Fig. 9.15. For a transmitting antenna, one end of the rhombus is connected to

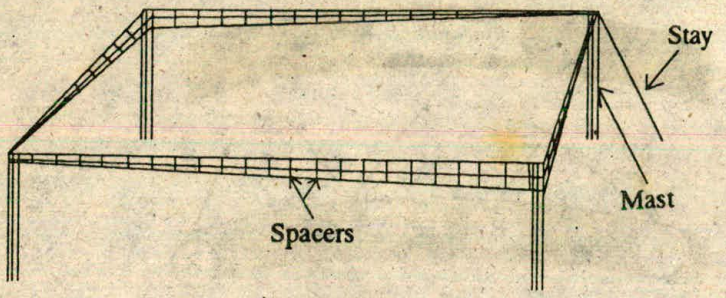


Fig. 9.15 Rhombic antenna structure.

the transmitter via a transmission line and the other end is terminated in a resistance. The antenna behaves like a slightly mismatched transmission line as the terminating impedance is slightly greater than the characteristic impedance to compensate for the loss of energy by radiation.

A log periodic (LP) antenna is more compact and has a good performance over a wider bandwidth than its rhombic counterpart. The basic arrangement of an LP antenna, shown in Fig. 9.16, has a number of radiating

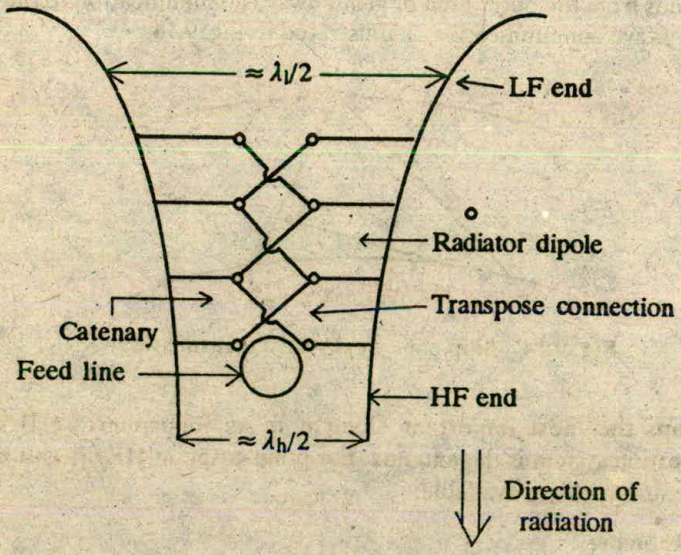


Fig. 9.16 A log periodic antenna.

elements (dipoles) with 180° transpositions between adjacent elements. The purpose of transposition is to enable the antenna to radiate from the smaller end. The physical lengths of the dipoles vary from a half-wavelength of the lowest frequency down to a half-wavelength of the highest frequency. The feed point for the antenna is at the end with the smaller dipoles. The relationship between dipole lengths and spacings results in a system in which for a given frequency three to four dipoles resonate and form a radiation centre. This centre moves steadily along the antenna as the frequency changes. The radiation pattern repeats periodically with the logarithm of the frequency and hence the name **log periodic antenna**.

Transmitting antennas concentrate radiation in wanted directions. As a result, a gain factor is achieved in the direction of radiation when compared to an isotropic source which radiates in all directions. An isotropic radiator is a hypothetical antenna and does not exist physically but represents a convenient reference for expressing the directive properties of practical antennas. Gain of a physical antenna is defined as

$$G = 20 \log \left(\frac{E_1}{E_2} \right) \tag{9.8}$$

where E_1 and E_2 are the field strengths given by the transmitting antenna and the isotropic source respectively at a given distance and total radiated power. A high gain antenna usually has subsidiary radiation lobes, known as **side lobes** in addition to the main lobe used for communication. Side lobes result in a loss of antenna gain, thereby reducing the useful power radiated. To take into account this loss, we define a parameter called **radiation efficiency** as

$$\eta = \frac{P}{P_i} \tag{9.9}$$

where

P = radiated power = $I^2 R$

P_i = input power = $I^2 (R + R_T)$

with

I = r.m.s. current flowing on the antenna

R = radiation resistance

R_T = equivalent terminal loss resistance

Therefore,

$$\eta = \frac{R}{R + R_T} \times 100\% \tag{9.10}$$

Antenna feed also introduces some loss. If the gain factor G is specified taking into account the radiation and antenna feed loss, it is referred to as **power gain**. The power gain is usually expressed in terms of **effective**

isotropically radiated power (EIRP), where the gain of an isotropic antenna is taken as 1 or 0 dB. To distinguish the gain factor specified by Eq.(9.8) from power gain, this factor is called **directive gain.** It is important to note that power gain is more relevant in the case of transmitting antenna and directive gain in the case of receiving antenna.

An antenna can be mounted either vertically or horizontally depending upon practical convenience. Accordingly, a radiation pattern is said to be either **vertically polarised** or **horizontally polarised.** Since in skywave communication, the signal reaches the receiver by one or more reflections from the ionosphere, the angle of launch of signal into the ionosphere assumes importance. The path length is determined by this launch angle which is known as the **take-off angle (TOA)** and is defined as the angle between the vertical and the line of maximum radiation. For a given path length and horizontally polarised antenna, the TOA decreases as the height of the antenna increases and is given by

$$\text{TOA} = \sin^{-1}\left(\frac{\lambda}{4h}\right) \tag{9.11}$$

where λ is the wavelength of radiation and h is the height of the antenna from the ground.

Skywave communication is prone to **fading** which is of two types: general fading, in which the whole signal fades and selective fading in which only some of the frequency components of the signal fade. General fading can usually be handled by automatic gain control (AGC) mechanism of the receiver. Selective fading occurs when the skywaves reach the destination via two or more paths. Different path lengths lead to phase distortion at the receiving end. Use of single side band transmission or frequency modulation, or restriction of the propagation to only one mode are the methods by which selective fading effects can be minimised. Selective fading is also known as **multipath fading.**

EXAMPLE 9.4 An antenna has a directive gain of 12 dBi, a radiation efficiency of 90% and a feeder loss of 3 dB. Determine its power gain. Spell out the significance of the power gain and directive gain values.

Solution

$$\text{Radiation loss} = 10 \log (0.9) = -0.5 \text{ dB}$$

Therefore,

$$\text{Power gain} = 12 - 3 - (0.5) = 8.5 \text{ dBi}$$

The Power gain of this antenna gives an increase in field strength of 8.5 dB over isotropic antenna and its directive gain yield an improvement of 12 dB in received (S/N) ratio.

9.4.2 Microwave Communication

Microwave communication is characterised by the following factors:

1. The useful ranges of frequencies lie between 150 MHz and 150 GHz.
2. It is a line-of-sight (LOS) communication and is limited by the horizon due to the curvature of the earth.
3. Signal propagation is affected by free space attenuation and precipitation.
4. Frequency or phase modulation (angle modulation) or spread spectrum and time sharing techniques are used.

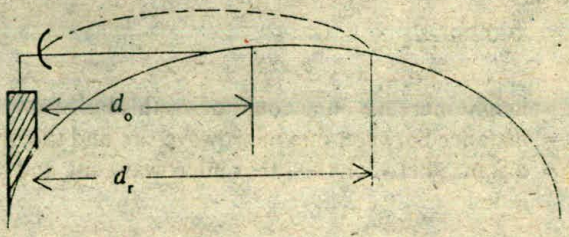
The useful frequency range of microwave spectrum is divided into a number of bands designated by letters. The bands and the associated radio frequencies (RF) are shown in Table 9.3. The bands are further divided into

Table 9.3 Microwave Bands

Band	Frequency range (GHz)	Band	Frequency range (GHz)
P	0.25 – 0.39	X	6.30 – 10.90
L	0.39 – 1.55	K	10.90 – 36.00
S	1.55 – 3.90	Q	36.00 – 46.00
C	3.90 – 6.20	V	46.00 – 56.00

a number of subbands. For example, K_u band covers the frequency range 10.9–14.5 GHz in the K band. For commercial microwave telecommunications, frequencies are largely allocated from the S, C and low-end K bands. Three carrier bands, 4, 6 and 11 GHz are used widely. Microwave systems are designed to have multichannel transmission capability of usually 1800 to 2700 nominal 4-kHz voice channels per radio carrier frequency.

Being a line-of-sight system limited by the horizon, the height of the antenna above the earth plays an important role in determining the transmission distance in microwave communication. Microwaves are usually bent or refracted beyond the optical horizon, i.e. the horizon visible to our eyes. Hence, the radio horizon is generally further away from the optical horizon as illustrated in Fig. 9.17. However, the distance to the radio horizon



d_o = optical horizon d_r = radio horizon

Fig. 9.17 Optical and radio horizons.

varies with the atmospheric refractive changes and can be even less than the optical horizon at times. As a rule of thumb, the optical and radio horizons may be expressed as

$$d_o = 0.46\sqrt{h} \tag{9.12}$$

$$d_r = 0.49h \tag{9.12a}$$

where

d_o = distance to the optical horizon in km
d_r = distance to the radio horizon in km
h = height of the tower in metres

We may define a correction factor K, taking into account the bending effect of the microwave beam. This factor is used to obtain any quantity related to radio horizon from the corresponding quantity related to optical horizon. We define K as

$$K = \frac{\text{value of the parameter for the optical horizon}}{\text{value of the parameter for the radio horizon}} \tag{9.13}$$

For the horizon distances, we have

$$K = \frac{d_o}{d_r} \tag{9.14}$$

If $K > 1$, the radio horizon is nearer than the optical horizon, and if $K < 1$, the radio horizon is further than the optical horizon.

The height of a microwave tower should be such that the radio beam is not obstructed by objects like buildings, trees, mountains etc. The height must obviously be more than the highest obstacle in the way such that the beam is clear of the obstacle. When seen from a tower, the effective height of an obstacle is more than its physical height. Two factors contribute to increasing the effective height:

- Earth's curvature bulge
- Fresnel diffraction.

The earth's bulge, which adds to the physical height of the obstacle, may be calculated as

$$h_{eb} = 0.078 d_1 d_2 \tag{9.15}$$

where

h_{eb} = height increase on account of earth's bulge
d_1 = distance between one microwave site and the obstacle in km
d_2 = distance between the other microwave site and the obstacle in km

The effect of ray bending can be accounted for by using the correction factor defined in Eq. (9.13).

The Fresnel phenomenon stems from the fact that electromagnetic wavefronts expand as they travel through space. The expanding properties result in reflection and phase transitions as the wave passes over obstacles. Hence, additional clearance is required above the obstacle to avoid such problems. The clearance required is expressed in terms of Fresnel's first zone, second zone etc. The radius of the first Fresnel zone R in metres is calculated as

$$R = 17.3 \left[\frac{d_1 d_2}{F(d_1 + d_2)} \right]^{\frac{1}{2}} \tag{9.16}$$

where F is the frequency of transmission in GHz and d_1 and d_2 have the same significance as in Eq. (9.15).

EXAMPLE 9.5 In the path profile of a microwave link of 25 km, a hill of height 70 m with trees is encountered at a distance of 10 km from the transmitting end. Carrier frequency is 6 GHz. Determine the tower height required. Assume a correction factor of 0.9 for ray bending.

Solution

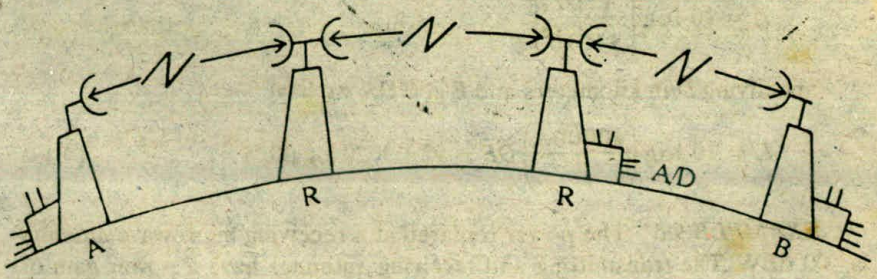
$$\text{Earth's bulge } h_{eb} = \frac{0.978h \times 10 \times 15}{0.9} = 13 \text{ m}$$

$$\text{Fresnel radius } R = 17.3 \left[\frac{10 \times 15}{6 \times 25} \right]^{\frac{1}{2}}$$

Allowing 15 m for trees and future growth, we have

$$\text{Height of the tower} = 70 + 13 + 17.3 + 15 = 122.3 \text{ m}$$

A microwave link is made up of two terminal sites and usually one or more repeater sites as shown in Fig. 9.18. At the transmitting terminal site voice channels are multiplexed (FDM) into a baseband signal which is then



A, B = terminal sites R = repeater A/D = add and drop links

Fig. 9.18 A microwave link.

modulated into the RF carrier. At the receiving site, the RF carrier is demodulated and the resulting baseband signal demultiplexed into individual voice channels. A repeater site is characterised by two antennas for the two directions. It receives, amplifies, and retransmits the RF signals to the next site in sequence. Some repeaters may add and drop a few lines serving some local connections.

Microwave systems may be either long-haul type or short-haul type. A long-haul system has only a small number of add and drop points but has many simple repeaters and covers a long distance end-to-end. A short-haul system consists of a relatively small number of repeaters with frequent add and drop points. Of the three commonly used microwave carrier bands, viz. 4, 6, and 11 GHz, the 4 GHz portion of the spectrum is used for long-haul transmission. The frequencies around 11 GHz carrier experience severe attenuation during rainfall and hence are not used in long-haul routes. They are suitable for short distance local communications. The 6-GHz band is useful for both long and short-haul communications.

Very high gain antennas are required for microwave communications. The free space path loss is high and the antenna gain is required to compensate for this. To get an idea of the magnitude of loss involved let us consider two isotropic antennas separated by a distance D. The power intercepted by the receiving antenna is given by

$$P_R = \frac{P_T \lambda^2}{(4\pi D)^2} \tag{9.17}$$

where

P_R = received power
P_T = transmitted power
λ = wavelength of operation
D = distance between two antennas

The power loss L when expressed in decibels is given by

$$L = 10 \log \left(\frac{4\pi DF}{3 \times 10^8} \right)^2$$

Specifying D in kilometers and F in GHz, we have

$$L = 20 \log \left(\frac{40000\pi}{3} \right) DF = 92.4 + 20 \log (DF) \tag{9.18}$$

EXAMPLE 9.6 The power required at a receiving microwave antenna is –60 dBW. The transmitting and receiving antennas have a power gain of 40 dB each. Determine the required transmitter power if the carrier frequency is 4 GHz and the distance between the antennas is 48 km. What happens to the free space power loss when the distance is doubled?

Solution From Eq. (9.18), we determine the free space loss as

$$L = 92.4 + 20 \log (48 \times 4) = 138 \text{ dB}$$

The power budget equation may be written as

$$P_T + G_T + G_R = L + P_R$$

where G_T and G_R are the transmitting and receiving antenna gains respectively. Substituting the values, we have

$$P_T + 40 + 40 = 138 - 60$$

Therefore, $P_T = -2$ dBW.

When the distance is doubled, the loss is given by

$$L = 92.4 + 20 \log (40 \times 96) = 144 \text{ dB}$$

Thus, doubling the transmission distance increases the path loss by 6 dB.

Microwave frequencies can be treated much the same way as light beams. You may know that parabolic reflectors are used in searchlights to produce a powerful beam of light by locating the light source exactly at the parabolic focus. Similarly, parabolic reflectors are used to realise highly directional high gain antenna concentrating microwave energy into a parallel beam. The antenna feed is located at the focal point of the parabolic reflector which is commonly called disk. The diameter of a disk varies from 20 cm for internal domestic TV antenna to 30 metres for ground stations working with geostationary satellites. Three versions of parabolic antenna are in use:

1. Simple parabolic
2. Horn
3. Cassegrain.

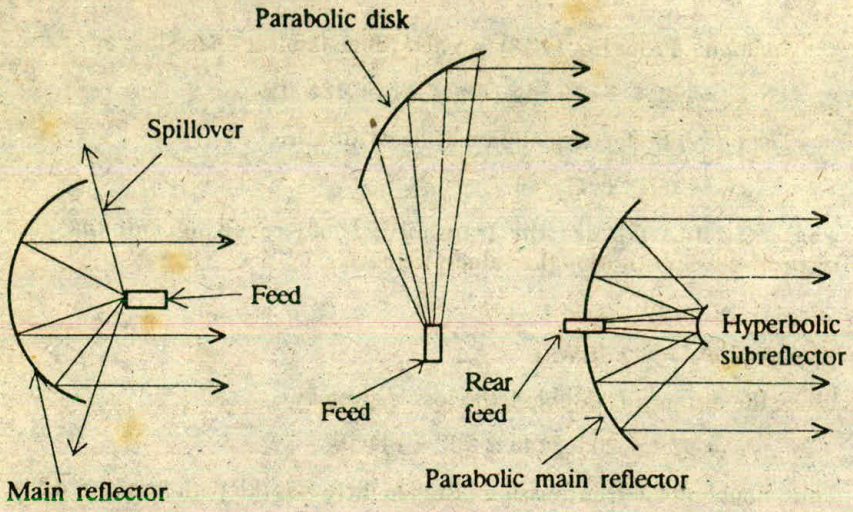
The three types are depicted in Fig. 9.19. All the three use parabolic disk as the main reflector. The reflector diameter is large compared to the wavelength of the signal. The larger the area of the parabolic surface, the greater the directivity, and hence the gain. The gain of the antenna and its effective reflecting area are related by the equation

$$G = \frac{4\pi A_e}{\lambda^2} \tag{9.19}$$

where G is the gain and A_e is the effective area of the antenna. The effective area is generally in the range of 0.5 to 0.6 of the physical area of the disk. The gain of a parabolic antenna relative to an isotropic antenna can be expressed in decibels as

$$G = 20 \log 7.8Fd \text{ dB} \tag{9.19a}$$

where F is the frequency of the carrier in GHz and d is the diameter of the antenna in metres. In Eq. (9.19a), the effective area is assumed to be 55% of



(a) Simple parabolic (b) Horn (c) Cassegrain

Fig. 9.19 Types of parabolic antennas.

the physical area. If G_T is the gain of the transmitting antenna and G_R the gain of the receiving antenna, then Eq. (9.17), applicable to isotropic antennas, can be modified to directive antennas as

$$P_R = \frac{P_T G_T G_R \lambda^2}{(4\pi D)^2} \tag{9.20}$$

Substituting for G_T and G_R from Eq. (9.19), we obtain

$$\frac{P_R}{P_T} = \frac{A_{eT} A_{eR}}{\lambda^2 D^2} \tag{9.20a}$$

where

A_{eT} = effective area of the transmitting antenna
A_{eR} = effective area of the receiving antenna

The horn reflector antenna, shown in Fig. 9.19(b), was invented by H.T. Friis in the mid 1940s and is in wide use today. It uses only a section of a very large parabolic surface as the radiator and is illuminated from the focal point by tapered horn of either round or square cross-section in such a way that the vertical-to-horizontal transition of the beam takes place within the antenna itself. This antenna has excellent radiation patterns, high gain and a low loss due to back reflection from the disk. It is therefore possible to mount several horn antennas on the same tower platform.

Cassegrain antenna, shown in Fig. 9.19(c), is based on the working principle of Cassegrainian telescope invented by Cassegrain, a 17th century

French astronomer. The antenna has two reflectors, a paraboloidal main reflector and a hyperboloidal subreflector. The subreflector is placed between a point source feed and the prime focus of the paraboloid. Microwave rays from the feed reflected by the subreflector appear to emerge from the focus of the paraboloid. The rays falling on the paraboloid are reflected parallel to its axis.

Blocking the aperture of a reflector antenna by its feed, or feed supports or subreflector generally degrades its directivity pattern. To eliminate aperture blocking, many high performance microwave systems employ **offset feed antennas**. In this configuration, the feed or the subreflector is offset from the focus and a relatively complicated mechanical geometry is used to illuminate the main reflector. The complicated offset geometry poses serious fabrication and assembly problems. Nevertheless, many high performance requirement specifications are met only by the offset geometry.

Every reflector antenna has an associated radiation or receiving feed system. Feeds can be designed to provide high aperture illumination efficiency, low sidelobes or a special beam contour. A widely accepted feed structure is the **corrugated horn** because of its large bandwidth, symmetric pattern, and relatively low sidelobes. The corrugated horn derives its name from the deep grooves in the inside wall of the horn.

Microwave energy is guided to the antenna feed from the transmitter system by means of waveguides. In the microwave frequency region, the electromagnetic energy propagates on the inner walls of the medium and hence the central conductor of a coaxial structure is no longer needed. Waveguides have come about because of this phenomenon. Depending on the shape of the cross-section, the waveguides are classified as

1. Rectangular waveguides
2. Circular waveguides
3. Elliptical waveguides
4. Square waveguides.

Rectangular waveguides usually have a cross-section with an aspect ratio of 1:2, i.e. the width being about twice the height. The waveguide size determines the cutoff frequency, i.e. a frequency below which satisfactory propagation cannot take place. Therefore, systems operating in different band use different sizes of guides. For example, for operation in the 3.7–4.24 GHz range, a rectangular waveguide has the larger dimension around 5.75 cm and for the frequency range 10.7–11.7 GHz around 2.25 cm. Waveguides are coded by a letter to indicate the shape of the cross-section and a number to indicate the larger dimension of the shape. For example, WR–5.75 means rectangular waveguide with the width being 5.75 cm in metric system.

Different waveguides support propagation in different modes. The dominant mode of propagation in a rectangular waveguide is TE_{01}, the

transverse electric mode. In this mode, the electric sector is perpendicular to the direction of propagation in the guide. Circular waveguides support several modes, depending upon the frequency of operation and how the energy is launched into the guide. (Recall the treatment on optical fibres in Chapter 7, which are nothing but circular waveguides suitable for optical frequencies.) Elliptical waveguides have some desirable mechanical properties that they can be easily bent and coiled. They are formed from a metal ribbon that is wound on a shaped mandrel in such a fashion that the finished product is semiflexible. Square waveguides are rarely used but have some applications which require propagation of two polarisations simultaneously.

As in the case of ionospheric communication, microwave communication also suffers from fading. Fading occurs mainly due to reflections from a stratified atmosphere or from surface land along the path. Such fading also causes multipath effects which result in constructive or destructive interference in the level of the incoming signal. Multipath fading varies with pathlength, frequency, climate and terrain. In dry, windy, mountainous areas, the multipath phenomenon is virtually nonexistent. Flat terrain along a path tends to increase the incidence of fading. Fading is prominent in hot, humid coastal regions.

Almost universally, fading is combated by using diversity techniques. Three types of diversity techniques are in use:

1. Space diversity
2. Frequency diversity
3. Polarisation diversity.

Diversity systems establish and use more than one path for signal transmission and work on the theory that fading does not occur simultaneously on all the paths.

A space diversity configuration is shown in Fig. 9.20. Here, two or more receiving antennas are used and the output is selected from the antenna

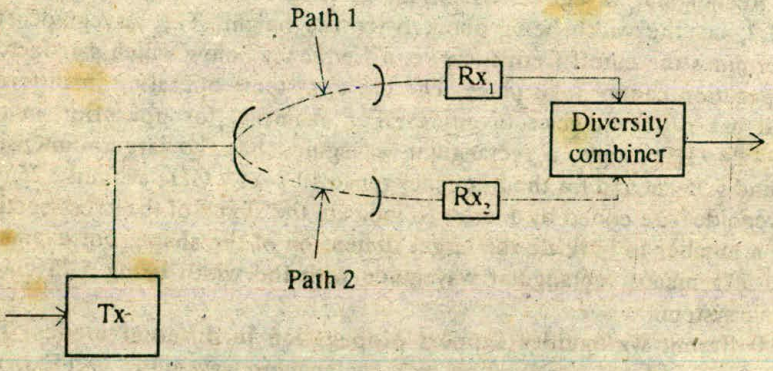


Fig. 9.20 Space diversity system.

receiving the strongest signal. The receiving antennas are physically separated, usually vertically on the same tower. The separation distance is at least 70 wavelengths.

Frequency diversity is more complex and expensive than space diversity. A frequency diversity set up is shown in Fig. 9.21. Here, the transmitters are

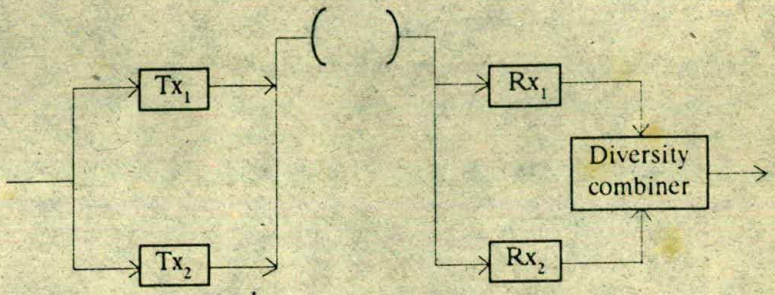


Fig. 9.21 Frequency diversity system.

modulated simultaneously by the same signal but they transmit on different frequencies. The frequency separation is at least two per cent. When fading occurs in one frequency, it may not probably occur on the other frequency as well. The more the frequency separation, the less is the chance that fading will occur on both paths. At the receiving end, there are two receivers, one for each frequency and a combiner to pick out the stronger signal.

An inherent advantage of frequency diversity system is the full equipment redundancy which gives a greater operational reliability and a very high probability of uninterrupted service. The primary disadvantage of frequency diversity is the requirement of double the frequency spectrum. In this age, when frequency spectrum is at a premium, frequency diversity schemes are not easily accepted by the licensing authorities.

As an alternative to frequency diversity, polarisation diversity is often resorted to. Here, two signals are transmitted, one horizontally polarised and the other vertically polarised. Two transmitting antennas are used. At the receiving end also, two antennas are used, each having dual polarised feed horns for receiving signals in both planes of polarisation. Polarisation diversity systems also offer a certain amount of redundancy as the frequency diversity systems.

9.4.3 Tropospheric Scatter Communication

Tropospheric scatter or simply troposcatter systems are single hop systems that operate over lengths that are greater than the line-of-sight distances. Atmospheric air turbulence and local variations or irregularities in the refractive index in the tropospheric region of the atmosphere cause scattering of electromagnetic waves. When a transmitter launches a narrow

beam of signals to its horizon, the beam 'illuminates' a region in the troposphere, which is well beyond the horizon of the transmitter. If a receiving station looks at this scattering region via a high gain antenna, it is able to receive the signals. This is illustrated in Fig. 9.22.

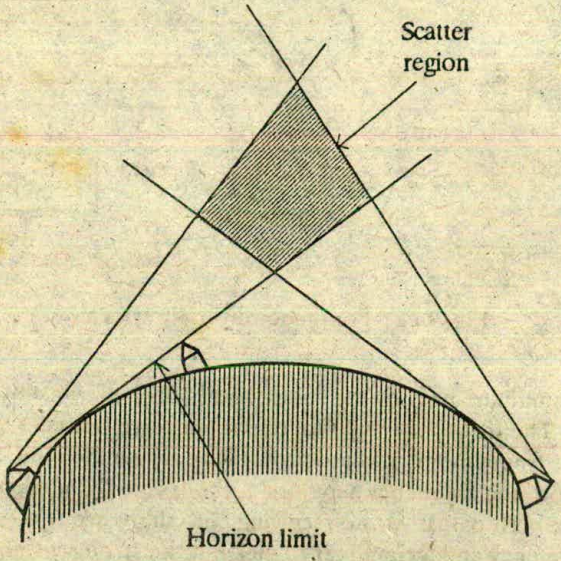


Fig. 9.22 Troposcatter phenomenon.

The level of the received signal in troposcatter communication is well below the level of signal obtained in free space communication for equivalent distance between the stations. Losses up to 260 dB are not uncommon. To achieve acceptable performance, very high gain antennas, high power transmitters, and receivers with low noise factors are required. Transmitters with 1 kW to 10 kW power output and parabolic antennas with 5, 10 or 20 m diameter dishes are common. Further, the received signal level varies considerably with time as the scattering is random. There are both long term and short term variations in the signals. In other words, troposcatter communication experiences both slow and fast fading. Slow fading can be corrected by the automatic gain control mechanism of the receiver. A sophisticated diversity system is required to reduce fast fading. Most present day operational systems employ quadruple diversity using both frequency and space diversity. Space diversity is almost universally used. The physical separation of antennas is normally in the horizontal plane with a separation distance greater than 100 wavelengths. Frequency diversity may not always be possible due to restrictions in the availability of multiple frequencies in the crowded RF band. As an alternative, polarisation diversity is used. The two polarisations together with the space diversity provided by the two antennas make up for four diversity paths.

A number of methods are available for estimating both long term and short term path losses in tropospheric communication. These include CCIR (International Consultative Committee for Radio) Report 238.3 and the US National Bureau of Standards (NBS) Technical Note 101. These methods are elaborate and complicated and a discussion on these is beyond the scope of this text.

Tropospheric scatter communication generally operates in the 400-MHz and 900-MHz, or 2-GHz and 4-GHz bands handling 12–240 FDM telephone channels or up to 120 digital channel. This form of communication is normally used only in thin-traffic routes over difficult terrain or water. Ministry of Telecommunications in India operates a troposcatter communication link between New Delhi and Moscow.

9.4.4 Satellite Communications

A communication satellite is like a big microwave repeater in the sky. It is placed in a circular orbit which lies in the plane earth's equator, at an altitude of about 36,000 km above the equator. At this altitude, the satellite has an orbital period equal to the duration of one sidereal day, i.e. 23 hours, 56 minutes and 4.1 seconds. With this orbital period, the satellite's relative angular velocity with respect to that of the earth is zero, and hence the satellite appears stationary when viewed from the earth, i.e. geostationary. The relationship between the angular velocity and the altitude is easily derived by balancing the two major forces acting on the satellite, viz. the centrifugal force and the gravitational force as

$$mr\omega^2 = \frac{MmG}{r^2} \tag{9.21}$$

where

M = mass of the earth, 5.98×10^{24} kg
m = mass of the satellite
r = distance of the satellite from the centre of the earth
G = gravitational constant, 6.67×10^{-11} Nm^2/kg^2

$$\omega = \text{angular velocity} = \frac{2\pi}{T} = \frac{2\pi}{24 \times 3600} = 7.27 \times 10^{-5} \text{ rad/s}$$

Substituting the known values in Eq. (9.21), we have

$$r = \left(\frac{6.67 \times 10^{-11} \times 5.98 \times 10^{24}}{7.27 \times 10^{-5} \times 7.27 \times 10^{-5}} \right)^{\frac{1}{3}} = 42{,}200 \text{ km}$$

Now, the radius of the earth, R, is 6400 km. Therefore, the altitude h of the satellite is given as $h = r - R = 35{,}800$ km.

A satellite in orbit experiences small disturbances due to some weak forces such as attraction from moon, sun and other planetary bodies, pressure of solar radiation and variations in the earth's gravitational force. These forces cause the satellite to drift from its orbit, retard or accelerate its

motion and affect its orientation (attitude) towards the earth. To overcome the effect of these forces, every satellite is provided with an attitude and orbit control system (AOCS) on board. The rate of fuel consumption in carrying out the AOCS functions and the quantum of fuel available on board are two factors that determine the lifetime of a satellite. There are also other factors such as battery life and solar cell degradation. It is expensive in terms of fuel consumption to maintain the exact attitude and orbit positions of a satellite at every instant of time. By allowing a satellite to have an orbital plane that is inclined at a small angle with respect to the earth's equatorial plane and drift marginally around its mean position, fuel can be conserved, thereby extending the useful life of the satellite. In this case, the satellite is said to be a **geosynchronous satellite** which has the same mean orbital period as the earth, but may have a small nonzero relative speed at any given instant, i.e. it is not truly stationary. A geostationary satellite is always geosynchronous, but the reverse is not necessarily true. Two Indian satellites, APPLE satellite and INSAT-1C, were both operated in geosynchronous mode to save fuel. (APPLE stands for Ariane passenger payload experiment and INSAT for Indian national satellite system. Ariane is a Franco-German satellite launch vehicle). In the case of geosynchronous satellites, ground station antennas must have capability for minor orientation manoeuvre to keep them pointed towards the satellite.

In any communication satellite, one or more transponders are used to receive, amplify and retransmit the incoming signals. The upward and downward transmission frequencies are chosen to be different to avoid interference. Three frequency bands are allocated for commercial satellite communication as shown in Table 9.4. The C band is widely used. Recent

Table 9.4 Frequency Bands for Satellite Communication

Band	Down link frequency range	Up link frequency range	Bandwidth
C	3.7 – 4.2 GHz	5.925 – 6.425 GHz	500 MHz
K_u	10.9 – 11.7 GHz	14.00 – 14.5 GHz	500 MHz
K_a	17.7 – 21.2 GHz	27.5 – 31.0 GHz	3.5 GHz

satellites are designed to operate in the K_u band. Heavy demand for bandwidth in the C and K_u bands is pushing up the frequency of operation to K_a band and above. Technology for operation in the K_a band is still in the development stages. The merits and demerits of the three bands are brought out in Table 9.5.

Starting with about 50 MHz bandwidth in the early systems, present day satellites have a bandwidth of 500 MHz to 2 GHz. The available bandwidth is generally split over a number of transponders which may be as many as about 50. INSAT-1 series of satellites have 12 transponders of 36 MHz bandwidth each. A 36-MHz transponder may carry a 500 Mbps data stream or 800 digital voice channels at 64 kbps.

Table 9.5 Satellite Frequency Band Characterisation

Band	Beam	Satellite spacing	Atmospheric effects
C band	Wide	4°	Less
K_u band	Narrow	1°	More
K_a band	Narrower	< 1°	Severe

Early satellites were designed with single spatial beam providing wide earth coverage. Theoretically, three such satellites properly placed in synchronous orbit could provide 100% coverage of the earth as shown in Fig. 9.23. Typically, these satellites had an EIRP of 20 to 28 dBW, requiring

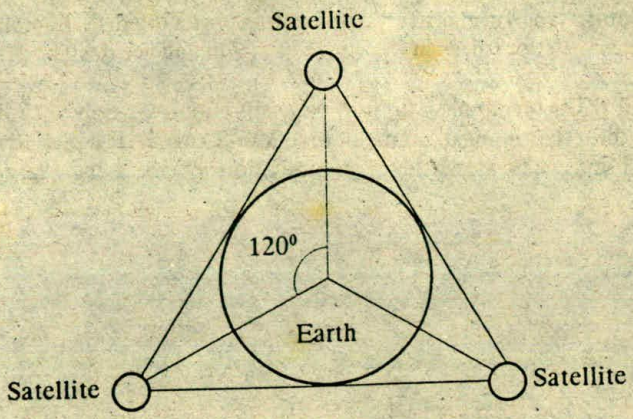


Fig. 9.23 Satellite positioning for 100% earth coverage.

large earth station antennas (10–30 m) for operation in C band. Technological advancements have brought about the use of multiple directional antennas on the present day satellites, each antenna providing a narrow spot beam, covering a small geographical area and an EIRP of 35 to 50 dBW. These satellites can operate with considerably smaller sized antennas of 3 to 6 m.

Spot beams give an added advantage of spectrum reuse. Same frequencies can be used for different geographical areas, thus achieving more effective bandwidth. Frequency reuse can also be achieved by using different polarisations in the beams. INTELSAT V reuses the 500-MHz C band four times and its successor INTELSAT VI is designed to reuse the spectrum six times. (INTELSAT is an international organisation set up in 1964 to establish and manage a global satellite communication system). Multiple beams, however, tend to isolate communities of earth stations removing the advantage of worldwide connectivity of geostationary sat-

By introducing processing and switching on board the spacecraft, dynamic time domain switching of traffic between multiple beams can be achieved, thus restoring full connectivity.

The cost of a satellite communication system is appropriated between the space and ground segments. The space segment comprises spacecraft and the equipment on board the spacecraft. The ground segment consists of the earth stations, antennas, ground control systems, user terminals etc. If the system is to serve a small population of ground terminals as in the case of intercontinental links, less investment can be made in the space segment and more in building large antennas and powerful ground stations. On the other hand, if the population of terminals to be served is large as in the case of direct broadcast systems, more investment is to be made in the space segment so that the cost of the terminals can be kept low. The present trend is to design more and more of **direct-to-user (DTU)** systems. This requires that we have satellite terminals that are affordable by every home. **Very small aperture terminals (VSATs)** are a step in this direction. Here, the nomenclature 'aperture' is used synonymously with the dish diameter. VSATs have dishes with apertures in the range of 1–2 m and are designed to be DTU terminals.

Small DTU terminals can, however, communicate only with large hubs and not directly amongst them. Direct VSAT-to-VSAT connectivity is not feasible. Figure 9.24 shows the communication scheme using a large hub and

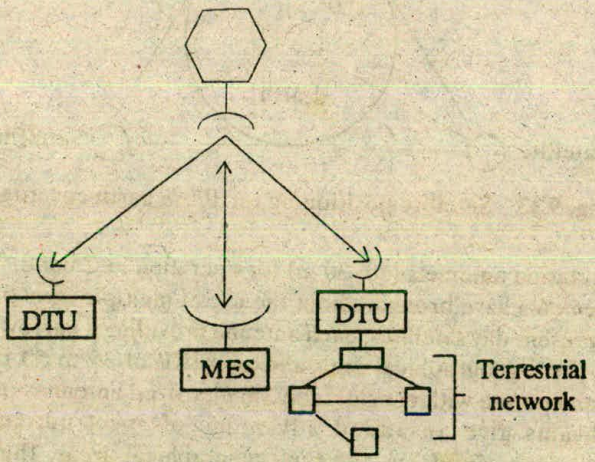


Fig. 9.24 Two-hop satellite network.

DTU terminals. There is a master earth station (MES) in the configuration which forms a vital communication link. The antenna of the MES acts as the large hub (typical size 10–30 m) with a high gain. A DTU terminal can communicate only with the MES antenna. Hence, information from one user to another is sent via the MES. The information hops twice before it reaches

the destination, once from the source user to MES and the second time from MES to the destination user. Hence, this scheme is known as a two-hop scheme. In India two-hop schemes are adopted in the government information network known as NICNET set up by the National Informatics Centre and the rural area business messaging network (RABMN) set up by the Department of Telecommunications (DOT). The NICNET operates in the K_u band using Intelsat VI or in C band using INSAT and the RABMN in the C band using INSAT. The main applications of VSAT systems are personal computer connectivity, low speed data communication and voice transmission.

Satellite communication differs radically from point-to-point terrestial links in many ways. Table 9.6 summarises these differences. By virtue of the altitude at which a geosynchronous satellite is placed, the distance between

Table 9.6 Comparison of Satellite and Terrestrial Communications

Satellite communication	Terrestrial communication
1. Star topology.	Mesh topology.
2. Satellite is a critical component; its failure results in total failure of the network.	Node failures do not affect the entire network. The network is fault tolerant.
3. Broadcast in nature.	Point-to-point in nature.
4. Large distances and hence large time delays are involved, typically 270 ms for one way communication.	Distances are relatively small and hence the time delay is small, typically 40 ms for intercontinental calls.
5. Microwave communication.	Microwave, cable or optical fibre communication.
6. Susceptible to weather conditions at higher bands.	Cable or optical fibre communication is insensitive to weather conditions.
7. Satellite life span is typically 7 years.	Life span is large, typically 30 years or more.

a ground terminal and the satellite transponder is usually over 40,000 kilometers, leading to a propagation delay of around 135 ms each in the uplink and the down link: A turnaround delay of over 0.5 s gets added to the response time which can be critical for some interactive applications. A two-hop system introduces a delay of one second, which renders even a telephone conversation to sound unnatural.

A satellite transponder is a common resource to be shared by a large number of ground stations for multiple simultaneous communications. Techniques are required to access and share the common transponder capacity. Many techniques have been devised and all of them are based on the frequency or time division concepts.

In a situation where the number of earth stations is limited and there is continuous traffic between a pair of originating and receiving stations, the traditional frequency division multiplexing (FDM) or time division multiplexing (TDM) techniques could be used where one slot is allocated to each station permanently. But, in practice, the traffic may be much less. Moreover, the number of earth stations is usually much larger than the number of FDM or TDM slots that can be made available. This is particularly so in systems using DTU terminals. In such a case, a free TDM or FDM slot needs to be allocated to an active station and deallocated as soon as the station traffic ceases. The same slot may be allocated to another station thereafter. Such a scheme is known as **frequency division multiple access (FDMA)** or **time division multiple access (TDMA)**, depending on whether frequency or time division is used. The term 'multiple access' is due to the fact that the same slot may be accessed by different stations at different times and a station may be allocated to different slots at different times. Multiple access may be **preassigned** or **demand assigned**. In preassigned schemes, a group of stations is identified with one or a group of FDM or TDM slots and given access only to these slots. In **demand assigned multiple access (DAMA)** scheme, there is a pool of space segment slots and any one of the available slots is assigned to a station when demanded by it.

One of the early demand assigned FDMA systems is SPADE, developed by COMSAT laboratories and used on early INTELSAT systems. The SPADE system provides a pool of up to 800 **single-channel-per-carrier (SCPC)** FDM channels in a transponder of 40 MHz bandwidth. Assignments may be made to any of the free SCPC channel on demand from any ground station in the system. Demand assignment is controlled by using a distributed DAMA controller. A common signalling channel is used by the DAMA controller to exchange messages relating to demand assignment. Each SPADE terminal maintains a complete channel assignment map for the entire system. A station requiring a channel first analyses the assignment map, chooses an unused channel and indicates the chosen channel on the signalling network. If the selected channel is still unused when the request is seen in the downward link, the channel is considered allocated. If two or more stations try to pick the same channel at the same time, a predetermined priority order among the stations is used to decide which one of them gets the channel. When a station finishes using the allocated channel, it sends a deallocation message on the common signalling network. All stations must receive all the assignment and release messages and update the assignment map continuously. DAMA controllers may also be centralised; central demand assignment involves a double-hop for every assignment, but is less complex and cheaper.

Similar to FDMA and TDMA, multiple access can be organised around code division multiplexing CDMA. In CDMA, individual bits of a message are encoded in terms of unique digital codes. Many stations, each having its own unique code, transmit their signals simultaneously. The signals are discriminated by means of a cross correlator, using the code of the desired transmission as the reference code. Since each transmitted symbol is

encoded as a sequence of higher rate symbols, the scheme demands a much larger bandwidth for transmission and the technique is called **code division multiple access**, and hence is referred to as **spread spectrum technique**. The number of higher rate coding symbols per bit, usually called chips, is an important parameter in CDMA, and is referred to as **processing gain** or **bandwidth spreading factor**. The spectrum spreading property significantly reduces interference from other transmissions. The codes used for CDMA are specially selected to have good cross correlation properties so that interference from other unwanted signals simultaneously may be minimised. A family of pseudo random (PR) codes, known as GOLD codes, have good cross correlation properties.

CDMA may be used in two ways. One way is that every transmitting station uses its own PR sequence encoding data and a receiver station uses a separate correlator for each station it wishes to receive. The other way is for the transmitting station to choose the PR sequence of the destination station to encode its message. In this case, each receiver station has only one correlator using its own PR code as the reference sequence. One of the applications, for which CDMA is very suitable, is to transmit low bit rate data from a central station to any of a large number of receive only small stations. The capability of CDMA to selectively address the remote stations by use of different codes comes in very handy here. In India, NICNET and RABMN use a CDMA access technique.

From Tables 9.3 and 9.4, it is seen that C and K bands are used for both terrestrial and satellite communication. In order to minimise interference with terrestrial microwave communication systems, the EIRP of the satellite systems is limited to 28 dBW in the 4-GHz band and 48 dBW in the 11-GHz band by international agreement. As a result, signals received by a satellite in the up link and by an earth station in the down link are very weak and are comparable to the thermal noise in the systems. Therefore, special considerations are required to minimise thermal noise in satellite communications systems. Thermal noise is conveniently expressed in terms of equivalent noise temperature as it is directly proportional to the temperature of the system for a given bandwidth. The thermal noise power P_n is given by

$$P_n = kT_sB \qquad (9.22)$$

where

T_s = absolute temperature of the system in degrees Kelvin
B = bandwidth in Hz
k = Boltzmann's constant expressed in watts per degree Kelvin per Hz

The value of k is 1.38×10^{-23} W/K/Hz. The **equivalent noise temperature** is defined as the temperature at the antenna of an ideal receiver which would produce a noise power equivalent to that measured in the

practical systems. The performance of the receiving system of the satellite and the earth station is expressed by a **figure of merit** which is the ratio of net antenna gain up to a reference point G and the equivalent noise temperature of the system T, expressed as

$$G/T = G_{dB} - 10 \log T \tag{9.23}$$

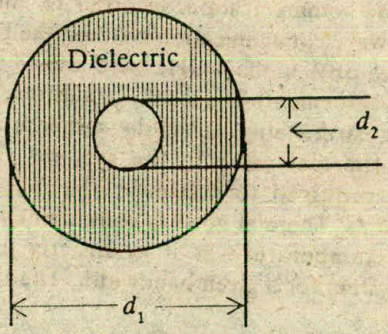
INTELSAT lays down a certain minimum figure of merit values for different classes of earth stations, for operation with INTELSAT satellites. For example, a standard A earth station must have the following minimum G/T value in the direction of the satellite for the polarisations chosen for the satellite series under clear sky and light wind conditions:

$$\frac{G}{T} \geq 40.7 + 20\log (F/4) \text{ dB/K} \tag{9.23a}$$

where F is the receive frequency in GHz in the C band range of 3.7–4.2 GHz.

9.4.5 Coaxial Cable Transmission

Coaxial cable is a 2-wire transmission line with a unique arrangement of the conductors which are held in a concentric configuration by a dielectric, with the inner conductor being surrounded by the outer conductor as shown in Fig. 9.25. The outer conductor which is grounded acts as an electromagnetic shield, almost eliminating the pick up of unwanted interference and reducing



d_1 = inner diameter of the outer conductor
d_2 = outer diameter of the inner conductor

Fig. 9.25 Coaxial cable structure.

the signal loss due to electromagnetic radiation. Since one of the conductors is grounded, coaxial cable is basically an unbalanced transmission line. The ungrounded inner conductor carries the signal voltage with ground as the reference.

For full duplex transmission, generally a pair of coaxial lines is used, although in some special cases a single line has been used for 2-way transmission. For long-haul systems, 4, 6, 8 or more lines (or tubes as the

single coaxial lines are called) are placed in a sheath. The characteristic impedance Z_0 of coaxial lines lies between 50 and 100 Ω. It is possible to estimate the line parameters of a coaxial line from its physical dimensions and the dielectric constant of the insulating material between the conductors. The line parameters of a transmission line have been discussed in Section 5.7, and we only present the values for a coaxial line here. Referring to Fig. 9.25, we have the following values for capacitance and inductance:

$$C = 24\epsilon / \log(d_1/d_2) \text{ pF/m}$$

$$L = 0.46 \log(d_1/d_2) \ \mu\text{H/m}$$

where ε is the dielectric constant of the insulating material. Therefore, the characteristic impedance Z_0 is given by

$$Z_0 = (L/C)^{\frac{1}{2}} = \frac{138}{\sqrt{\epsilon}} \log (d_1/d_2) \ \Omega \tag{9.23b}$$

The attenuation constant A_c for a coaxial cable at an operating bandwidth of F MHz is given by

$$A_c = a + b + \sqrt{F} + cF \text{ dB/km} \tag{9.24}$$

where a, b and c are constants dependent upon the physical parameters of the cable. For long haul transmission, two standard sizes are used. The dimensions of the standard cables and the associated values of a, b and c are presented in Table 9.7. The values of phase velocities V_p are also indicated for the two cables in Table 9.7 for high frequency (see Section 5.7) assuming PVC as the insulating material.

Table 9.7 Standard Coaxial Cables

	d_2 (mm)	d_1 (mm)	a	b	c	V_p (m/s)
Size 1	1.2	4.4	0.066	5.15	0.0047	1.8×10^8
Size 2	2.6	9.5	0.013	2.305	0.003	1.8×10^8

Coaxial cables are usually buried at a depth of 90–120 cm, depending on frost penetration in a given locality. Repeaters for coaxial cables are placed at uniform intervals along the route. To facilitate this, cable lengths are factory cut so that the splice occurs right at repeater locations. The repeaters are of two types: secondary or dependent repeaters and primary or main repeaters. Main repeaters have independent power sources and are installed in surface housing. Secondary repeaters derive their power from the cable itself and are usually buried. Power to the dependent repeaters is supplied by the main repeaters. A main repeater supplies power to about 15 secondary repeaters in each direction. The repeater spacing is dependent upon the size of the cable and the frequency of operation. For 2.6/9.5 mm cable at 12 GHz, the dependent repeater spacing is 4.5 km.

The design of a coaxial system involves the following:

1. Repeater design
2. Repeater spacing
3. Equalisation
4. Temperature considerations
5. Supervision and fault location
6. Power feed
7. Right of way.

Consider a coaxial cable system 100 km long using 9.5 mm cable and capable of transmitting up to 2700 voice frequency (VF) channels totalling to a bandwidth of about 12 MHz. At this frequency, the cable attenuation is about 8.0 dB/km as calculated from Eq.(9.24) using the data given in Table 9.7. For the 100-km length, the total attenuation works out to 800 dB. The first choice is to design a single amplifier of 800 dB gain and plant the same at the front end of the 100-km section. However, it is not practicable to design an amplifier with a gain of 10^{80}. Obviously, one or two repeaters would not do the job. A practical approach is to use 25–30 repeaters of 26–32 dB gain, each placed at an interval of 3–4 km. Each repeater contributes to thermal noise in the system and hence the number of repeaters is restricted by the upper limit for noise. The system can be designed with one main repeater which is capable of powering about 15 dependent repeaters on either side. For this purpose, a power feed unit at the repeater usually uses a supply voltage of 650 V d.c. between the inner and outer conductors and delivers a direct current of 110 mA. The crosstalk performance of the coaxial cables is very good. Both far end crosstalk and near end crosstalk attenuation is greater that 140 dB. Crosstalk poses no problem in coaxial systems unlike shielded wire pairs.

Equation (9.24) indicates that the cable attenuation is proportional to the square root of the frequency predominantly. As a result, voice channel signals, assigned to the high frequency segments of the line frequency, are heavily attenuated compared to the ones assigned to the low frequency carrier segments. Consequently, the voice channels at the high carrier frequency end are considerably affected by noise. Equalisation ensures that the signal-to-noise ratio in each voice channel is the same no matter where it is assigned in the line frequency spectrum. In other words, equalisers attempt to make the amplitude response of the cable uniform throughout the frequency range and reduce amplitude distortion. Amplitude equalisers may be broadly classified as fixed and variable equalisers. Three types of coaxial cable amplitude equalisers are commonly used:

1. Basic equalisers
2. Line build-out (LBO) networks
3. Design deviation equalisations.

Base equalisers are used when cable sections are of uniform length. The equalisation function is built into the repeaters. The gain of the repeater amplifier is made proportional to the square root of the frequency, thus compensating for the attenuation which is also proportional to the square root of the frequency. The LBO networks are used when the cable sections are not uniform. They are designed to cater to small variations in the nominal length of the cable. The third type of equalisers compensate for design deviations of nominal design parameters of standard dependent repeaters. A design deviation equaliser is installed, one for each 10, 15 or 20 repeaters to compensate for gross design deviations over that group of repeaters.

Variable or adjustable equalisers are designed to deal with gain frequency variations with time caused by temperature variations and component aging. Gain variations are around 0.2% per °C. Automatic gain regulations are achieved by sending a pilot tone at a fixed frequency periodically. Any variation in gain noticed for the pilot tone is exactly compensated. This results in automatic regulation for other frequencies as well although there may be small errors in compensation for these frequencies. Another method of regulating the gain for temperature variations is to use a thermistor buried in the ground near the repeater to monitor the temperature variations and adjust the gain accordingly. Phase vs. frequency characteristics of the coaxial cables are well defined and the amplitude equaliser may well be designed to perform delay equalisation as well. Usually, in practical systems, a common equaliser is used. Sometimes, delay equalisers are designed separately using tapped delay lines to ensure minimum intersymbol interference.

Supervision in coaxial cable system is concerned with monitoring the conditions of repeaters remotely at a manned location. Different pilot tones are transmitted to different repeaters which respond thereby giving the status.

Coaxial cable systems are manufactured for transmission rates ranging 34–565 Mbps presently. A practical design includes multiplexing/demultiplexing, line coding/decoding, scrambling/descrambling and pulse shaping units. A typical 140-Mbps systems multiplexes from 34-Mbps streams and uses 7/9 stage scrambler unit and 4B3T line code. With this line code, the symbol rate on the line works out to be 104 M bauds.

9.5 Numbering Plan

The objective of numbering plan is to uniquely identify every subscriber connected to a telecommunication network. In the early stages of development, a numbering scheme was confined to a single local exchange, and exchanges were identified by the names of the towns in which they were located. This scheme works well as long as there is only one exchange per town. But as the subscriber volume grew, it became necessary to introduce

more than one exchange in a town. Generally, a large centrally located exchange called the **main exchange** serving the main business centre of the town, and a number of smaller exchanges known as **satellite exchanges** serving different residential localities, were used to cope with the growing traffic in a large area. The area containing the complete network of the main exchange, and the satellites is known as **multiexchange area**. A common numbering scheme was then required for the area so that the digits dialled to identify a given terminating exchange do not vary with the exchange originating the call. The importance of this aspect has been discussed in Section 3.1. For calls originating from a location outside the multiexchange area, there is a need to identify the area by a common code. The common numbering scheme is sometimes called a **linked numbering scheme**. In this scheme, all exchanges in a town were collectively identified by the name of the town.

The introduction of **subscriber trunk dialling (STD)** or **direct distance dialling (DDD)** for intercity and intertown long distance connections called for a national numbering plan, where multiexchange areas are identified uniquely by numbers. Subsequent development of **international subscriber dialling (ISD)** makes it necessary to have an international numbering plan and to have the national numbering plan conform to the international one.

A numbering plan may be open, semi-open or closed. An **open numbering plan**, also known as **nonuniform numbering scheme**, permits wide variation in the number of digits to be used to identify a subscriber within a multiexchange area or within a country. This plan is used in countries equipped extensively with non-Director Strowger switching systems. In such cases, the numbering scheme is usually an exact image of the network structure and requires to be changed if the network structure changes (see Section 3.1). This scheme is now almost extinct. A **semi-open plan** permits number lengths to differ by atmost one or two digits. Today, this scheme is the most common and is used in many countries including India, Sweden, Switzerland, and the United Kingdom. In **closed numbering plan** or the **uniform numbering scheme**, the number of digits in a subscriber number is fixed. This scheme is used by a few countries which include France, Belgium and the countries in the North America (USA, Canada, Hawaii, etc).

An **international numbering plan** or **world numbering plan** has been defined by CCITT in its recommendations E.160–E.163. For numbering purposes, the world is divided into zones as shown in Fig. 9.26. Each zone is given a single digit code. For the European zone, two codes have been allotted because of the large number of countries within this zone. Every international telephone number consists of two parts as shown in Fig. 9.27(a). The country code contains one, two or three digits, the first digit being the zone code in which the country lies. For example, in zone 3, France has the country code '33' and Albania '355', and in zone 9, India has the country code '91' and Maldives '960'. In cases where an integrated

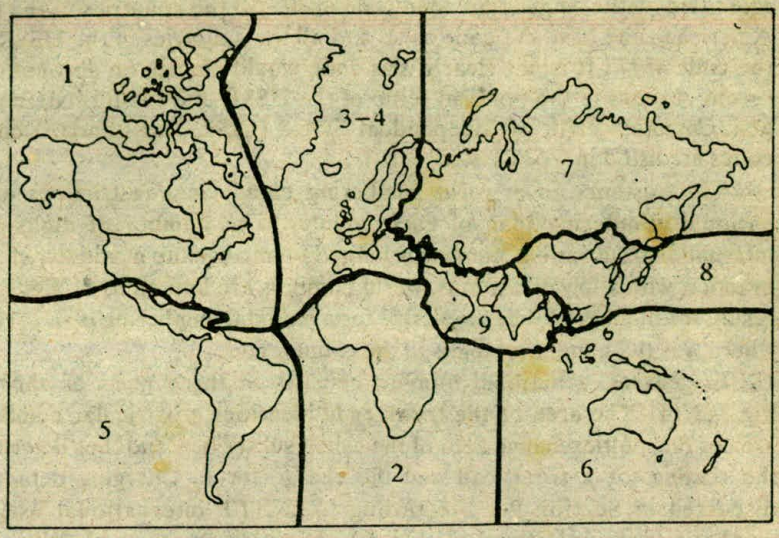
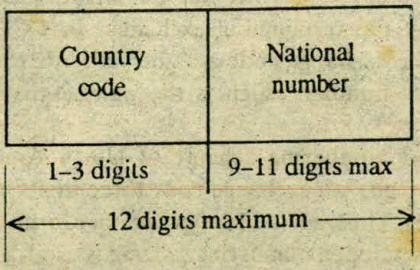
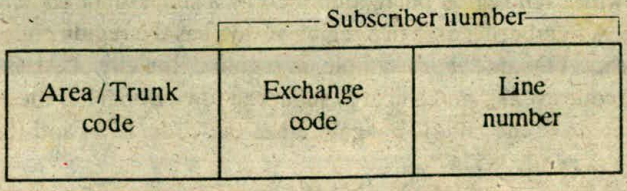


Fig. 9.26 World numbering zones.



(a) International telephone number



(b) National telephone number

Fig. 9.27 Telephone number structure.

numbering plan already covers an entire zone, the countries in that zone are identified by the single digit zone code itself. All the countries in the North American zone have the code as '1' and all the countries in the USSR have the code as '7'. It is not clear if this code would change on account of the recent changes in the political setup of the USSR and the formation of the new Commonwealth of Independent States (CIS). The standard country codes are listed in CCITT Rec. E.163.

The existence of a world numbering plan places restrictions on the national numbering plan of each country. The number of digits in an international subscriber number is limited to an absolute maximum of 12. In practice, with a few exceptions, world numbers are limited to 11 digits. As a result, the number of digits available for a national numbering plan is $11 - N$, where N is the number of digits in the country code.

In general, a national number consists of three parts as shown in Fig. 9.27(b). The **area** or **the trunk code** identifies a particular numbering area or the multiexchange area of the called subscriber, and thus determines the routing for a trunk call and the charge for it. Charging details are discussed in Section 9.6. According to CCITT international usage, a numbering area is defined as that area in which any two subscribers use identical dialling procedure to reach any other subscriber in the network. An **exchange code** identifies a particular exchange within a numbering area. It determines the routing for incoming trunk call from another numbering area or for a call originating from one exchange and destined to another in the same numbering area. **Subscriber line number** is used to select the called subscriber line at the terminating exchange. In CCITT terminology, the combination of the exchange code and the subscriber line number is known as the **subscriber number** which is the number listed in the telephone directory.

In the closed numbering scheme of North America, an area code consists of three digits, an exchange code three digits, and a subscriber line number four digits. Thus, a fixed 10-digit number forms the national number. In India where a semiopen numbering scheme is used, a national number has 7–9 digits. But the apportionment between the STD code and the subscriber number varies widely. STD codes have 2–6 digits and the subscriber numbers 3–7 digits. The exchange codes are 1–3 digits long and the subscriber line numbers 2–4 digits. Broadly, the country is divided into eight regions with each region being identified by a single-digit code as shown in Fig. 9.28. A numbering area in a region always has the region code as the first digit of the STD code. For example, in region 2, the city of Ahmedabad has the STD code as '272' and a nearby suburban town Bopal has the code '2707', and in region 8, the city of Bangalore has the code '812', and the hill town Tirumala the code '85747'.

In addition to dealing with the numbering structure, a national numbering plan must also deal with appropriate dialling procedures for



Fig. 9.28 Indian telecommunication regions.

local, national and international calls. The term 'local call' here implies a call within a numbering area and the term 'national call' a trunk call between two different numbering areas within the same country. Basically, there are four possible approaches to dialling procedures:

1. Use a single uniform procedure for all calls, viz. local, national and international calls.
2. Use two different procedures, one for international calls and the other for local and national calls.
3. Use three different procedures, one for international calls, second for national trunk calls, and the third for local calls.

4. Use four different procedures, three procedures same as given in 3 above and a fourth procedure for calls in the adjacent numbering areas.

Approach 1 demands that 11 or 12 digits be dialled uniformly for any number to be obtained. This places an unnecessary burden on the subscriber and the digit processing subsystems of the exchanges and hence this approach is not resorted to. In approach 2, a common procedure of dialling the full national number for both local and national trunk calls is adopted. This is justified only if the national trunk traffic represents a high proportion of the total traffic which generally is not the case. Hence, this approach is also not favoured.

Approach 3 is commonly adopted by most of the nations. Here, a subscriber directory number is dialled for calls within a numbering area, a national number for national trunk calls, and an international number for foreign calls. A need to distinguish one class of numbers from the other now arises. A standard technique that has been adopted for this purpose by most of the countries, is the use of a set of one or more **prefix digits**. Since the first digit of the country codes can be any of the digits 1–9, the prefix digit conveniently starts with a zero. Usually, a single digit '0' prefix is used to distinguish between a local call and a national call. A two-digit '00' or a three-digit, e.g. '010' prefix is used to differentiate between national and international calls. The first prefix digit '0' routes a call to a trunk exchange and the following prefix digit(s), if any, causes the call to be switched to the international gateway exchange. The dialling procedure calls for the required prefix to be dialled followed by the appropriate (national or international number). Approach 4 suggests special short dialling codes for adjacent numbering areas. This approach tends to reduce the number space available to users (see Example 9.7), hence is not extensively used.

Size and delineation of numbering areas require careful consideration. Large numbering areas necessitate long and expensive junctions from satellite exchanges to central trunk switching centres. A numbering plan must make a generous allowance for growth in the number of subscribers for up to 50 years ahead. The numbering area should not be so large that the number of subscribers it will ultimately serve is beyond its numbering capacity. A large number of subscribers within a numbering area implies long subscriber numbers and cumbersome dialling. A maximum size of seven digits for local numbers appears to be at the acceptable limit. If the numbering area is small, the number of such areas is large and so is the number of trunk switching centres. In such a case, the number of area codes available in the national numbering plan may run out too soon. If the numbering areas are small, nearby exchanges having considerable traffic between them may belong to different numbering areas, thus complicating the dialling procedure and entailing higher charges.

Any special provision in numbering scheme has the effect of reducing the number space available for identifying unique subscribers. This is well illustrated by an example.

EXAMPLE 9.7 Consider a 11-digit international number with 2-digit country code, 2-digit area code and a 7-digit subscriber number. Prefixes '0' and '00' are used for identifying national and international numbers respectively. A few special services like fire, ambulance and police are given short subscriber numbers in the range 100–199. The network is designed for remote maintenance and the maintenance exchanges are identified by the number range 900–999. Determine the number space available for subscribers and the fraction of the space lost.

Solution

Maximum possible number space = $10^2 \times 10^2 \times 10^7$

Due to prefixes, area codes and subscriber numbers cannot start with zero. Due to special services, the subscriber numbers cannot start with a 1. Due to maintenance, the number range starting with 9 is not available for subscribers. Hence, we have the total available number space as

$$10^2 \times 90 \times 7 \times 10^6 = 63 \times 10^9$$

$$\text{Fraction of space lost} = \frac{10^{11} - 63 \times 10^9}{10^{11}} = 0.37$$

Another important aspect of the numbering plan is the **direct inward dialling (DID)** facility for PABXs. With the advent of electronic switching, most of the PABXs provide direct dialling access to the public networks, i.e. **direct outward dialling (DOD)**. from extensions. However, the incoming calls are routed through the operator. Increasingly, customers now desire that it should be possible to dial a PABX extension directly from a public network without having to go through the operator. The DID facility is desirable particularly in the context of long distance dialling when a caller has to pay for the time that elapses before the operator establishes a connection to the desired extension. There are basically two approaches to providing DID facility.

1. Use a set of suffix digits to the national number to identify PABX extensions.
2. Allot a national number to each PABX extension.

Under the restriction of a maximum of 12 digits in the international number, approach 1 has the effect of increasing the length of the subscriber number and thereby reducing the digits available for exchange and area codes. Approach 2 reduces the number of subscribers who can be accommodated within the numbering scheme of the local exchange. Both

approaches reduce the effective number space available; approach 1 more severely than approach 2. A better approach would be to add suffix digits over and above the international, national or local number. This calls for extensive modifications in exchange systems. An elegant solution seems to lie in the use of DTMF signalling once the connection is established to the PABX system. A data in voice answer (DIVA) feature is required in the PABX for this purpose. Here, as soon as a PABX is dialled, the PABX enables a digit receiver and a recorded voice message prompts the dialler to dial additional extension digits. These digits are analysed by the PABX digits receiver and an appropriate connection established. The DIVA scheme has been discussed in Section 3.2.

Yet another facility demanded by the customers is likely to have a major impact on numbering schemes. In some countries, particularly in Northern America, common carriers are provided by private companies. Some are more efficient than others. As a result, users would like to select a long distance carrier of their choice. Three or more digits have to be added to the dialling plan for this purpose and extensive changes and additions are required to implement this capability in tens of thousands of switching systems. Once again, an elegant answer to this problem seems to lie in exploiting the DIVA facility.

9.6 Charging Plan

Providing a telecommunication service calls for investment in capital items as well as meeting operational expenses. The capital cost includes that of line plant, switching systems, buildings and land. Operating costs include staff salaries, maintenance costs, water and electricity charges and miscellaneous expenses. A telecommunication administration receives its income from its subscribers. A charging plan provides for recovering both the capital costs and the operating costs from subscribers. The cost of shared resources like the switching equipment is amortised among a large number of subscribers over a period of time. The cost of dedicated resources like the telephone instrument and the subscriber line must be recovered from individual customers. The operating costs must be worked out depending on the quantum of resources used in providing a service and the duration for which these resources are used. Taking all these factors into account, a charging plan for a telecommunication service levies three different charges on a subscriber:

1. An initial charge for providing a network connection
2. A rental or leasing charge
3. Charges for individual calls made.

A subscriber's share of the capital costs of the common resources is generally covered in the initial connection charge and the rental component.

The rental may be levied on a monthly, bimonthly, quarterly, half yearly or annual basis. Certain operating costs are incurred even if the network carries no traffic. These are covered by the rental. Charges for the individual calls include the operating costs in establishing and maintaining the calls, and a component for the capital resources used. In practice, exact apportionment of costs to different heads is rather difficult, if not impossible. There are also other factors like marketing policy and government regulation. For example, often government regulations demand that revenue from a trunk network be used to subsidise the cost of local networks. The technical progress in trunk transmission has resulted in significant cost reductions in trunk networks whereas the local network services still continue to be expensive. By feeding revenue from one service to another, the subscribers are given reasonable tariff structures for both local and long distance services.

Charging for individual calls is accounted for by using either a metering instrument connected to each subscriber line or a metering register assigned to each subscriber in the case of electronic exchanges. We use the term 'meter' in the ensuing discussions to denote the instrument or the register. The count in the meter represents the number of charging units. A bill is raised by assigning a rate to the charging unit. The count is incremented by sending a pulse to the meter. Charging methods for individual calls fall under two broad categories:

- Duration independent charging
- Duration dependent charging.

Local calls within a numbering area are usually charged on a duration independent basis. The charging meter is incremented once for every successful call, i.e. whenever the called party answers. In the past, some systems made a distinction between the calls within an exchange and the calls across exchanges within a numbering area. Depending upon the number of exchanges involved in setting up a call, more than one pulse is sent to the charging meter. The scheme of sending more than one pulse for a call is known as **multimetering**. In the multiexchange case the control system in the originating exchange sends as many pulses as the number of exchanges in the connection, to the calling subscriber meter. Today, it is more usual to apply one unit charge to all the calls within a numbering area irrespective of the number of exchanges involved.

In the olden days when STD and ISD facilities were not available, the trunk calls were established with the help of operators who were also responsible for the call charging. The subscriber meters are then useful only for local calls. To avoid the capital cost of providing meters and the operating costs of reading them at regular intervals and preparing the bills, some administrations have adopted a **flat rate tariff system**, where some fixed charges for an estimated average number of local calls are included in the rental. This scheme is advantageous to subscribers who make a large number of calls but unfair to sparing users. To reduce the disparity, business

subscribers are charged a higher flat rate compared to domestic subscribers. When flat rate charging is used, the subscribers naturally tend to make more calls. This necessitates local exchanges to be designed for a higher traffic level. Some administrations combine the features of flat rate and call rate charging. The rental covers a certain number of free calls per rental period and only calls above this number are charged for. India uses this scheme. This method is usually adopted from the marketing angle. It may be noted that this scheme offers no particular advantage in terms of reducing the capital or operating costs. With the introduction of STD and ISD, automatic charging demands that subscriber meters be installed. Moreover, with the advent of data transmission, local calls tend to be longer in duration. Due to these reasons, flat rate charging is likely to be discarded soon.

In the case of duration dependent charging, a periodic train of pulses from a common pulse generator operates the calling subscriber's meter at appropriate intervals. This method is called **periodic pulse metering**. In this case, the charge for a call is proportional to its duration.

As indicated in Fig. 8.1, the traffic carried by a telecommunication network varies throughout the day. The quantum of switching equipment and junction plant provided in the network is based on the estimated busy hour traffic. A large part of this hardware remains idle during offpeak hours. In order to restrict the peak demand and encourage offpeak demand, it is common to make the metering rate vary with the time of day. This is done by suitably changing the pulse repetition frequency under the control of a time-of-the-day clock. Table 9.8 shows the tariff variation during a 24-hour

Table 9.8 Tariff Variation in India

Period of day	Metre pulse repetition rate
08.00 hours – 19.00 hours	X
19.00 hours – 22.00 hours	$X/2$
22.00 hours – 06.00 hours	$X/4$
06.00 hours – 08.00 hours	$X/2$

period as in vogue in India at the time of writing. The meter pulse repetition rate is reduced to half and one-fourth of the normal rate during offpeak hours. In India, on Sundays and other national holidays, the one-half rate applies during 08.00 to 19.00 hours as well.

Trunk calls are almost invariably charged on a duration dependent basis. In addition, the charges also depend on the radial distance between the calling and the called stations. That is, the trunk call charging is based on the distance-time product. When a trunk call is established through an operator, certain minimum charges are levied to cover up the labour cost. A typical tariff is based on a minimum charge for up to three minutes together with a

charge of less than a third for every additional minute. When DDD or STD facility is used to establish a long distance call, charging is usually accounted for by pulsing the meter at an appropriate rate. Depending on the time of day and the distance involved between the stations, the meter pulsing frequency varies. Table 9.9 shows the pulse repetition rates adopted for different

Table 9.9 Distance vs. Metering Pulse Rate in India

Distance (km)	Metering pulse rate (pulses/min)
20 – 50	1.67
50 – 100	5.00
100 – 200	7.50
200 – 500	15.00
500 – 1000	20.00
> 1000	30.00

distances in India during the normal-rate hours, i.e. from 08.00 to 19.00 hours on working days including Saturdays. For other times, the rate is deduced as per Table 9.8.

In countries that employ a flat rate tariff for local calls, subscribers do not have meters and an automatic method known as **automatic toll ticketing** or **automatic message accounting** is used for trunk call charging. Here, a supervisory circuit records information regarding the call on a storage medium such as magnetic disk or punched tape, which then is processed by a computer and the bill prepared. When pulse metering is used, the subscriber receives a bulk bill giving a single total charge covering both trunk and local calls. The use of automatic toll ticketing enables an itemised bill to be prepared. Present day electronic exchanges are capable of providing a detailed account of both the local and long distance calls made by a subscriber.

When ISD is used, call charging is carried out by the same methods as for STD. However, pulse metering may encounter two difficulties. First, on the longest and most expensive international calls, the pulse rate is often more than one per second that a sluggish electromechanical meter may fail to operate properly. Second, most electromechanical meters have 4-digit capacity, and a heavy user may incur more than 9999 units of charge in the interval between meter readings, and thus fail to be charged correctly. It is necessary to equip such subscriber lines with 5-digit meters. For international connections, it is necessary to associate call metering with incoming and outgoing international circuits in addition to subscriber lines. This is to enable operating administrations to apportion charges for calls passing through more than one country and to collect charges from other administrations as required.

In countries where electronic exchanges are not yet prevalent, operator assistance is required to obtain special trunk services like person-to-person calls and reverse charging. These services are expensive. In cases where pulse metering is used for charging, a subscriber may organise such special services on his own at a minimal cost. Pulse metering enables short calls to be made costing only a single unit fee. By this means, a subscriber can ascertain the availability of the called person or request the called party to return the call instead of using operator assistance for such services and thereby save on expenses significantly.

Public telephone booths use **coin operated boxes** which are of two types: prepayment and post payment type. A user inserts coins before dialling in prepayment type and in the latter after the called party answers. In these cases, special signalling provisions are required between the exchange and the coin box. These include 'refund' signal in the case of prepayment coin boxes and 'open coin slot' signal for postpayment type coin boxes.

For subscribers having heavy point-to-point traffic, dedicated or leased lines may be more economical than STD lines. Such nonexchange lines are provided by the administrations on round-the-clock basis. Charging is based on the distances involved. As the user traffic increases, there occurs a break-even point beyond which the leased line becomes more economical.

EXAMPLE 9.8 A telephone administration provides leased lines at the rate of Rs. 600 per km for a minimum rental period of 3 months. A heavy point-to-point traffic user has his offices located 600 km apart and is confronted with the choice of using STD or leased lines. At what traffic volume per day, should he move over to leased line? Assume 20 working days per month and a rate of Re. 1 per unit recorded by the meter.

Solution

Cost of renting the leased line = 600×600 = Rs. 360,000

Using the data in Table 9.9, we have

Cost of STD calls per hour = 60×20 = Rs. 1200

Let the break-even point occur when the STD line is used for x hours in three months. Then we have

$1200x = 360{,}000$, or

x = 300 hours/3 months = 100 hours/month = 5 hours/day

If the subscriber uses the STD line for more than five hours a day, obtaining a leased line is economical.

9.7 Signalling Techniques

Signalling systems link the variety of switching systems, transmission systems and subscriber equipments in a telecommunication network to enable the

network to function as a whole. As mentioned in Section 1.3, three forms of signalling are involved in a telecommunication network:

1. Subscriber loop signalling
2. Intraexchange or register signalling
3. Interexchange or interregister signalling.

In a telephone network, subscriber loop signalling depends upon the type of telephone instrument used. The details of subscriber loop signalling for a rotary dial telephone have been discussed in Section 2.1 and the signalling details for a dual tone multifrequency telephone in Section 3.2. Multifrequency signalling has brought about new services like data-in-voice-answer, which fall in the class of user-to-user signalling facilities. Such facilities are more common in the context of data communication and integrated services. Signalling for data terminals is discussed in Chapter 10 and for the ISDN end equipments in Chapter 11.

The intraexchange signalling is internal to the switching system and is heavily dependent upon the type and design of a switching system. It varies from one model to another even with the same manufacturer. Some aspects of this signalling have been considered in Chapter 4 while discussing stored program control. This signalling does not involve signalling systems of the type required on the switching network and hence is not discussed any further. This section mainly deals with interexchange and network-wide signalling. When interexchange signalling takes place between exchanges with common control subsystems, it is called **interregister signalling.** The main purpose of interregister signalling is the exchange of address digits which pass from exchange to exchange on a link-by-link basis. Network-wide signalling also involves end-to-end signalling between the originating exchange and the terminating exchange. Such a form of signalling is called **line signalling.**

Signalling techniques fall under two major classes: **inchannel signalling** and **common channel signalling** (CCS). Different types of signalling techniques exist in each of these two classes. These are summarised in Fig. 9.29. Inchannel signalling, also known as **per trunk signalling,** uses the same channel which carries user voice or data to pass control signals related to that call or connection. In contrast, CCS does not use the speech or the data path for signalling. It uses a separate common channel for passing control signals for a group of trunks or information paths. Inchannel signalling has the merit that no additional transmission facilities are required for signalling whereas this is not the case with CCS.

Clearly, inchannel signalling is influenced by the various factors associated with the information path. This brings about different forms of inchannel signalling. When an unamplified audio line plant which permits an individual metallic signalling path per speech circuit is used, d.c. signalling can be applied. The d.c. signalling is relatively simple, reliable and cheap.

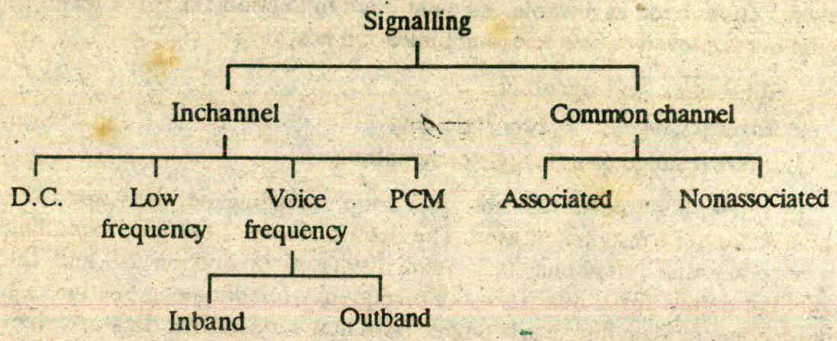


Fig. 9.29 Signalling techniques.

However, where amplified audio circuits are in use, low frequency a.c. signalling may be adopted. Low frequency a.c. signalling and d.c. signalling cannot be provided when FDM transmission systems are used. In such cases, voice frequency signalling is used. Obviously, a digital signalling scheme is desired for TDM systems.

Voice frequency (VF) signalling may be **inband** or **outband**. Inband VF signalling uses the same frequency band as the voice, i.e. 300–3400 Hz. Inband VF signalling must be protected against false operation by speech. An interesting story prevails to illustrate the significance of this statement. A woman subscriber is supposed to have complained to a leading telephone company in the United States that her line gets disconnected often during a conversation. The company being conscious of its reputation checked out thoroughly all the subsystems but found nothing abnormal and the problem continued to persist. The company then decided to analyse the lady's voice and found that the lady has a tendency to generate a tone at around 2600 Hz lasting for a duration of 100 ms or more which was being detected by the inband signalling system as the line disconnect signal. Such occurrences preclude the use of inband signalling during speech phase. An important advantage of inband signalling is that the control signals can be sent to every part where a speech signal can reach. It is also independent of the transmission systems as the signals would be carried along the route like the speech signals. Even A–D and D–A conversion processes do not affect them. Inband VF signalling is the most widely applied signalling system presently in the long distance telephone networks owing to its flexibility of operation.

Outband signalling uses frequencies above the voice band but below the upper limit of 4000 Hz of the nominal voice channel spacing. Signalling during the speech period is possible, thus allowing continuous supervision and control of a call. However, only a very narrow bandwidth is available for signalling. Extra electronics is required to handle the signalling band. As a result, this scheme is not used widely.

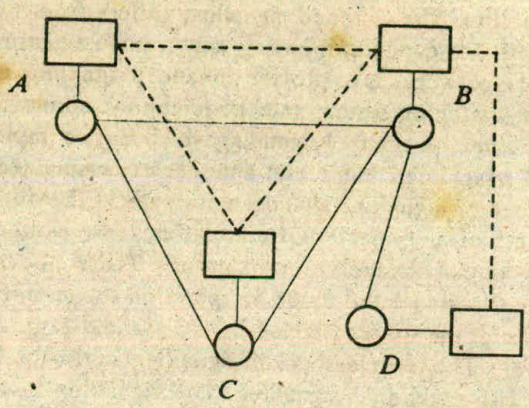
Both inband and outband signalling schemes have limited information transfer capacity. Preclusion of use of inband signalling during speech phase

limits its capability. The outband signalling suffers from the very limited bandwidth. Both are capable of having only a small signal repertoire. The trend in modern networks is to provide enhanced signalling facilities for the subscriber, the switching system and the telephone administration. Such a need is met by common channel signalling which may be implemented in two ways: **channel associated mode** and **channel nonassociated mode.** In the former, the common signalling channel closely tracks the trunk groups along the entire length of a connection. In the latter, there is no close or simple assignment of control channels to trunk groups. These modes are illustrated in Fig. 9.30. In the associated mode of operation shown in Fig. 9.30(a), the signalling paths for the speech paths *A-B, A-C-B* and *B-D* are *A-B, A-C-B* and *B-D* respectively. The term 'associated signalling' in the the CCS should not be confused with inchannel signalling. The signalling in CCS associated mode is still done on a separate signalling channel; only that the signalling path passes through the same set of switches as does the speech path. Network topologies of the signalling network and the speech network are the same. The advantages of the scheme are the economic implementation and simple assignment of trunk groups to signalling channels.

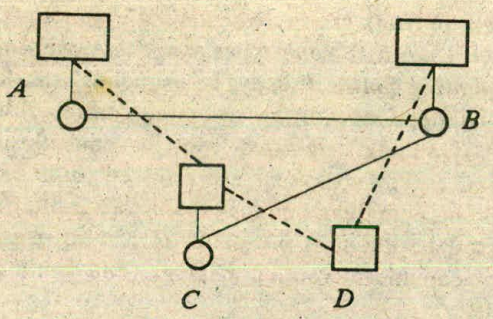
A CCS network consists of two types of nodes: **signalling transfer points (STPs)** and **signalling points (SPs)**. The signalling transfer points (STPs) usually have a connection with the switching centres although this is not essential. Since signalling originates from the control subsystems of the switching centres, these subsystems are referred to as SPs from the CCS signalling viewpoint. Signalling point is capable of handling control messages directly addressed to it but is incapable of routing messages. Signalling transfer point is capable of routing messages and could also perform the functions of a SP.

In the nonassociated mode of operation, the signalling information may follow a route that is different from the one taken by the speech path as shown in Fig. 9.30(b). The signalling paths for the speech paths *A-B* and *B-C* are *A-C-D-B* and *B-D-C* respectively. The network topologies for the signalling and the speech networks are different. There is no switching centre at the point *D* and only a STP is present. This approach offers flexibility as far as the CCS network is concerned. The signal messages may be transferred between the two end switching systems via any available path in the CCS network according to its own routing principles. This, however, demands a more comprehensive scheme for message addressing than is needed for channel associated signalling.

A CCS network may use associated signalling in some constituent parts and nonassociated signalling in other parts. The term 'quasiassociated signalling' is used to indicate such an operation. For example, if a signalling path exists between *B* and *C* in Fig. 9.30(b), the segment *B-C* may use associated signalling, whereas the signalling for the connection *A-B* is nonassociated. Quasiassociated signalling may also imply simplified signal



(a) Channel associated signalling



(b) Nonassociated signalling

—— speech or information path – – – – signalling path

switching system ☐ signalling transfer point (STP)

Fig. 9.30 Modes of operation of CCS.

routing over predetermined paths. The signalling paths are not 'associated' but are fixed for given speech connections.

From the network point of view, inchannel signalling and common channel signalling differ in many ways. These differences are summarised in Table 9.10. A few of the differences need elaboration. Consider a situation when the called subscriber is busy. In the case of inchannel signalling, the busy tone originates from the terminating exchange and the switched path is held up as long as the calling subscriber does not go onhook. On the other hand, in the case of CCS, the busy tone can be returned by the originating exchange, relieving all the intermediate trunks immediately. With regard to

Table 9.10 Inchannel vs. Common Channel Signalling

Inchannel	Common channel
Trunks are held up during signalling.	Trunks are not required for signalling.
Signal repertoire is limited.	Extensive signal repertoire is possible.
Interference between voice and control signals may occur.	No interference as the two channels are physically separate.
Separate signalling equipment is required for each trunk and hence is expensive.	Only one set of signalling equipments is required for a whole group of trunk circuits and therefore CCS is economical.
The voice channel being the control channel, there is a possibility of potential misuse by the customers.	Control channel is in general inaccessible to users.
Signalling is relatively slow.	Signalling is significantly fast.
Speech circuit reliability is assured.	There is no automatic test of the speech circuit.
It is difficult to change or add signals.	There is flexibility to change or add signals.
It is difficult to handle signalling during speech period.	There is freedom to handle signals during speech.
Reliability of the signalling path is not critical.	Reliability of the signalling path is critical

signal repertoire, the common channel can be configured with the bandwidth required to carry control signals for a rich variety of functions. Consider the sequence of events for call set up with inchannel signalling when more than one switch is involved. A control signal is sent from one switch to the next in the intended speech path. At each switch, the signal can be transferred to the next only after the associated circuit is established through the switch. With CCS, forwarding of control information can progress as soon as the speech path availability is marked. The actual path set up can progress independently and perhaps in parallel, in different segments of the complete path. Since the signalling information is not actually sent over speech paths in CCS, the integrity of the speech path is not assured. It is possible to mark a faulty speech link as available and establish a path using the same. To avoid this, certain checks such as routine testing of idle paths and the continuity test of an established path become necessary in CCS. Many speech circuits are served by one CCS link and an error occurring in the link, or failure of

the link may affect an appreciable number of speech circuits. Therefore, error control and redundant paths become mandatory in CCS systems. In general, an error rate of 1 in 10^8 to 1 in 10^{10} is desired. Automatic diversion of signalling traffic to an alternative circuit is necessary if the error rates become excessive or a link fails totally. In contrast, only one speech circuit is affected if an error occurs in the inchannel signalling system.

Most of the present day telephone networks are equipped with inchannel signalling systems. There is, however, little doubt that networks will gradually evolve to common channel signalling. At the international level, signalling standards are specified by CCITT. So far, seven international standards known as **signalling system number 1-7 (SS1-SS7)** have been adopted. There is also a variation of SS5 known as **SS5** ***bis***. In addition, two regional standards **R1** and **R2** pertaining to North America and Europe respectively have also been adopted by CCITT. The signalling systems SS1-SS5 deal with inchannel signalling. The systems SS6 and SS7 are CCS systems. In Section 9.8, we discuss some of the important features of inchannel signalling systems, and in Section 9.9, we deal with CCS systems.

9.8 Inchannel Signalling

The range of CCITT specified inchannel signalling systems reflects the evolution of international signalling requirements to meet the continually changing conditions of the international network. The CCITT inchannel signalling systems and the signalling techniques used in each of them along with the envisaged applications are presented in Table 9.11. The early systems, SS1, SS2 and SS3 are of historical interest only. At present, interest

Table 9.11 CCITT Inchannel Signalling Systems

Signalling system	Signalling technique and application
SS1	500/20 Hz signalling for manual international service
SS2	600/750 Hz two voice frequency system for semi-automatic service
SS3	2280 Hz single voice frequency system with limited application in semiautomatic and automatic services
SS4	2040 and 2400 Hz two VF compound end- to-end analog signalling
SS5	2400 and 2600 Hz two VF compound analog line signalling and 2/6 multifrequency inband analog interregister signalling with TASI equipment
SS5 *bis*	Same as SS5 but for circuits with or without TASI equipment

in the international inchannel signalling is confined to SS4, SS5 and SS5 *bis*, and in the regional systems to R1 and R2. The other signalling system of interest is the PCM signalling.

The international signalling systems SS4, SS5 and SS5 *bis* adopt inband signalling using a combination of two voice band frequencies, or a single voice frequency. In addition, systems SS5 and SS5 *bis* use multifrequency (MF) signalling for interregister signalling. In SS4, there is no separate interregister signalling. In Section 3.2, we have discussed MF signalling in detail. Here, we present some details of SS4 and SS5 systems. As discussed in Section 3.2, a compound signal of two voice frequencies is less likely to be imitated by speech than a single frequency of equal duration and thus provides a better talk-off performance.

In system SS4, the 2-VF compound signal is usually used as a preparatory or prefix signal element to the control element which comprises single frequency pulse(s) of either of the two signalling frequencies. The transmission of single frequencies may be short or long. The transmitted durations and recognition times at the receiver end for the signalling elements in SS4 are given Table 9.12.

Table 9.12 Timings for SS4 Signalling Elements

Element	Transmitted duration (ms)	Recognition times (ms)
Compound	150 ± 30	80 ± 20
Single-short	100 ± 20	40 ± 10
Single-long	350 ± 70	200 ± 40

Some sample control signals and their associated codes in SS4 are shown in Table 9.13. The digits of the dialled number are transmitted as binary codes of four elements. One of the two frequencies, 2040 Hz, represents a binary '1' and the other, 2400 Hz, represents a binary '0'. The pulse durations are 35 ± 7 ms with a gap of similar time between the pulses.

Table 9.13 Some Control Signals in SS4

Control signal	Code
Terminal seizure	PX_s
Transit seizure	PY_s
Clear forward	PX_l
Forward transfer	PY_l

P = prefix element
X_l = 2040 Hz long
Y_l = 2400 Hz long
X_s = 2040 Hz short
Y_s = 2400 Hz short

In SS5, the line signalling comprises either a compound of the two VF frequencies or a continuous single frequency. Interregister signalling uses 2-out-of-6 MF code. Initially, the system was jointly developed by UK Post Office and the Bell Laboratories for dialling over time assigned speech interpolation (TASI) equipped transAtlantic cables. This was the first application of intercontinental dialling and of TASI equipment. The system was subsequently specified by the CCITT as a standard in 1964 and has since found increasing applications in other parts of the world. Most of the Atlantic, Pacific and Indian Ocean circuits use SS5 at present. In view of the high cost of transocean cables and the consequent need to utilise the cables as efficiently as possible, TASI was considered essential for these cables. The design features of SS5 were dictated by TASI requirements.

The concept of TASI has been briefly explained in Section 8.5.4. We elaborate this now a little more. The normal speech activity of a subscriber on call is about 35 per cent. As a result, full duplex 4-wire speech transmission circuits are less than half utilised. The TASI technique attempts to improve trunk utilisation by assigning a circuit to a speech channel only when there is speech activity. In this way, a given number of circuits can support more than double the number of speech channels.

In TASI, each channel is equipped with a speech detector which, on detecting speech, arranges for a circuit to be assigned to that channel. Since this process of speech detection and establishment of trunk-channel association takes definite time, the speech burst is clipped for that duration. Typical clip duration is about 15 ms when a channel is available. It increases under busy traffic conditions when a free channel may not be available immediately. In order to reduce the extent of interpolation, a circuit is not disassociated from the channel for short gaps of speech. For this purpose, the speech detectors are provided with a short hangover time and a circuit is disconnected only when the speech gap is longer than the detector hangover time. The digital counterpart of TASI is known as **digital speech interpolation (DSI)**.

As with speech bursts, inchannel signalling information also experiences clipping in a TASI environment. This calls for special consideration in designing signalling systems for TASI environment. Unless signals are of sufficient duration to permit trunk-channel association and reliable recognition at the receiving end, there is the likelihood of the signal being lost partially or fully. With pulse signalling, it has been determined that a 500-ms duration is required to account for the extreme trunk-channel association condition. Allowing for reliable recognition, a pulse of 850 ± 200 ms duration is considered suitable. But, pulse signals of such lengths would slow down the signalling process considerably. The pulse gaps would result in the channel being disassociated, thus leading to unnecessary TASI activity. Moreover, fixed length pulses cannot take advantage of lightly loaded conditions when the channel assignment time is low. For these

reasons, the line signalling in SS5 and SS5 *bis* is chosen to be continuously compelled to ensure that the trunk-channel association is maintained throughout the period of signalling. However, continuous compelled signalling precludes the efficient use of trunks under heavy load conditions as TASI is effectively disabled for the trunk carrying the signalling information.

Interregister signalling carrying address information would be far too slow if continuous compelled signalling is adopted. Pulse signalling is preferred here and two different techniques are used to maintain trunk-channel association during the signalling period:

1. The address information is transmitted *en bloc* after gathering all the address digits, and the gaps between the pulses are ensured to be less than the speech detector hangover time.
2. Address digits are transmitted as and when they arrive and a lock tone is transmitted during the gaps.

The SS5 adopts *en bloc* transmission scheme whereas SS5 *bis* uses the lock tone method. The *en bloc* method facilitates checking of the validity of the address by digit count and avoids expensive intercontinental circuits being ineffectively taken during incomplete dialling. However, this method increases the post dialling delay because the digits are accumulated before the signalling begins. The lock tone method permits overlapped operation of digits being received and transmitted. It reduces the postdialling delay. The trunk, however, is not used as efficiently as in *en bloc* transmission.

A standard method of transferring signalling information between the switching equipment and the signalling equipment is required for all signalling systems. One such standard is provided by ear and mouth (E and M) control. This method was developed and adopted by Bell System for the toll networks in Northern America. E and M leads constitute a standard interface for delivering and accepting uniform signal conditions. There are three types of E and M interfaces: types I, II and III. Type I interface has two leads, one for each direction of transmission. This interface was primarily developed for electromechanical exchanges. The M-lead carries d.c. signals from the switching equipment to the signalling equipment to operate the outgoing part of the signalling terminal as illustrated in Fig. 9.31. The E-lead carries d.c. signals from the incoming part of the signalling terminal to the switching equipment. Thus, signals from exchange *A* are carried on M line and received on E lines at exchange *B* and similarly from exchange *B* to *A*. Type II E and M interface is a 4-wire fully looped interface and is the preferred one for electronic switching systems. Type III is a compromise between Type I and II and uses a 3-wire partially looped arrangement.

The CCITT R2 signalling system combines an outband line signalling system and an interregister MF signalling system. Conceptually, outband

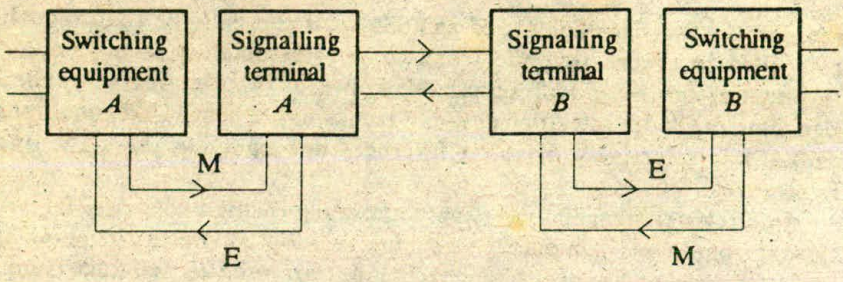


Fig. 9.31 E and M signalling control.

signalling includes d.c., low frequency a.c., inslot PCM, and signalling above the effective speech band. But in practice, the term 'outband' is applied to systems based on a.c. signalling using frequencies above the upper frequency limit of the effective speech band, i.e. above 3400 Hz and within the 4-kHz speech channel spacing. Outband signalling is generally performed only in the FDM transmission environment.

In FDM systems incorporating outband signalling, the channel bandwidth is divided into a speech subchannel and a signalling subchannel using suitable filters. Usually, a single frequency signalling is used. A 4-wire circuit is used for forward and backward signalling paths and for full duplex operation. The frequency is chosen to lie approximately midway between two adjacent channels and the CCITT recommends 3825 Hz.

E and M signalling control can be applied to outband signalling. A typical outband configuration using E and M control is shown in Fig. 9.32. Since d.c. is used for signalling, a.c. to d.c. and d.c. to a.c. conversions are required for incoming E lead and outgoing M lead respectively. A low pass filter (LPF) is placed in the speech path before the modulator (M) at which the signal tone is injected into the main transmission path. Similarly, a LPF is placed after the demodulator (D) and the signal tone extraction point at the receiving end. The LPF at the transmitting end removes any energy in the signalling band which may be present in the incoming speech or produced by the preceding stages of the equipment. This is to ensure that there is no interference with the signalling at this stage. The LPF at the receiving end prevents signal tone from being audible during speech.

As the signalling frequency must not be extended to the switching equipment, outband signalling is done on link-by-link basis and end-to-end signalling is precluded. Since the signalling is independent of speech, there is virtually complete freedom in the choice of signalling mode. Signalling may be two-state (on/off) continuous tone or pulse signalling. The former is simple to implement but its potential to support a large signal repertoire is limited. Moreover, the signal level must be relatively low to avoid overloading of transmission systems. A higher level is permissible with pulse signalling and a larger signal repertoire can be supported. Despite the

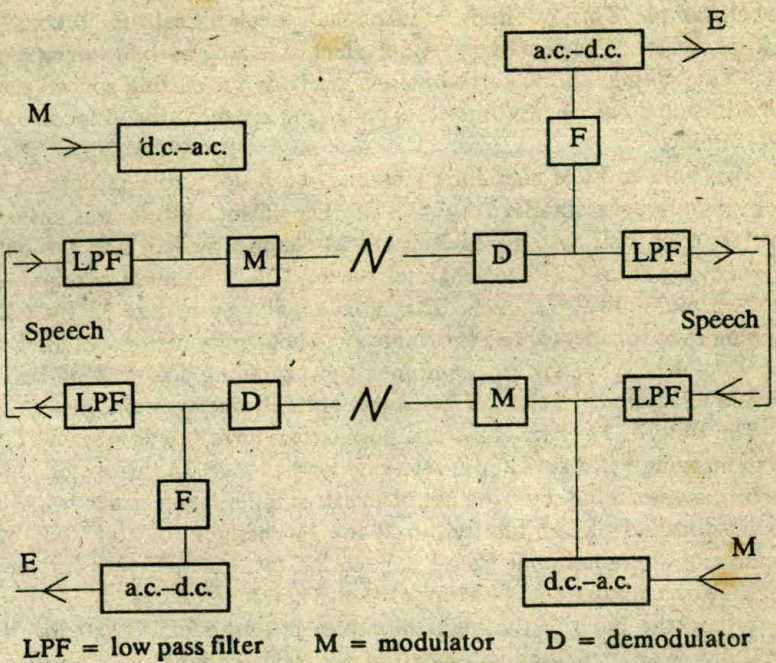


Fig. 9.32 Outband signalling scheme with E and M control.

flexibility and simplicity, outband signalling is not widely prevalent mainly due to the large segment of existing carrier, coaxial and FDM systems which do not have facilities for supporting outband signalling.

Now a days, PCM systems are widely used, and signalling in these systems is of considerable interest. In PCM systems, signalling can be achieved by sampling and coding signalling information in the same way as for speech. At present, this is the case for some signalling systems. For example, the SS4 and SS5 and SS5 *bis* are to be used in this manner in PCM systems. The inherent nature of PCM, however, allows a convenient way of transmitting signalling information as a built-in signal. In this method, advantage can be taken of the fact that signalling information need not be sampled as frequently as speech signals.

With built-in PCM signalling, in addition to the bits required to transmit speech and ensure frame synchronisation, bits are required to carry signalling information which is usually binary coded at the transmitting end and reconverted to original form at the receiving end. The signalling information pertaining to a particular speech channel may be carried in the same time slot as the speech or in a separate time slot. The former is known as **inslot signalling** and the latter as **outslot signalling**. Since the time slot size is fixed at eight bits, inslot signalling implies reduced bandwidth for speech transmission. Usually one bit is used for signalling, leaving seven bits for

speech sample. This, in effect, reduces the speech bit rate to 56 kbps from 64 kbps. Inslot PCM signalling is equivalent to analog outband signalling. In outslot signalling, all the eight bits are available for coding speech sample and additional one or two time slots per frame are introduced for signalling and framing.

Two built-in PCM signalling systems, one inslot and one outslot, have been specified as standard by CCITT. The inslot system was originally developed by Bell Systems as Bell D2 24-channel system and the outslot system by CEPT as CEPT 30-channel system. The 24-channel system uses all the 24 time slots for speech. The 30-channel system has 32 time slots, 30 being used for speech, one for frame alignment and one for signalling. The Bell D2 24-channel system combines the signalling features of Bell D1 24-channel system and the UK 24-channel multiframe system.

The Bell D1 24-channel system, popularly referred to as system T1, was the pioneering PCM system put into extensive service. In this system, frame synchronisation is based on the use of an extra bit per frame. Each frame uses 193 bits instead of 192 bits required for 24 channels of eight bits each. Because of the additional frame delineation bit, the line rate for the T1 system works out to be 1.544 Mbps (193 × 8 kHz) instead of 1.536 Mbps. In T1 system, the bit 8 of the eight bits in each time slot is used for inslot signalling. The signalling information for each channel is identified by observing the sequence of signalling bit in consecutive frames.

The UK post office 24-channel system is organised on a 4-frame multiframe basis with bit 1 of the eight bits in each time slot being used for frame and multiframe synchronisation and signalling. The 4-frame structure is shown in Fig. 9.33. Bit 1 of each 8-bit time slot in frames 1 and 3 contains the signalling information for the respective speech channels. Bit 1 of every time slot in frame 2 is unused. First bit positions of the first 16 time slots in frame 4 are used for frame synchronisation with the last eight first bit positions being unused. The synchronising pattern is 1101010101010101. Since two bits per multiframe are used for signalling per speech channel, the signalling repertoire is larger in this case than T1 system. The signalling rate, however, is 2000 signals per second instead of 8000 per second as in the case of T1 system.

The Bell D2 24-channel system retains the additional 193 bit as in D1 system for frame synchronisation, adopts a multiframe structure, uses two bits per frame for signalling and improves the speech transmission performance. There are 12 frames in the D2 multiframe. Frames 6 and 12 are designated as signalling frames. The time slots in all other frames, i.e. 1–5 and 7–11, carry speech using all the eight bits. The 8-bit speech encoding reduces the quantisation noise and thus secures better speech transmission performance, particularly over multilink connections involving several conversions. In frames 6 and 12, only seven bits per time slot are used for speech transmission, leaving the eighth bit for signalling. The frame

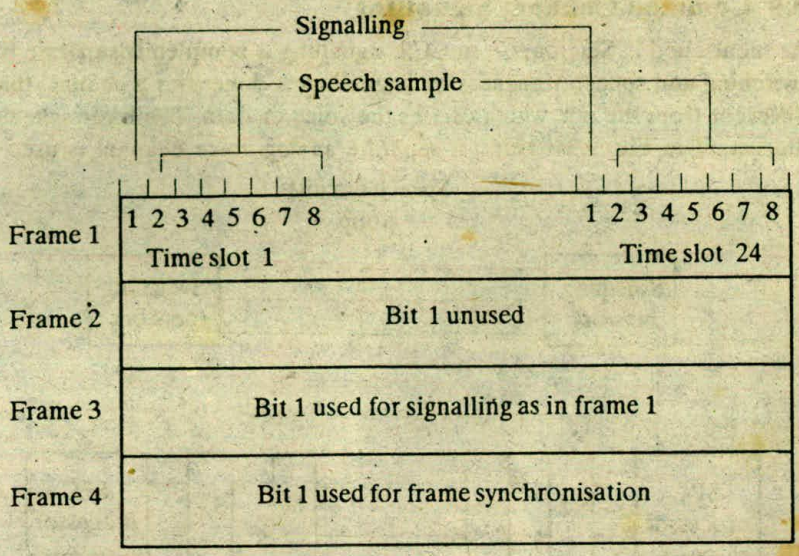


Fig. 9.33 Bell D2 24-channel multiframe PCM signalling structure.

alignment bits, i.e. 193rd bit in each frame, in the odd numbered frames are used for frame synchronisation and the ones in the even numbered frames for multiframe synchronisation. The signalling rate is 650 signals per second with two bits per signal.

The bit stealing in frames 6 and 12 in the Bell D2 system denies the full gain of 8-bit speech encoding. The outslot 30-channel PCM signalling system realises the full potential of 8-bit encoding as all the speech samples are encoded using eight bits. Of the 32 time slots per frame in this system, time slot 0 is used for frame synchronisation and the time slot 16 for signalling. Obviously, one time slot with eight bits cannot support signalling for 30 speech channels in every frame. Hence, a multiframe structure with 16 frames is adopted for signalling purposes. With this structure, time slot 16 in each frame except frame 0, carries signalling information for two speech channels. Thus, 15 time slots in frames 1–15 carry signalling information for 30 channels in every multiframe. With 15 time slots for 30 channels, four bits are available for signalling information per channel. The signalling rate works out to be 500 signals per second with each signal represented by four bits. The first four bits of time slot 16 in frame 0 are used for multiframe alignment with the bit pattern 0000, with the remaining four bits unused. To enable easy recognition, this all-zero pattern is prohibited as a combination of signalling bits in time slot 16 of other frames. This limits the number of signalling patterns to 15 per channel. Sometimes, this CEPT signalling scheme is referred to as **bunched signalling.**

9.9 Common Channel Signalling

As mentioned in Section 9.7, in CCS, signalling is completely separate from switching and speech transmission. Signalling is done over a channel that is different from the one which carries the voice or data. The basic scheme is illustrated in Fig. 9.34. Here, a separate analog voice channel is used for

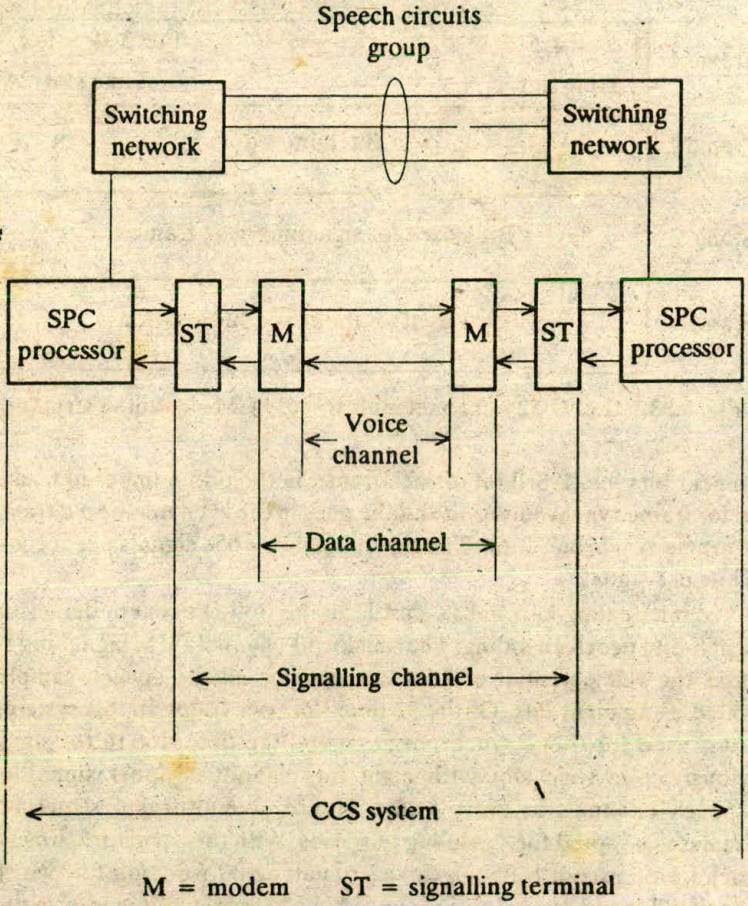


Fig. 9.34 Basic scheme for CCS.

signalling. Since signalling in CCS is digital in nature as explained in Section 9.7, modems (modulator/demodulator) are used for carrying digital data over analog lines. Transmission of digital data on analog telephone channels using modems is fully described in Section 10.1. In Fig. 9.34, two signalling channels, one for each direction, are used in a dedicated manner to carry signalling information. Since the channels are dedicated for signalling they are capable of carrying signalling information for a group of circuits. The

group size is determined by the capacity of a signalling channel. A phase-equalised voice channel is capable of supporting a bit rate of 2.4 kbps with acceptable error rates for signalling. At this bit rate, one CCS link can carry signals for 1500–2000 speech circuits.

The CCS network is basically a store and forward (S&F) network where signalling information travels on a link-by-link basis along the route. When the signalling information is received at a node, it is stored, processed and forwarded to the next node in the route. Store and forward networks are discussed in detail in Section 10.2. In CCS, signalling information is transferred as messages of varying length usually defined as one or more fixed length signalling units (SUs). A message of one signal unit length is called single unit message (SUM) and the one with multiple signal units as multiunit message (MUM). A signal unit is divided into a number of constituent bit fields each having its own function. Typical structures of a SUM and a MUM are shown in Fig. 9.35. The format of the first SU of a

Header	Signalling information	Circuit label	Error check

(a) Single unit message

Header	Signalling information	Circuit label	Error check
Sub header	Length	Other signalling information	Error check
Sub header	Length	Address digits	Error check

(b) Three-unit message

Fig. 9.35 Typical CCS signalling message formats.

MUM is identical to a SUM. In SS6, each SU is of size 28 bits of which the last eight bits are check bits. In SUM, 5, 4 and 11 bits are allocated for header, signalling information and circuit label respectively. In MUM the SUs subsequent to the first SU, have 2, 2 and 16 bits for subheader, length field and additional signalling information respectively. Routing details, address digits etc. form additional signalling information. Each address digit is four bits long. A signalling message unit may be an address unit, centralised service message unit, acknowledgment unit, synchronisation or idle unit, or

management message unit. In SS6, the SUs are grouped into blocks of 12 each, the last SU being an acknowledgment unit. An acknowledgment unit carries the number of the block being transmitted, the number of the block being acknowledged and information as to whether or not each of the 11 SUs of the block being acknowledged was received without error. Blocks are numbered sequentially but the individual SUs in the block are identified by their position in the block. The first 11 SUs of a block consist of message or synchronisation SUs. The latter is transmitted in the absence of other signalling traffic to facilitate block and SU synchronisation. It carries a number which indicates its position within the block.

The concept of common channel signalling has been further expanded in SS7 which was first defined in 1980, with revisions in 1984 and 1988. SS7 has been designed to be an open ended CCS standard that can be used over a variety of digital circuit switched networks. While the network being controlled is circuit switched, the control signalling itself uses packet switching. The functions in SS7 are defined assuming packet switched operation, but the actual implementation can be in circuit switched nodes as additional functions. This, in fact, is the case when channel associated signalling mode is chosen for implementation.

Specific attention has been given to the requirements of ISDN while designing SS7. The internal control and network intelligence essential to an ISDN are provided by SS7. Although SS7 is suitable for operation over analog channels and at speeds less than 64 kbps, it is primarily optimised to work with digital SPC exchanges utilising 64-kbps digital channels. SS7 is suitable for operation over both terrestrial and satellite links.

The protocol architecture of SS7 has four levels as shown in Fig. 9.36. The lower three levels referred to as the **message transfer part (MTP)** provide a reliable service for routing messages through the SS7 network. The lowest layer, signalling data link, is concerned with the physical and electrical characteristics of the signalling links between STPs and between STPs and SPs (see Section 9.7). All signalling data links in SS7 are full duplex links dedicated to SS7 traffic.

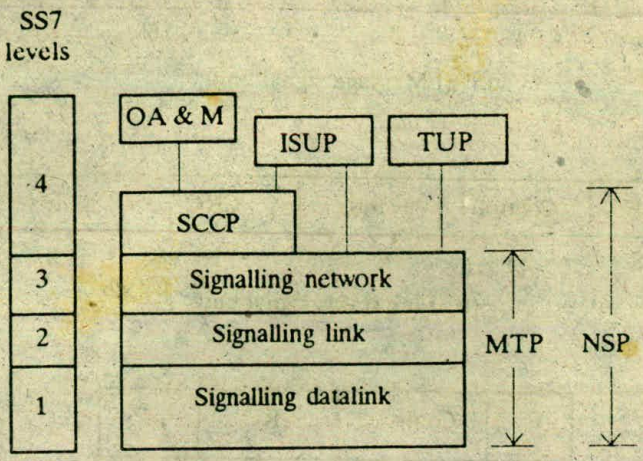
The main purpose of the second layer, viz. the signalling link, is to turn a potentially unreliable physical link into a reliable data link. The signalling link layer must ensure that

1. there are no losses or duplication of control message,
2. messages are delivered in the same order in which they originate, and
3. there is a match between the receiver capacity and the transmission rates. (This function implies that the receiver is capable of exercising flow control over the sender.)

There are three types of signalling units (SUs) defined in SS7:

1. Message signal unit (MSU)

2. Link status signal unit (LSSU)
3. Fill-in signal unit (FISU)



MTP = message transfer part NSP = network service part
TUP = telephone user part ISUP = ISDN user part
SCCP = signalling connection control part
OA&M = operations, administration and maintenance

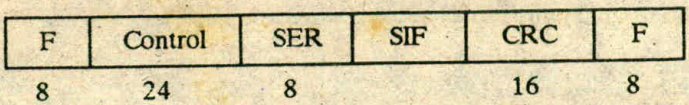
Fig. 9.36 Architecture of SS7.

The formats of these SUs are shown in Fig. 9.37. All the SUs begin and end with a flag field which has the unique bit pattern 01111110. The flags act as delimiters for the SUs. A common flag may be used as the closing flag for one SU and the opening flag for the next if the SUs are transmitted in continuum.

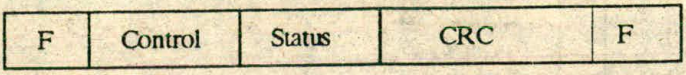
It is possible that the unique flag bit pattern appears inside the SU, thus destroying synchronisation. To avoid this problem, a technique known as **bit stuffing** is used. This technique was mentioned in Section 5.7 and is discussed in detail in Section 10.4.2 in the context of data networks. In brief, in this technique, the transmitter inserts an extra 0 whenever it comes across five consecutive 1's in the data part of the SU. The receiver on detecting five 1's deletes the zero following it. Thus only the flag pattern can contain six 1's.

All the SUs in SS7 contain a 16-bit error checking field. Cyclic redundancy code (CRC) is used for error checking and the technique is described in Section 10.4.2.

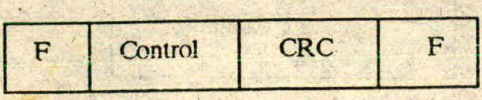
The control field consists of five subfields as shown in Fig. 9.37(d). The backward sequence number (BSN) and the backward indicator (BI) bit together permit piggybacked acknowledgment of the SUs received from the other side. A negative acknowledgment is indicated by inverting the BI bit which remains unchanged for all subsequent positive acknowledgments. The



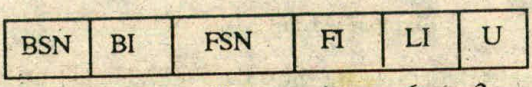
(a) Message signal unit



(b) Link status signal unit



(c) Fill-in signal unit



(d) Control subfields

F = flag CRC = cyclic redundancy code SIF = signalling information
SER = service information field BSN = backward sequence number
BI = backward indicator FSN = forward sequence number
FI = forward indicator LI = length indicator U = unused

Fig. 9.37 Formats of signalling units.

forward sequence number (FSN) identifies the SU uniquely using modulo 128 count. A retransmission is indicated by inverting the forward indicator (FI) bit which remains unchanged until another retransmission occurs. The length field specifies the length of the following information fields in octets. This provides a crosscheck on closing flag. It also serves as a SU type indicator. A value of 0 indicates a FISU, a value of 1 implies a LSSU and a value of 3 to 63 specifies a MSU.

The service information octet specifies the type of message and whether the message relates to a national or international network. This signalling information field (SIF) contains level 3 and 4 information. This consists of an address label and user data from the telephone user part (TUP), ISDN user part (ISUP) or operations, administration and maintenance (OA&M) part.

The address label field is 32 bits long with a 14-bit signalling link selection subfield. The link selection subfield is used to select a path from among alternative routes.

The level 3 signalling network functions relate to message handling and net. ork management. Message handling involves discrimination, routing and distribution of messages. The discrimination function analyses the destination code in the address label to decide whether a message is to be routed to another node or distributed to one of the user parts in the local node. The particular user part to which a message is to be distributed at the destination node is decided based on an analysis of the type of message information in the SER field. The discrimination function is needed only in STPs. Routing function may be invoked by the discrimination function or by a local level 4 entity. The routing decision is based on the value in the signalling link selection subfield. With a 4-bit link selection subfield, a total of 16 different routes may be defined through the network. In general, all of the control signals associated with a single call will follow the same route. This guarantees that the control messages arrive in sequence at the destination.

The main objective of network management function is to monitor signalling links and overcome link failures or degradations. The goal is to achieve an unavailability figure of not more than 10 minutes per year per route. This is attempted through redundancy of links and dynamic rerouting. The emphasis on the network management in SS7 stems from the fact that the functioning of the common signalling network is critical and a failure in the signalling network may affect all the subscribers. Since failure recovery and restoration actions may involve multiple networks across national boundaries, it is essential to lay down failure and congestion recovery procedures.

The **signalling connection control part** (SCCP) residing in level 4 was added to SS7 specifications in 1984. The SCCP and MTP together are referred to as the **network service part** (NSP). The main purpose of SCCP is to enhance the limited routing, distribution and addressing capabilities of the third layer, to meet the needs of those user parts requiring enriched information transfer facilities. The other modules in level 4 are invoked as per the requirements of a service. The TUP is invoked in response to actions by a subscriber at a telephone. The control signalling associated with TUP deal with establishment, maintenance, and termination of telephone calls. Similarly, ISUP deals with ISDN subscriber calls and related functions. Finally, the OA&M module deals with messages relating to network management, operations and maintenance.

The SS7 architectural principles are similar to the ones in ISO open system interconnection (OSI) architecture proposed for data networks. Many of the functions in each of these layers are common in these two architectures. In fact, SS7 architecture came into existence much after the ISO-OSI model and is largely based on the work done for ISO-OSI model.

The ISO-OSI reference model is discussed in detail in Chapter 10. Readers are referred to Sections 10.2–10.4 in order to get a better appreciation of SS7 architecture.

9.10 Cellular Mobile Telephony

Cellular mobile radio telephony is a significant step forward in the provision of communication services. The aim of cellular systems is to provide personal telephone services to subscribers on the move. The largest demand for this service at present is the telephone in the car used by the business executives, senior government officials and field personnel like doctors and sales representatives. This service could be extended to any mobile subscriber provided the communication terminal can be carried in a brief case or in the pocket. Another cellular type radio telephone system currently under development, based on the principles of cordless telephones, is likely to spread cellular communications on a larger scale.

Cellular systems have evolved from military radio design. But their practicability for civilian applications has become possible only due to the recent technological developments. In mobile telephony, it is necessary to change the operating frequency as one moves from one transmitter area to another. In traditional military mobile radio, change of operating frequency is tedious involving manual retuning of circuits and readjustment of antennas. In modern cellular systems, operating frequencies can be changed many times a second under the control of microprocessors and pseudo random number generators.

From the subscriber point of view, a mobile telephone must function much the same way as the normal line-fed telephone. It should be possible to dial calls not only to other mobiles in the same city but also to ordinary lines or mobiles anywhere in the world using the usual long distance and international dialling procedures. Each mobile set must have its own individual number so that it can be addressed from any other telephone in the network.

In cellular communications, a region is geographically divided into cells as shown diagramatically in Fig. 9.38. In theory, the cells are hexagons, but in practice, they are less regular in shape. Each cell has a base station. Mobile terminals in a cell communicate via the base station of that cell. All calls are routed through the base station which functions like a local exchange for the mobile telephones. Each base station is equipped with a low power transmitter whose power output is such that satisfactory reception is achieved throughout the cell area but no interference takes place with other cells having the same frequency of operations. Clearly, neighbouring cells cannot operate at the same frequency and the cells with the same frequency of operation must be separated by a minimum distance in which the signals

are attenuated to an insignificant level. Figure 9.38 shows four frequency allocations with one cell separation before the frequencies are reused. For a

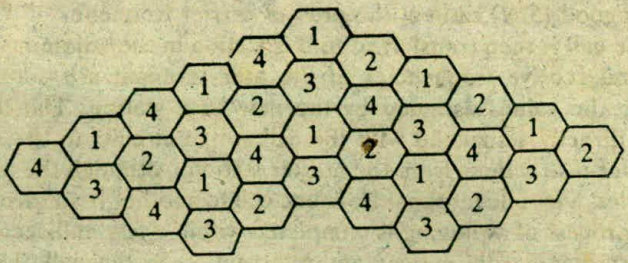


Fig. 9.38 Cell structure for mobile communications.

greater separation, a larger number of carrier frequencies is required. Assigning frequencies to cells is a form of graph colouring problem: how can one colour the cells such that the adjacent cells do not have the same colours? The problem is made more complex by irregular cell shapes and sizes.

The most common form of modulation used in cellular communications is the narrow band frequency modulation. In this scheme, carrier frequencies are separated by about 25 kHz. The total radio bandwidth allocated for cellular mobile communications is sufficient to support about 280 channels. For a seven-cell cluster, this gives 40 channels per cell. In this case, the maximum busy hour traffic to achieve a grade of service not more than 0.02, works out to be 15 erlangs using Erlang B formula. For typical holding times and calling probabilities, a busy hour traffic of 15 erlangs represents the average traffic generated by about 1500 subscribers. Thus, the cell sizes have to be chosen based on the limit of subscribers per cell area. The cell sizes need not be uniform throughout a cellular range. For example, the cell sizes may be small in the crowded central part of a city and become larger and larger towards the peripheral part.

When a subscriber initiates a call at the mobile terminal, the instrument sends a message to its base station announcing its telephone number. The base station then allocates an available frequency to the telephone for communication. When the call is finished, the frequency is deallocated and made available for another caller in the same cell.

The process of changing over the carrier frequency when a subscriber moves from one cell area to another is somewhat complicated. In addition to the base station in each cell, there is a master station in the system which keeps track of the carrier frequency allocations by the different cells and manages the overall functioning of the system. All the base stations and the central station are processor controlled. As a mobile terminal moves around a cell, the signal-to-noise ratio of its received signal at the base station of the cell varies. If the (S/N) ratio becomes worse than a predetermined level, the

microprocessor at the base station informs the master processor of this, which checks with the adjoining cell stations to find out which one of them is receiving a good (*S/N*) ratio at the present carrier frequency of the mobile station. The call is then transferred to this station in the adjoining cell. The transmit and receive frequencies of the user terminal are automatically changed to the values assigned by the new base station. The telephone exchanges directly associated with the call switch circuits to the new base station under instructions from the master station. Although the process of 'handing over' a mobile terminal from one cell to another is an involved one, the entire process of switching is completed within a few milliseconds such that a subscriber hardly notices any interruption in the call. There can, however, be difficulties if data is sent over cellular links as even a brief break in the link can affect data transfer.

Cellular communication service is more expensive than regular telephone service. The cellular mobile units are more complex in design than the regular telephone instruments as they need to have automatically tunable transmitters and receivers covering the entire frequency bandwidth. The design of the base stations and the master station is also complex.

A car cellular unit has three pieces of equipment: the transmission unit, the control unit and the antenna. The transmission unit is usually placed in the trunk of the car, the control unit is mounted on the dash board and the antenna is either roof-mounted or wind-shield mounted. The control unit resembles an ordinary telephone in design. Researchers are still looking for technological advancements to reduce the risk attempting to dial while driving a vehicle. Voice-activated dialling appears to be promising.

The development currently underway in the field of cordless phones is likely to further promote the cellular communications. The second generation cordless phones, called CT2, use digital technologies and are capable of working with very many CT2 stations which are to be established in railway stations and other public places. It is expected that by mid 1990s many CT2 stations will be in operation, variously called **phone zone, phone point, telepoint** etc. In working, CT2 phones send out their own identification code to the nearest phone point before each call to enable the exchange to bill the owner for the call. In order to reduce the complexity associated with the cellular systems, two restrictions are placed on CT2 systems. Firstly, no incoming calls are permitted to CT2 phones, so that the network does not have to keep track where a particular subscriber is at any point of time. Secondly, once a call is established, a subscriber is not permitted to move from one phone zone to another so that there is no need for the intercell handing over arrangements which complicate the cellular communications. Although research and development in cellular systems have been driven by the urgent need to improve mobile communications, the use of digital radio links on a very large scale with ordinary static telephone systems, replacing local distribution cable networks, may well turn out to be the most significant outcome of this research in the near future.

FURTHER READING

1. Bylanski, P. and Ingram, D. G. W., *Digital Transmission Systems*, Peter Peregrinus Ltd., Herts (UK), 1979.
2. Combes, P. E., *Microwave Transmission for Telecommunications*, Translated by Duffin, W. J., John Wiley & Sons, Chichester (UK).
3. Das, J., *Review of Digital Communications*, Wiley Eastern., New Delhi, 1988.
4. Flood, J. E. (ed.), *Telecommunication Networks*, Peter Peregrinus Ltd., Herts (UK), 1977.
5. Freeman, R. L., *Telecommunication System Engineering*, 2nd ed., Wiley-Interscience, New York, 1989.
6. Freeman, R. L., *Telecommunication Transmission Handbook*, 2nd ed., Wiley-Interscience, New York, 1981.
7. Freeman, R. L., *Reference Manual for Telecommunication Engineering*, Wiley-Interscience, New York, 1985.
8. Welch, S., *Signalling in Telecommunications Networks*, Peter Peregrinus Ltd., Herts (UK), 1979.

EXERCISES

1. A telephone company is required to give connections to a few subscribers located at a distance of 15 km from its nearest exchange. Discuss the different design options available for this purpose.
2. A subscriber loop of 18 km is to be supported from an exchange that uses a 40 V battery with a 400 Ω short-circuit protection resistance. Electronic telephones are used as the subscriber instruments. Determine the wire gauge that needs to be used.
3. Determine the system gain and fade margin of a microwave repeater given that

Bandwidth	= 5 MHz
Antenna gains	= 30 dB each
Path length	= 50 km
Noise figure for the receiver	= 7 dB
Branching and coupling losses	= 5 dB

4. An ionospheric communication uses a carrier frequency of 10 MHz. If the height of the antenna is 15 m, determine the take off angle.
5. A 3.2-m parabolic antenna of a satellite earth station has an effective area of 55% of the physical area. The frequency of operation is 4 GHz. If this antenna is to be suitable for a standard A earth station of INTELSAT, what is the equivalent noise temperature for this antenna?
6. A satellite is placed in a circular orbit at an altitude of 1600 km. Determine its orbital period. If an earth station antenna is able to track

the satellite through an angular space of 75°, what is the time duration for which the satellite is visible in a day?

7. An Indian company proposes to lease the required bandwidth from INSAT-1D satellite to establish a voice communication network among 100 remote users. About 10 per cent of the users are expected to be active at any point of time. Voice is transmitted using standard ADPCM. Determine the bandwidth required in terms of fraction of a transponder capacity per single-hop and two-hop operations if demand assigned TDMA scheme is used for access.

8. A spread spectrum system operates at the same carrier frequency and power as an AM system. It, however, uses a bandwidth of 1 MHz instead of 10 kHz used by AM system. Determine the improvement in the noise immunity in the case of spread spectrum operation.

9. A spread spectrum system transmits PCM voice using a chip rate of 1024 kbps. What is the bandwidth spreading factor?

10. A coaxial cable has a capacitance of 66 pF/m. The commercial characteristic impedance of the line is 75 Ω. What is the inductance per metre of the line? What is the cable velocity factor (V_p/speed of light) for high frequency range operation?

11. If the coaxial cable in Exercise 10 has a series resistance of 0.05 Ω/m and a leakage resistance of 10^{10} Ω, determine the frequency above which high frequency approximations apply (see Section 5.7).

12. Two central offices are 100 km apart. They are to be connected by a coaxial cable which has a loss of 22 dB/km. There is a coupling and leakage loss of 4 dB at the central office. The area has wide temperature variation and to compensate for this a factor of 1.1 is to be provided as far as losses are concerned. From the noise point of view, loss in any section should not exceed 25 dB. Determine the
    (a) distance between the central office and the first repeater,
    (b) distance between two repeaters, and
    (c) number of repeaters required.

13. What is the maximum number of subscribers that can be given telephone connections in Maldives? Compare this with the population of the country.

14. How many telephone subscribers can be there in Kolar Gold Fields (K.G.F) in India? If the town is to be a multiexchange area with a maximum of six exchanges, what is the maximum line capacity from each exchange?

15. A telephone company proposes to introduce a simplified charging scheme where charging is directly proportional to the time a user

remains off-hook. Discuss the merits and demerits of such a charging scheme.

16. In a TASI system, determine the probability of clipping if 50 speech channels are supported on 25 speech circuits. Assume an activity factor of 0.4 per user and a Poisson arrival.

17. A telephone conversation is routed through four switching systems. Compare the times required to establish the call in the case of inchannel signalling and CCS systems given that

    Circuit switching time between exchanges = 2 s
    Signalling time between exchanges = 0.2 s
    Dialling time = 8 s

    Circuit links can be set up in parallel in the case of CCS.

18. A PCM system uses a multiframe structure with 18 frames. One time slot is dedicated for signalling and synchronisation in each frame comprising 32 time slots. Four bits are dedicated for frame synchronisation in each frame and four bits are used for multiframe synchronisation in frames 1 and 10. Determine the signalling rate and the size of the signalling repertoire in this system if every multiframe is to carry at least one signal for each voice channel.

19. Draw a cell structure that has a frequency allocation such that there is a separation of atleast two cells between cells that reuse the same frequency. Ensure that the total number of frequencies used is minimum.

10

Data Networks

In the last few decades, computer and communication technologies have been coming together to support many new applications and developments. One of the early examples of such a development is data networks. Data networks have their beginning in remote computing which came about for two reasons in late 1950s and early 1960s. On the one hand, the computer users found it increasingly inconvenient to go to a central and possibly far away computer centre to use the computer, and desired computer access in their own premises. On the other hand, computer manufacturers and managers were keen on increasing the clientele for their computers for economic reasons and found an opportunity for this in remote computing. A suitable communication network is required to realise remote computing, which would carry data from the computer to the remote unit and vice versa.

In the beginning, data transmission was organised using telegraph or telex networks as they could carry digital signals directly. There were, however, two limitations. First, the speeds were limited to 50 bauds in the case of switched circuits or to 110 bauds in the case of leased lines. Secondly, these networks did not provide a wide coverage of the population. The possibility of carrying signals at higher speeds in telephone networks and their wide coverage led to the serious consideration of public switched telephone network (PSTN) as a candidate system for data transmission. As remote computing was maturing, the idea of sharing data, information and other resources among computers emerged. The focus shifted from terminal-to-computer communication to computer-to-computer communication. Early efforts were confined to interconnecting homogeneous computer systems from the same vendor. Soon, the advantages of interconnecting computers from different vendors were realised and the progress made in this direction has led to evolution of modern **public data networks** (PDNs).

Data networks are classified according to their geographical coverage:

- Wide area networks (WANs)
- Metropolitan area networks (MANs)
- Local area networks (LANs).

Intercity, intercountry and intercontinental networks are known as WANs. Based on the communication infrastructure used, they may be classified as **terrestrial data networks (TDNs)** or **satellite based data networks (SBDNs)**. In TDNs, data communication is organised using cables, fibre optic lines or radio links. A geostationary or a geosynchronous satellite is used for communication in SBDNs.

A metropolitan area network interconnects computers within a metropolitan city. Community antenna television (CATV) cables, twisted pair wires or shielded lines, optical fibres, radio links or line-of-sight (LOS) optical communication links provide the communication medium for MAN. The broadband capability of CATV cables permits carrying voice, data and video simultaneously. Thus MANs are usually multimedia networks.

Local area networks are confined to a single building or a group of buildings generally belonging to the same organisation. Optical fibres, twisted pair and coaxial cables are used as the communication media for LANs. Fibre optic networks (FONs) are suitable for both LANs and MANs. Synchronous optical networks (SONET) are designed to operate at high speeds.

LANs, MANs and WANs are generally interconnected in a hierarchical manner to form a global network as shown in Fig. 10.1. LANs are often

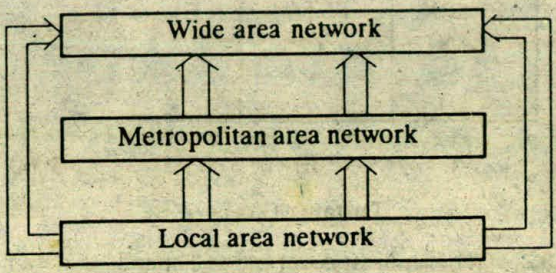


Fig. 10.1 Data network hierarchy.

connected directly to WANs, particularly in places where MANs are not installed or have not developed well. Apart from the different geographical coverages, the range of data rates supported on these networks also differs widely. Figure 10.2 summarises the typical data rates and geographical coverages for these networks.

10.1 Data Transmission in PSTNs

Public switched telephone networks and electronic PABXs are designed to carry analog voice signals. They can, however, be used for data transmission by employing suitable interfaces. As seen from Fig. 10.2, LANs can be designed around PABXs, and MANs around PSTNs. In these cases, the data rates are usually limited to a maximum of 64 kbps. Terrestrial data networks

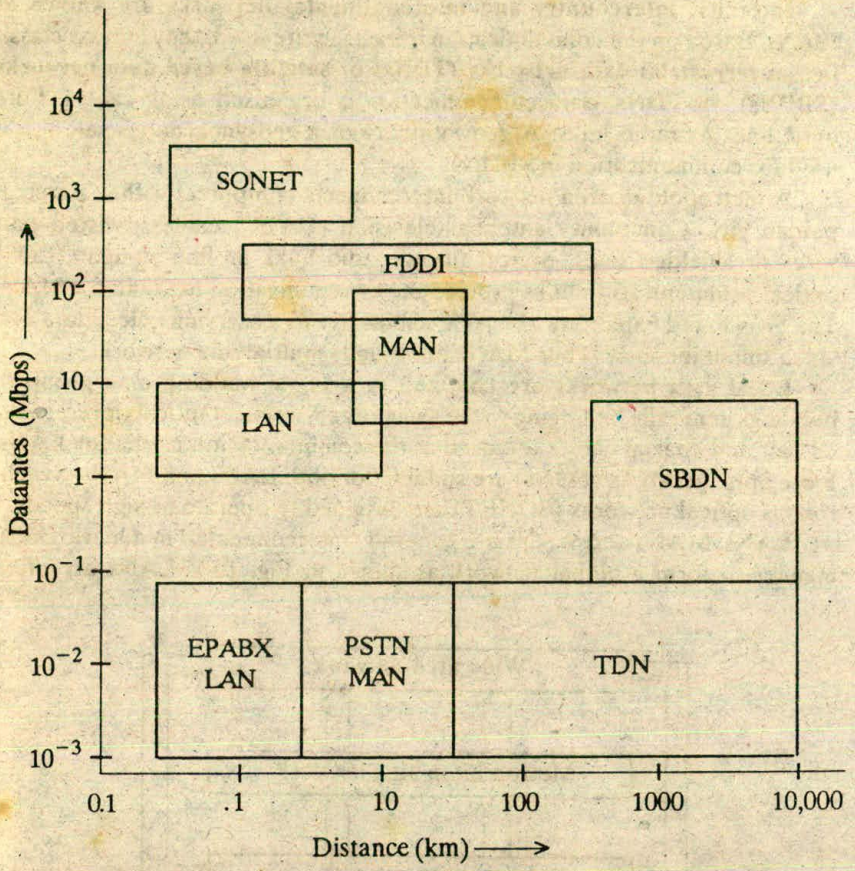


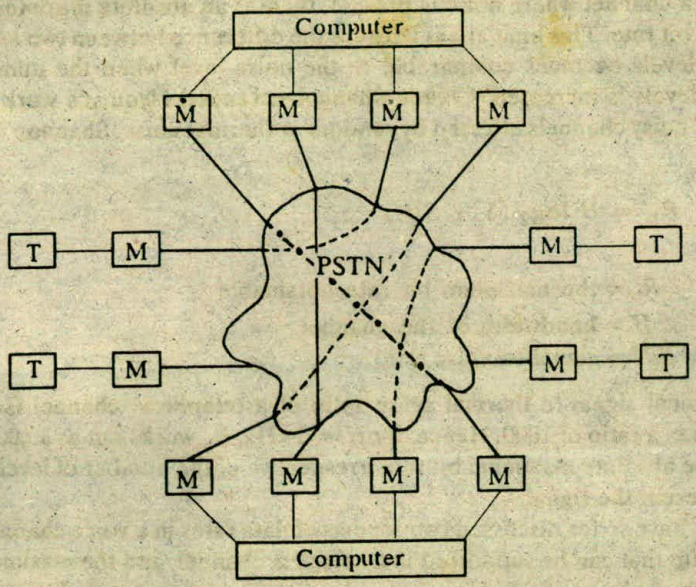
Fig. 10.2 Geographical coverage and speeds of data networks.

and the upcoming integrated services digital networks can, however, support data rates of 1.544 or 2.048 Mbps.

Transmission of digital data signals over PSTN networks demands that the digital signals be converted to analog form at the transmitting end and vice versa at the receiving end. A modulator translates the data pulses into voice band analog signals at the transmitting end. At the receiving end, the analog signals are demodulated to recover the digital information. A combined modulator/demodulator unit is called a **modem**.

Initially, modems were used to connect terminals, located in remote places, to a central computer. Later, computer-to-computer communication was established using modems and PSTNs. A data communication scheme using modems and a PSTN is shown in Fig. 10.3. The digital interface of a modem is connected to the computer and the analog interface to the telephone network. In addition to data exchange, the digital interface permits control signals to be exchanged between the modem and the computer.

In a PSTN, a connection may be established using a dial-up facility or a dedicated nonswitched leased line. A 2-wire or 4-wire leased line may be used. The modems usually have automatic dialling and answer facility which is useful when working with dial-up line.Modems are driven by a communication software package that runs on the terminal or the personnel computer of the user. The transmission may be half-duplex or full-duplex. In Fig.10.3, both terminal-to-computer and computer-to-computer connections are shown.



M = modem T = terminal —— 4-wire leased line
– – dial-up line –·– 2-wire leased line

Fig. 10.3 Data communication using PSTN.

10.1.1 Data Rates in PSTNs

A voice channel in a PSTN is band limited with a nominal bandwidth of 3.1 kHz. What is the maximum data rate that a voice channel can support with its limited bandwidth? A first-cut estimate of this can be obtained from Nyquist's theorem which applies to noiseless channels and states:

$$R = 2H \log_2 V \text{ bps} \qquad (10.1)$$

where

R = maximum data rate
H = bandwidth of the channel
V = number of discrete levels in the signal

For a 3-kHz channel, and a binary signal, the maximum data rate works out to be 6000 bps, if the channel is ideal. In a practical channel, the maximum rate would come down. By increasing the number of levels used to represent the signal, the bit rate may be increased arbitrarily in a noiseless channel. It is important to recognise that the actual number of signal transitions is still limited to the binary level limit; the effective bit rate goes up with more than two signal levels as each signal level can now represent a group of two or more bits. The maximum rate of signal transitions that can be supported by a channel is known as **baud rate** or **symbol rate.**

In a channel where noise is present, there is an absolute maximum limit for the bit rate. This limit arises because the difference between two adjacent signal levels becomes comparable to the noise level when the number of signal levels is increased. Claude Shannon extended Nyquist's work to the case of noisy channels affected by random or thermal noise. Shannon's result states:

$$R_b = H \log_2 (1 + S/N) \tag{10.2}$$

where

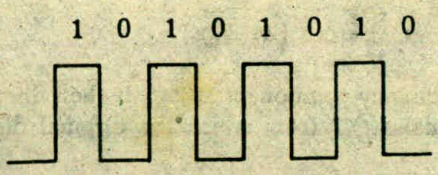
R_b = the maximum bit rate obtainable
H = bandwidth of the channel
S/N = signal-to-noise ratio

Typical signal-to-thermal noise ratio of a telephone channel is about 30 dB, i.e. a ratio of 1000. Hence, for H = 3 kHz, R_b works out as 30,000 bps. R_b is the absolute maximum bit rate irrespective of the number of levels used to represent the signal.

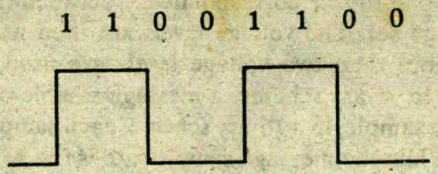
We have so far discussed two aspects of data rates in a voice channel: the baud rate that can be supported in a noiseless channel, and the maximum bit rate that can be supported in a noisy channel. The baud rate and the bit rate are related as

$$R_b = R \times n$$

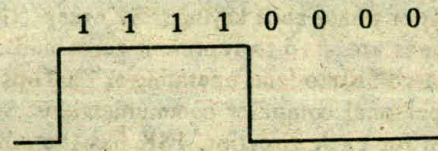
where n is the number of bits required to represent signal levels. For binary signals, $n = 1$ and $R_b = R$. The bandwidth required of the channel is $R/2$ Hz as given by Eq. (10.1). The next thing that we need to estimate is the baud rate that can be supported in a practical environment. Here, there are two opposing factors: on the one hand, the channel is noisy, which goes to reduce the maximum baud rate; on the other hand, the data is rarely an alternating pattern of ones and zeros so that the effective transition rates (baud rate) in the data are much lower than the bit rate. This is illustrated in Fig. 10.4. In Fig. 10.4(a) the baud rate is equal to the bit rate, and in Figs. 10.4(b) and (c), it is one-half and one-fourth respectively. Taking such factors into account and by performing experiments, it has been observed that baud rates up to 2400 bauds may be transmitted reliably through a PSTN voice channel. By



(a) $R = R_b$



(b) $R = R_b/2$



(c) $R = R_b/4$

Fig. 10.4 Baud rates and bit rates.

using additional signal levels, say 16, the effective bit rates may go up to 9.6 kbps. The techniques of using additional signal levels to increase the bit rate are the ones used in the design of modems.

10.1.2 Modems

Amplitude, frequency and phase modulation are all used in the design of modems. In amplitude modulation, zeros and ones are represented by two different voltage levels. A signal waveform $s(t)$, called baseband signal, is generated from the digital data. This is then multiplied by a sinusoidal carrier, say $\cos(2\pi f_0 t)$, to generate a modulated signal

$$m(t) = s(t) \cos(2\pi f_0 t) \tag{10.3}$$

At the receiver end, the modulated signal is again multiplied by $\cos(2\pi f_0 t)$, yielding a received signal

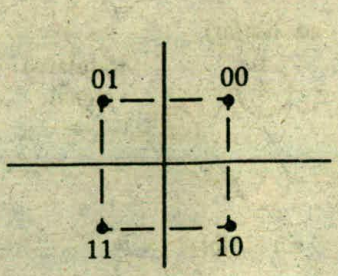
$$r(t) = s(t)\cos^2(2\pi f_0 t)$$
$$= \frac{s(t)}{2} + \frac{s(t)}{2}\cos\left(2\pi(2f_0)t\right) \quad (10.4)$$

The high frequency component at $2\pi f_0$ is then filtered out, leaving the demodulated signal $s(t)/2$ from which the original digital signal may be attained.

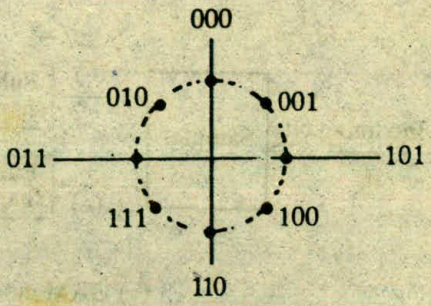
As mentioned in Section 10.1.1, it is not necessary that only two voltage levels are used when generating $s(t)$. A new digital message may be chosen at every symbol interval, which is mapped to one of a set of discrete voltage levels. There is a one-to-one correspondence between the discrete message set and the voltage set. The voltage levels are then used to modulate the carrier. When more than two voltage levels are used, the modulation is known as M-ary. In M-ary scheme, a message sample contains k bits, such that $2^k = M$. For example, in a 16-ary scheme, each sample contains four bits of information. While the message sample rate is the baud rate, the bit rate is four times the baud rate in this example. This technique of varying the amplitude of sinusoidal carrier using the voltage levels of the baseband signal is known as **amplitude shift keying (ASK).**

If the carrier is sinusoidal and the baseband voltage level is used to vary the frequency or phase of the carrier, the modulation is known as **frequency shift keying (FSK)** or **phase shift keying (PSK)** respectively. In FSK, two or more different tones are used to represent zeros and ones, or a group of message bits. Binary FSK modems operating at 1200 bps on dial-up lines are widely used for personal computer communications. Separate half-duplex channels are used for each direction. FSK has the advantage of constant amplitude signal, which makes it robust against the nonlinearities of power limiting devices in the channel.

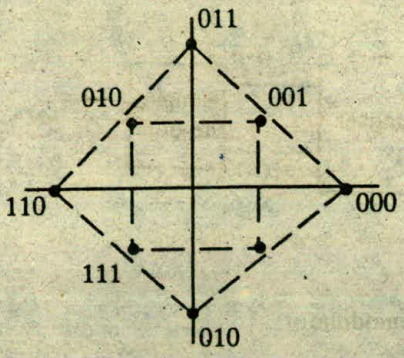
In the most common form of PSK, the carrier wave is symmetrically shifted by 45°, 135°, 225°, or 315°. Four different phase shifts permit two bits to be transmitted per shift. The scheme, known as quadrature phase shift keying (QPSK), is as shown in Fig. 10.5(a). Figure 10.5(b) shows an 8-ary PSK, where each shift transmits three bits. A variation of binary PSK is known as differential phase shift keying (DPSK). In an M-ary scheme, when the value of M is increased, the minimum distance between signal points decreases, thus making the scheme susceptible to noise, phase jitter, frequency jitter and other impairments. A combination scheme of two or more modulation techniques would result in an M-ary scheme that is more robust. One of the widely used schemes is **quadrature amplitude modulation (QAM)**, where phase and amplitude modulation techniques are combined. In QAM, incoming bits are mapped into two baseband signals $s_1(t)$ and $s_2(t)$. Then $s_1(t)$ is multiplied by $\cos(2\pi f_0 t)$ and $s_2(t)$ by $\sin(2\pi f_0 t)$. The sum of these products forms the transmitted QAM signals as shown in Fig. 10.6. At the receiver, the incoming waveform is separately multiplied by $\cos(2\pi f_0 t)$ and $\sin(2\pi f_0 t)$. When the multiplied waveforms are filtered, we get $s_1(t)/2$



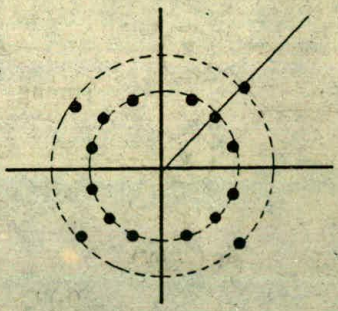
(a) 4-ary PSK: 2 bits/baud



(b) 8-ary PSK: 3 bits/baud



(c) 8-ary QAM: 3 bits/baud

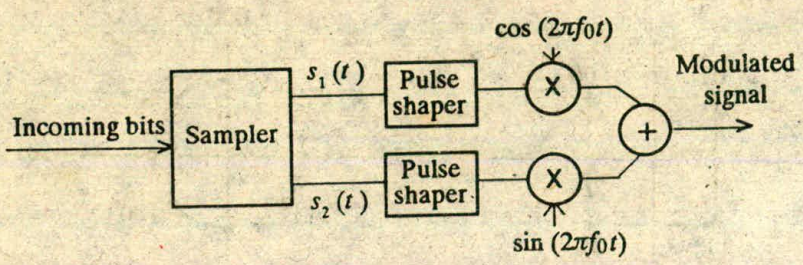


(d) 16-ary QAM: 4 bits/baud

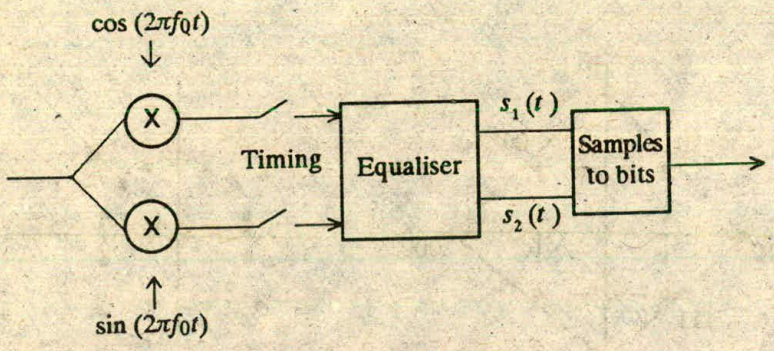
Fig. 10.5 *M*-ary PSK and *M*-ary QAM state diagrams.

and $s_2(t)/2$, from which the bit stream is reconstructed. QAM is widely used in high speed modems. The signals $s_1(t)$ and $s_2(t)$ are usually band limited to 1200 Hz, allowing 2400 samples per second for each baseband waveform. Variations in the channel frequency response is compensated by using adaptive equalisation at the receiver. The 8-ary and 16-ary QAM schemes are shown in Figs. 10.5(c) and 10.5(d) respectively. These state diagrams depict the operating states of the modems. A shift along the x- or y-axis represents amplitude variation and a circular movement phase variation.

Recent modem designs use sophisticated error detection and correction mechanisms for reliable data transmission. Many protocols are available for this purpose in the public domain and the most widely used one is Microcom's Microcom networking protocol (MNP), which may be considered to have become a *de facto* standard. There are no standards as yet from the international bodies in this regard. Since MNP conforms to datalink layer of ISO-OSI reference model (see Section 10.3), its developers are pushing it for standardisation.



(a) QAM modulator



(b) QAM demodulator

Fig. 10.6 QAM system.

There are at present eight classes of MNP protocols, which are implemented in hardware for efficiency considerations. Class 1 is seldom in use any more. Class 2 is a character oriented protocol where complete asynchronous characters are transmitted along with the protocol overhead bits. This class is also falling into disuse. Class 3 is a bit oriented protocol which converts asynchronous characters into synchronous data streams for transmission and thereby achieves a protocol efficiency greater than 100%. Class 4 is similar to class 3, but yields a better throughput. MNP class 5 uses data compression techniques to improve throughput. Class 6 uses the features of classes 3 through 5 and, in addition, implements a pseudo duplex operation to better the throughput further. Classes 7 and 8 are further improvements over the earlier classes.

Modems are generally provided by network operators (Department of Telecommunications in India) or by vendors who are not necessarily the manufacturers of computer systems. Hence, there is a need to standardise the interface between the modem and the computer equipment. CCITT terminology for the modem is data circuit terminating equipment (DCE),

and for the computer it is data terminal equipment (DTE). DCE is often referred to as data communication equipment, outside CCITT. It is important to recognise that DCE and DTE are generic terms and may be applied to a variety of equipments not necessarily only to modems or computers. For example, a DTE may be a terminal, workstation or a computer, and a DCE may be a modem or a computer based node in a data network.

A series of standards, known as V-series, has been defined by CCITT for interfacing DTEs to DCEs operating with PSTNs. The series also defines a variety of DCEs using different modulation techniques and operating at different speeds using either a leased PSTN or dial-up line. A list of some of the important V-series modem standards is given in Table 10.1. Examples of DTE-DCE interface standards and other standards related to data transmission on PSTN are:

- V.5 Standardisation of data signalling rates for synchronous data transmission in the PSTN
- V.24 DTE-DCE interface and control signals
- V.28 DTE-DCE electrical characteristics for unbalanced double-current interchange circuits
- V.53 Limits for the maintenance of telephone type circuits used for data transmission.

Table 10.1 Some V-Series Modem Recommendations

V-series	Speed	Modulation	Application
V.22	1200 bps	4-ary DPSK	FD, D-up, 2-W
V.22 bis	2400 bps (1200 bps fall back)	16-ary QAM	FD, D-up, 2-W
V.23	600/1200 bps	FSK	HD, D-up
V.29	9600 bps	16-ary QAM	FD, 4-W
V.32	up to 9600 bps	32-ary QAM	FD, D-up, 2-W
V.33	14,400 bps	128-ary AM-PM	FD, 4-W

HD = half duplex
FD = full duplex
4-W = 4-wire leased circuit
D-up = dial-up connection
2-W = 2-wire leased circuit

10.2 Switching Techniques for Data Transmission

In Section 10.1 we saw how telephone networks are used to carry data. They are basically designed to carry voice traffic and there are some significant

differences in the nature of voice and data traffic. Voice traffic is generally continuous (except for the silence periods in normal speech), whereas data traffic is bursty in nature. When a user sits at a terminal and works with a computer, interactively, he/she spends time thinking, keying in the query or command to the computer, and waiting for a response from the computer before proceeding further. In essence, the user is in one of three states; thinking, keying-in or waiting. During thinking and waiting states no data transmission takes place. During the waiting state the computer is busy processing the user command. In order to have a low waiting time, the user input must be transmitted to the system expeditiously soon after the keying-in is over. Similarly, once the response is computed by the machine, the same must be transmitted to the user at a fast rate. Thus, the data traffic is generally bursty in nature and calls for high bandwidth for short durations. This is true of the data traffic between two computers also. In contrast, voice traffic needs low bandwidth (3.4 kHz) for long durations. Typically, the transmission line is idle for 85–95% of the holding time in the case of data transmission and is busy for a similar period in the case of a telephone conversation. (Pauses in normal speech are considered as active transmission periods.) While voice traffic is half-duplex, data traffic may be half or full-duplex. Another important difference lies in acceptable error and loss rates. No errors or losses are acceptable in data transmission whereas a small amount of speech loss is often not noticeable. While the speech traffic always takes place in real time, the data traffic may or may not occur in real time. Interactive use of computers requires real time or near real time response, whereas file transfer between two computers need not be in real time. Table 10.2 summarises the differences between voice and data traffic.

Table 10.2 Differences between Voice and Data Traffic

Voice traffic	Data traffic
Continuous	Bursty
Low bandwidth for long duration	High bandwidth for short duration
Typical line utilisation 85–95%	Typical line utilisation 5–15%
Half duplex	Half or full duplex
Real time	Nonreal time or near real time
Loss acceptable	Loss unacceptable
Error tolerable	Error unacceptable

The recognition of the diverse characteristics of voice and data traffic has led to the development of a switching technique other than the one used for voice transmission. This technique is better suited for transmitting data traffic. Hence, two switching techniques are prevalent for data trans mission:

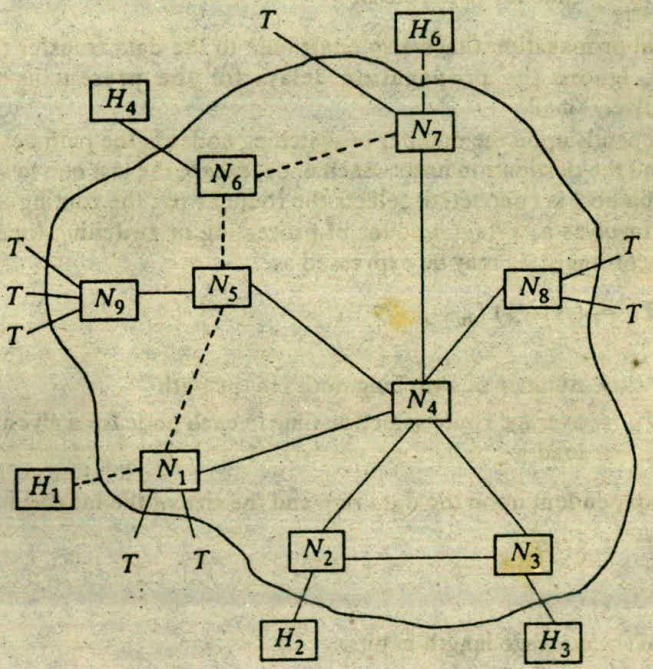
- Circuit switching
- Store and forward (S&F) switching.

Circuit switching is entirely analogous to the telephonic switching. S&F switching functions in a fashion that is analogous to the postal or telegraph system. In fact, a comparison of the characteristics of the postal system and the telephone system applies equally well to the comparison of S&F switching and circuit switching.

10.2.1 Circuit Switching

In circuit switching, an electrical path is established between the source and the destination before any data transfer takes place. The electrical path may be realised by physical wires or coaxial cables or radio or satellite links. It remains dedicated to the communicating pair for the entire duration of the transmission irrespective of whether data is actually transferred or not. No other potential user can use the path even if it is idle. The connection is released only when specifically signalled so by either of the communicating entities. Data transmission using a PSTN connection is a typical example of a circuit switched data transfer.

Figure 10.7 illustrates the principle of circuit switching. When the host computer H_1 wants to transfer data to the host computer H_6, a connection



H = host N = node T = terminal

Fig. 10.7 Circuit switched network.

request is made to the switching node N_1 which, in turn, selects a suitable neighbouring node through which the desired connection may be established, say in this case node N_5. Node N_5, in turn, selects a suitable onward path and so on until an electrical path is established between H_1 and H_6. The path selection is generally based on a routing algorithm that may take into account the network traffic, path length etc. Once a path is established, data transfer begins. There are three explicit phases involved in circuit switched data transfer:

1. Connection establishment
2. Data transmission
3. Connection release.

Accordingly, time taken for data transfer in a circuit switched connection T_{cs} has three components:

$$T_{cs} = T_e + T_t + T_r \tag{10.5}$$

where

T_e = time for path setup or connection establishment
T_t = time for data transmission
T_r = time for path release

Signal propagation times also contribute to the data transfer time. We, however, ignore the propagation delays for the present as they are comparatively small.

T_e depends upon the number of switching nodes in the path between the source and the destination hosts. Each node, except the last one to which the destination host is connected, selects the route as per the routing algorithm and this involves a certain amount of processing or switching time in each node. Accordingly, T_e may be expressed as

$$T_e = (N - 1)T_m \tag{10.6}$$

where

N = number of switching nodes in the path
T_m = average route selection time in each node for a given network load

T_t is dependent upon the data rate and the size of the message

$$T_t = M/R$$

where

M = message length in bits
R = data rate in bps

T_r is also dependent upon the number of switching nodes in the path. Connection release is generally initiated by a release signal which propagates

from one end to the other. On receiving 'connection release' signal, each node in the circuit performs certain housekeeping operations such as making entries in routing tables.

T_r may be expressed as

$$T_r = NT_h \tag{10.7}$$

where T_h is the time taken by a node to make housekeeping entries. Therefore, we have

$$T_{cs} = (N-1)T_m + M/R + NT_h \tag{10.8}$$

Propagation time is dependent upon the electrical distance between the source and the destination and the medium of transmission. The free space propagation speed is close to the velocity of electromagnetic waves in vacuum, whereas the speed in typical coaxial cables is of the order of 200 m/μs. While the propagation time one way may be determined by knowing the distance and medium properties, the number of times the propagation time needs to be counted for each data transfer phase depends on the actual protocols used in each phase. For example, in a simple protocol for connection establishment phase, successful connection is indicated by a signal sent from the destination switching node to the source switching node. In this case, twice the propagation time T_p would figure in T_e. Assuming no acknowledgement during data transmission and connection release phases, only one T_p needs to be accounted for in each T_t and T_r. Accordingly, Eq. (10.8) may be rewritten as

$$T_{cs} = (N-1)T_m + M/R + NT_h + 4T_p \tag{10.9}$$

where T_p is the one way propagation time.

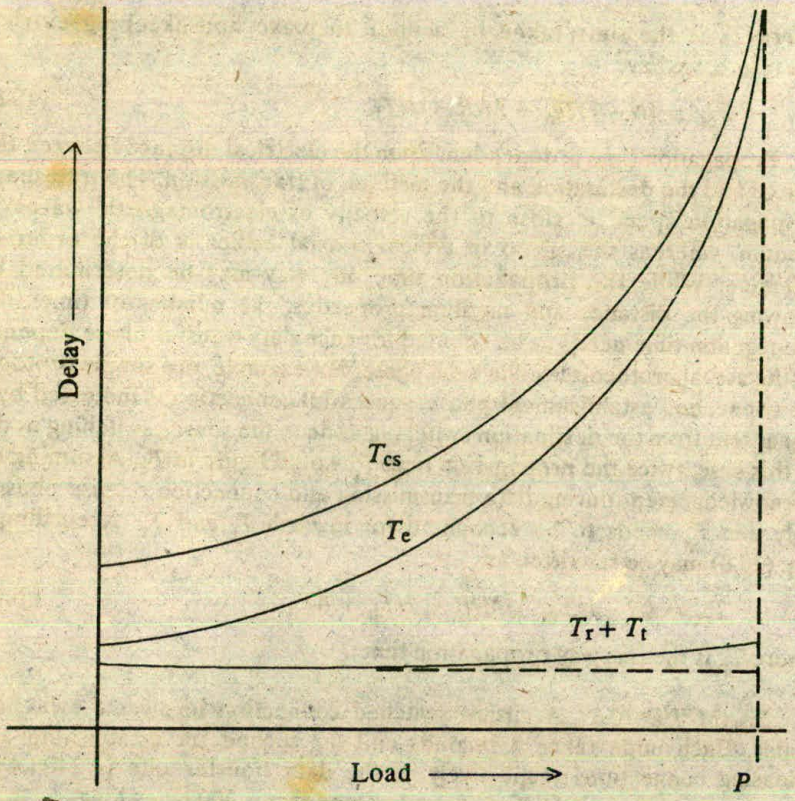
EXAMPLE 10.1 A circuit switched connection involves 5 switching nodes. Each node takes 2 seconds and 0.2 second for establishing and releasing connections respectively. If the data transfer rate is 2400 bps, compute the data transfer time for a message that is 300 bytes long.

Solution Neglecting the propagation time, we have from Eq. (10.9),

$$T_{cs} = 4 \times 2 + 300 \times 8/2400 + 5 \times 0.2 = 10.0 \text{ s}$$

In this example, nine seconds out of a total time of 10 seconds are spent on connection establishment and release and the actual data transfer lasts for only one second. Circuit switching is inefficient for small messages where the total data transfer time is dominated by the path setup time. Consequently, circuit switching is preferred where large volume data transfer is involved. Connection establishment time T_e is often a function of load on the network. As the load increases, each of the concerned node may have to look for a number of alternative routes before one is chosen. Also, the paths chosen may, in general, have a larger number of nodes than when the network is

lightly loaded. As a consequence, the path set up time T_e increases with load. Connection release and data transmission times are not significantly affected by the load. The variation of different times with respect to load is shown in Fig. 10.8. Blocking occurs when the network is fully loaded, as indicated by the local value *P*.



T_{cs} = total circuit switching delay T_e = connection establishment time
T_r = connection release time T_t = data transmission time

Fig. 10.8 Variations of delay components in circuit switching.

Circuit switching has certain disadvantages for transmitting data traffic. The path set up time which is typically of the order of 20–30 s or more turns out to be an excessive overhead for bursty computer traffic which typically lasts for a few seconds or less. The entire line quality is affected if there is one bad link in the circuit. In fact, this is a typical problem faced in many PSTN circuit switched connections. The intercity traffic in PSTNs is carried by high quality coaxial or microwave or satellite links whereas the subscriber end links are of poor quality with the result that the entire circuit is of poor quality although the major portion (typically 95% or more) of the circuit uses high quality links. This problem is often termed as 'last mile' problem.

Similarly, the speed of operation of the circuit is limited by the slowest link in the circuit. This leads to poor utilisation of high capacity lines. In circuit switched connection, the required bandwidth is statically allocated and the unused bandwidth is wasted. The network provides no error control facilities which are to be handled by the end systems.

10.2.2 Store and Forward Switching

A store and forward (S&F) network configuration is shown in Fig. 10.9. In S&F switching, the switching nodes have the ability to store user messages

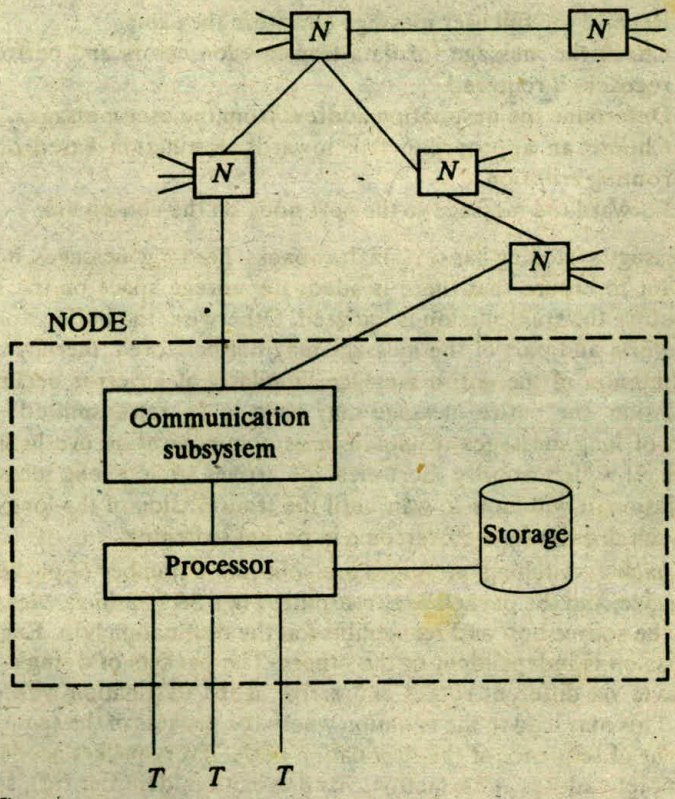


Fig. 10.9 Store and forward network.

and forward the same towards the destination as and when the links become available. For this purpose, each node is equipped with a processor and some buffer storage. No end-to-end link is set up prior to data transmission. The user deposits his/her message to the nearest switching node and then on, the network takes the responsibility for delivering the message to the destination user or host. (Observe the analogy to the postal system.)

The network moves the user information from node to node. One such movement is called a hop. Since the communication links are used one at a time between any two nodes, line speeds can be utilised efficiently. S&F switching may be classified as:

- Message switching
- Packet switching.

In message switching, once the transmission is initiated, a message is transmitted in its entirety without a break from one node to another. The node processor performs the following functions:

1. Receive the full user message and store the same.
2. Check the message for data transmission errors and perform error recovery if required.
3. Determine the destination address from the user message.
4. Choose an appropriate link towards destination based on certain routing criterion.
5. Forward the message to the next node on the chosen link.

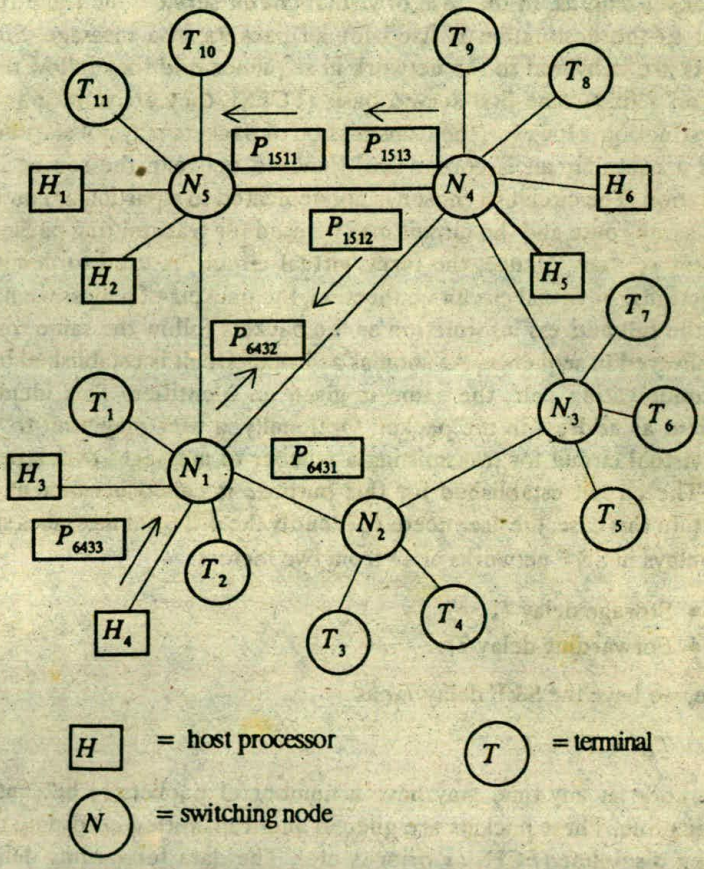
Message switching has certain drawbacks. For long messages, it becomes important to ensure that there is adequate storage space on the receiving node before the transmission is initiated. Otherwise, the buffer storage may become full, and part of the message may not be stored, thereby requiring retransmission of the entire message. Similarly, if an error occurs during transmission, the entire message may have to be retransmitted. Retransmission of long messages results in large communication overheads in the network. If a high priority short message arrives while a long message is in transmission, it will have to wait until the transmission of the long message ends. Such drawbacks are overcome in packet switching.

In packet switching, messages are split into a number of packets, often fixed in size, and the packets are transmitted in a S&F fashion. Messages are split at the source host and reassembled at the destination host. Each packet transmission is independent of the others. The packets of a single message may travel via different routes and arrive at the destination with different delays. This may lead to the situation where the packets of the same message arrive out of sequence at the destination node. Every packet needs to carry the complete address information, viz. destination identifier (id), source id, message id, and packet id, in addition to the actual user data. A typical packet format is shown in Fig. 10.10(a).

A packet switching schematic is shown in Fig. 10.10(b). In this schematic, a packet is numbered using a four subscript quantity where the first subscript is the destination host id, the second the source host id, the third the message id, and the fourth the packet id of the message. For example, P_{6432} indicates that this is the second packet of the third message originating from the host 4 and destined to host 6. It may be observed that the source host delivers the packets of a message in sequence to the network node and it is natural to

Header					
Destination id	Source id	Message id	Packet id	Control	User data

(a) A typical packet format



(b) Packet switching network (PSN)

Fig. 10.10 Packet switching.

expect that the packets are delivered to the destination host in proper sequence. However, as the packets may arrive out of sequence at the

destination node, it becomes the responsibility of the network to resequence the packets before delivery to the destination host. This calls for considerable overhead in terms of buffer storage and processing power at the network nodes and hence turns out to be somewhat an expensive service. In order to be cost effective, packet networks offer two different forms of services:

- Datagram service
- Virtual circuit service.

In these services, no resequencing is required to be done by the network. Datagram service is normally used for transmitting short messages of one or two packet lengths. In the case of virtual circuit service, the route from the source to the destination is fixed for all packets of a message. Since the packets are delivered to the network in sequence and they follow the same route on a first-come-first-served basis (FCFS), they arrive in sequence at the destination. However, the transmission of packets may not start until and unless a route/circuit is chosen and finalised between the source and the destination. The circuit so chosen is not dedicated to a particular connection, as the same route and the circuit may be used for transmitting packets from different sources. Hence, the term 'virtual circuit' is used to denote such connections. In virtual circuit connection, the packets of a message need not carry the full address information as the packets follow the same route and are delivered in sequence. As soon as a virtual circuit is established between a communicating pair, the same is given an identifier. This identifier is sufficient as address in the packet. Optionally, a user may want to use the same virtual circuit for transmitting a number of messages over a period of time. The circuit established for this purpose is called **permanent virtual circuit**. In this case, the user needs to identify the different messages suitably.

Delays in S&F networks arise from two factors:

- Storage delay T_s
- Forwarding delay T_f

Hence, we have the S&F delay T_{sf} as

$$T_{sf} = T_s + T_f \tag{10.10}$$

A node at any time may have a number of packets to be sent on an outgoing line. These packets are queued and transmitted according to some queuing discipline (FCFS or priority etc). The data forwarding delay T_f is largely due to this queuing delay. A small component comes from the processing that needs to be performed by the node processors for routing etc. T_f may be expressed as

$$T_f = (N-1)(T_q + T_m) + T_t \tag{10.11}$$

where

T_q = average queuing delay in each node
T_m = processing delay in each node
T_t = message or packet transmission delay
N = number of nodes involved in the transfer

The destination node is assumed not to contribute to any delay as the destination host is directly attached to it and no other network traffic is on this link. If the transfer rate R is uniform on all the links, then the data transmission delay T_t is given by

$$T_t = (N + 1)M/R + T_p \tag{10.12}$$

as there are $(N + 1)$ data links or hops from the source to the destination. M is the message length in the case of message switching and is the packet size in the case of packet switching. T_p is the total propagation delay from source to destination. If the data transfer rates are nonuniform, T_t is given by

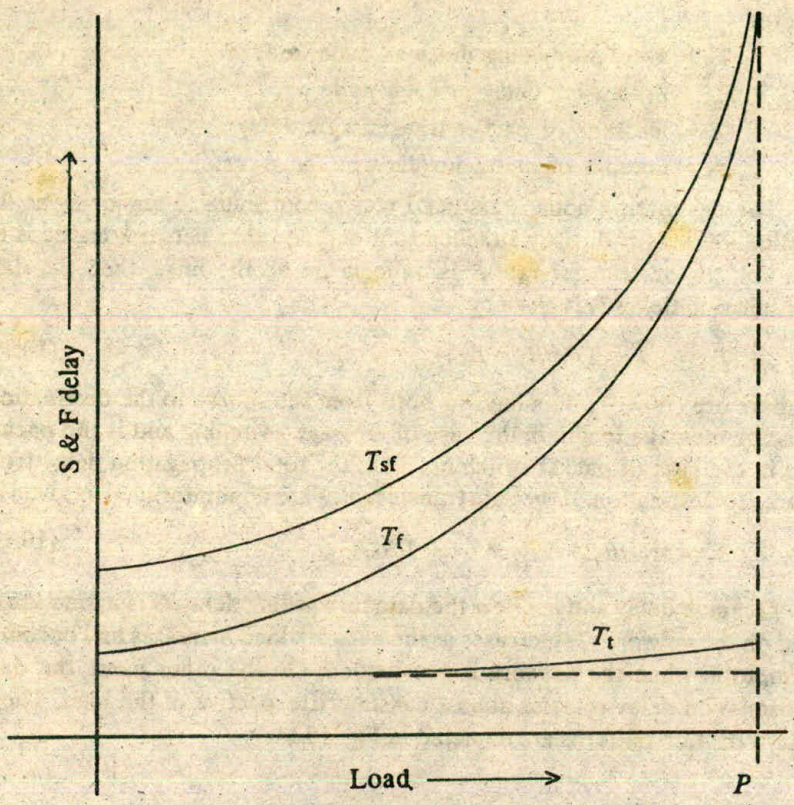
$$T_t = M(1/R_1 + 1/R_2 + \cdots + 1/R_{N+1}) \tag{10.13}$$

Queuing delay and in effect the data forwarding delay T_f is a function of load on the network. T_f increases as the network load increases and becomes asymptotic when the network is overloaded. On the other hand, the data transmission delay remains almost constant irrespective of the load. These delay variation patterns are depicted in Fig. 10.11.

10.3 Data Communication Architecture

Data communication among computers involves a number of functions such as physical transmission of bits, error control, routing and session establishment. In order to efficiently implement these functions, vendors of computer systems evolved their own architectures. Examples of vendor specific architectures are System Network Architecture (SNA) of IBM and Digital Network Architecture (DNA) of Digital Equipment Corporation (DEC). Such architectures permit interconnection of computers from the same vendor but not from different vendors. Systems or networks, which are not open to other vendor systems for networking, are known as 'closed' systems or networks. In order that heterogeneous computer systems from different vendors be interconnected as a network, an architecture which is used as a standard by all the vendors is required. The heterogeneity covers the following aspects:

1. Systems of different vendors
2. Systems under different managements
3. Systems of different complexities
4. Systems of different technologies.



T_f = forwarding delay T_{sf} = store and forward delay
T_t = transmission delay

Fig. 10.11 Variation of delay components in S&F switching.

ARPANET, the network project supported by Advanced Research Projects Agency of the Department of Defence, United States, is one of the pioneering efforts in interconnecting heterogeneous systems. The efforts put in and the experience gained in the project have significantly contributed to the emergence of a world standards architecture for computer communication, largely pursued and set out by International Standardisation Organisation (ISO). These standards, now well known as ISO-Open System Interconnection (ISO-OSI) standards, are widely accepted. The standards are based on a reference architecture which is described in the ISO standard IS 7498. CCITT has also adopted this standard under its own number X.200. The architecture is considered 'open' as any vendor's system conforming to this reference model is capable of organising information transfer with any other vendor's system which also conforms to the same architecture.

10.3.1 ISO-OSI Reference Model

Before we discuss this model, a few definitions are in order.

System: A system is one or more autonomous computers and their associated software, peripherals and users, which are capable of information processing and/or transfer.

Subsystem: A logically independent smaller unit of a system. A succession of subsystems make up a system as shown in Fig. 10.12.

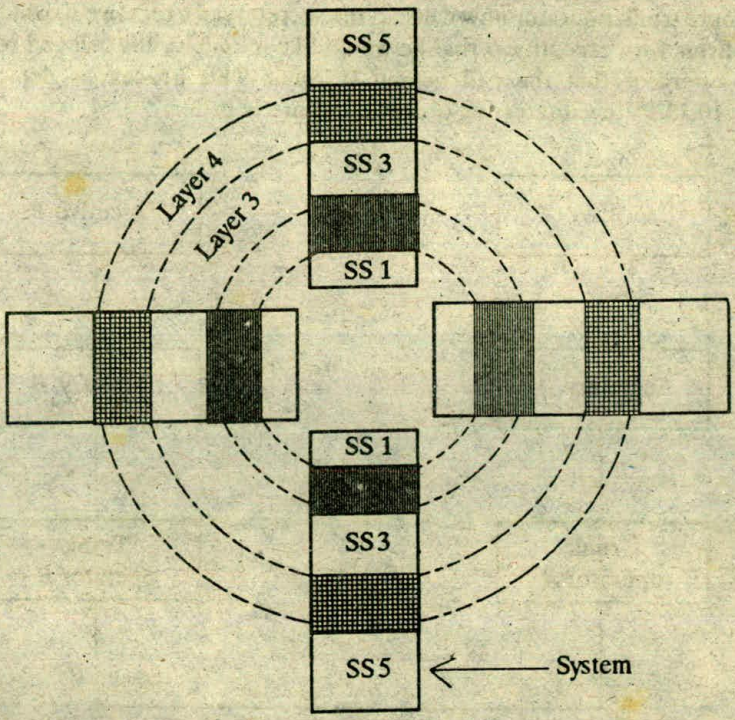


Fig. 10.12 Systems, subsystems and layers in ISO-OSI model.

Layer: A layer is composed of subsystems of the same rank of all the interconnected systems. The concept of a layer is illustrated in Fig. 10.12, which shows a five-layer network. The subsystems and the layers are numbered starting with one at the bottom level.

Entity: The functions in a layer are performed by hardware subsystems and/or software packages. These are known as entities.

ISO-OSI architecture is a layered one. Layering is a natural choice for communication architectures. This is well illustrated by an example.

Consider the activities that are involved when executives *A* and *B* of two companies in different cities want to converse over a trunk telephone connection. For the sake of illustration, let us assume that there is no subscriber trunk dialling (STD) facility between the two cities. Let executive *A* be the calling party. As a first step, he requests his secretary to connect him to executive *B*. His secretary in turn calls up the trunk operator and communicates the calling number, called number, nature of the call, the name of the particular person called etc. Then, the local trunk operator calls up the other trunk operator in the other city and communicates the details. Remote trunk operator now calls up the secretary of executive *B*, who in turn confirms with executive *B* that he would like to receive the call and requests the operator that the call be put through. This process is depicted in Fig. 10.13. A few interesting observations are in order:

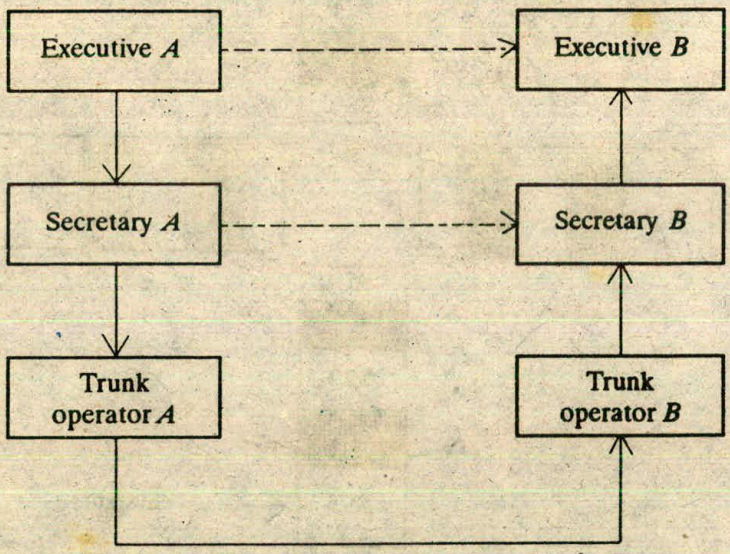


Fig. 10.13 A three-layer structure for a trunk call connection.

1. A three-layer structure is used in this communication process.
2. The conversation between an upper and lower layer is strictly business like.
3. There is generally a little private conversation between the trunk operators and the two secretaries on account of their familiarity. In other words, entities in the same level or layer exchange information using their own private protocols.
4. A layer obtains services from its immediate lower layer and provides services to its immediate upper layer. In this sense, a layer acts both as a user as well as a service provider.

5. There are fairly well defined functions to be performed by each layer.
6. It is immaterial as to how the functions of each layer are implemented. For example, the secretary may ask his assistant to book the call and as far as the executive is concerned, it is immaterial who books the call as long as the call is booked.

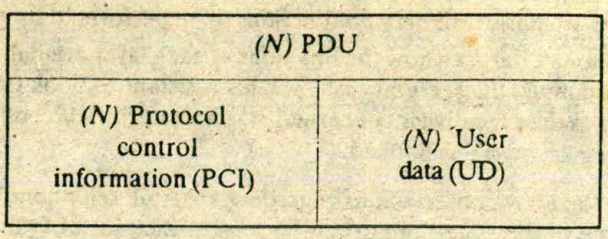
In fact, the above observations regarding a simple telephone conversation are some of the stated important layering principles of OSI reference model. IS 7498 recommends standard reference conventions when discussing layers and their functions. Some of the important reference conventions are as follows:

1. A layer is referred to as the (N) layer, N being the layer number.
2. The layer immediately below (N) layer is referred to as ($N-1$) layer.
3. The layer immediately above (N) layer is referred to as ($N+1$) layer.
4. Services offered by the (N) layer are termed (N) services.
5. The entities of (N) layer are referred to as (N) entities.
6. Entities in the same layer, but not in the same subsystem, are known as **peer entities.**
7. Peer entities communicate using what are known as **peer protocols.**
8. Data exchange between peer entities is in the form of **protocol data units (PDUs).**
9. Data exchange between entities of adjacent layers is in the form of **interface data units (IDUs).**
10. IDUs exchanged between (N) and ($N+1$) layers are referred to as (N) IDUs.

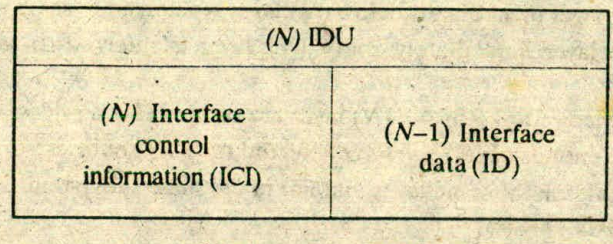
The structures of the PDUs and the IDUs are shown in Figs. 10.14(a) and 10.14(b) respectively. Both the data units have two parts each: control information and data. Normally (N) PDU becomes ($N-1$) interface data as shown in Fig. 10.14(c). Depending upon the complexity of implementation, only a part of (N) interface data may be sent to the receiving end of a connection, in which case this part is known as (N) **service data unit (SDU).** In simple implementations, (N) PDU, ($N-1$) interface data, and ($N-1$) SDU are all the same.

Entities of a layer interact with entities of adjacent layers via service **access points (SAPs).** SAPs at the interface of (N) layer and ($N+1$) layers are referred to as (N) SAPs. There are certain restrictions with regard to the use of SAPs:

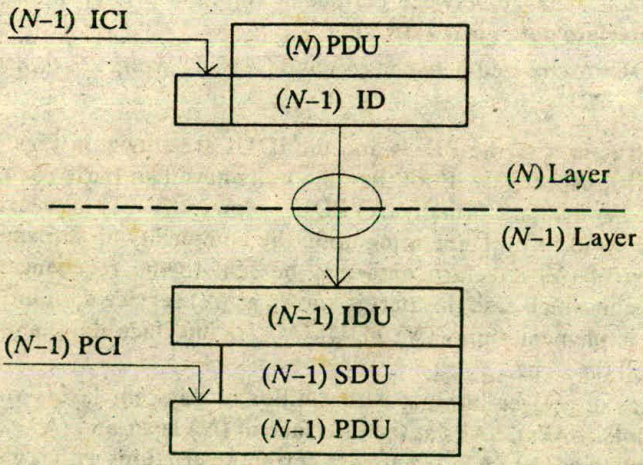
1. An ($N+1$) entity may obtain services via one or more (N) SAPs
2. An (N) entity may serve one or more (N) SAPs
3. An (N) SAP may be served by only one (N) entity and may serve only one ($N+1$) entity.



(a) Structure of PDU



(b) Structure of IDU

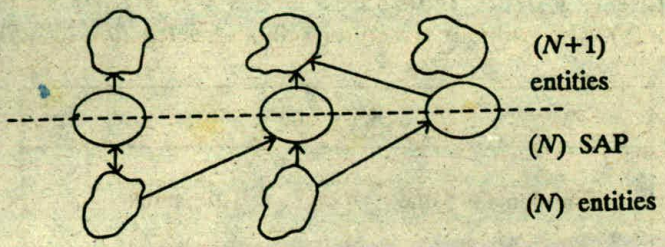


(c) PDU, IDU, SDU relationship

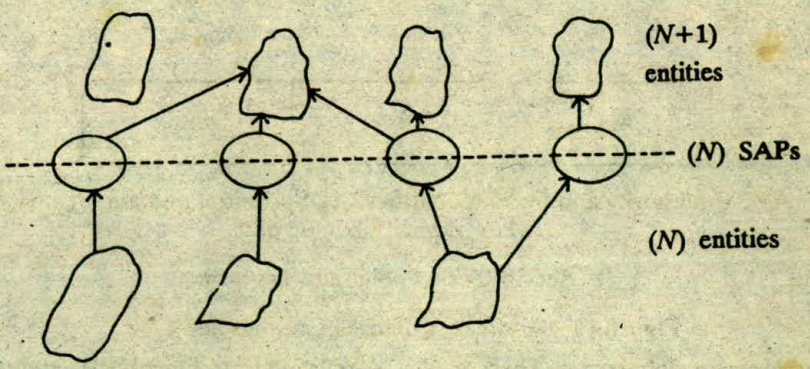
PDU = protocol data unit IDU = interface data unit
ICI = interface control information
PCI = protocol control information SDU = service data unit

Fig. 10.14 Structures of PDU and IDU.

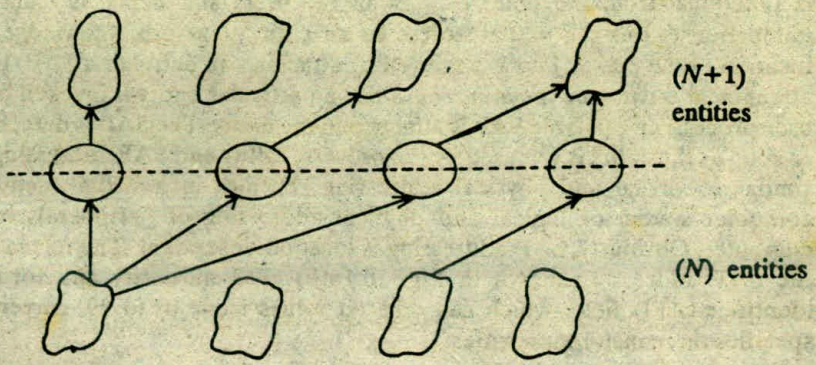
These restrictions are illustrated in Fig. 10.15. The connection shown in Fig. 10.15(a) is not permissible as the (*N*) SAP in the middle is being serviced by two (*N*) entities. Similarly, the connection in Fig. 10.15(b) is also wrong as



(a) One SAP is served by more than one entity, which is not permissible



(b) One SAP serves more than one entity, which is not permissible

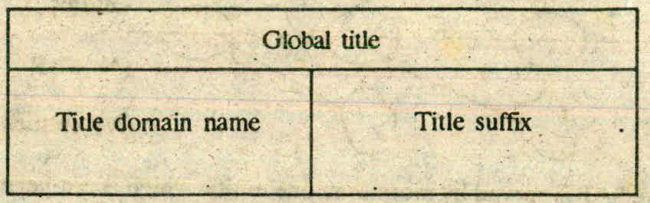


(c) Permissible connection

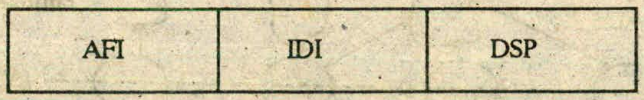
Fig. 10.15 SAP-entity connection modes.

one of the (N) SAPs is serving more than one ($N + 1$) entity. The connection in Fig. 10.15(c) is valid.

Entities in the OSI environment are identified by unique global titles which are valid regardless of the location and movement of the entities. The global title is a structured address and has two parts as shown in Fig. 10.16(a). The two-part structure facilitates delegation of authority to



(a) Entity identification



AFI = authority and format identifier IDI = initial domain identifier
DSP = domain specific part

(b) Service access point address format

Fig. 10.16 Address structure of entities and SAPs.

assign new global titles. There is also a provision to use local titles within a limited context as specified by the title domain. This scheme of local and global titles is very similar to the scheme of local and global variables in structured programming languages like Pascal or ALGOL. Entity addressing is functional from the point of view of the users and hence is location independent. But the actual access to an entity demands knowledge of location of the entity. This is achieved by attaching an entity to a SAP. This attachment is dynamic in nature, i.e. when a particular entity is invoked by a user process, an available SAP is attached to the entity. The SAP is detached as soon as the entity is released by the process. Entity and SAP addressing is similar to virtual and physical addressing schemes in a virtual memory computer system or logical and physical addressing of peripherals in a computer. Obviously, SAP addressing is location dependent. The format of the SAP address is shown in Fig. 10.16(b). The authority and format identifier (AFI) field, which can take on values from 10 to 99, currently specifies three major authorities:

1. CCITT
2. ISO member bodies
3. ISO international organisations.

In addition, AFI also identifies the format and type of address present in the initial domain identifier (IDI) field and the domain specific part (DSP). Binary and packed decimal formats are generally supported. IDI and DSP types may include PSTN numbering, packet switched network numbering, ISDN numbering, telex numbering and other similar numbering schemes. The full SAP address is of variable length, up to 40 decimal digits or 20 bytes. An important objective of SAP addressing format is to enable convenient internetworking.

OSI reference model proposes a general layered concept, with provision for adding or deleting layers as demanded by factors like service complexity, technology options etc. Taking into account the functions required of the architecture for organising computer communication in the present day context, Sub Committee 16 of the Technical Committee 97 (TC97/SC16) of ISO, has recommended a 7-layer architecture, which is shown in Fig. 10.17. In the figure, two end systems that communicate with each other via two intermediate systems are shown. Only the first three layer functions come into action in an intermediate node. Entities in these layers always communicate with peer entities in the adjacent system. In other words, in the first three layers, the communication proceeds on a **link-by-link basis**. In contrast, entities in layers 4–7 communicate with peer entities in the end systems. There is no communication with entities in the intermediate systems. In this sense, layers 4–7 are often called **end-to-end layers**. The 7-layer architecture has been arrived at after a careful application of a broad set of layering principles. The important principles are:

1. Create layers to handle functions which are manifestly different in the process performed or technology involved.
2. Collect similar functions into the same layer and create a boundary at a point where the number of interactions across the boundary are minimised.
3. Create a layer of easily localised functions so that the layer could be totally redesigned and its protocols changed in a major way to take advantage of new advances in architectures, hardware and software technology without changing the services offered or the interfaces with the adjacent layers.

Sections 11.4 and 11.5 describe the functions performed, services offered, and the protocols used by each of the seven layers.

10.4 Link-to-link Layers

The first three layers, viz. physical, data link and network layers, form the link-to-link layers of OSI reference model. Entities in an OSI layer perform certain functions to fulfill the stated purpose of the layer. They obtain services from the immediate lower layer and provide services to the

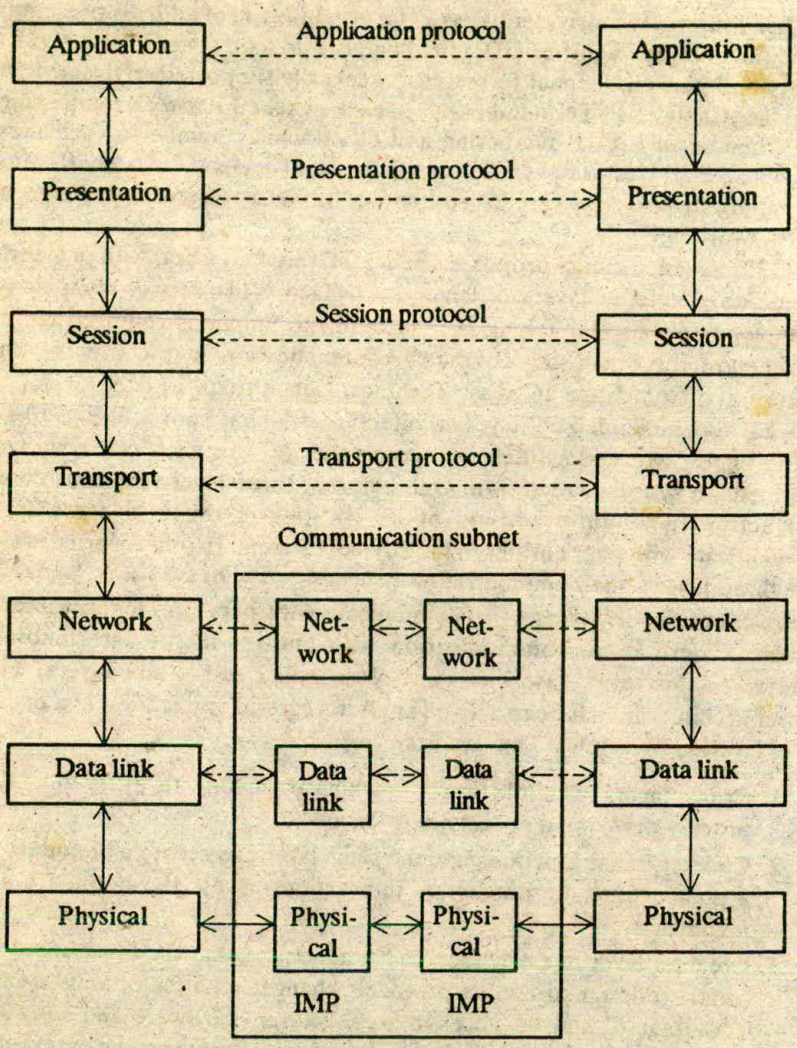


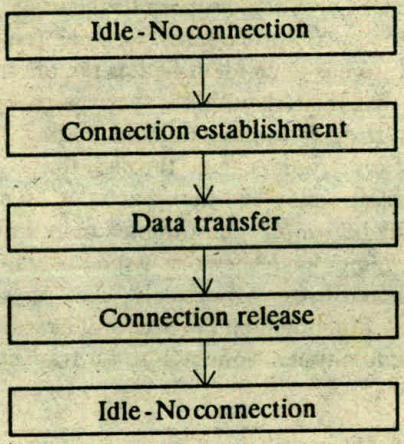
Fig. 10.17 ISO OSI reference model.

immediate upper layer. OSI services may be placed under two broad categories:

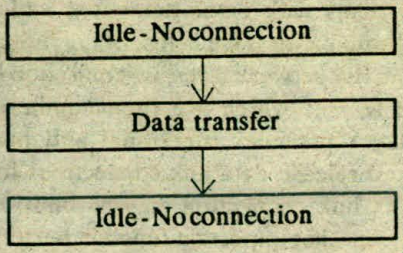
- Connection oriented services
- Connectionless services.

In connection oriented services, a connection is first established between the sender and the receiver before data transfer can commence. The connection may be virtual (logical) or physical, depending upon the network

capabilities and facilities. The essence of a connection oriented service is that a connection acts like a tube or a pipe delivering the data to the receiver strictly in the same order in which the data was put into the connection by the sender. Connection oriented service is modelled after the telephone system. In contrast, connectionless service is modelled after the postal system. Each submission by the sender is treated independently of others and is self-contained with the full address of the destination and the source indication which may be the full address too. In connectionless service, when two messages are sent to the same destination one after another, it is possible that the first one is delayed and the second one arrives first. Datagram service and the virtual circuit service discussed in Section 10.2 are examples of connectionless and connection oriented services respectively. The operation of the two categories of services is shown in Figs. 10.18(a) and (b).



(a) Connection oriented service



(b) Connectionless service

Fig. 10.18 Operation of connection oriented and connectionless networks.

A connection oriented service has provision for acknowledgements, flow control and error recovery, whereas a connectionless service does not generally have such provisions.

Peer entities of OSI layers communicate using peer protocols. Protocols are strict procedures and sequence of actions to be followed in order to achieve orderly exchange of information among peer entities. Corresponding to the two categories of services, there are two sets of protocols; one set for the connection oriented services and the other for connectionless services. It is important to recognise that layer protocols relate to the implementation of services of the layer and therefore are not visible to the users or other layers. This separation of services and the protocols provides complete freedom to change protocols at will without affecting the services.

10.4.1 Physical Layer

It is essential that the OSI architecture permits the usage of a realistic variety of physical media and control procedures. Keeping this in mind, the lowest layer of the architecture has been identified as the physical layer. This layer performs functions associated with the activation and deactivation of physical connections. It deals with encoding/decoding of signals and the bit level transmission of electronic signals through the available transmission medium. The transmission may be synchronous or asynchronous. Mode of data transmission may be simplex, half duplex or full duplex. Transmission and data encoding schemes have been discussed in detail in Chapters 5 and 7 for cables and optical fibres respectively. The physical layer provides mechanical, electrical, functional and procedural characteristics to activate, maintain and deactivate physical connections for transmission of bits.

10.4.2 Data Link Layer

Some physical communication media like telephone lines have error rates that are not acceptable for the great majority of data network applications. Special techniques are required to ensure error free transmission of data. The data link layer deals with error detection, and automatic recovery procedures required when a message is lost or corrupted. For this purpose, a user of this layer, i.e. the network layer, is required to break up the data to be transmitted into frames which are then numbered and transmitted sequentially. The layer provides functional and procedural means to establish, maintain and release data link connections for the entities in the network layer. A data link connection may be built upon one or several physical connections. Another important function performed by the data link layer is the link level flow control of frames. Flow control is essentially a traffic regulation mechanism that will have to be enforced when the receiver is unable to accept frames as fast as the transmitter is able to send.

A data link may be of point-to-point type as in the case of terrestrial networks or broadcast type as in the case of SBDNs, LANs or MANs. In the case of broadcast type channels, the data link layer will have to perform an additional function of acquiring or accessing the channel before data transmission can begin. Media access techniques for satellite channels are discussed in Section 10.6, for LAN configurations in Section 10.7, and for MANs in Section 10.8. In this section, we confine ourselves to point-to-point links. In this case, the media access is an insignificant operation as there are no contenders for the channel. Hence, our discussions mainly concentrate on efficient channel utilisation, error recovery and flow control mechanisms. These discussions are also applicable for data transmission in broadcast type channels, once the channel is acquired.

The main sources of error in a system are:

1. Thermal noise which is internal to the system
2. Impulse or spike noise which originates from man-made sources like automobiles or signalling in telephone systems, and from natural sources like lightning
3. Crosstalk which occurs through electromagnetic radiation from parallel and adjacent wires which behave like an antenna.

Errors are of two types: isolated and bursty. Errors due to thermal noise are generally isolated, whereas errors due to spike noise or crosstalk are bursty in nature. Error control mechanisms are chosen depending upon the type of error that is predominant in a given system. There are three error control mechanisms that are commonly used:

1. Echo checking
2. Forward error correction (FEC)
3. Automatic repeat request (ARQ).

Echo checking is normally used only when the source is intelligent. In echo checking, the data is echoed back to the sender and the data is retransmitted if an error is detected by the sender. Since all the data is echoed back, the scheme requires double the transmission bandwidth required for one way transmission. The scheme works well in a full duplex connection with a large channel capacity. It is typically used for terminal-to-computer communication and is not suitable for large volume data transfer.

Forward error correction is based on the hypothesis that it is possible to detect and correct errors at the receiving end by adding appropriate redundant information to the data at the sending end. FEC scheme is useful where isolated errors occur and the reliability of the channel is low. The most popular error correcting code is the **Hamming code.**, where a code set is formed by adding r number of redundant bits to the legal non redundant information words. For example, an 8-bit character set has 256 legal

combinations. By adding three bits to each of the characters, we generate a code set which has a total 2048 combinations of which only 256 are considered legal.

In a code set, the minimum number of bit changes required to produce one legal word from another is called the Hamming distance between the two code words. A code set in which at least one pair of legal code words has a distance d, but no pair has a distance less than d, is called a distance-d code set. Once a distance-d code set is formed we get two important results:

- Bit changes up to $d-1$ bits would result in illegal code words. By recognising the illegal code words at the receiving end, we are in a position to detect up to $d-1$ bit errors.
- By applying a simple nearest neighbour principle, an illegal codeword can be corrected to the nearest legal codeword, i.e. errors of up to $(d-1)/2$ bits can be corrected.

It is interesting to note that a parity bit scheme is a distance-2 Hamming code and hence can detect single bit errors. The error, however, cannot be corrected as the resulting illegal codeword is equidistant from the legal codewords and hence the nearest neighbour principle fails.

In automatic repeat request (ARQ) scheme, the information word is coded with adequate redundant bits so as to enable detection of errors at the receiving end. If an error is detected, the receiver asks the sender to retransmit the particular information word. ARQ scheme is useful where the expected errors are bursty in nature or error rate of the channel is low, i.e. the channel is fairly reliable. Most often, the errors encountered in data communication systems are bursty in nature. Hence, ARQ schemes are used extensively in data networks.

Error detection by using one parity bit is not satisfactory for burst errors. The error detection method which is used widely in ARQ protocols is known as **polynomial coding** or **cyclic redundancy coding (CRC)**. Here, a checksum of r bits is computed and appended to the message stream, such that the combined stream, when treated as a polynomial, is exactly divisible by a predefined generator polynomial $G(x)$ of degree r. An error is known to have occurred if $G(x)$ does not divide exactly the received bit stream. Three standard generator polynomials are used widely in WAN applications:

$$\begin{aligned} \text{CRC-12} &= x^{12}+x^{11}+x^3+x^2+x+1 \\ \text{CRC-16} &= x^{16}+x^{15}+x^2+1 \\ \text{CRC-CCITT} &= x^{16}+x^{12}+x^5+1 \end{aligned} \qquad (10.14)$$

CRC schemes are very effective in detecting errors. Typical CRC schemes detect the following errors:

1. All single and two bit errors
2. All odd bit errors

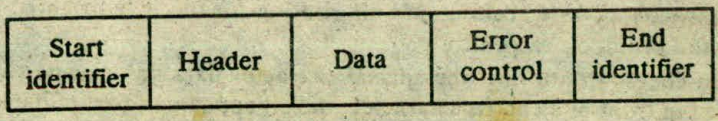
3. All burst errors of length less than r
4. Over 99.9% of all burst errors of length equal to or greater than r.

A detailed treatment of checksum generation in CRC schemes is beyond the scope of this text. Readers are referred to Further Reading [11] for more details.

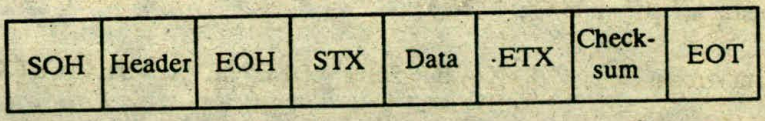
An ARQ protocol is characterised by four functional steps:

1. Transmission of frames
2. Error checking at the receiver end
3. Acknowledgement
   (a) negative if error is detected (NAK)
   (b) positive if no error is detected (PAK)
4. Retransmission if acknowledgement is negative (NAK) or if no acknowledgement is received within a stipulated time.

It may be noted that ARQ protocols require two way communication even if the information transfer is simplex, i.e. one way only. Information is exchanged in the form of frames, the beginning and the end of which are identified by means of flags or special characters. The general structure of a frame is shown in Fig. 10.19(a). The header information may contain

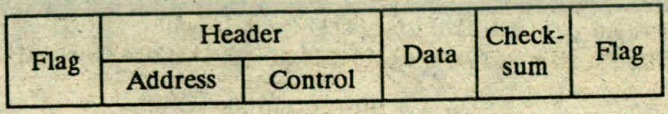


(a) General structure of a frame



EOH = end of header ETX = end of text EOT = end of transmission
SOH = start of header STX = start of text

(b) structure of a character oriented frame



(c) Structure of a bit oriented frame

Fig. 10.19 Data link frame structure.

destination address in the case of multidrop lines or command/response identifier in a point-to-point link. In addition, some control information is carried in the header. Data may be in the form of characters or a continuous sequence of bits. Accordingly, we have two types of frames, character or byte oriented, and bit oriented, with structures as shown in Figs. 10.19(b) and (c) respectively. ARQ protocols that deal with two types of frames are known as **byte oriented** and **bit oriented protocols** respectively. Examples of the former include the binary synchronous communication (BISYNC) protocol of IBM and of the latter the synchronous data link control (SDLC) protocol of IBM. SDLC was modified/improved by a number of standardising agencies and adopted under different names as follows:

Advanced data communication control procedure (ADCCP) : ANSI
High level data link control (HDLC) : ISO
Link access procedure (LAP) : CCITT
Link access procedure-balanced (LAPB) : CCITT

It is possible that the control characters like STX, SOH or ETX which convey special meaning to the receiver, may appear as part of the data in the character oriented frames. Similarly, the flag pattern may appear in the binary stream of data in the bit oriented frames. A mechanism is required so that the receiver does not act on the special characters or flag patterns if they appear as part of the data. The mechanisms used for this purpose are known as **character** or **byte stuffing** and **bit stuffing** for the byte oriented and bit oriented protocols respectively.

In character stuffing, the special character data link escape (DLE) is stuffed in front of a control character when it appears as part of data. When a DLE character is encountered by the receiver, it deletes the character from the main stream of data and ignores the special significance, if any, of the character following DLE. A data stream and the associated character stuffed stream are shown in Fig. 10.20(a).

Data stream	STX A B ··· SOH ··· DLE ··· ETX ··· STX ··· ETX
Byte stuffed stream	STX A B ··· DLE SOH ··· DLE DLE ··· DLE ETX ··· DLE STX ··· ETX

(a) Character stuffing

Data stream	0101 ··· 011111r110 ··· 001111100 ··· 0101
Bit stuffed stream	0101 ··· 01111101110 ··· 00111110000 ··· 0101

(b) Bit stuffing

Fig. 10.20 Stuffing techniques for data frames.

In Fig.10.20(a), the first STX and the last ETX are the only characters recognised by the receiver as control characters and the significance of all other special characters SOH, ETX and STX are ignored. Similarly, the flag pattern, which is usually 01111110, is treated by stuffing a zero when five consecutive 1's are encountered in the data. This is illustrated in Fig. 10.20(b)

A simple ARQ protocol is known as **stop-and-wait** or **idle RQ protocol**. In this protocol, the sender transmits a frame and waits for an acknowledgement. The operation of the protocol is described in a flowchart in Fig. 10.21. A retransmission count is used to avoid endless retrans-

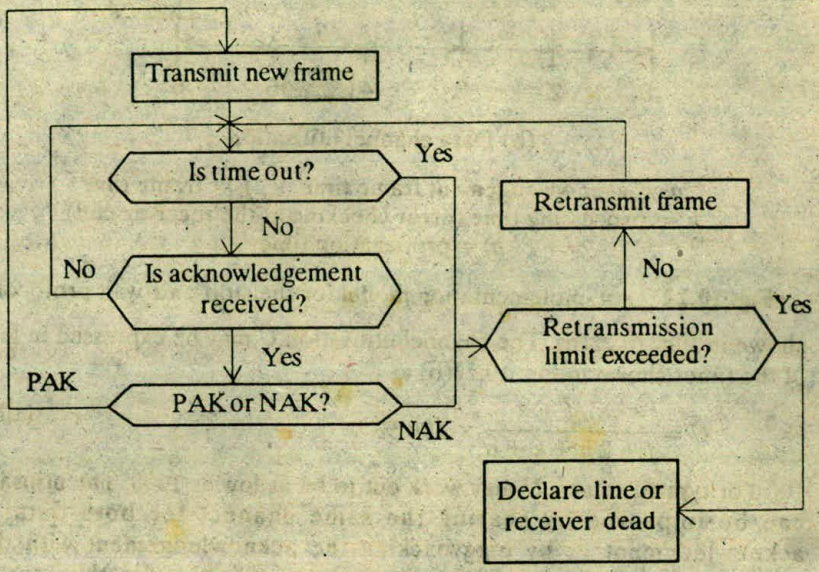
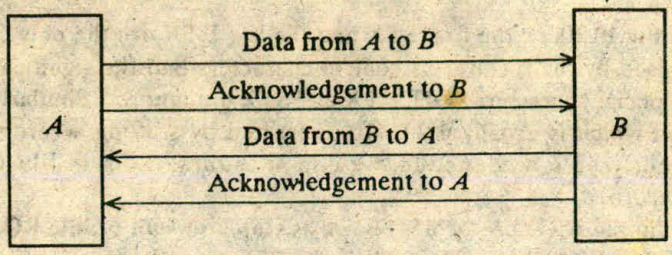


Fig. 10.21 Operation of the stop-and-wait protocol.

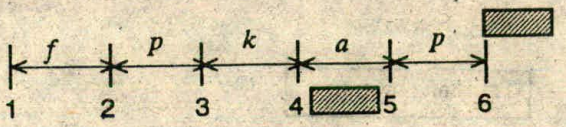
missions. If a number of NAKs are received, one may conclude that the line quality is poor, and if no acknowledgement is received, one may conclude that the receiver is dead or the link is broken. In idle RQ protocol, data transmission may be organised in a number of ways:

1. Full duplex (2 or 4 wire) or half duplex
2. Separate or a common link for both data and acknowledgement packets
3. Separate packet for acknowledgement, or acknowledgement as part of data flowing in the reverse direction, known as **piggybacked acknowledgement**.

In order to get a feel for the performance of the protocol, let us consider the model shown in Fig. 10.22(a) and the data channel utilisation details



(a) A model for idle RQ protocol



(b) Data channel utilisation

a = acknowledgement frame time f = frame time
k = processing time (error checking at the receiver end)
p = propagation time

Fig. 10.22 An implementation model for the stop-and-wait protocol.

shown in Fig. 10.22(b). The channel utilisation U may be expressed in terms of the times shown in Fig. 10.22(b) as

$$U = \frac{f}{f + 2p + k + a} \times 100\ \% \tag{10.15}$$

For typical values, U may work out to be as low as 10%. The utilisation can be improved by sharing the same channel for both data and acknowledgment or by piggybacking the acknowledgement with data. Readers are urged to work out the channel utilisation for the improved schemes. In the piggybacking scheme, it may be necessary to send a separate acknowledgement frame in case there is no reverse traffic, lest a time out error may occur at the sender's end.

Generally, all ARQ protocols require that a sequence number be used to identify a frame. Consider the sequence of steps in stop-and-wait ARQ protocol:

1. Error-free frame arrives at the receiver's end.
2. Receiver transmits PAK frame.
3. The PAK frame gets lost and does not reach the sender.
4. The sender times out and retransmits the frame.

How would the receiver know that the frame that is received is a new one or a retransmitted one? To bring about this distinction, a 1-bit sequence number is used in stop-and-wait protocols. If the sequence bit is the same as

the previous frame, the frame is a retransmitted one; otherwise it is a new one. In a piggybacked scheme, whenever a frame (new or retransmitted) is sent, it must reflect the latest acknowledgement status. Otherwise, **deadlocks** may occur in the system. Readers may like to consider as to what would happen if this is not the case when both stations A and B initiate frame transmissions simultaneously.

Utilisation of the channels can be considerably improved by using what are known as **continuous RQ** or **pipelined ARQ protocols**. In pipelined protocols, a sender may transmit frames continuously one after another without waiting for acknowledgements. However, the number of frames that may be transmitted in this manner is limited by the buffer capacity at the receiver. The acknowledgement traffic on the channel may be reduced by acknowledging the latest correctly received frame. This would imply that all the other frames sent before the one acknowledged have also been received correctly. When the receiver finds a frame in error, a retransmission is requested in either one of the following modes:

- Selective repeat
- Go back n.

In the former, the receiver requests the transmission of a specific frame, whereas in the latter, the retransmission of all the frames after the frame number n is requested. The first case demands a large buffer capacity at the receiver than the latter case. On the other hand, transmission overheads are higher in the latter case. In both cases, there is an obvious need for numbering sequentially the frames. It is also important to recognise that the frames must be delivered to the network layer in the same order in which they are received at the data link layer and the sequence numbering helps this process.

Considering the limited buffer availability at the receiving ends, the sequence numbering may be a modulo of some number m. Usually modulo-8 or modulo-128 is used. The buffer can then be organised in a **sliding window** fashion such that frames may be received and buffered depending upon the window size. The maximum window size w is limited by the retransmission strategy used:

For selective repeat : $w \leq m/2$

For go back n : $w \leq m - 1$

The restriction on the window size is necessary to know correctly the status of transmission. For example, if we use a window size of 6 in a modulo-8 scheme with selective repeat retransmission strategy, wrong interpretations can occur under the following conditions:

1. Frames 0–5 are sent and successfully received.
2. PAK 5 is sent and the window is set to receive frames 6, 7 or 0–3.
3. PAK 5 is lost, sender times out and retransmits frame 0.

In this case, the receiver would interpret this frame 0 as a new one. If the window size were restricted to four or less, this anomaly would not occur. Similar situations can be visualised in the case of go back *n* strategy, and readers are encouraged to work out conditions to be convinced of this.

In ARQ protocols, although the window size is predetermined and agreed upon by both the sender and the receiver, it is possible that the receiver is unable to accept frames due to various other reasons such as lack of buffer space. In such cases, a flow control mechanism is invoked whereby the receiver requests the sender to 'stop' until further advice to 'go' is given. This flow control mechanism is known as **stop-and-go strategy**. Sequence number, acknowledgement and flow control information are all part of the control field of a frame as shown in Fig. 10.23. The field **'kind'** is used to

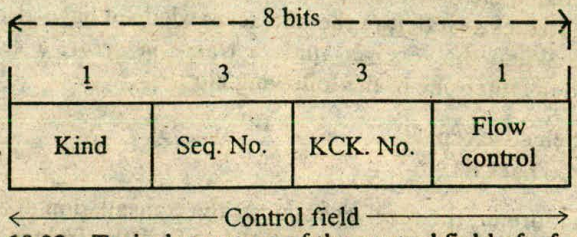


Fig. 10.23 Typical structure of the control field of a frame.

indicate the type of frame; frame with data or an acknowledgement frame or a command frame etc.

10.4.3 Network Layer

The highest link-to-link layer in the OSI model is the network layer. Although this layer functions on a link-to-link basis, it is concerned with transmission of packets from the source node to the destination node. It deals with routing and switching considerations that are required in establishing a network connection which may involve the use of several transmission resources in tandem, including a number of intermediate switching nodes of different subnetworks. The network layer makes invisible to the transport layer, the details of the underlying communication media and the different characteristics of the transmission and network technologies. It only assures a certain quality of service to the upper layers. Since an end-to-end connection may involve routing through a number of different networks, internetworking is an important function of the network layer. Addressing schemes, network capabilities, protocol differences, and accounting and billing are all issues to be handled in internetworking. Network congestion, which may occur due to too many messages on a particular route, is also tackled by the network layer.

There is a wide variety of routing and congestion control algorithms. However, OSI network layer specifications do not recommend and discuss

any routing or congestion control algorithms and they are left as implementation dependent features. We briefly discuss some of the generic routing and congestion control strategies here. Routing strategies may be classified as shown in Fig. 10.24. Routing algorithms may be deterministic or

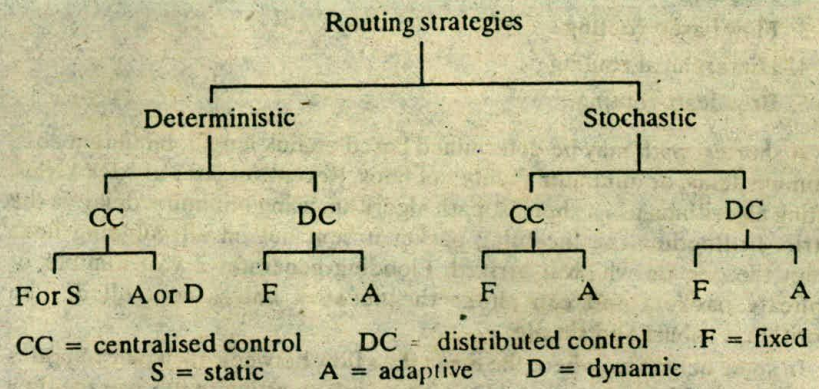


Fig. 10.24 Classification of routing algorithms.

stochastic in nature. Routing decisions may be taken centrally by collecting global information from the network or in a distributed fashion based on the local conditions. Finally, each strategy may use a fixed criterion or adapt itself to varying network traffic conditions. A routing algorithm that uses a precomputed route from a given node to another is a deterministic, distributed, static algorithm. If the route changes, say with time of the day, one may call the algorithm dynamic but still deterministic. If the routing information is computed by a central node based on the knowledge of network topology etc. and distributed to other nodes, then it is centrally controlled. If the route is chosen based on some probability calculation, say a random number, the algorithm is stochastic in nature.

A number of measures may be used in assessing the performance of a routing algorithm:

1. Minimum delay
2. Minimum number of intermediate nodes or hops
3. Processing complexity
4. Signalling capacity required on the network
5. The rate of adaption in the case of adaptive algorithms
6. Fairness to all types of traffic
7. A reasonable response time over a range of traffic intensities
8. Robustness: the ability to reach the destination even when parts of the network fail
9. Stability: the ability to reach the destination quickly without wandering.

Some examples of the widely used algorithms are:

1. Shortest path
2. Flooding
3. Flow based routing
4. Hierarchical routing
5. Broadcast routing.

A shortest path may be determined based on link length, minimum cost, minimum delay, or minimum number of hops. Recent versions of ARPANET routing algorithms use a shortest path algorithm using minimum delay as the metric. In flooding, an incoming packet is sent out on all outgoing links except the one on which it arrived. Flooding generates a vast number of duplicate packets and can choke the network unless controlled. The algorithm is robust and simple.

In some networks, where the mean data flow between each pair of nodes is relatively stable and predictable, it is possible to estimate the average flow on each line and then the mean packet delay for the whole network. Flow based routing attempts to minimise the overall mean packet delay. When networks grow large, it is not possible for each node to keep track of the best route for every other node. Hierarchical routing segments the network to different levels and a node is then only concerned about routing to the appropriate level to reach the destination.

Congestion occurs when there are too many packets in the network. One method of congestion control is to preallocate buffers throughout the route before starting the packet transmission from a source node to a destination node. The method actually avoids congestion by taking a preventive measure and is easily implementable in virtual circuit connections. Another method is to limit the total number of packets that can exist on the network at any time. This technique is known as **isarithmic control**.

Internetworking is unfortunately not well defined in OSI model. **Internet protocol** (IP) of ARPANET is far more popularly used as it has been in use for a number of years and has stood the test of time. This, in conjunction with the **transmission control protocol (TCP)**, is still being widely used. More details of internetworking are presented in Section 10.12.

10.5 End-to-End Layers

Many data networks were operational before the OSI model was designed. These networks were well thought out up to the network layer, but little had been done about the transport layer and above. As a consequence, the design of the bottom three layers of OSI was highly influenced by pre-OSI developments. In contrast, OSI design is expected to dominate the upper four layers in the years to come except for the influence of transmission control protocol (TCP) of ARPANET at the transport layer, which has come to be accepted widely.

10.5.1 Transport Layer

Transport layer is the first end-to-end layer in the OSI architecture. It is responsible for matching user message characteristics and service requirements with that of the network capabilities. In a packet switched network, the transport entity breaks up a long user message into packets to match the network capabilities. The packets are reassembled at the destination transport subsystem to reconstruct the user message. The transport entities may invoke sophisticated error control protocols to provide a reliable session service on an unreliable network. Similarly, a number of low rate user services may be multiplexed to use efficiently a single network connection, or a high data rate requirement from the user may be split into a number of network connections. Multiplexing and splitting, performed by the transport layer are transparent to the sessions layer. Like every other layer, transport layer is also concerned with the establishment, control and release of the transport connections between peer entities in the source and destination systems. End-to-end flow control and error recovery are also functions of the transport layer. The need for flow control arises when the speed or the buffer space of the destination machine does not match with that of the source machine. Similarly, end-to-end error recovery becomes necessary when the source or destination system fails or when the network becomes disconnected due to link failures.

As far as a user is concerned, it is the transport layer that offers transport services regardless of the underlying subnetwork. The user makes his service requests to this layer by specifying certain 'quality of service' (QOS) parameter values. Some QOS parameters that are of direct interest to the users are:

1. Transit delay
2. Residual error rate
3. Protection
4. Transfer failure probability
5. Priority
6. Throughput.

A user may specify a desirable and a maximum value for the transit delay. Residual error rate specifies the undetected errors or irrecoverable losses that a user may be willing to accept. Protection deals with data security. Transfer failure probability is a measure of how well the transport layer is able to meet the user demands in terms of other parameters. Priority enables the user to give more importance to some of the connections than others. There are other QOS parameters like connection establishment and release delays, and the probability of a connection being terminated abruptly due to some internal network problems. These influence the confidence level of a user on the network. When a user makes a request to the transport layer for a connection, one of the following three things may happen:

1. Connection may be established as requested.
2. Options may be negotiated and a connection established with changed parameter values. This happens when the network capabilities are limited and the network is unable to meet the user requirements.
3. Connection is rejected as the network is unable to handle even the minimal requirements of the user.

Once a connection is granted, it is the responsibility of the transport layer to provide the necessary QOS demanded by the user. To ensure this, different transport protocols are designed, which take into account the quality of service available from the networks. For this purpose, networks are placed under three categories A, B and C:

A : Networks that provide reliable error-free service

B : Networks that provide reliable packet delivery but where occasional network wide resets occur

C : Networks that provide unreliable service characterised by packet loss and duplication, and high residual error rates.

OSI transport protocol has five variants to deal with different categories of networks and to provide varying degrees of service sophistication to the user:

Class 0 : Simple protocol with no error recovery or flow control

Class 1 : Basic error recovery protocol

Class 2 : Protocol with multiplexing ability

Class 3 : Protocol with error recovery and multiplexing

Class 4 : Protocol with elaborate error detection and recovery mechanisms.

Classes 0 and 2 are suitable for category A networks, classes 1 and 3 for category B and class 4 for category C networks.

10.5.2 Session Layer

The main function of the session layer is to organise different sessions between cooperating entities and perform all related functions like synchronisation, failure management, control, etc. for the successful execution of a session. Online search of databases, remote job entry, remote login to a time sharing system and file transfer between two systems are all examples of sessions. Different sessions have different requirements. For example, a dialogue session may be two-way simultaneous, or one-way alternate. A large file transfer session may call for rollback points being established in order to recover from system crashes. An online transaction processing session calls for semaphore management, and file, record and sometimes even item level

lock mechanisms. A quarantine service is one which enables a specified number of presentation layer SDUs to be transported to the destination system but not actually deliver them unless explicitly so requested by the sender. The session layer may also offer a directory service.

In a dialogue session, the interacting entities will have to communicate in an orderly fashion. Generally, it is a half duplex communication process with waits. For example, a user sends a query to a database system and waits for a response. A dialogue session in OSI environment is managed by using a data token. Only an entity which has a token can communicate. After the communication is over, the data token is transferred to the other entity. At the time of session establishment, a decision is taken as to which of the two entities gets the token first.

Synchronisation is the operation in which rollback points, otherwise called synchronisation points, are set up in a session in order to recover from failures. Two types of synchronisation points, major and minor, are permitted. Minor synchronisation points fall in between major synchronisation points as shown in Fig. 10.25. A rollback can go beyond one or more minor

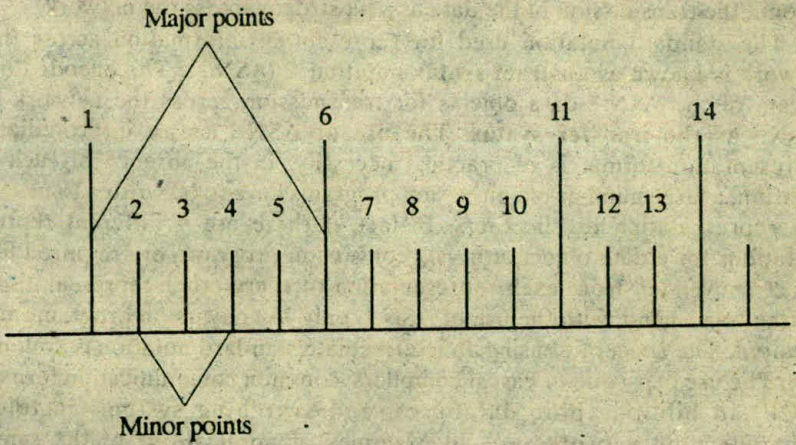


Fig. 10.25 Synchronisation points in session layer.

synchronisation points but cannot go beyond a major synchronisation point. For example, a failure between 10 and 11 in Fig. 10.25 may roll back to any of the points 10, 9, 8, 7 or 6, but not beyond 6. Rolling back is called resynchronisation.

Yet another facility offered by the session layer is known as **activity management**. Here, a user is permitted to divide his total session job into a number of smaller, independent, parallel asynchronous and perhaps event oriented activities. These activities are somewhat like tasks in a multitasking computer system.

Session layer functions are traditionally connection oriented. However, extensive research has been done in connectionless form of dialogue services in the context of distributed systems. This work, which is generally known as **remote procedure call (RPC)**, is now being made available in the OSI environment. RPC is often considered as an application layer function in the OSI model.

10.5.3 Presentation Layer

The purpose of the presentation layer is to represent information to the communicating application entities in a way that preserves the meaning while resolving syntax differences. Syntax differences are resolved by encoding application data into a standard abstract notation that is valid throughout the network. Thus, file format differences (e.g. IBM or DEC format), data representation differences (e.g. ASCII or EBCDIC) or data structure differences are all resolved by using a standard notation. Data transformation and formatting may include data compression, encryption etc. There are two aspects associated with network wide handling of a variety of data in a standard form. First, the representation of the data in a standard form, and second, the transmission of the data as a bit stream across the network.

The standard notation used for representing information across the network is known as **abstract syntax notation 1 (ASN.1)**. The encoded bit stream of the ASN.1 data objects for transmission across the network is known as the **transfer syntax**. The use of ASN.1 as an intermediate representation format is of practical necessity. In the absence of such a notation, a large number of conversion programs have to be written to match one representation to all others. In fact, if there are n different representations for a data object, $n(n-1)$ conversion programs are required for direct conversion from each representation to every other representation. On the other hand, with the use of ASN.1, only $2n$ conversion programs are required. The concept of using an intermediate standard notation is nothing new. The use of p-code in Pascal compilers, common communication format (CCF) in bibliographic databases, and switching systems in telecommunication networks are all examples of applications of the same concept.

ASN.1 data representation technique is somewhat similar to the ones used in programming languages like Pascal and C. There are a few predefined primitive data types and a set of constructors which can be used to build complex data types from the primitive ones. The primitive data types and the constructors are given in Table 10.3. The first four data types are well known. The primitive data type 'ANY' enables a user to defer type definition of a data object to a later instance like run time. Any of the valid data types can be assigned to the data object defined as ANY. Initially, 'NULL' is a data item of no value. Object identifier is a name used to identify objects or subroutines or libraries associated with the processing of ASN.1 or transfer systems. The meanings of the ASN.1 constructs are given in Table 10.3.

Table 10.3 Primitive Data Types and Constructs in ASN.1

ASN.1 data types	Constructs	Meaning
Integer	Sequence of	Ordered list of a single data type
Boolean	Sequence	Ordered list of a mix of data types
Bit string	Set of	Unordered collection of a data type
Octet string	Set	Unordered collection of various data types
ANY	Choice	Any one type taken from a list
NULL		
Object identifier		

The transfer syntax for ASN.1 consists of four fields as shown in Fig. 10.26. ASN.1 has a flexible syntax notation permitting a large number of

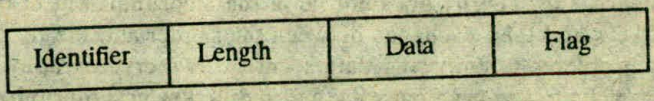


Fig. 10.26 ASN.1 transfer syntax.

optional parameters. If the optional parameters are not used, a NULL value may be transmitted in place of these fields. But this would result in excessive overhead in transmission and subsequent processing of the PDU. A better scheme is to omit the optional fields that are not used, in which case a method of identifying the actual parameters that are transmitted is required. This is achieved by 'tagging' the parameters. The tags in ASN.1 are placed under four broad categories:

1. Universal
2. Application
3. Private
4. Context specific.

Each tag under a category is specified by a number. The identifier field of the transfer syntax has three sub-fields; the first two bits specify the tag type, the next bit specifies whether the data object is of primitive type or the constructed type, and the last five bits the tag number. The second field specifies the length of the data field and the last field 'flag' indicates the end

of data in cases where the length is not specified. A provision exists for specifying tag numbers greater than 5-bit quantities and lengths greater than 8-bit quantities.

As in the case of routing, data compression and encryption techniques are not standardised in the ISO model and are left as implementation dependent subjects. Widely used data compression techniques such as run length coding and Huffman coding are briefly discussed in Section 11.2 in the context of emerging facsimile services. Here, we consider some aspects of data encryption. The main purpose of data encryption is to protect user data from intruders. There are two kinds of intruders: passive and active. A passive intruder listens to user messages, analyses the traffic in terms of length of messages, frequency of the message and the communicating pair of users. An active intruder modifies, delays, corrupts or takes out user messages and sometimes insert his own message to confuse the communicating pair. Passive intrusion often precedes active intrusion. Encryption techniques should cater for both types of intruders. A general encryption scheme is shown in Fig. 10.27. The user text U is operated upon by the encryption algorithm E using the encryption key K to obtain the ciphered text C which is generally not understandable by the intruder. At the receiving end, the ciphered text is operated upon by the decryption algorithm using a decryption key K_1 to obtain the user text back. There are not many encryption/decryption algorithms, and the ones available are well known. As a consequence, the security does not lie in the algorithms but in the keys. Design of keys and their exchange between the sender and the receiver are important problems in enforcing data security. Two cryptography systems are known: **private key cryptography** and **public key cryptography**. In the former, the encryption and decryption keys are the same, i.e. $K = K_1$ in Fig. 10.27. In the latter, the keys are different.

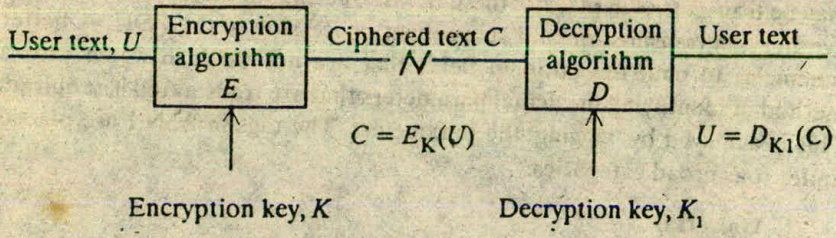


Fig. 10.27 A general encryption/decryption scheme.

Encryption algorithms may be broadly classified as

- Substitution algorithms
- Transposition algorithms.

In simple words, substitution algorithm replaces every character in the text by one or the other character according to a rule based on a key.

Transposition algorithms simply jumble the characters according to a rule based on a key. Complicated encryption algorithms can be designed by using a combination of substitution and transposition algorithms. The idea of such algorithms is to make the process of deciphering very difficult even if the algorithm is known. Data Encryption Standard (DES) adopted by US National Bureau of Standards, is one such complicated algorithm involving 19 distinct stages of processing, using a 56-bit key and processing 64-bit user text segments.

Key distribution is a major problem in any private key cryptography scheme. Public key cryptography system, first proposed in 1976 and considered to be a major breakthrough in cryptography, eliminates the problem of key distribution. However, useful keys are very long (100 digits or more) and considerable computer time is required to work out the keys. Hence the scheme is yet to become practical. But the scheme is indeed elegant and we describe the same briefly. In this scheme, a user has a pair of keys; one which is public p, made known to all others in the network and the other secret s, which is known only to the particular user and is never distributed. The two keys have properties such that

$$D_s[E_p(U)] = U \tag{10.16}$$

$$D_p[E_s(U)] = U \tag{10.17}$$

$$D_{p'}[E_p(U)] \neq U \tag{10.18}$$

$$D_{s'}[E_s(U)] \neq U \tag{10.19}$$

where

$p' = p$ or any other key but not s

$s' = s$ or any other key but not p

Equation (10.16) states that a message once encrypted using the public key of the user can be deciphered by using the secret key of the same user. Equation (10.17) states the vice versa condition. Equations (10.18) and (10.19) bring out that the same key cannot be used for both encryption and decryption. They also state that any key other than the corresponding secret or public key cannot decode the message. The scheme is not only capable of maintaining secrecy but also capable of communicating the authenticity or the digital signature of the sender. The latter aspect is extremely important in many transactions relating to banking, military operations etc. Figure 10.28 shows how the scheme works between the sender A and the receiver B.

The sender message is first encrypted using his own secret key, sA to obtain C' which is further encrypted using the public key of B, pB. The ciphered text thus obtained is transmitted to B, which can only be deciphered by using the secret key of B as governed by Eqs. (10.16) and (10.18). It cannot be read by anyone else other than B. Thus, secrecy is maintained. Having

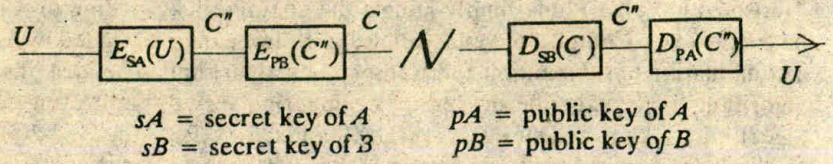


Fig. 10.28 Digital signature and secrecy in public key cryptography.

obtained C', the only way to obtain U is to decrypt using the public key of A, as governed by Eqs. (10.17) and (10.19). The fact that a meaningful message is obtained using pA for decryption implies that the message could have been sent only by A, thus authenticating the message. The only hitch in the entire scheme is the determination of $(p-s)$ pair. Many researchers have proposed algorithms for the same, but it is beyond the scope of this text to discuss them.

10.5.4 Application Layer

As the highest layer in the OSI reference model, the application layer provides services to the users of OSI environment. The layer provides all services that are directly comprehensible by the users, which include

1. Electronic mail or message handling services
2. Directory services
3. Cost allocation
4. Determination of quality of service
5. File transfer and management
6. Editors and terminal support services
7. Telematic services like videotex.

In general, every application requires its own software which in turn uses a number of generic supporting packages. These generic packages and the application specific packages are part and parcel of the application layer. What is listed above is a set of generic applications and support packages. For example, file transfer or remote file access may be used by airlines reservation system, banking applications etc. Similarly, electronic mail may be used in order placement and processing systems. There is yet another level of support packages that are used by some of these generic applications packages. They include connection establishment, transaction processing, database crash recoveries etc. Two such packages are:

- Association control service element (ACSE)
- Commitment, concurrency and recovery (CCR).

In the following paragraphs of this section, we discuss file transfer, access and management (FTAM) and virtual terminal support functions. Electronic mail and videotex are described in Section 11.2. Directory

services in OSI environment provide the presentation SAP addresses of the users. The access is based on attributes which may be the name, address, profession, organisational affiliation, telephone number etc.

File transfer and remote file access are two of the most common applications in a network environment. In file transfer, an entire file, irrespective of its nature such as binary, textual, or graphical etc., is copied from one system to another. Remote file access is similar to file transfer except that only parts of files are read or written. The techniques used for remote file access and file transfer are similar. FTAM involves three aspects:

1. File design
2. Concurrency control
3. Replication.

File design deals with structure, attributes and operations on the file. OSI file design is based on the concept of a **virtual file store**. The virtual file store presents a standardised interface and provides a set of standardised operations. It enables a user to access and operate on files without having to know the internal machine structures etc.

Like most file systems, OSI virtual file store supports hierarchical file structure. The attributes of the OSI virtual file are given in the Table 10.4.

Table 10.4 Attributes of OSI Virtual File

Attribute name	Data type
File name	String
Allowed operations	Bit Map
Access control	List
Account number	Integer
Date and time of the file creation	Time
Date and time of last file modification	Time
Date and time of last file read	Time
Date and time of last attribute modification	Time
Owner	Userid
Identity of last file modifier	Userid
Identity of last attribute modifier	Userid
Identity of last reader	Userid
File availability	Boolean
Contents type	Objectid
Encryption key	String
Size	Integer
Maximum future size	Integer
Legal qualifications	String
Private use	String

Most of the attributes names are self-explanatory. The field, 'legal qualifications' is a string field stating some legal aspects regarding the information contained in the file, e.g. copyright. The field 'private use' is again a descriptive field giving information that is not covered by the other attributes.

The standard operations of the OSI virtual filestore are given in Table 10.5. The operations are placed under three categories; file level, attribute level and record level.

Table 10.5 OSI Virtual Filestore Operations

File level	Attribute level	Record level
• Create	• Read	• Locate
• Delete	• Modify	• Read
• Select		• Insert
• Deselect		• Replace
• Open		• Extend (Append)
• Close		• Erase

Concurrency control is handled using 'locks' of two kinds: shared and exclusive locks. Shared locks are used for reading and the exclusive locks for writing. Locking may be provided at three levels: file level, sub-tree level of a hierarchical structure, and record level.

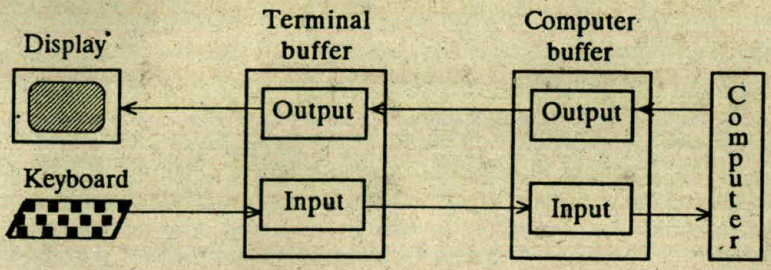
File replication is important from the redundancy and failure points of view. A simple solution, but one not without problems, is to have a master file and treat all other copies as slaves. Updates are done only on the master file, which are then propagated to other slave files. This scheme fails if the master file is not available for some reason. A more sophisticated scheme is based on voting quorum. Here, a specified number of file servers are invoked simultaneously for carrying out an operation like update or modify.

A major problem encountered in networks is the incompatibility between the user terminals and the computer systems. Unfortunately, attempts to standardise terminals have not met with success. OSI approach to solving the problem is to define a **virtual terminal** which presents a uniform interface network wide. The virtual terminal is essentially an abstract data structure that reflects the present display status which can be manipulated from the keyboard as well as the computer. The current contents of the data structure are displayed on the screen. Although there are a number of minor variations in terminal designs, the terminals can be placed under three broad categories:

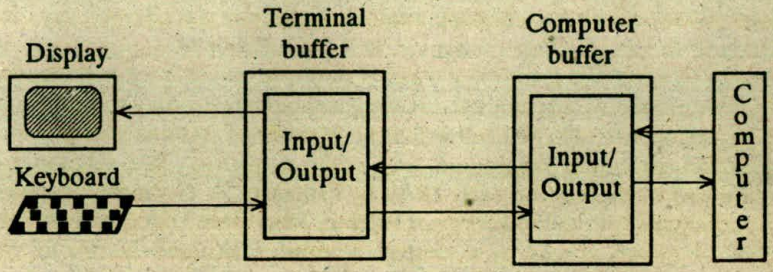
1. Scroll mode or dumb terminals
2. Page mode or semi-intelligent terminals
3. Form mode or intelligent terminals.

Dumb terminals are connected to the network via an interface unit known as Packet Assembler/Disassembler (PAD), which provides a standard interface to the network on the one side and on the other side the RS 232-C serial interfaces to the terminals.

Virtual terminal feature of OSI uses two models for interaction with intelligent terminals. The models, known as asynchronous and synchronous models, are shown in Figs. 10.29 (a) and (b) respectively. In the asynchro-



(a) Independent buffers for input and output: asynchronous mode



(b) Common buffer for both input and output: synchronous mode

Fig. 10.29 Virtual terminal data structure models.

nous model, input and output can take place simultaneously. In the synchronous model, specific conventions are needed so that only one update can take place at a time; i.e. either the terminal processor or the computer can update the buffer but not both, as the buffer is common. The asynchronous model is more popular. In both the models, replica of data structures exists at the terminal and computer ends.

10.6 Satellite Based Data Networks

In the context of data networks and the OSI reference model, there are some important aspects of satellite communication systems which require consideration:

1. Satellite network topology and configurations, modulation schemes and bandwidth utilisation; these are aspects related to the physical layer functions of the OSI reference model.
2. Being a common communication resource accessible by all or a group of earth stations simultaneously, media access becomes a nontrivial function in the data link layer.
3. Satellite communication being broadcast in nature, routing becomes a trivial function; however, organising point-to-point or point-to-multipoint connections in a broadcast environment needs some consideration.
4. Since a geostationary communication satellite is placed at an altitude of about 36,000 km above the equator, the signal will have to travel a distance of 72,000 km or more between the source and the destination, resulting in a significant propagation delay of 250–300 ms. Session and transport layers have to be concerned about this factor.

A satellite based network has a 'star' topology, the satellite being the central hub. If the hub fails, the entire communication network comes to a standstill. In view of this, reliability and redundancy considerations are important when designing satellites.

Demand assigned multiple access (DAMA) techniques described in Section 9.4.4 function well when the number of stations is moderate, say a few hundreds. The schemes become unwieldy if the number of stations runs into thousands. In such cases, another class of multiple access protocols, known as ALOHA protocols, becomes attractive, particularly for data transmission. These protocols can also be used when the number of stations is small.

ALOHA protocols originated with radio networks. The scheme was proposed and studied in the early 1970s by Professor A. Norman Abramson and his colleagues at the University of Hawaii. They were concerned with the problem of giving access to a central computer situated in one of the Hawaian islands from a number of terminals located in other islands using a single common radio communication channel. Satellite channels also present a similar scenario. ALOHA means 'hello' in Hawaiian language.

In the simplest form of ALOHA, often called 'pure ALOHA' to distinguish from other forms of ALOHA protocols, a station or terminal starts transmission as soon as it is ready. In this case, anyone of the following three things may happen:

1. There may be an ongoing transmission already in the channel and the new transmission may collide with the existing one, thereby affecting both transmissions.
2. The channel is free and this transmission goes on successfully.
3. The channel is free to start with, but another station gets ready and starts transmission when this transmission is in progress. The transmissions collide with each other and no useful data transfer takes place.

These events are illustrated in Fig. 10.30. Part of the frame *A* collides with an ongoing transmission. The end part of the frame *B* is corrupted by

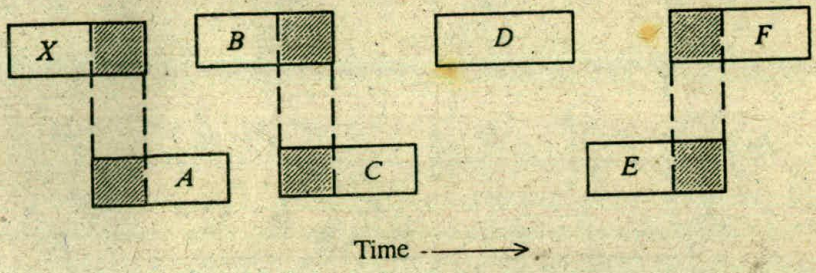


Fig. 10.30 Collision in pure ALOHA access.

transmission *C* with the result that both the frames are not transmitted successfully. Frame *D* is safely transmitted. Frames *E* and *F* collide once again. Whenever a collision is experienced, the stations concerned wait for a random amount of time and retransmit the same frame once again. As one may intuitively feel, the throughput obtainable from this scheme is likely to be very low and, in fact, is limited to a maximum of 18.4 per cent if the traffic is Poisson. This can be shown to be so by a simple analysis. The analysis is done in the context of fixed frame sizes and Poisson traffic. Poisson arrival is given by the equation

$$P_k(t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{10.20}$$

where

$P_k(t)$ = probability that there are k arrivals in time t

λ = mean arrival rate in packets per second

Let

τ = the frame time, i.e. time required to transmit a frame

G = load offered to the system per frame time, including the retransmitted frames

$$G = \lambda\tau \tag{10.21}$$

Consider a frame *F* being placed on the channel at any instant of time. The frame will be transmitted successfully only if two conditions are met:

1. There is no ongoing transmission at the time of placement of this frame.
2. No new station should begin transmission during the transmission of this frame.

Condition 1 will be true if and only if no transmission has originated during the time period τ seconds prior to the placement of the present frame

on the channel. Condition 2 implies that no transmissions should occur during the period τ seconds after the frame is placed on the channel. This situation is depicted in Fig. 10.31. The time duration 2τ is called the

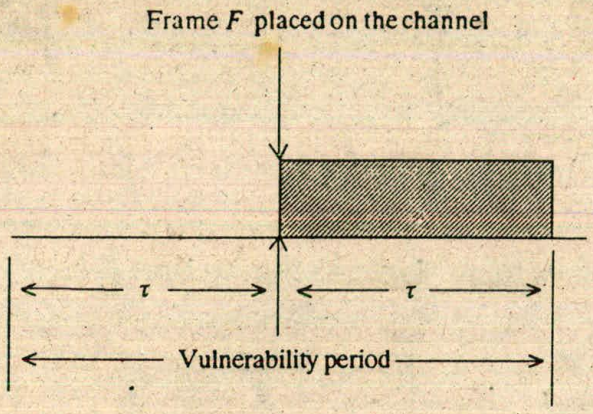


Fig. 10.31 Vulnerability period in pure ALOHA scheme.

vulnerability period. We are interested in the probability of zero arrival during the vulnerability period. From Eq. (10.20) we have

$$P_0(2\tau) = \frac{(\lambda 2\tau)^0 e^{-2\tau\lambda}}{0!}$$

Using Eq. (10.21), we get

$$P_0(2\tau) = e^{-2G} \tag{10.22}$$

Equation (10.22) is the probability of successful transmission of one packet. As there are G packets being offered to the system per frame time, the throughput of the system in terms of frames per frame time is given by

$$S = Ge^{-2G} \tag{10.23}$$

The maximum throughput is obtainable from the system at some optimum loading value G. Lower values of G mean reduced load on the system and higher values of G would result in excessive collisions. To determine the maximum throughput, we differentiate Eq. (10.23) with respect to G:

$$\frac{dS}{dG} = e^{-2G} - 2Ge^{-2G} \tag{10.24}$$

Setting Eq. (10.24) to zero, we get the value of G for maximum throughput as 0.5 frame per frame time. Therefore, the maximum throughput in the pure ALOHA access scheme is

$$S_{max} = 1/2e = 18.4\% \tag{10.25}$$

The throughput of the pure ALOHA scheme can be doubled by introducing time slots in the channel and permitting a new transmission to start only at the beginning of a time slot. Such a scheme is known as **slotted ALOHA**. A time slot is equal to the frame time τ. The vulnerability period for a slotted ALOHA scheme is shown in Fig. 10.32. A frame F is placed on the

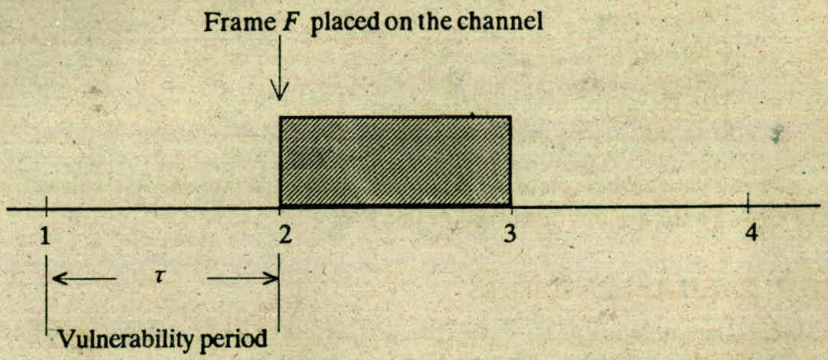


Fig. 10.32 Vulnerability period in slotted ALOHA.

channel at the beginning of the time slot period 2–3. The frame will be transmitted successfully only if no other transmission commences during the beginning of the same slot. Another transmission can exist only if any other station got ready during the slot period 1–2. A station getting ready during the slot period 2–3 does not affect the transmission of F as it can commence its transmission only at the beginning of the slot period 3–4. Hence, the vulnerability period in slotted ALOHA slot is the period preceding the slot in which the frame is placed on the channel. Therefore, the probability of successful transmission of the frame is the same as the probability of zero arrival in the time period τ:

$$P_0(\tau) = e^{-G} \tag{10.26}$$

The probability of success for the full load G gives the throughput in frames per frame time:

$$S = Ge^{-G} \tag{10.27}$$

The maximum throughput is obtained by differentiating Eq. (10.27) with respect to G, setting the result equal to zero, and getting the optimum value of G. In this case, the value for G works out to be 1 frame per frame time. The maximum throughput works out to be

$$S_{max} = 1/e = 36.8\% \tag{10.28}$$

Apart from the throughput, there are two other parameters that are of interest in ALOHA systems: percentage of empty slots and percentage of

collisions. The values of these parameters for the optimum load conditions are shown in Table 10.6 for both pure and slotted ALOHA. Readers are urged to work out these values.

Table 10.6 Performance Parameters of ALOHA Protocols

Parameter	Pure ALOHA (%)	Slotted ALOHA (%)
Maximum throughput	18.4	36.8
Empty slots	36.8	36.8
Collisions	44.8	26.4

There is another class of ALOHA protocols known as **reservation ALOHA**. These are not discussed in this text.

10.7 Local Area Networks

A local area network (LAN) typifies a distributed environment and finds applications in a number of areas. Some examples are:

1. Office automation
2. Factory automation
3. Distributed computing
4. Fire and security systems
5. Process control
6. Document distribution.

Of these, the first two applications have evolved to the extent that world standards for these are emerging. Office automation standard, known as **Technical and Office Protocol (TOP)**, and the factory automation standard, known as **Manufacturing Automation Protocol (MAP)**, are discussed in Section 10.11.

Being distributed in nature, LANs offer a number of benefits and for the same reason suffer from certain disadvantages too. The advantages offered by the LANs are:

1. Unlike a large centralised system, a LAN may evolve with time. It may be put into operation with a small investment, and more systems may be added as the need arises.
2. Since LAN is a set of multiple interconnected systems, it offers a good back up capability in the event of one or two systems failing in the network. This, in turn, enhances the reliability and availability of the systems to users.
3. LAN provides a resource-sharing environment. Expensive peripherals, hosts and databases may be shared by all the LAN users.

4. A LAN adhering to a certain standard permits multivendor systems to be connected to it. Thus, a user is not committed to a single vendor.
5. In LAN, the systems are generally so chosen as to meet most of the user requirements locally and the network is used only for resource and information sharing purposes. Due to this, each user gets a better response than would be the case in a centralised system. LAN tends to exhibit an improved performance.
6. LAN offers flexibility in locating the equipments. Most computers on a LAN are physically placed at the user table, which is most convenient for working and improves the productivity significantly.

LANs also suffer from certain disadvantages:

1. The incremental growth may lead to uncontrolled growth and may ultimately result in more investment than would have been the case with a centralised system.
2. LAN standards offer many options. Although systems from different vendors conform to the same standard, options exercised may vary from vendor to vendor with the result the interoperability is not guaranteed. Incompatibility may arise at the hardware, software or data organisation level.
3. A distributed system environment raises problems of security, privacy and data integrity. For example, a computer virus introduced in one system may very quickly spread to other systems.

10.7.1 LAN Technologies

There are three major aspects in a LAN:

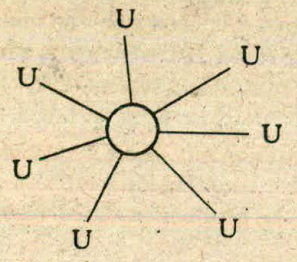
1. Medium of transmission
2. Topology
3. Access method.

Three important media of transmission are in use: twisted pair wire, coaxial or CATV cable, and fibre optic cable.

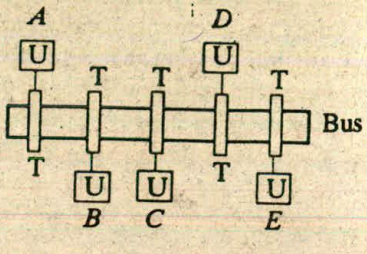
Twisted pairs are used in low speed LANs using baseband transmission. In this mode of transmission, data is transmitted as simple electrical levels often without any modulation. There is no multiplexing and the entire bandwidth of the medium is used for transmitting signals from one station. Baseband transmission links are sometimes termed 'wire only' links as there are no other equipments used for transmission.

In contrast, broadband transmission uses modulation techniques and is suitable for transmitting high speed and multiplexed data. CATV and coaxial cables are used for broadband transmission at speeds of 10 Mbps or more. Fibre optic cables carry data at rates up to 100 Mbps. Radio communication may also be used for transmitting data in LANs. Although some laboratory systems have been built using radio links, such systems are not in use.

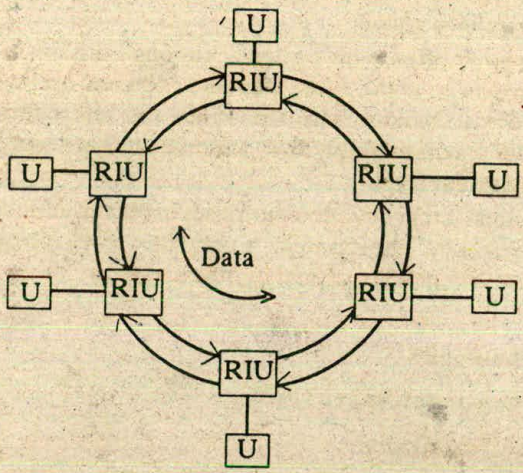
Three topologies are widely used in LANs: star, bus and ring. These topologies are shown in Fig. 10.33. In star topology, the LAN systems are



(a) Star topology



(b) Bus topology



RIU = ring interface unit

(c) Ring topology

Fig. 10.33 LAN topologies.

connected to a central switch which establishes connection between pairs of systems. Such LANs are generally designed around electronic PABXs or computerised branch exchanges (CBX). A bus, in the bus topology shown in Fig. 10.33(b), is a single pair or a bunch of wires that carry the electrical signals. The LAN systems are connected to the bus using a passive tap such that all systems are able to monitor the signals on the bus simultaneously. The bus acts as a broadcast medium. In a ring topology, the LAN systems or their interfaces take an active part in the transmission. A ring is formed by connection of a number of point-to-point links between pairs of ring

interface units (RIU) as shown in Fig. 10.33(c). RIU being an active device, its failure puts normal ring operation out of order. It becomes necessary to consider fault tolerant techniques such as dual rings or folded ring operation for handling RIU failures in a ring topology. In contrast, being a passive unit, the failure of an interface in a bus topology does not affect the operation of the LAN. There are a number of other topologies that have been studied in the context of LANs. These include tree, hypercube, multistage store and forward networks. However, LANs based on these topologies are not widely used at present.

Three access methods are predominantly used in LANs: switched access, contention or multiple access, and token passing access. Switched access is used in LANs that are designed around CBXs. Electronic switching systems have been extensively discussed in Chapters 4 and 6. In this chapter, we are concerned only with contention and token passing access techniques. There are also other access techniques like reservation access, load adaptive access, and tree structure based access. Discussion of such schemes is beyond the scope of this text.

Different types of LANs are obtained by choosing different combinations of medium, topology, and access method. If we consider three media, viz. twisted pair, coaxial and optical fibre, three topologies, viz. star, bus, and ring and two access methods, viz. contention and token passing, we would have a total of 18 different types of LAN. But not all of them are technically feasible or practically possible. For example, there are still unsolved engineering problems with regard to the use of optical fibres as bus media. Token passing access is a natural choice on the ring topology although contention access is technically feasible. As a result, only three combinations of access techniques and topologies are popularly used:

1. Multiple access bus
2. Token passing ring or token ring
3. Token passing bus or token bus.

We discuss these LANs in the following sections. While twisted pair, coaxial cable and optical fibres are all suited for ring topology, only twisted pair and coaxial and CATV cables are suited for bus topology. The choice of a particular medium would determine the physical layer techniques such as signal encoding. Media access is a function of the data link layer. Being a nontrivial function, media access in LANs is conceived as a sublayer of the data link layer as in the case of satellite networks. The current trend in ISO-OSI model is to place media access sublayer as part of the physical layer. Use of optical fibre as a medium calls for special considerations in medium access and data transmission. Hence, we discuss fibre optic LANs separately in Section 10.9.

10.7.2 Multiple Access Bus LAN

Consider Fig.10.33(b) which shows a bus LAN. A bus is a broadcast medium and only one data transmission can take place at any instant of time. If the stations A, D & E get ready to transmit simultaneously and want to get hold of the bus at the same time, who should be given access first? Obviously, there is a contention to access the bus and this needs to be resolved. The problem is similar to that of accessing a satellite channel. All protocols, DAMA and ALOHA protocols, discussed in the context of satellite networks, could be used for the bus LANs. But, a LAN environment differs from satellite environment in one very important respect. When a transmission is initiated by a station in a satellite network, it is heard by all other stations after a delay of one hop (or two in the case of 2-hop network) of 270 ms. A satellite station is unable to know whether the channel is occupied or not at a given instant, for whatever it hears is something that originated 270 ms ago. In contrast, being in a localised environment with not more than 5–15 μs of end-to-end propagation delay, a LAN transmission can be heard by all other LAN stations almost instantaneously (in relative terms). In other words, a LAN station is able to know whether the channel is busy or not at any instant. This carrier sense capability of a LAN station can be advantageously used to design more efficient multiple access schemes than DAMA or ALOHA schemes. Such schemes are known as **carrier sense multiple access (CSMA) schemes**.

In CSMA, whenever a station gets ready for transmission, it first listens to the bus to see if there is any ongoing transmission. If there is one, the new transmission is not initiated until the bus becomes free. This ensures that an ongoing transmission is not corrupted by a new transmission. If the channel is found free, the new transmission is started immediately. Since the channel is sensed before transmission, the scheme is sometimes referred to as **listen-before-talk scheme**. If more than one station gets ready during an ongoing transmission, all of them may attempt to transmit immediately after the bus becomes free and a collision would occur. The collided stations wait for random times and then sense the channel again. It may be noted that this scheme is similar to the slotted ALOHA scheme discussed in Section 10.6. The scheme, however, performs better for two reasons. First, if the channel is found free, the transmission is initiated immediately. In slotted ALOHA, even if a station gets ready when a slot is free, the transmission cannot start until the beginning of the next slot. Second, variable length frames may be transmitted accommodating both short and long frames. In ALOHA, the slots are of fixed time duration. Closed form analytical throughput expressions can be derived for CSMA protocol using a line of reasoning similar to the one used for slotted ALOHA. However, such a derivation is beyond the scope of this text. Interested readers are referred to Further Reading [1]. We, however, give the expression for maximum throughput:

$$S_{max} = 1/(1 + 2B) \qquad (10.29)$$

where B is the time, expressed as a fraction of the frame time, required for all stations to detect an idle channel after a transmission ends. The throughput is expressed in frames per unit time. Another performance measure is the maximum channel utilisation which is given by

$$U_{max} = \frac{t_f}{t_f + t_p} = \frac{1}{1 + a} \tag{10.30}$$

where

U_{max} = maximum channel utilization
t_f = frame time
t_p = end-to-end propagation time
$a = t_p/t_f$

Between the transmission of two frames, a time gap of t_p is required.

Three variations are possible in CSMA protocols:

1. 1-persistent
2. Nonpersistent or zero-persistent
3. p-persistent.

When a station finds the channel busy, it may continue to sense the channel and transmit the frame immediately after the channel becomes idle. In this case, the CSMA protocol is known as 1-persistent. Instea he station may decide to sense the channel again after a random time when it finds the channel busy. This case is known as nonpersistent CSMA. The p-persistent CSMA makes a decision to persist for immediate transmissio with a certain probability p. This is the generalised case. The operation of rsistent scheme is illustrated by means of a flowchart in Fig.10.34. When annel is busy, a random number (RN) is generated between 0 and 1. If the RN is less than or equal to p, then the station persists until the channel becomes idle and then transmits. For example, if $p = 0.75$, the station would persist in 75% of the attempts. Otherwise, the station waits for a random time before making another attempt. Obviously, $p = 1$ means the station always persists, and $p = 0$ means that the station never persists.

There are other writers who define p-persistence somewhat differently. They consider it as the case where different probabilities are used for new and retransmission frames. There are yet others who consider p as the probability with which a station transmits when the channel is idle. When you want to take a taxi and find one waiting at the stand, would you toss a coin to decide whether to take the taxi or not? On the other hand, when you do not find one in the stand, you may like to toss the coin to decide whether you should wait at the stand until the next taxi arrives or walk back to your office, do some useful work and try again a little later.

A few words about the delay performance and the throughput of the CSMA schemes are in order. In 1-persistent CSMA, everyone is waiting to

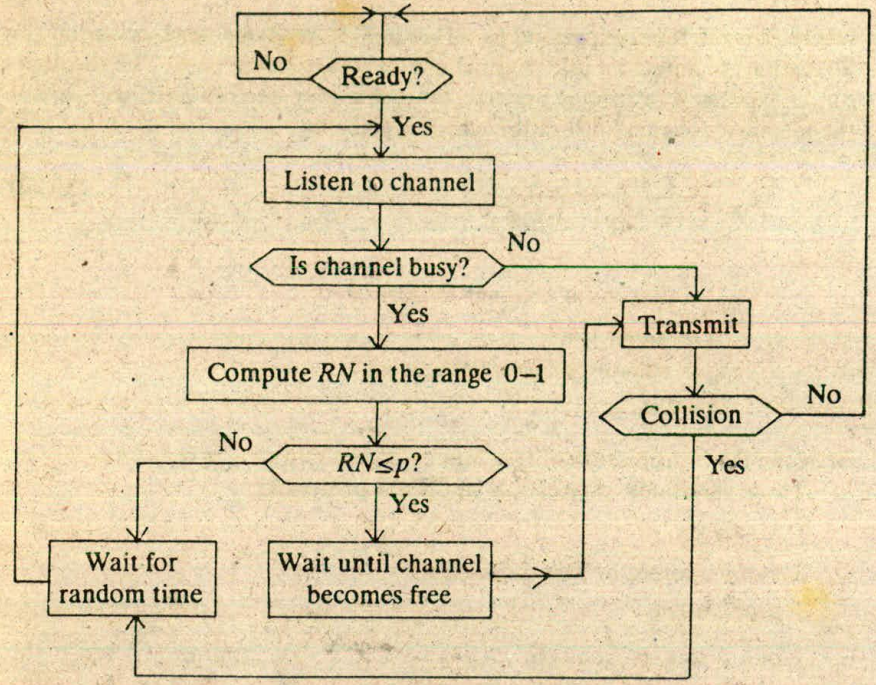


Fig. 10.34 *p*-persistent CSMA protocol.

seize the channel at the earliest opportunity. One would expect more collisions and this, in fact, is the case. The throughput is affected due to excessive collisions. In 0-persistent case, the station at the 'slightest provocation' (channel being busy) withdraws from an attempt to transmit. While the collisions are low, the channel remains idle more often. The throughput is low and the mean delay experienced by a frame, before it is delivered to the destination, is large.

The ability to sense the channel also enables an efficient collision detection mechanism to be implemented in bus LANs. A transmitting station can continue to sense the channel to see what appears on the bus is the same as what is being transmitted. This is sometimes termed as **listen-while-talking**. If what the bus 'listens-to' is not the same as what it is 'talking', a collision is said to have occurred. In a CSMA scheme, a collision occurs when two or more stations sense the channel free and start transmission. Consider the two end stations A and B in a bus LAN shown in Fig.10.35. Let station A sense the channel free at time t and start a transmission. The signal reaches B after t_p seconds of propagation delay. If B senses the channel at time $t + t_p - \Delta$, it would find it free and could start a transmission. Station B would detect a collision after Δ seconds, whereas station A can detect a collision only after $2t_p$ seconds. Two important points emerge from this:

1. Transmission from a station must last at least for $2t_p$ seconds. Otherwise a collision may go undetected.
2. If no collision is detected within the first $2t_p$ seconds, the frame is assumed to be safe.

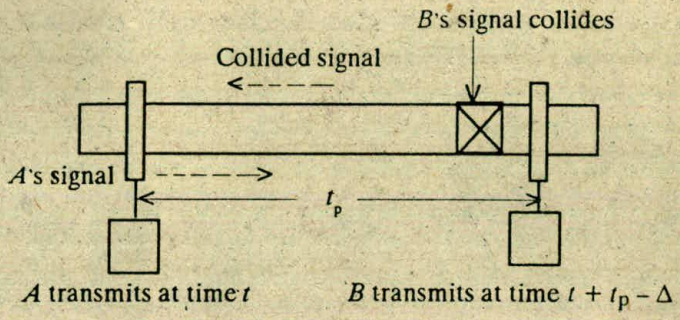


Fig. 10.35 Collision detection in CSMA LAN.

Point 1 above places a lower bound on the size of the frame for a given data rate in a CSMA LAN with collision detection (CD) feature as above. A LAN operating on this principle is known as **CSMA/CD LAN**. The commercial **Ethernet LAN** originally proposed by Xerox, Intel and DEC jointly is a CSMA/CD LAN. Obviously, CSMA/CD LAN performs better than the pure CSMA LAN, where a collision is detected only after the full frame has been transmitted. It may be noted that a collision is detected only by the stations which are participating in the transmission. It is necessary for other stations, in particular the destination station, to know about a collision. Therefore, as soon as a station detects a collision, it sends out a special signal, usually known as **jamming signal**, on the bus to let know all the stations about the collision.

EXAMPLE 10.2 A CSMA/CD bus spans a distance of 1.5 km. If the data rate is 5 Mbps, what is the minimum frame size?

Solution

Typical propagation speed in LAN cables = 200 m/μs

End-to-end propagation delay t_p = 1500/200 = 7.5 μs

Minimum frame size = $2 \times 7.5 \times 10^{-6} \times 5 \times 10^6$ = 75 bits

Minimum frame size for CSMA/CD LAN may be expressed as

$$F_{min} = 10dR \qquad (10.31)$$

where

d = length of the cable in km
R = data rate in Mbps

In arriving at Eq. (10.31), propagation delay is taken to be 5 μs/km.

In all our discussions so far regarding multiple access, we have said that a station waits for a random time on detecting a collision before it retransmits the frame. Wait time may be randomised according to one of many strategies. One strategy which is commonly used in CSMA/CD LANs is known as **binary exponential back off**. Without loss of generality, the random wait time may be discretised in terms of basic time units (BTUs). Usually, a BTU is taken to be the average frame time. When a station experiences the first collision in transmitting a frame, it generates an integer 1 or 2 randomly and depending on the outcome, attempts a retransmission after 1 or 2 BTUs respectively. If it experiences a collision for the second time, random integer is generated in the range 1–4 and the retransmission is attempted after the corresponding number of BTUs. Table 10.7 shows the BTU range used after every collision and the corresponding probability of access in any BTU. Since the BTU range is doubled after every collision, and the probability of access is halved, the scheme gets the name exponential backoff. As may be apparent, the scheme is adaptive to load conditions. Successive collisions imply a large number of active stations and hence the need to lower the probability of immediate reattempt.

Table 10.7 Exponential Back-off

Collision number	1	2	3	4	5
BTU range used	1–2	1–4	1–8	1–16	1–32
Probability of attempt in any BTU within the range	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}

It is important to note that in ALOHA, CSMA and CSMA/CD schemes, there is a definite probability, however small it may be, that a station may experience indefinite collision. Therefore, the frame delay is said to be **unbounded** in these schemes. In practice, the probability of long delays may be reduced to an insignificant level by limiting the maximum load on the system. However, one will have to be careful before adopting these schemes for time critical applications. A typical application for CSMA/CD LANs is in the office environment. Similar to reservation ALOHA schemes, contention free access techniques are possible in bus LANs.

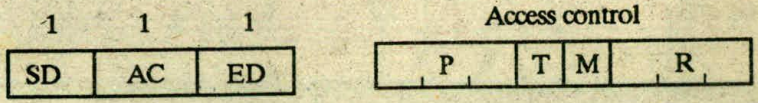
10.7.3 Token Passing Ring LAN

The physical aspects of a ring LAN are described in Section 10.7.1. The RIUs (see Fig.10.33(c)(p.452)) receive, regenerate and retransmit the data bit by bit (serial transmission) on a ring. In the process, an RIU is capable of

copying and examining every bit that passes through it. By storing a group of bits received in sequence, an RIU may also examine bit patterns that travel on the ring. These functions of RIU consume time and hence it is normal for each RIU to introduce a 1-bit delay in data transmission. When a user station is not participating in the ring operation, the corresponding RIU may be bypassed, in which case the 1-bit delay from that RIU does not exist.

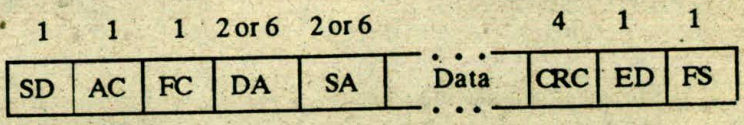
Ring being a continuous structure and the RIUs being active repeaters, data once placed on the ring would go round the ring indefinitely, unless removed specifically. In contrast, data placed on a bus dies down automatically after a shortwhile. Thus, both data placement and removal require consideration in a ring.

The token passing access mechanism enables a station to transmit its data on the ring. In this mechanism, one or more tokens are used to give ring access to the stations. In this text, we limit our discussion to a single token protocol. Incidentally, the current internationally standardised token ring is also based on single token access. A token is usually a 3-byte pattern as shown in Fig. 10.36(a). Starting delimiter (SD) is a unique pattern that does



SD = starting delimiter byte AC = access control byte
ED = end delimiter byte T = token bit M = monitor bit
P = priority bits R = reservation bits

(a) Token format



FC = frame control DA = destination address
SA = source address CRC = cyclic redundancy code
FS = frame status

(b) Frame format

Fig. 10.36 Token and frame formats in a token ring LAN.

not appear as part of data. This is ensured by using Manchester encoding for transmitting data and using a known invalid Manchester code pattern for SD. Figure 10.36(a) also shows the details of the access control (AC) byte. When the token bit $T = 0$, the token is called a free token, otherwise a busy token. The other bits in the AC byte are discussed later. End delimiter (ED) is also a unique pattern chosen along the lines of SD.

When the ring is silent, i.e. all stations are idle and have no data to transmit, a free token circulates on the ring indefinitely. Any station which has data to send can do so by recognising a free token from the SD pattern and the T bit, and converting it into a busy token by setting bit $T = 1$. It sends its data immediately following the AC byte as per the frame format given in Fig.10.36(b). All other stations in the ring would now see a busy token, examine the destination address (DA), and copy the data if it is destined to them. The originating station reintroduces a free token at the end of the data. In order to prevent any one station from occupying the ring for too long, a limit is set on the time for which a station can hold a token. This time limit is known as **token holding time (THT)**. Once the limit is reached, the station is required to introduce a free token in the ring even if it has more data to transmit. Within the limit of THT, a station may transmit multiple frames of data.

Token passing provides a round robin access as the free token passes from one station to another in sequence. All stations get a chance to transmit data within a specified maximum time, known as **token rotation time (TRT)**, which is given by

$$\text{TRT} = N \times \text{THT} + w \tag{10.32}$$

where

N = No. of stations in the ring
w = walk time

Walk time is the time taken by a bit to travel around the ring and is given by

$$w = t_p L + (N/R) \ \mu s \tag{10.33}$$

where

t_p = propagation delay in μs/km
L = physical length of the ring in km
R = data rate in Mbps

The term (N/R) gives the delay introduced by the stations in the ring with each station contributing 1-bit delay. The round robin access and the bounded delay make ring LANs suitable for real time applications.

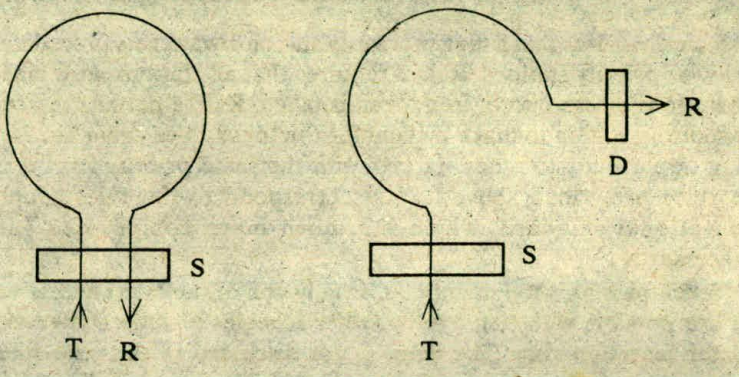
The walk time in a ring must be at least as large as to accommodate all the bits of a free token which circulates indefinitely when the ring is idle. In other words, the first bit of the free token should not return to a RIU before the last bit of the token leaves that RIU. Interestingly, this requirement places a minimum bound on the operating speed of a ring for a given physical length and number of stations.

$$R \geq S_{FT}/w \tag{10.34}$$

where S_{FT} is the size of the free token in bits.

Another interesting issue arises when a station completes its transmission even before its first bit has had time to walk around the ring. If the station introduces the free token immediately following its transmission, multiple frames can exist on the ring simultaneously. To avoid this, one may impose a restriction that a station should reintroduce a free token on the ring only after a lapse of at least one walk time after the frame transmission begins. This is so in the standard token ring LAN.

In a ring where each RIU introduces 1-bit delay, the data in the ring is removed by the source station as shown in Fig. 10.37(a). Data may also be



T = transmit R = receiver S = source D = destination

Fig. 10.37 Data removal in ring LANs.

removed by the destination station as shown in Fig. 10.37(b). In this case, each RIU must delay the data until the complete DA is seen by it. If the DA is its address, the data is copied and removed from the ring; otherwise it is sent out on the ring. For the frame format shown in Fig. 10.36(b) with two bytes of DA, the delay required is 40 bits. Data removal by the source station is more commonly used.

Occasionally, it may be necessary to skip round robin access in the ring to permit a station with a priority message to gain control of the ring quickly. This is achieved by using priority bits in the AC byte as shown in Fig. 10.36(a). The priority bits are set through the reservation bits. The 3-bit fields permit eight levels of priority and reservation. The priority and reservation scheme typically works as follows. Let

P_m = priority of the message to be sent
P_r = priority level set in the AC byte
R_r = reservation level set in the AC byte
R_m = level at which a station wishes to make reservation

Rule 1 : A station can convert a free token into a busy token only if $P_m > P_r$.

Rule 2 : When a token is taken as per rule 1, the reservation bits are set to zero. This has the effect of reinitialising reservation process throughout the ring.

Rule 3 : A station can make a reservation either in the busy token or free token only if $R_m > R_r$. In the case of free token, rule 1 takes precedence over rule 3.

Rule 4 : When a free token is introduced at the end of a transmission or after THT, P_r is set to R_r or R_m whichever is greater. If $R_r < R_m$, the reservation bits are set to R_m or else to zeros.

Rule 1 ensures that a higher priority station always gets precedence over the lower priority stations. Rule 2 requires that all stations must make their reservations afresh once a free token is taken. Rule 4 permits the currently transmitting station to make a claim for the token once again but only after giving opportunity to other stations with the same priority level messages. The rules prescribed above do not correspond to the rules adopted for international standard, which are much more complicated (perhaps unnecessarily).

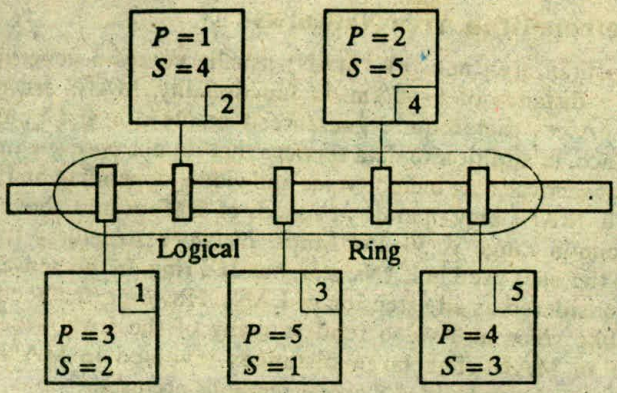
Token passing, though popular, is only one of the many access schemes that are possible with ring LANs. Other schemes include **slotted ring** and **register insertion ring**. The latter is not discussed in this text. Interested readers may refer to Further Reading [10, 11]. The former is discussed in the context of fibre optic networks in Section 10.9.

10.7.4 Token Passing Bus LAN

Token passing bus LAN is an attempt to combine the strengths of the bus architecture and the token passing access mechanisms. The structure of the LAN is a passive bus and the stations form a logical ring for token passing as shown in Fig. 10.38. Being a logical ring, the physical location of a station does not matter. Each station knows its predecessor P and its successor S. The token passes from a station to its successor. The token and frame structures are very similar to the ones used in the token ring. The strengths of this LAN are:

1. Robustness of the passive bus
2. Bounded delay of token passing access
3. No data removal considerations
4. No constraint on the operating speed, frame size or physical length.

A problem that is unique to the token bus LAN, is the formation of the logical ring. In token ring, a station can withdraw from participation by simply bypassing the RIU and join the operation by activating the RIU. In token bus, the logical ring is affected every time a station withdraws or joins



P = predecessor S = successor

Fig. 10.38 Token passing bus LAN.

the ring. When a station withdraws, it informs its neighbours of its intention giving its P and S values which are used to change the P and S values in the successor and predecessor stations respectively. When a station wants to join the ring, it may do so in one of two ways. It may transmit a special signal jamming the present transmission on the bus. This signal is heard by all stations which then go through the cycle of establishing the logical ring. This method disrupts the operation of the bus. Alternatively, each active station periodically broadcasts a special frame to find out if there is a new station that has come up between itself and its successor. This frame is known as 'solicit successor' frame. When a station goes down without formally withdrawing from the ring, its predecessor recognises this fact by observing no response on the bus when the token is passed on. It then transmits a frame known as 'who follows' frame to determine the next successor.

Both token ring and token bus suffer from the potential danger of token being lost. A station which is currently holding the token may go down without passing the token on. A token may be corrupted by noise disturbing the token pattern which is used to recognise its presence. The stations sense the loss of the token when no token arrives for a time duration of TRT or more since the last token. They then enter the 'enter token' phase in which a new token is introduced in the system. Priority and reservation schemes similar to the ones used in token ring also apply to token bus LAN.

A variation of token passing bus scheme is what is known as **implicit token** or **carrier sense multiple access/collision avoidance (CSMA/CA) scheme**. Here, when a station completes a transmission, it just goes idle. On detecting the idle channel, the next station in sequence starts transmission if it has a frame to send or else remains idle. Successive stations wait for successively longer times and decide to transmit or remain idle.

10.8 Metropolitan Area Networks

A metropolitan area network (MAN) usually covers a geographical area spanning a distance of 5–50 km. In functionality, MANs are expected to support services that require guaranteed bandwidth and bounded delay performance, in addition to data services that do not pose such restrictions. Additional services may include voice and video, depending on the available bandwidth. MANs may operate at speeds of 1 Mbps and above although a more common range is 50–150 Mbps. Network structures to MAN are similar to the ones used in LANs: star, bus and ring. In this sense, the MANs may be considered as an extension of LANs. However, the need to support services like voice and video renders many of the LAN access methods unusable in MAN. The large distances involved in MAN affect the performance adversely if LAN access methods are used.

EXAMPLE 10.3 Compute the maximum channel utilisation for a MAN which uses CSMA mechanism and has a length of 50 km, and operates at 50 Mbps with a frame length of 2000 bits.

Solution We determine the maximum channel utilisation from Eq. (10.30). Assuming coaxial cable as the medium, the propagation delay is 5 μs/km. Therefore,

$$t_f = \frac{2000}{50} = 40\ \mu s$$

$$a = t_p/t_f = 250/40 = 6.25$$

$$U_{max} = 0.14$$

The transmission media considered suitable for MAN applications are broadband coaxial cable and optical fibres. Networks based on optical fibres are discussed in Section 10.9. In this section, we discuss briefly the coaxial cable networks.

Coaxial cable technology is well developed for unidirectional transmission, and hence, MAN configurations use two unidirectional buses as shown in Fig. 10.39. Headend is a transponder which receives, amplifies and retransmits the upstream signal at a different frequency on the downstream cable. All stations transmit on the upstream cable and receive on the downstream channel.

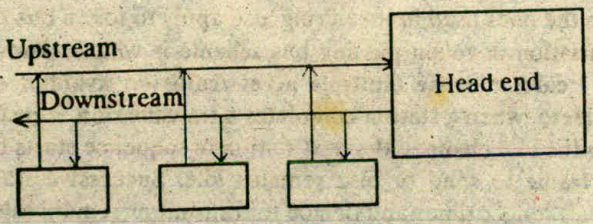


Fig. 10.39 Dual bus MAN architecture.

Configuration with passive headends are also possible and are discussed in the context of fibre optic networks in Section 10.9. Media access on MAN is concerned with providing access to the upstream channel to one of the many contending stations. Two schemes are generally considered:

- Polling by headend
- TDMA with reservation.

Polling is like a token passing scheme. The polls sent by headend act as tokens to the stations. Conventional polling implies that a poll signal is sent to a station, its response received and then the next poll sent to the next station and so on. The polling delay here is quite large and such a scheme is not useful in MANs. A scheme which cuts down delay is known as **hub polling** or **daisy chain polling**. Here, the poll is passed from one station to the next in sequence after use without the headend having to transmit a poll everytime. Headend transmits a new poll only after a complete cycle.

Schemes using TDMA require a synchronisation process by which the stations can identify the beginning of the slots allocated to them. This is done by observing the beginning of one's slot on the downstream and accessing the upstream after a fixed time:

$$t_a = t_0 + 2(t_k - t_e) \tag{10.35}$$

where

t_a = time of access of the upstream
t_0 = time of observation of the slot on the downstream
t_k = some fixed arbitrary time which is greater than the one way propagation time from headend to the end of the bus or vice versa
t_e = time required for the signal to travel from the station to headend or vice versa

The estimation of t_e is done as follows: A special synchronising slot is transmitted by the headend to a station on the downstream channel. On seeing this, the station immediately transmits a slot on the upstream and waits for the slot to appear on the downstream. Half of the time elapsed between the two events is t_e.

Another important structure that one encounters in MAN environment is the tree structure used by CATV systems. Here, the transmissions are broadcast from the root to the leaves. A separate frequency is used for communication from the leaves towards the root. There is the problem of collisions at the root when the frames are sent towards it. An access method, known as **Homenet**, has been proposed to tackle this problem. (Refer to Further Reading [4]). Sometimes, a MAN may be designed only for data transmission. In such cases, packet radio networks may turn out to be a cost effective alternative. Line-of-sight optical communication systems are also suitable for MANs.

10.9 Fibre Optic Networks

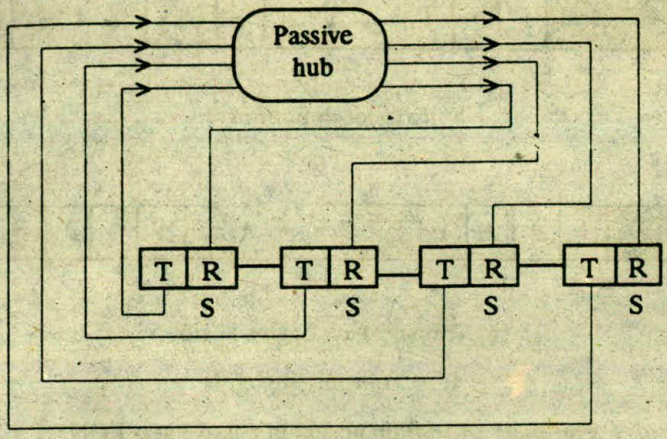
Optical fibre networks are characterised by

1. High speed operation (typically 100 Mbps or more)
2. Ability to span large distances (100–200 km)
3. Ability to support a moderate number of stations; typically 10 to a few hundred stations are supported with a maximum limit around 1000.

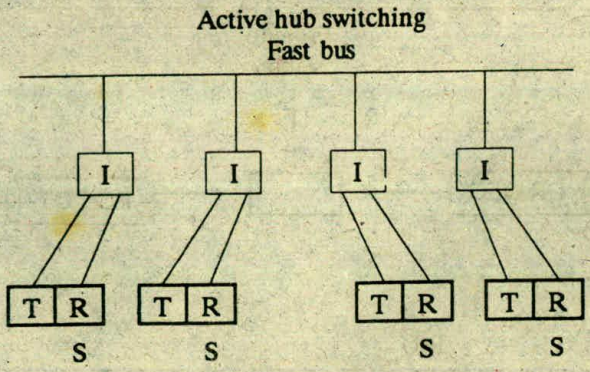
These characteristics make the fibre optic networks suitable for high speed LANs and MANs with a limited number of stations. Fibre networks may be configured around a star, ring or bus structure. The number of stations that can be supported in a star or a bus structure is relatively low compared to that in a ring configuration. Optical fibres are inherently unidirectional and this influences the way in which the network structures are realised, and the consideration for medium access. Some of medium access considerations discussed in the context of MANs in Section 10.8 are also applicable to fibre optic networks (FONs).

The central hub in a star FON may be passive or active as shown in Fig. 10.40. The passive hub is usually a silica cylinder. The incoming fibres are fused to one end of the cylinder and the outgoing fibres to the other end. The light emitted by the transmitters (LEDs) of the stations diffuses inside the passive hub illuminating all the receivers (photodiodes). The incoming energy is divided among all the outgoing lines. The output power of the transmitter and the sensitivity of the receiver would determine the number of nodes that can be connected to the passive hub. The limited number of stations in passive-star FONs permits a simple medium access protocol to be used. Following the end of a transmission, a fixed number of time slots usually equal to or twice the number of stations, is defined as illustrated in Fig. 10.41. Each station is assigned one or two slots, as the case may be, in which it may start transmission provided no other station prior to it has started transmission. In Fig. 10.41(b) the station assigned to slot 5 (station 5) starts transmission and all stations after slot 5 will have to wait until the end of transmission by station 5, when the cycle repeats. Obviously, stations allotted to earlier slots attain priority over other stations, resulting in an unfair access scheme. To bring about fairness in access, $2N$ slots are used where a station is allotted two slots, one in the lower half, i.e. in the range 1 to N and the other in upper half, i.e. in the range $(N + 1)$ to $2N$. A station is required to transmit alternately in the lower and upper half. Carrier sensing is possible in a passive star as the transmitted signal would return to the station after the round trip propagation delay. This opens up the possibility of contention protocols with collision detection feature.

In an active-star FON shown in Fig. 10.40(b), the optical signals are converted to electrical domain and a fast switching is organised within the hub. Medium access mechanism is not related to the switching mechanism in



(a) Passive star FON



(b) Active star FON

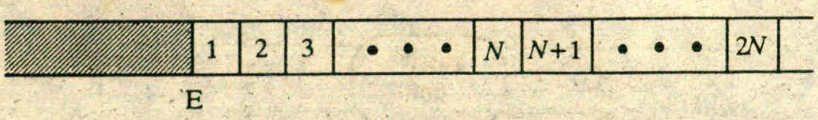
T = transmitter R = receiver S = station I = interface

Fig. 10.40 Star FON configurations.

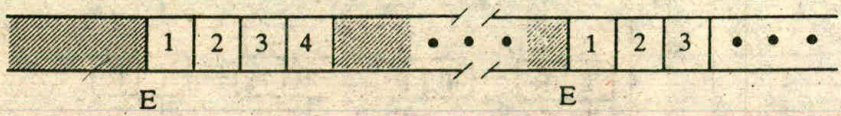
the hub. Figure 10.40(b) shows a fast switching bus, which uses parallel paths for data transfer.

The working principle of a fibre optic ring network is very similar to the cable based ring LAN discussed in Section 10.7.3. We only highlight the differences here. The ring interface units (RIUs) in a FON perform all the functions carried out by the RIUs in the cable LAN. In addition, they are required to convert optical signals to electrical domain for processing and reconverting them to optical domain for transmission as shown in Fig. 10.42.

Token passing is popularly used as the fibre access scheme in the ring FONs. The physical length of the fibre optic ring being large (up to 200 km), the token reintroduction requires consideration.



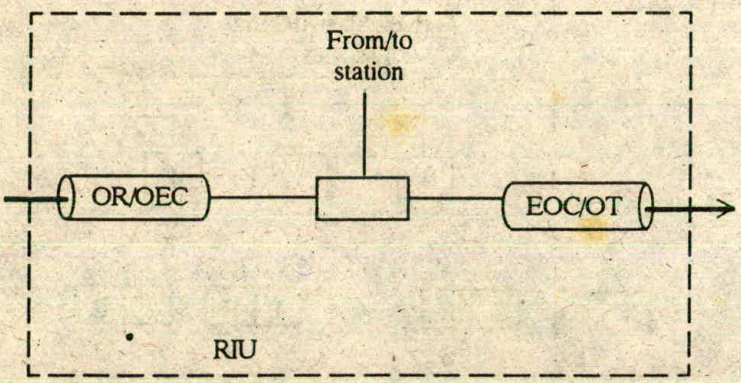
(a) Slot definition



(b) Transmission begins at slot 5

E = transmission ends

Fig. 10.41 Medium access in passive star FON.



OR = optical receiver OEC = optical to electrical converter
EOC = electrical to optical converter OT = optical transmitter

Fig. 10.42 RIU for ring FON.

EXAMPLE 10.4 Consider a fibre optic token ring with a physical length of 100 km, an operating speed of 100 Mbps and having 100 stations each introducing 1-bit delay. Compute the maximum ring utilisation if the free token were to be reintroduced after the first bit arrives at the source. Assume a frame length of 1000 bits and the stations are equally spaced.

Solution From Eq. (10.33), the walk time is given by

$$w = t_p L + (N/R)$$

$t_p = 3.3\,\mu s$ per km for optical fibre

Hence,

$w = 333 + 1 = 334\ \mu s$

Frame time $t_f = 1000/100 = 10\ \mu s$

There is a transmission for $10\ \mu s$ for every $334\ \mu s$. Therefore,

$$\text{Maximum ring utilisation} = \frac{10}{334} = 0.03$$

In order to improve the channel utilisation, the free token is introduced immediately after a transmission. As a consequence, multiple frames may be present on the ring at any point of time.

Token passing FON is being standardised under an IEEE project (see Section 10.10) and the standard is known as fibre distributed data interface (FDDI). The FDDI proposes a dual ring configuration as shown in Fig. 10.43

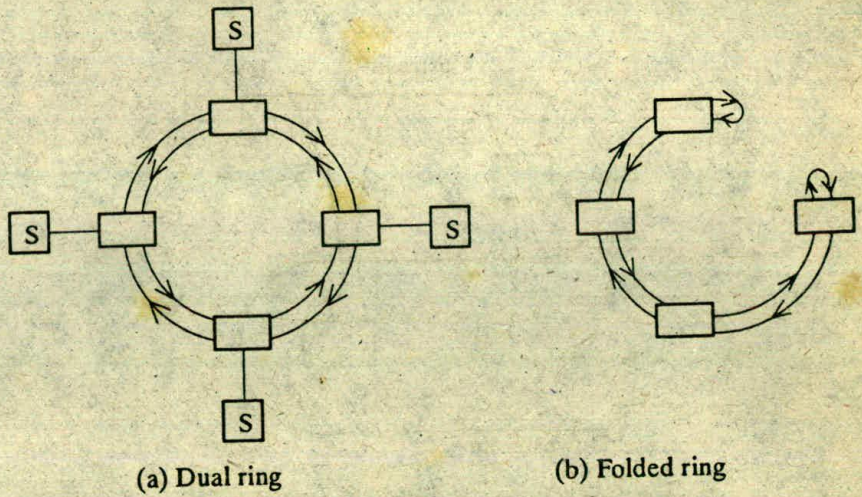


Fig. 10.43 Fault tolerant dual ring FON.

for better fault tolerance. A station may connect to only one of the rings or both the rings depending on the criticality of operation. If either one of the rings breaks, the other can be used as a back up. If both of them break at the same point due to fire or accident in the cable duct etc., the rings can be looped back as shown in Fig. 10.43(b) to form a single ring of approximately twice the length.

A fibre optic bus network may use two independent buses with an active headend similar to the MAN configuration shown in Fig. 10.39 or without a headend. The latter configuration is used in a network called FASNET (see Further Reading [5]). Alternatively, a single bus may be used in a folded fashion, either U or S shaped as shown in Fig. 10.44. All stations transmit on the upstream and receive on the downstream. There is a provision to sense the

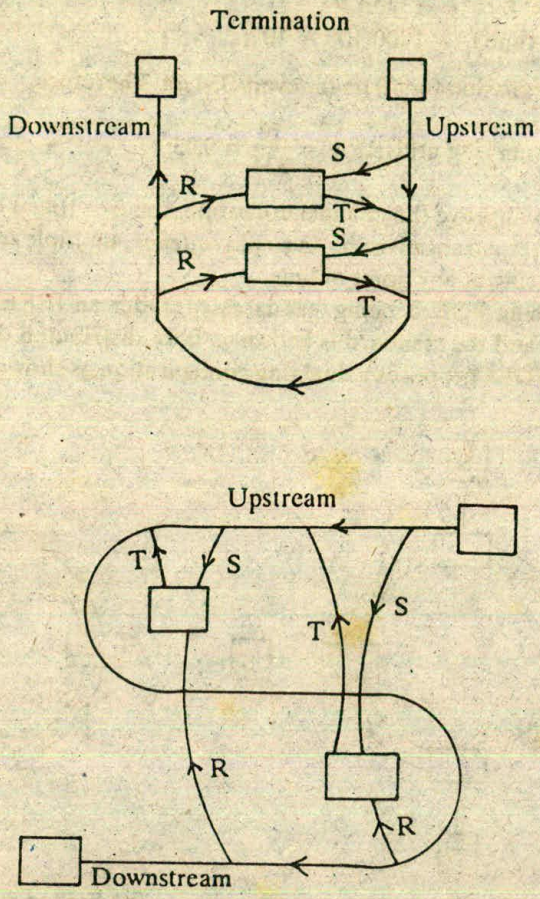


Fig. 10.44 Folded bus FONs.

presence of a transmission on the upstream. Taps on the bus drain away the optical power and hence the bus configuration is not capable of supporting a large number of stations. The main difference between the U and the S shaped configurations is that in the S configuration the stations have uniform propagation delay for the signals to travel from their transmit taps to their respective receive taps. This fact can be gainfully employed in medium access strategy. In the S configuration, transmission of frames takes place in a cyclic order. The cycle is initiated by the station that is ready to transmit and is closest to upstream termination end. A new cycle is started by observing the end of the previous cycle on the downstreams. This will be first observed by the station that is closest to the upstream termination end. Once

initiated, other stations follow one after another by sensing the end of a frame on the upstream. If a station senses incoming traffic on the upstream while transmitting its own frame, it immediately aborts its transmission and lets the transmission which must have originated from a station closer to upstream termination end than itself, to proceed. Its own packet is sent after the upstream becomes quiet. An example network which uses S-shaped configuration is expressnet (see Further Reading [12]).

10.10 Data Network Standards

Three major international bodies have been significantly contributing to the data network standards. The bodies, their major area of concentration, and the standard series identifiers are shown in Table 10.8.

Table 10.8 International Bodies and Their Main Standardisation

Body	Main thrust areas	Series
ISO	OSI reference model; end-to-end layers (4–7)	7xxx 8xxx 9xxx 10xxx
CCITT	Link-to-link layers (1–3) of wide area networks: PSTN, PDN and ISDN; electronic messaging and directory services	V series X series I series
IEEE	Link-to-link layers (1–3) of local and metropolitan area networks	802. x

Standards laid down by ISO and CCITT have international legal standing, whereas IEEE standards have to be adopted by ISO to attain this legal standing. Standards work at IEEE is sponsored by ANSI which is affiliated to ISO. Standards evolved by CCITT are adopted by ISO under its own series number and vice versa. For example, ISO basic OSI reference model IS 7498 is adopted by CCITT as X.200. Similarly, X.400 message handling system of CCITT is adopted by ISO under the name "message oriented text interchange system (MOTIS)" as IS 10021. The X.400 message handling system is described in Section 11.2. While adopting standards, these bodies may introduce minor changes.

As explained in Section 10.3, a layer in OSI reference model offers services to the layer above. Services are of two types: connectionless and connection oriented. The peer entities in a layer communicate using peer protocols. Accordingly, standards for a layer deal with the two types of services and their associated protocols as shown in Fig. 10.45. Some of the ISO standards for the end-to-end layers are shown in Table 10.9. ISO standards

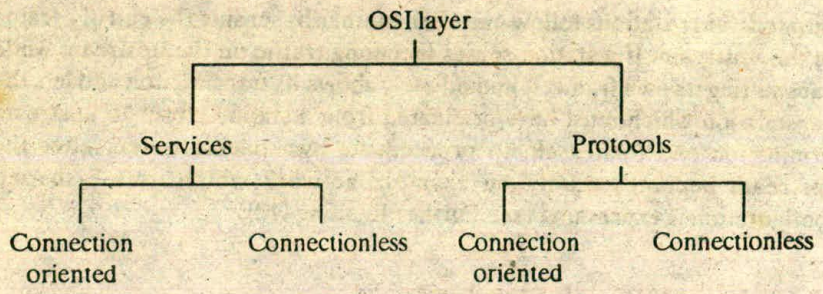


Fig. 10.45 OSI-layer standards.

Table 10.9 Some ISO Standards for End-to-End Layers

Number	Description
IS 8072	Connection oriented transport service definition
IS 8072 DADI	Connectionless transport service definition
IS 8073	Connection oriented transport protocol specifications
IS 8602	Connectionless transport protocol specifications
IS 8326	Connection oriented session service definition
IS 8327	Connection oriented session protocol specification
DIS 8822	Connection oriented presentation service definition
DIS 8823	Connection oriented presentation protocol definition
DIS 8824	Specification of Abstract Syntax Notation 1 (ASN.1)
Application Layer standards	
DP 9579	Remote database access
DP 9596	Management information protocol
DP 9594	Directory services
DIS 8571	File transfer access and management
DIS 8613	Office document architecture and interchange format
DIS 9040	Virtual terminal service definition
DIS 9041	Virtual terminal protocol specification
DP 10026	Distributed transaction processing
DP 10160	Interlibrary loan service definition
DP 10161	Interlibrary loan protocol specification
DP 10162	Search and retrieve service definition
DP 10163	Search and retrieve protocol specification

go through three stages: draft proposal (DP), draft international standards (DIS), and international standards (IS or ISO). The subject of distributed applications over networks has been the main concern of ISO subsequent to the evolution of the basic OSI reference model.

Functionalities defined in the lower three layers of OSI reference model have traditionally been the responsibility of telecommunication network providers or common carriers as they are called in the US. It is only natural that CCITT, whose only voting members are the national telecommunication carriers of different countries, has been concentrating on the standards for the lower three layer functionalities. In the context of data communication, CCITT is concerned with three classes of networks: GSTN, PDN and ISDN. Although the standards relating to ISDN are discussed in Chapter 11, some aspects are discussed here with a view to placing data communication standards in perspective. Three types of connections are supported on the networks: circuit switched, leased circuit and packet switched. All the three types are supported on PDNs and ISDNs, but only the first two are supported on GSTNs. CCITT standards relate to these three types of connections. A major aspect of the CCITT standards is the interfacing of the data terminal equipment (DTE) and the data circuit terminating equipments (DCE). In the context of standards, DTEs may be placed under four categories:

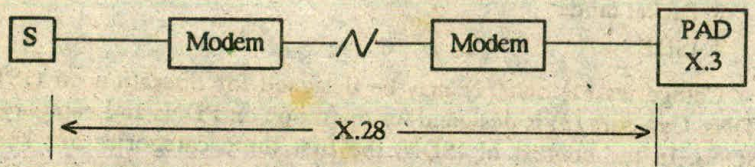
1. Asynchronous or start/stop type
2. Synchronous
3. Packet mode
4. ISDN.

Categories (1) and (2) may be designed for operation on GSTNs or PDNs. Category (3) is designed for operation on PDNs and category (4) on ISDNs. In the context of ISDN, the first three categories are known as nonISDN terminals. Maintaining compatibility with earlier systems is an important requirement of any telecommunication network operation. It should be possible to use GSTN terminals on PDN or ISDN, and PDN terminals on ISDN. Standards take care of this aspect. DTE-DCE interfaces are specified in the V-series for PSTNs, in the X-series for PDNs, and in the I-series for ISDNs. The V-series details have been presented in Section 10.1. Five different types of DTEs are required to be supported on PDNs as shown in Table 10.10. The table also shows the applicable X-series standard along with the relevant network connection type.

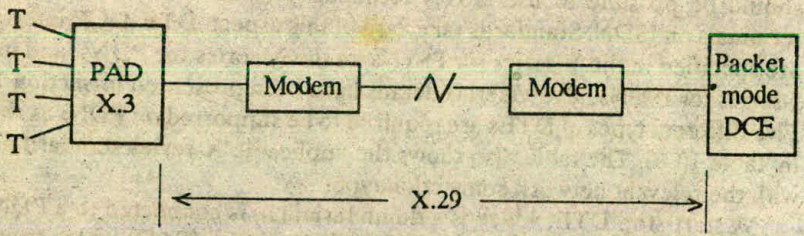
A start-stop DTE, which is a dumb terminal, is connected to a PDN via an interface known as packet assembler/disassembler (PAD) which converts the character stream from the dumb terminal into a X.25 packet by assembling the characters for transmission through the packet switched network. At the receiving end, a PAD converts the X.25 packet back into an asynchronous character stream for passing on to the receiving terminal. Three separate X-series CCITT recommendations, popularly known as triple X standard, are associated with the PAD interface. The standard X.3 defines the PAD parameters, X.28 defines the terminal-to-PAD interface and X.29 defines PAD-to-packet mode DCE interface. These are illustrated in Fig.10.46. Terminals may be directly connected to the PAD or via modems

Table 10.10 DTE-DCE Interface Standards for Operation in PDNs

DTE type	Standard	Relevant network connection
Asynchronous designed for GSTN	X.20 bis	Circuit switched or leased
Synchronous designed for GSTN	X.21 bis	Circuit switched or leased
Asynchronous designed for PDN	X.20	Circuit switched or leased circuit
Synchronous designed for PDN	X.21	Circuit switched or leased circuit
Packet mode	X.25	Packet switched (dedicated circuit between DTE and DCE)
Asynchronous	X.3, X.28, X.29	Packet switched (circuit switched or leased circuit between DTE and DCE)



(a) Terminal connection to PAD



(b) PAD-to packet mode DCE connection

Fig. 10.46 Triple X protocol set.

as shown in Fig.10.46(a). Figure 10.46(b) shows the applicability of X.29 standard, which also defines PAD-to-PAD communication protocols.

When a user of a terminal initially establishes contact with a PAD, he may set up certain parameters describing the conversion between the PAD and the terminal. There are 18 PAD parameters in CCITT X.3 and these are

numbered. For setting up a parameter, a user specifies the parameter number and its value. Some sample parameters and their permitted values are as follows:

1	Should the PAD echo the characters back to the terminal?	0 = No 1 = Yes
2	Is the PAD allowed to send service signals to the terminal?	0 = No 1 = Yes
3	Should PAD add filler characters after line feeds? If so, how many?	0 = No 1 = Yes
4	Does PAD allow editing?	0 = No 1 = Yes

A PAD may be stand-alone or integrated as part of the packet mode DCE as shown in Fig. 10.47.

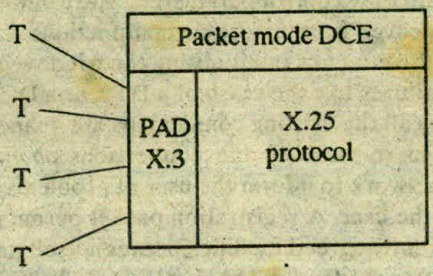


Fig. 10.47 An integrated PAD and X.25 DCE.

The standards X.20 and X.21 were defined by CCITT in the mid 1970s with the expectation that computers would be able to interface directly to DCEs associated with the digital networks. This did not happen. As a result, there are very few networks that support X.21. As an interim measure, and to permit existing DTEs to interface to digital networks X.20 bis and X.21 bis were defined. These have the same interface definitions as the V-series but are capable of working with digital networks and include call establishment and control procedures.

The standard X.25 defines the DTE/DCE interface for a packet mode DTE to communicate with a packet switching exchange (PSE) or a DCE in a PDN. The standard is an integrated set of three levels of protocols corresponding to layers 1–3 of OSI reference model. At the physical layer level, it uses X. 21 and X. 21 bis protocol. At the data link layer level, it uses a balanced link access procedure (LAPB) which is similar to the bit oriented sliding window protocol discussed in Section 10.4. At the network level, X. 25 supports virtual circuit (VC) and permanent virtual circuit (PVC) connections discussed in Section 10.2. The connectionless datagram service was initially introduced in X.25 but was dropped in the 1984 recommendations and reintroduced in the 1988 recommendations. For managing virtual

circuit connections, X.25 uses five different types of PDUs, which are known as the network PDUs (NPDUs) or simply as packets:

1. Call setup and clear
2. Data and interrupt
3. Flow control and reset
3. Diagnostic
5. Registration.

Call setup and clear packets are used for exchanging messages like call request, call accept, clear request, clear confirmation etc. Interrupt packets are used to cope with unusual situations that may arise during normal data transfer, e.g. a terminal user wants to stop data transmission from a host. These packets are given priority over data packets and delivered in an expedited manner. An interrupt may be generated by a DTE or a DCE. Flow control packets give information regarding the readiness of the receiver and are used to negotiate flow control parameters, viz. window and packet sizes. Reset is used to recover from errors or malfunctions related to a specific connection. It has the effect of reinitialising the window parameters to zero. For more serious failures like the crash of a DTE or a DCE, a restart facility is available whereby all the existing connections are abandoned and a 'clean' state is provided to re-establish the connections *ab initio*. A diagnostic packet allows the network to inform the user of problems including errors in the packet sent by the user. A registration packet permits a DTE to request changes to the currently agreed facilities between itself and the network.

Several LAN standards and MAN standards have been produced by IEEE under a project known as **IEEE 802**. The standards are summarised in Fig. 10.48. IEEE 802 recommendations propose a 3-layer architecture which

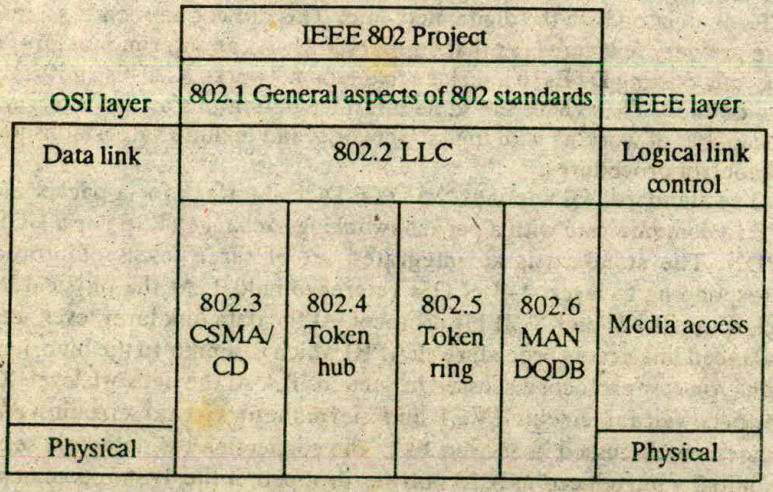


Fig. 10.48 IEEE LAN and MAN standards.

corresponds to the first two layers of ISO-OSI reference model. In the original version, logical link control (LLC) and media access control (MAC) layers are considered as sublayers of OSI data link layer. Recently, there is a trend among many researchers to consider MAC layer as a sublayer belonging to the OSI physical layer. The MAC layers and the physical layers are described in standards IEEE 802.3 through 802.6. All these standards are compatible at the 802.2 LLC level. Unacknowledged and acknowledged connectionless service and connection oriented service are defined at the LLC level. The 802.1 gives an overview and the architectural aspects of the set of standards. The 802.3 defines 1-persistent CSMA/CD, the 802.4 defines token passing bus, and the 802.5 specifies single token ring access mechanisms. At the physical layer 802.3 standard permits two versions of baseband, coaxial cables, commercially known as thin and thick ethernet cables, both operating at 10 Mbps. The thin cable version is also called the cheapernet. The thick cable can have a maximum segment length of 500 m and the thin cable 185 m. In addition, unshielded twisted pair baseband cables operating at 1 or 10 Mbps covering a distance of 250 m or 100 m respectively, and a 75-Ω broadband coaxial cable operating at 10 Mbps covering a distance of 3600 m are supported. A star wiring topology, referred to as star LAN, is recommended in the case of 1-Mbps unshielded twisted pair medium. As long as the operating speed is the same, the cables can be mixed freely in the same network.

The 802.4 recommendations specify two options for the physical medium. A 75-Ω carrier band cable operating at 1, 5 or 10 Mbps speed or a 75-Ω broadband cable operating at the same speed but covering larger distances (760 m) can be used. The 802.5 standard specifies a shielded twisted pair operating at 1 or 4 Mbps as the physical medium.

The IEEE 802.5 has been extended by the ANSI committee X3T9.5 to use optical fibre as a medium to operate at 100 Mbps. This standard is known as fibre distributed data interface (FDDI). In the original allocation of business, LAN standards with data rate up to 20 Mbps is placed under the purview of IEEE, and LAN standards with data rates above 50 Mbps under X3T9.5 committee whose primary responsibility is computer input/output (I/O) interface standards. FDDI has been designed to serve two primary areas of applications:

- High performance interconnection among main frames, and main frames and their associated peripheral subsystems.
- Backbone network for use with low speed LANs.

FDDI was originally specified as a packet switched network. A recent version, FDDI-II, specifies a circuit switched mode of operation. Here, the bandwidth is allocated in increments of 6.144 Mbps isochronous channels, supporting up to a maximum of 16 channels. Each of the isochronous channels may be dynamically reassigned as three 2048 kbps or four 1536 kbps

channels corresponding to European and North American telephone channels respectively.

IEEE 802.6 committee which has been charged with the responsibility of evolving standards for metropolitan area networks has had several setbacks. When the 802.6 committee was charted in 1981, it was supported by satellite communication industry which was looking for an economical high speed medium to link its ground stations with customers. However, the growth of satellite data communication fell far short of expectations and the industry support to IEEE 802.6 died down within a year or two. The committee then turned to CATV industry for support which did not come through either. In late 1984, some companies interested in fibre optic networks lent support to the committee. A fibre optic slotted ring proposal operating at 43 Mbps was finalised by the committee in 1986. When the standard documents for physical and media access layers were ready for letter ballot, the companies withdrew funding. In 1987, the committee endorsed a proposal from Telecom Australia as the standard for MAN. This protocol, officially known as **distributed queue dual bus (DQDB)** and commercially known as the **queued packet and synchronous exchange switch (QPSX)**, is based on a fibre optic dual bus architecture. It is expected that IEEE 802.6 DQDB proposal will be adopted as MAN standard by ISO and CCITT shortly.

An interesting scenario is emerging in the field of data network standards. A number of new physical media and transmission methods are being introduced for data and multimedia transmission. Additional physical and datalink layer standards are being specified to support these new arrivals. Internetworking is becoming complex, calling for additional specifications at the network layer. Side by side, more and more distributed applications involving heterogeneous data types are being supported in the OSI environment. This is bringing about a host of service and protocol specifications at the application and presentation layers of the OSI model. However, the session and transport layer service and protocol definitions are more or less static. This phenomenon is changing the shape of the OSI reference model as shown in Fig. 10.49. The model is now being called 'OSI wine glass model' after its shape.

10.11 Protocol Stacks

It may be recalled that OSI model is only a framework for interoperability and not a standard in itself. As discussed in Section 10.10, a number of protocols standardised by different agencies exist for each layer. It then becomes necessary that a specific protocol be chosen in each layer for a given application. A collection of such protocols, one per layer, is known as a **protocol suite** or a **protocol stack**. A chosen protocol is called a functional protocol for the given application. Two such well known protocol suites are **manufacturing automation protocol (MAP) and technical and office**

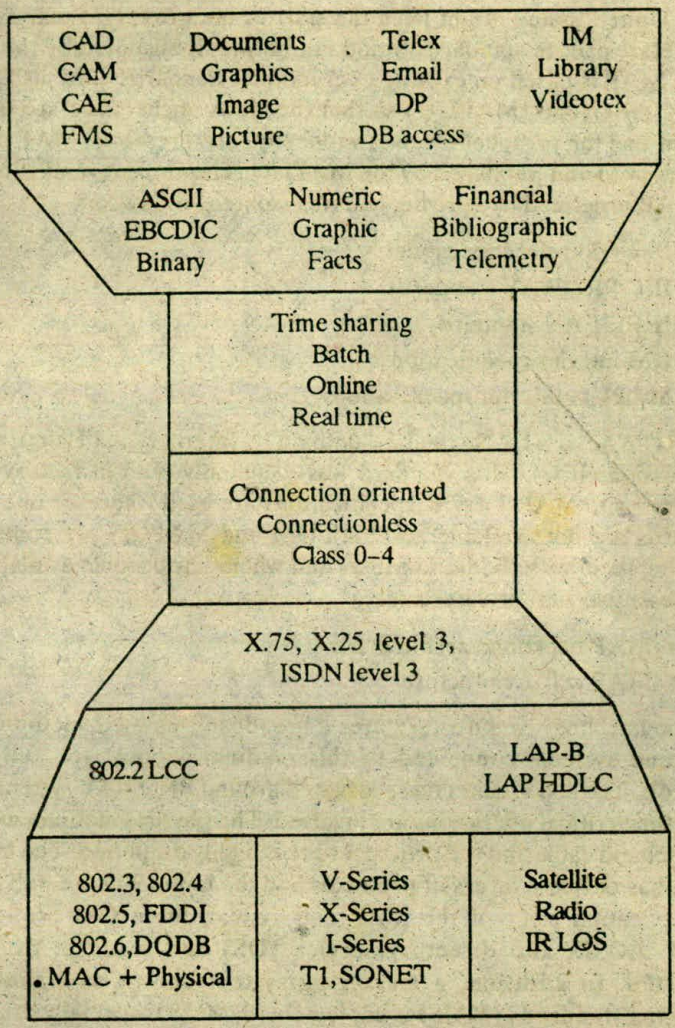


Fig. 10.49 OSI wine glass structure.

protocol (TOP). We discuss these two protocol stacks in this section. In this chapter, we also present a protocol stack that may be suitable for library automation purposes, and we call this suite **library automation protocol (LAP)**.

10.11.1 Manufacturing Automation Protocol (MAP)

Manufacturing automation protocol (MAP) is a standard for information transfer among computer systems in a manufacturing plant. Work on standardising MAP was initiated and spearheaded by General Motors (GM)

of the United States. Right from the start of the work, GM was aware that proprietary private standards would not be widely accepted by the vendors. Hence MAP is based on existing international standards. GM also formed a MAP User Group (MAPUG) so that the views of the users are taken into account and the protocol receives wider support. Presently, MAP standards are approved and published by the MAPUG. The current version number is 3.0. MAP is defined by deriving features from the following:

1. ISO OSI reference model
2. IEC PROWAY standards
3. IEEE LAN standards
4. GM interim specifications
5. MAPUG interim specifications.

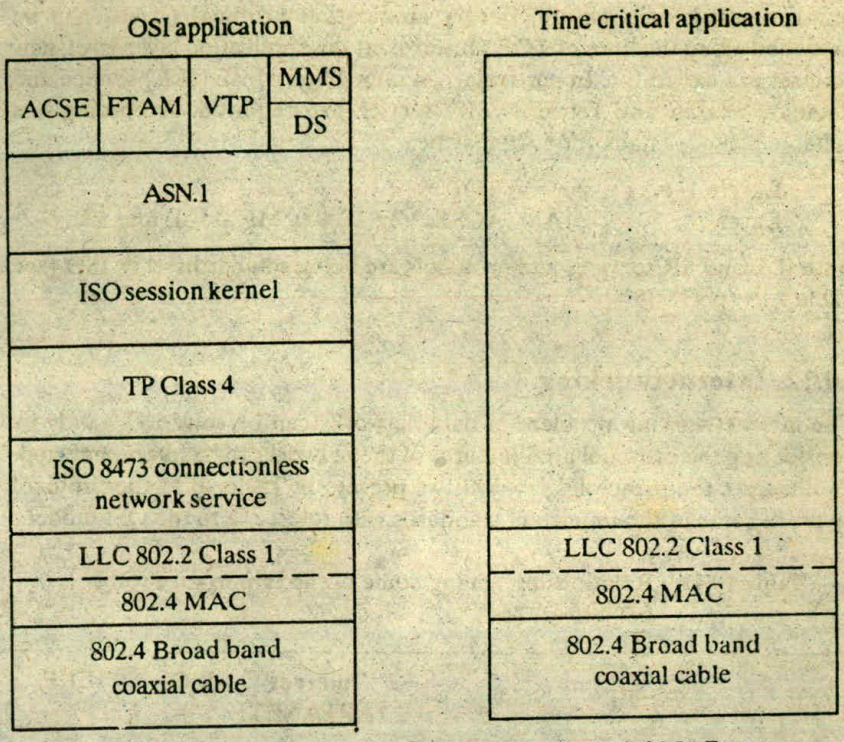
PROWAY is a LAN standard defined by International Electrotechnical Commission (IEC). This standard was commonly used in factory environments prior to MAP. Interim specifications are used wherever international standards are not available and both GM and MAPUG are committed to replacing them with the standards as and when they become available. MAP specifies a two-part architecture:

- MAP backbone architecture
- MAP cell architecture.

The backbone architecture uses a broadband physical medium and the MAP end systems connected to this medium support the full suite of protocols. The cell architecture, designed around PROWAY systems, is used where time critical operations are involved. The physical medium in this case is a baseband cable and the protocol suite is highly simplified. The mini MAP system has only two layers of protocols and the layers 3–7 are null layers. In the application layer of the MAP end system, three OSI protocols, viz. FTAM, ACSE, and directory service (DS) and a subset of VTS are supported. In addition, a messaging system known as **manufacturing messaging standards** (MMS) is implemented. MMS is specially designed for machine-to-machine communication, e.g. messaging between engineering workstations and robots. MAP also defines a combination system, MAP/PROWAY system, which is capable of functioning either as a PROWAY system or as a MAP end system on the baseband cable. MAP protocol suites are shown in Fig. 10.50. Other MAP systems include bridges, gateways and routers whose functions are described in Section 10.12.

10.11.2 Technical and Office Protocol (TOP)

The Boeing company spearheaded the development of technical and office protocol (TOP) for office automation applications. Since office functions do not call for real time responses, IEEE 802.3 CSMA/CD has been chosen for



(a) MAP end system (b) PROWAY or mini MAP system

Fig. 10.50 MAP protocol suites.

TOP. In addition, IEEE 802.5 token ring is also included in TOP specifications. Layers 3–6 of TOP use the same protocols as MAP. At the application layer, TOP supports FTAM, DS, VTS and X.400 MHS.

TOP defines five types of systems that may be present in a TOP network: an end system which implements the full TOP protocol suite, a repeater, a bridge, a router and a gateway. The last four systems are described in Section 10.12.

In MAP and TOP environments, small computers dominate. It is envisaged that many companies may produce plug-in boards to the small computers, containing full TOP or MAP implementation. It is then essential that a standard application service interface be defined so that a user has the freedom to choose the MAP or the TOP board from any of the vendors who offer them, for use in his computer. MAP and TOP define such an interface.

10.11.3 Library Automation Protocol (LAP)

Library functions differ from office functions or factory requirements. Library automation is a potential application area for data network. Being

non-time-critical in nature, library automation protocol (LAP) can be modelled along the lines of TOP. However, at the application layer level, new services are called for. In particular, an inter-library loan (ILL) service and database search and retrieve (SR) service are required. Accordingly, a suitable protocol stack for LAP may be:

Layers 1–6 : same as TOP
Layer 7 : FTAM, ACSE, VTP, X.400, MHS, ILL and SR.

Both ILL and SR services and protocols are being standardised by ISO (see Table 10.9 (p.472)).

10.12 Internetworking

The internetworking problem in data networks can be solved elegantly by considering the protocol architectures of the networks to be interconnected. To illustrate the principles involved we present in Table 10.11 the protocol layers of some of the important standards with reference to the OSI model.

Table 10.11 Relationship among Some of the Important Protocol Architectures

OSI	CCITT	SNA	Internet (ARPANET)	ISO	IEEE
Application		End user	SMTP Telnet	ASN.1	
Presentation			FTP		
Session		NAU services Data flow control		SPM	
Transport		Transmission control	TCP UDP	TP classes 0–4	
Network	X.75, X.25 level 3	Path control	IP ARP ICMP	Internet sublayer	
Data link	LAP-B	SDLC Token ring			LLC
Physical	X.21				Various MAC and physical

In Table 10.12, ARP denotes address resolution protocol, ASN.1 abstract syntax notation.1, FTP file transfer protocol, ICMP internet control message protocol, IP internet protocol, LAP-B link access procedure-B, NAU network addressable unit, SDLC synchronous data link control, SMTP simple mail transfer protocol, SPM session protocol machine, TCP transmission control protocol, Telnet virtual terminal protocol and TP transport protocol.

It is seen from Table 10.11 that protocol suites may be formed by using protocols from different standards. For example, an IEEE 802.3 LAN may run Internet (ARPANET) protocol in layers 3–7, or a X.25 network may run SNA protocols in layers 4–7. It is the commonality of protocols in the upper layers that decides on the scheme to be used for internetworking. For example, a X.25 network and a LAN running SNA in layers 4–7 may be internetworked by using an interworking unit that resolves the protocol differences in layers 1–3. An interworking unit is termed as a repeater, a bridge, a router or a gateway, depending on the level at which interworking is achieved, as shown in Fig. 10.51.

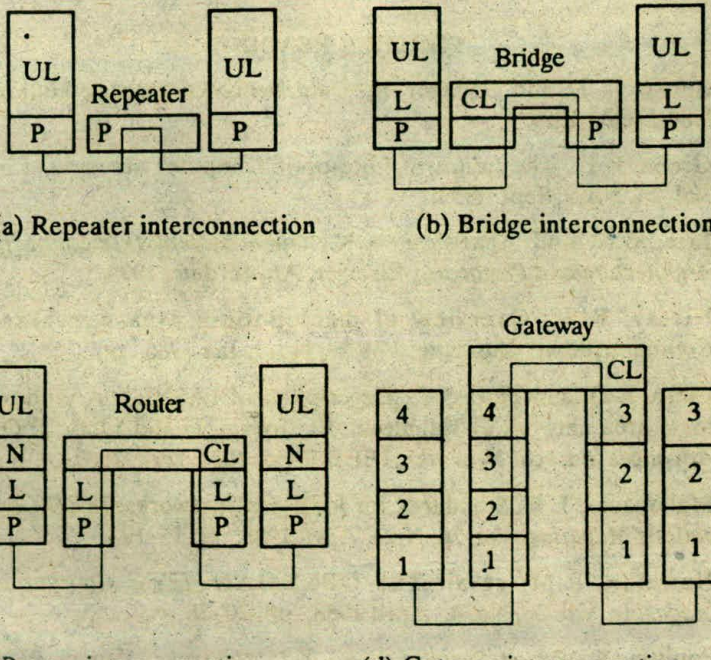


Fig. 10.51 Internetworking structures.

First let us consider two networks which use identical protocols in all the layers. These may be interconnected using repeaters at the physical layer level as shown in Fig. 10.51(a). A repeater boosts the signal level and transmits the bits from one network to another making the two networks look like one single network. This kind of connection is useful in large computers where a campus LAN is realised using multiple identical LANs. Two networks which differ in physical layers but not in data link layer and above, e.g. CSMA/CD and token ring LAN, are interconnected using a bridge as shown in Fig. 10.51(b). A bridge is an intelligent device capable of recognising addresses and transmitting only the internetworking frames across the networks. Bridges may also be used in interconnecting identical networks from the traffic considerations point of view. Routers are used as shown in Fig. 10.51(c) when the two networks to be interconnected use the same protocols in layer three and above. Example of interconnection is a X.25 WAN with a 802.6 MAN. An interworking unit which implements a combination of bridge and a router functions may be used to interconnect two LANs via a WAN. Networks with differing communication architectures use a gateway as shown in Fig. 10.51(d). Networks which differ in some nonadjacent layers may be interconnected using suitable gateway structures.

FURTHER READING

1. Bertsekas, D. and Gallager, R., *Data Networks,* Prentice-Hall of India, New Delhi, 1989.

2. Green, P. E., The future of fibre-optic computer networks, *Computer,* Vol. 24, No. 9, Sept. 1991.

3. Jain, B.N. and Agarwala, A.K., *Open System Interconnection: Its Architecture and Protocols,* Elsevier, Amsterdam, 1990.

4. Klessy, R.W., Overview of metropolitan area networks, *IEEE Communications Magazine*, Vol. 24, No. 1, Jan 1986, pp 9–15.

5. Limb, J. O. and Flores, C., Description of FASNET: A unidirectional local area network *in* Kunmerle, K., Tobagi, F. and Limb, J. O. (eds), *Advances in Local Networks*, IEEE Press, New York, 1987.

6. Mallenauer, J. F., Standards for local area networks. *IEEE Communications Magazine,* Vol. 26, No.4, April 1988, pp 15–19.

7. Newman, R.M. et al. The QPSXMAN. *IEEE Communications Magazine,* Vol. 26, No. 4, April 1988, pp 20–28.

8. Poulton, S., *Packet Switching and X-25 Networks,* Pitman Publishing, London, 1989.

9. Ross, F. E., FDDI – A tutorial. *IEEE Communications Magazine,* Vol. 24, No. 5, May 1986, pp 10–17.

10. Stallings, W., *Local Networks*. 3rd ed., Macmillan, New York, 1990.

11. Tanenbaum, A. S., *Computer Networks*, 2nd ed., Prentice-Hall of India, New Delhi, 1990.

12. Tobagi, F.A, Borgonova, F. and Ferrta, L., Expressnet: A high performance integrated services local area network, *in* Kunmerle, K., *et al*, 1987, *op. cit.*

EXERCISES

1. There is a saying that "If Mohammad does not go to the mountain, the mountain comes to him". Discuss the analogy of this statement to remote computing.

2. Depict by means of a tree structured diagram the different classes of data networks.

3. A noiseless communication channel has a bandwidth of 6 MHz. What is the maximum data rate that can be supported in this channel if four levels are used to represent digital data?

4. What is the maximum data rate achievable, if a binary signal is sent over a 3-kHz channel whose *S/N* ratio is 40 dB?

5. If the CCITT standard 2048 kbps PCM stream is to be carried on a 50-kHz channel, what is the maximum tolerable *S/N* level?

6. The operating state diagram of a modem has two operating points on the *x*-axis. What modulation is used by the modem?

7. The operating state diagrams of a modem has states at the coordinates (1,1), (1,–1), (–1,–1) and (–1,1). What is the data rate achieved by the modem at 1200 baud?

8. A DCE converts an asynchronous transfer from the computer to synchronous transmission on the line. Asynchronous transmission uses one stop bit. If 100 characters are packed into a synchronous block, determine the transmission efficiency achieved by the DCE. Assume four bytes of overhead for every block.

9. List at least six differences between the postal and telephone systems and bring out the analogy between S&F and circuit switched connections.

10. A packet switching network has N nodes fully connected. What are the best, the average and the worst case transmission path lengths in hops?

11. A network offers both circuit and packet switching facilities. Given that:

    T_s = circuit set up time per station

    T_q = processing and queuing delay in each node for packets

p = data bits in a packet
d = overhead bits in each packet
R = data rates in bps over all links, and
k = number of hops or links for a connection in both circuit and packet switched modes

Compare the delay in sending a message of M bits in circuit and packet switched modes. Under what conditions the packet switching mode has a lower delay?

12 In Exercise 11, given that $T_s = 2$ s, $T_q = 0.02$ s, $p = 200$ bits, $d = 20$ bits, $R = 1200$ bps and $k = 5$, compute the packet and circuit switching delays for (a) $M = 1000$ bits, and (b) $M = 10{,}000$ bits.

13. In a packet switched network, determine the packet size that minimises the total delay.

14. A network offers datagram, virtual circuit (VC) and permanent virtual circuit (PVC) facilities. A user has 50 messages of 9600 bytes each to be transmitted over a period 10 hours in a day. Determine which one of the services is best suited for his requirements, given the following tariff:

Call setup charges per call = Rs. 5
One time special charges for setting up PVC for one day = Rs. 300
Packet transmission charges = Re. 1 per packet

The packet size is 120 bytes. Overhead bytes are eight and four for datagram, and VC and PVC, respectively.

15. An interactive computer user generates messages at an average rate of two messages per minute. The messages are on an average 40 characters long. If the line speed is 9600 bps, calculate the percentage utilisation of the line due to user generated traffic alone.

16. A circuit switched network is used for transmitting datagrams. If typical path setup time is 20 s what is the percentage overhead due to setup time if the average datagram length is 600 characters and the speed of the line is 2400 bps.

17. An intercity circuit switched connection has a total length of 450 km from subscriber to subscriber. The subscriber link to the city central exchange is 10 km on either side. The intercity link (between the two central exchanges) is capable of carrying 9600 bps and the subscriber lines are limited to 1200 bps. If the tariff structure is such that the charges are Re.1 per 10 km of intercity link per minute of holding time and Re.1 per intracity subscriber link for the entire duration, calculate the monetary loss to the subscriber due to 'last mile' problem when he uses the link to transmit a file of 36 kbytes.

18. State whether the statements given below are true or false. Justify your answers with reasons.
    (a) If an $(N + 1)$ entity obtains services from four (N) entities, the services have to be routed through four (N) SAPs.
    (b) The user data of (N) PDU is nothing but $(N - 1)$ PDU.
19. Can you think of analogies for OSI entities and SAPs, in radio broadcasting?
20. A communication channel which is affected by burst errors of length 16 bits or more, is known to give an error performance of 1 in 10^6 bits. If 100,000 frames of size 1000 bits each are transmitted through this channel, what is likely to be the overhead due to retransmission? Assume that an ARQ protocol is used and no errors occur during acknowledgments or retransmissions.
21. If the bit string 01101111110011111011 is subjected to bit stuffing, what is the output string? Assume a flag pattern of 01111110 for which bit stuffing is required.
22. A point-to-point channel uses a coaxial cable of 2000 km length. If the data rate is 9.6 kbps, for what frame lengths does idle RQ protocol give a channel utilisation of at least 50% ? The ACK frame size is 24 bits, and the error checking time is 1 ms.
23. Frames of 10,000 bits are sent over 2-Mbps satellite channel. The ACKs are always piggybacked. What is the maximum achievable channel utilisation for stop-and-wait protocol? Ignore time required for error detection.
24 In Exercise 23, estimate the channel utilisation if a pipeline protocol is used with 3-bit sequence number and selective repeat strategy.
25. A fully connected network with four nodes uses controlled flooding strategy for routing to ensure robustness. Flooding is controlled in the following way:
    1. All frames are uniquely numbered.
    2. Every node remembers the frame it has once forwarded.
    3. If a frame returns to a node, it is discarded and not forwarded again.

    Determine the number of times a frame is transmitted when originated by one of the nodes, and the number of copies received by the destination.
26. State with reasons which of the seven layers of the OSI reference model will play a role in the following circumstances:
    (a) A user with a line mode terminal wants to communicate with a user having screen mode terminal.

(b) The network supports a packet service with packet lengths of 1024 bytes. A user wants to send short messages of 128 bytes each.

(c) A user wants to acquire real time data from a remote site.

(d) A user would like to use a preferred path for communications.

(e) Traffic through a particular node is excessive and congestion occurs in that part of the network.

(f) A user wants to have secure communication for selected messages.

(g) A network offers secure communication facility.

27. A group of terminals are to be connected to a central site using ALOHA schemes over a 9600 bps channel. Frame length is 200 bits. Terminals generate, on an average, one frame every two minutes. For maximum throughput condition, estimate the number of terminals that can be supported if the scheme used in (a) pure ALOHA and (b) slotted ALOHA.

28. Show that a mean number of times a frame is transmitted in a pure ALOHA system is 2.7 under maximum throughput condition.

29. Measurements in a slotted ALOHA channel show that 10% of the slots are idle. What is the channel load G? What is the throughput S?

30. An ethernet LAN has a cable distance of 2 km between extreme nodes. If the data rate is 10 Mbps, what is the lower bound for the frame size? What is the channel utilisation at this frame size if there are no collisions?

31. A token ring LAN of length 10 km operates at a data rate of 10 Mbps. There are 50 RIUs each of which introduces a 1-bit delay. What is the token rotation time if

(a) all stations have data to send, and

(b) each station generates one frame of data every two seconds?

Assume a frame size of 5000 bits.

32. In a 50-station token ring, the frame format has 8-bit busy token, 8-bit destination address, 8-bit source address and the data in that order. The ring is designed such that the destination removes the frame from the ring. If the physical length of the ring is 5 km and the data rate is 20 Mbps, what is the walk time?

33. A MAN configuration uses two buses with active headend. The length of each bus is 20 km. There are 20 stations located at an interval of 1 km each. The station located at 1 km from the headend senses its downstream frame at time t_0. When should the station access the upstream for transmission?

34. A fibre optic token ring MAN is 200 km long and operates at a speed of 100 Mbps. The data from the ring is drained by the source. A token is reintroduced only after draining the full data. There are 100 stations in the ring. What is the ring utilisation if the frame size is (a) 2000 bits (b) 4000 bits?
35. In Exercise 34, determine the maximum number of simultaneous transmissions that may exist on the ring, if the token is reintroduced immediately after the transmission of the frame is over.
36. Suggest a protocol stack for a banking network.

11

Integrated Services Digital Network

Integrated Services Digital Network (ISDN) has been perhaps the most important development to emerge in the field of Computer Communications in the 1980s, and it will probably continue to dominate the developments in the 1990s too. Unlike many other developments, ISDN is a well conceived and planned area of development in the field of telecommunications. CCITT has been pioneering and guiding the efforts towards the development of ISDN. The process of digitalisation of telecommunication networks started in early 1960s when PCM was introduced for digital transmission of voice signals. Arrival of transistorised minicomputers opened up the possibility of digital switching in the late 1960s. CCITT was quick to recognise the feasibility of digital telecommunication networks and set up a study group called Special Study Group D in 1968 to look at a variety of issues related to the use of digital technology in telephone networks. This study group is the forerunner of today's Study Group XVIII set up in 1976, and has the responsibility for all ISDN related activities within the CCITT. The first formal definition of ISDN was given by the Study Group D in its recommendations, G.702, issued and adopted by CCITT in 1972:

> **Integrated Services Digital Network** An integrated digital network in which the same digital switches and digital paths are used to establish different services, for example, telephony and data.

It is interesting to note that a study, which started with the objective of exploring the use of digital technology, quickly shifted its emphasis to an Integrated Digital Network (IDN) and within a period of four years even foresaw the emergence of ISDN. From then on, the progress towards ISDN has been steady and unrelenting. In 1980, the first set of ISDN standards, G. 705, emerged, which laid down six conceptual principles on which ISDN should be based:

1. ISDN will be based on and will evolve from the telephony IDN by progressively incorporating additional functions and network features including those of any other dedicated networks so as to provide for existing and new services.

2. New services introduced into the ISDN should be so arranged as to be compatible with 64 kbps switched digital connections.
3. The transition from the existing networks to a comprehensive ISDN may require a period of time extending over one or two decades.
4. During the transition period, arrangements must be made for the interworking of services on ISDNs and services on other networks.
5. The ISDN will contain intelligence for the purposes of providing service features, maintenance and network management functions. This intelligence may not be sufficient for some new services and may have to be supplemented by either additional intelligence within the network or possibly compatible intelligence in the customer terminals.
6. A layered functional set of protocols appears desirable for the various access arrangements to ISDN. Access from the customer to ISDN resources may vary, depending upon the service required and on the status of evolution of national ISDNs.

The above principles clearly bring out the systematic and evolutionary approach to the development of ISDN. Pursuance of these conceptual principles led to a refinement of the definition of ISDN in 1984:

> An ISDN is a network, in general evolving from a telephony IDN, that provides end-to-end digital connectivity to support a wide range of services, including voice and nonvoice services, to which users have access by a limited set of standard multipurpose user-network interfaces.

The key points of the above definition have to be noted. First, the ISDN is an infrastructure to support a wide variety of services and is not a network designed for any specific service. Secondly, the end-to-end digital connectivity implies that the digitisation process begins right at the user premises. Thirdly, it should be possible to support every conceivable service on ISDN, for any such service is either a voice or nonvoice service. Finally, a small set of carefully chosen interfaces should enable the support of all possible services. The users of ISDN should not be burdened with too many specialised interfaces, but at the same time, an expensive universal interface should be avoided.

11.1 Motivation for ISDN

Three factors are responsible for the developments towards ISDN:

1. Sociological or societal needs
2. Economic necessities
3. Technological developments.

The rapid developments in various facets of society call for increasing and complex communication facilities. A biotechnologist today would like to examine a blood sample remotely, simultaneously compare the analytical results of other samples stored in a centralised database, consult his assistant who is presently in a laboratory some distance away, and report the findings as the investigation progresses, to his superior who is in another building. To meet such a demand, we need to electronically transmit the microscopic image of the blood sample and reproduce the same graphically on the computer screen of the biotechnologist, at a rate fast enough to faithfully reproduce the movements of living cells, etc. Simultaneously, the same communication medium is used to access a remotely located database and hold an audio conference with his assistant and superior.

As another example, a senior executive of a company, who often takes important decisions at home late in the evening or while on a holiday, would like to give instant effect to his decisions. This may call for access to different computer systems connected in the form of a network, electronic banking facilities, facsimile transmission and desktop image processing facilities, all in the place where he is at present. In other words, automobiles, homes, boats, hotel rooms and railway compartments will need to house electronic workstations. Automatic calling party identification and guaranteed transaction privacy and security are important signalling requirements of systems that can support such applications.

In effect, the society is demanding a telecommunication system that can support universal access to a host of services. In such a system, it should be possible for a user to attach to the network anywhere in the world the equipment of his choice to obtain a particular service. The user will be allotted a permanent identification number or code, like the income tax permanent account number or the social security number, which would be valid for his lifetime. No matter where a user lives, or how often he shifts residence, dialling the number assigned to the user would always ring his telephone, computer system or any other equipment. Similarly, irrespective of wherefrom a user obtains his services, the related charges are debited to his account. The society is looking for a telecommunication infrastructure that can carry voice, data, image, graphics, video, etc., simultaneously supported by sophisticated signalling systems. Society may demand many things, but are these technologically feasible, and if so, are these economically viable? We now discuss these points.

Traditionally, network providers have put up separate and independent networks to support different services. Telex network, data network, telephone network and CATV networks are examples of such a development. Can the network providers continue the same path of evolution for the plethora of new services demanded by the society? Independent networks call for separate administration, maintenance staff, and building for housing switching systems. The independent and duplicate infrastructural facilities

lead to high capital cost, low maintenance efficiency and high management cost. In addition, the network facilities are never fully utilised as the services are independently supported on different networks. The net result is that the overheads turn out to be excessive, leading to become economically unviable network services. Obviously, the independent network approach is not viable for future services. One should consider the possibility of sharing resources to keep the costs down. Thus, sheer economic necessity is forcing network providers to look for solutions where many different services may be integrated and supported on common network resources.

Searching for new solutions is of no avail unless technology developments make possible such solutions. In fact, it is the technology factor that brought about the independent network solutions earlier. The end equipments for different services were analog in nature and had different electrical, electronic, signal and communication characteristics. It was necessary to design different networks to suit each of these devices. For example, the voltage, current and bandwidth requirements of a teleprinter, a computer terminal and a telephone instrument are vastly different. Accordingly, the switching systems, signalling systems and transmission media have to be different. The desire of the network providers to use a common network infrastructure can fructify only if there are uniform electrical, signal and communication requirements for all types of services. Such a common substratum is provided by the digital technology. If the end equipment, switching, transmission and signalling systems can all function in the digital domain, the interface and communication requirements can be uniform for all the services. Today, the digital technology has matured to a level where all the above mentioned functions of a telecommunications network can be realised in the digital domain. It is interesting to trace the history of the process of digitalisation in the field of telecommunications.

Traditionally, digital technology always appeared first in the field of computers. In the late 1930s and the early 1940s, while the theory of digital communication (pulse code modulation, sampling theorem, etc.) was coming of age, digital computers were already being built. The first digitalisation step in telecommunications appeared when slow channel PCM technique was introduced for long distance transmission in the early 1960s. Transistorised or integrated circuit based minicomputers opened up the possibility of digital switching in the late 1960s. These computers were used in the exchanges, and switching was organised in space domain or time domain. Developments in the field of packet switching networks led to the possibility of digital signalling in the mid 1970s. Since late 1970s, digital end equipments such as digital telephone and digital facsimile are being developed.

Each Plenary Assembly of CCITT, which takes place once every four years, assigns specific topics related to ISDN to the concerned study groups for an in-depth study during the ensuing four-year period. The result of such an organised study in the last 20 years is that today we have a clear picture of

ISDN architecture and a plethora of services that can be supported in ISDN. However, the ISDN technology, design approaches and standards are still evolving and will remain the focus of development in the 1990s.

Thus, we are now at a point of time when there exists a synergy of social demands, technological feasibility and economic viability with regard to ISDN. Consequently, an era of ISDN will be heralded all over the world during the next twenty years or so. As must be obvious by now, ISDN is the apt nomenclature for the telecommunications network of the future in which all network functions will be handled in digital domain and a single network will support a set of integrated services.

11.2 New Services

ISDN will support a variety of services including the existing voice and data services and a host of new services. A short list of some of the important new services is:

1. Videotex
2. Electronic mail
3. Digital facsimile
4. Teletex
5. Database access
6. Electronic fund transfer
7. Image and graphics exchange
8. Document storage and transfer
9. Automatic alarm services, e.g. smoke, fire, police and medical
10. Audio and video conferencing.

A few of the services are described in the following sections.

11.2.1 Videotex

Videotex is a generic term for systems that provide easy to use, low cost computer based services via communication facilities. Three forms of videotex exist:

1. View data
2. Teletext
3. Open channel teletext.

View data is fully interactive videotex. This means that requests for information or service from a user are actually sent to, received by, and acted on by a centralised computer.

Teletext is broadcast or pseudo-interactive videotex service. Teletext users may select the information to be seen, the pace at which the information is to be displayed and, often, the sequence of display. The information is cast in the form of frames and a set of frames which is called a

magazine is recycled continuously. Teletext is a one-way communication system and there is no real interaction between the user and the computer.

Open channel teletext is totally noninteractive oneway videotex. With this form of videotex, the user receives preselected information in a predetermined order. There is no interaction, either real or apparent. The user has no control over the pace or sequence of display. He just has to watch what appears on the screen and pick up the information that is of use to him, as and when it appears on the screen. Open channel teletext may be classified into three categories according to the way the preselected information is displayed and the way the display channel is used:

1. Dedicated open channel
2. Open captioning
3. Closed captioning.

In dedicated open channel teletext, a separate transmission channel is dedicated for the display of preselected information. Open captioning shares a normal display channel and the teletext display appears at fixed intervals along with other programmes on the channel.

11.2.2 Electronic Mail

Electronic mail, popularly known as email, may be defined as the communication of textual messages via electronic means. Even the telex communication is electronic in nature. Then, what is the difference? There are many. While telex communication is terminal-to-terminal, electronic mail communication is user-to-user. In telex, messages destined to a number of users are sent to the same terminal from where it is distributed by an operator or a messenger. On the other hand, electronic mail is delivered to the mail boxes of individuals. Telex works on a circuit switched mode, whereas electronic mail is a store and forward (S&F) service. Electronic mail is a computer based messaging system, whereas the telex is generally not. Being a person-to-person communication system, electronic mail turns out to be a cheaper alternative to telephone conversation and eliminates the time spent in establishing phone calls. For a telephone conversation to materialise, both the calling and the called party must be present simultaneously. It is not unusual that Mr *X* calls *Y* and *Y* is not present. Some time later, when Mr *Y* is in, he returns the call only to find Mr *X* is not at his desk. In fact, some studies indicate that as much as 70 per cent of the business phone calls during business hours do not succeed in the first attempt due to unavailability of the called party. Electronic mail permits communication between two parties without the parties actually being present simultaneously. Privacy is ensured as the mail is delivered to individual's mail box, which can be opened only by the intended recipient.

Another important advantage of electronic mail is its ability to reduce the consumption of paper in the office. Internal memos and reports can be exchanged electronically without using paper. Being a computer based messaging system, files prepared using office automation packages like word processor, database manager and spreadsheet package can be easily exchanged as electronic mail. This facility has the potential of improving office efficiency considerably.

Early electronic mail systems were organised around a single time sharing or multiuser computer system, wherein electronic mail was exchanged among the users of the system. Since the late 1970s, electronic mail facility is being extended over networks. A typical configuration of an electronic mail system is depicted in Fig. 11.1. There are two major

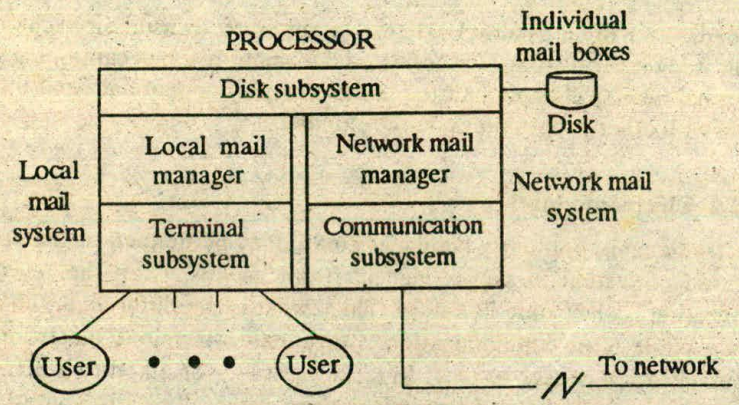


Fig. 11.1 A typical configuration of an electronic mail system.

components of the system: one to handle mail within the system and another to handle mail over the network. Both share a common disk storage where mail boxes are maintained. The terminals connected to the system may be distributed throughout the organisation and kept on the tables of individuals. Every registered user is provided with a mail box in the system. Any user may log on to the system from any of the terminals and send mail to another user on the same system or on the network. If the recipient is currently logged on, a message may be flashed on his terminal about the arrival of mail. Otherwise, he is informed about the mail waiting in his mail box as soon as he logs on to the machine. If the mail arrives when a recipient is logged on to the machine, he may suspend the present work and get into a conversational mode with the sender. Thus, real time exchange of messages is possible in an electronic mail system, if the two concerned parties are logged on to the same machine at the same time. Electronic mail, being a S&F service on a network, real time exchange may not be possible.

Users of homogeneous computer systems were the first to introduce electronic mail over networks. There are a number of computer networks in

the world which support electronic mail service. Some of the well known networks that offer electronic mail service are UUNET, BITNET, CSNET and JANET. As a next step in the development, internetwork electronic mail has been receiving the attention of the world network communities. An important problem in internetwork messaging is the highly varying mail address structures used by different networks. Users of a particular network are familiar with their own addressing schemes which need to be modified and recast into another format when sent over to another network. This is conveniently carried out by using gateways without burdening the user with the knowledge of different addressing schemes and the associated conversion process. Most of the networks in the world have established a gateway to one or more networks for messaging purposes. Thus it is possible now to send electronic mail to almost anyone on any network in the world. There is also an attempt to evolve a uniform addressing scheme for all networks.

In the context of Open System Interconnection (OSI) networks, electronic mail is considered as an application process running on the seventh layer. Standard electronic mail service components have been defined and approved by 1984 CCITT Plenary Assembly. These are known as X.400 family of standards for message handling systems (MHS). Presently defined standards do not deal with the user interface or the services available directly to the user. They do, however, specify a set of base services that can be used for building suitable user level interfaces. Table 11.1 gives a list of X.400 family of standards. X.400 service components are a part of the application layer of the OSI reference model.

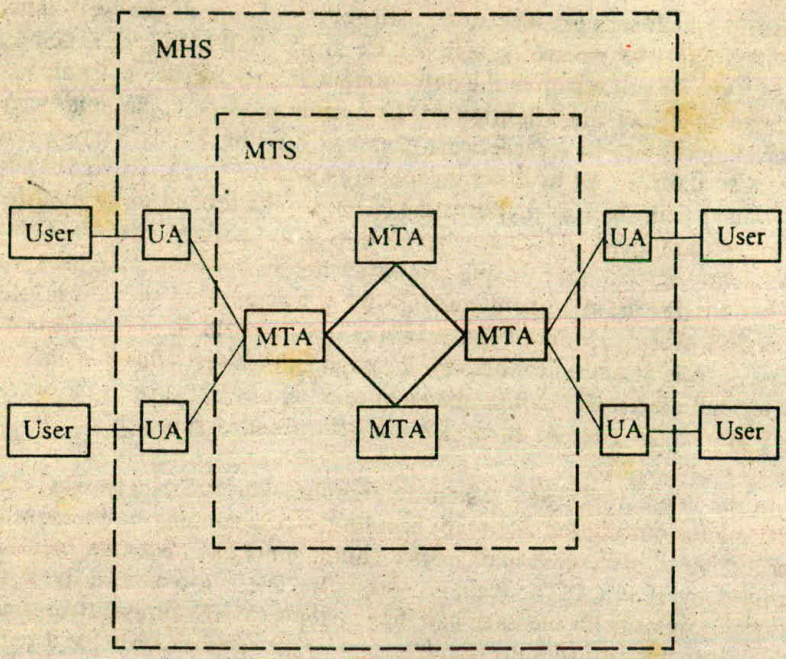
Table 11.1 X.400 Family of Standards

Number	Subject dealt with
X.400	System model — service elements
X.401	Basic service elements and optional user facilities
X.408	Encoded information-type conversion rules
X.409	Presentation transfer syntax and notation
X.410	Remote operations and reliable transfer server
X.411	Message transfer layer
X.420	Interpersonal messaging — user agent layer
X.430	Access protocols for teletex terminals

The MHS model as defined in X.400 has two types of entities:

- User agent entity (UAE)
- Message transfer agent entity (MTAE).

The X.400 model is shown in Fig. 11.2. User agent (UA) performs functions relating to the preparation, submission, and delivery of messages.

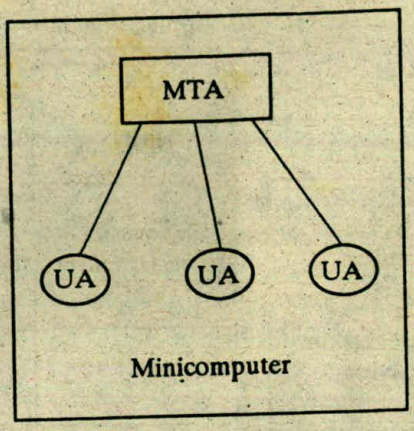


MHS = message handling system MTA = message transfer agent
MTS = message transfer system UA = user agent

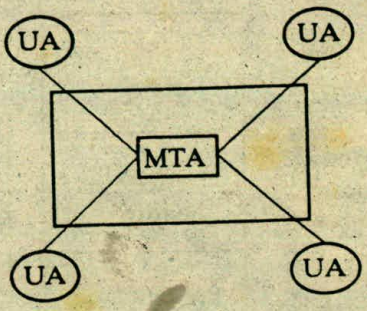
Fig. 11.2 X.400 message handling system model.

It also assists the user in other message functions such as filing, replying, retrieving and forwarding. Message transfer agent (MTA) is concerned with transfer of messages across the network and functions in an environment designated as message transfer system (MTS). It obtains messages from the source UA and delivers the same to the destination UA. On accepting a message, the message transfer agent (MTA) performs either a delivery function or a routing function. If the destination UA is in the same system as the MTA or is attached to the MTA directly, then the MTA performs a delivery function; otherwise it performs a routing function.

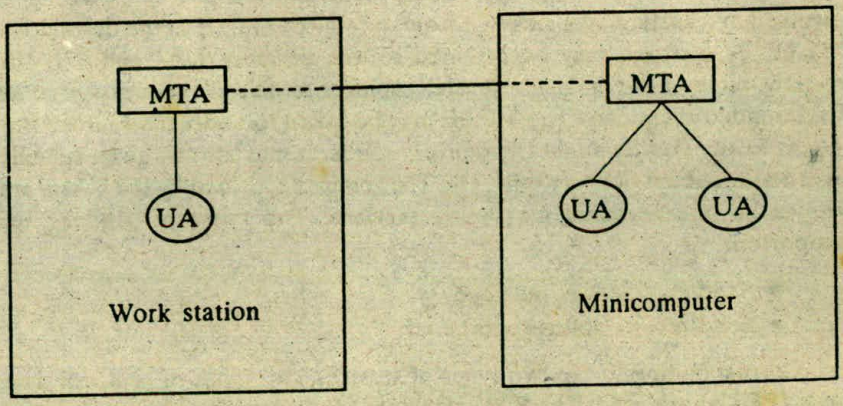
Depending on the physical location of the MTAs and UAs, a number of different physical realisations of email systems are possible. Some typical realisations are shown in Fig. 11.3. In general, in a multiuser system, a single MTA and many UAs are present. In a personal computer no MTA is usually implemented and there is only one UA. In a workstation, one MTA and one UA are present. In a computer dedicated to mail transfer functions, there is only one MTA which interacts with a number of UAs outside the machine. Figure 11.3(a) shows an electronic mail configuration around a mini-computer system, Fig. 11.3(b) a number of personal computers attached to



(a) Minicomputer based



(b) Dedicated mail transfer system



(c) Workstation minicomputer configuration

MTA = message transfer agent UA = user agent

Fig. 11.3 Email system configuration.

the dedicated mail transfer system, and Fig.11.3(c) a workstation attached to a minicomputer.

In the context of the OSI reference model, X.400 MHS is a part of the application layer. The UA and the MTA functions are treated as sublayers as shown in Fig. 11.4. In systems where only the UA function is implemented, the message transfer sublayer comprises a submission and delivery entity (SDE). The entity acts as an interface between a local UA and a remote MTA. In this sense, SDE implements a remote procedure call protocol. In the intermediate systems, only the message transfer agent entity (MTAE)

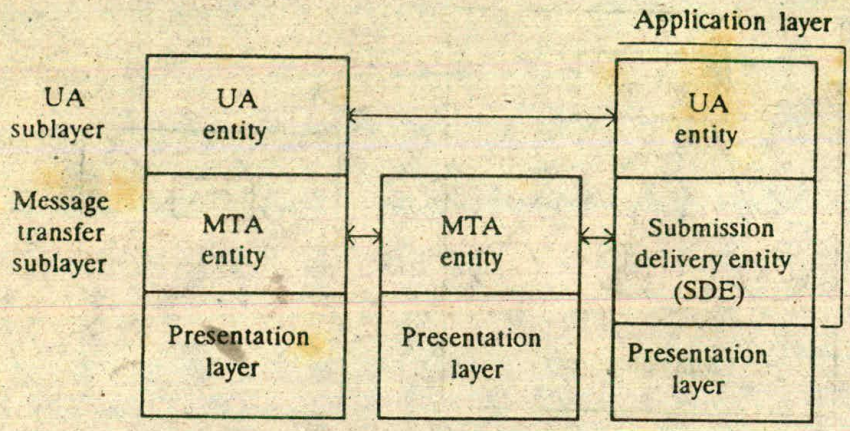


Fig. 11.4 X.400 in the context of OSI model.

plays a role and acts as a relay. Zero or more intermediate systems may be involved in transferring a message from the source MTAE to the destination MTAE. A message may be destined to one or more UAs and it is the responsibility of the message transfer sublayer to deliver the message to all the intended recipients. The MTAE may be asked to notify the source User Agent Entity (UAE) about the delivery status. It may also be given specific instructions about delivery time, etc. The services provided by the UAEs are known as interpersonal messaging services. Two types of services are supported:

- Send/receive user message
- Send/receive status reports.

Corresponding to the two types of services, two types of protocol data units (PDUs) are used. The PDU structure for exchanging user messages is shown in Fig. 11.5. The message structure consists of three parts. The body of the message is the actual user information, which may be considered as the user data unit in the OSI parlance. To this is added the control unit or the

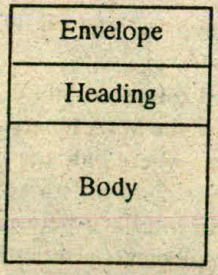


Fig. 11.5 X.400 user message format.

header of the user agent sublayer to form the UA layer PDU. This PDU, which forms the data unit for the message transfer sublayer, is then passed on to that sublayer. Here, the message transfer control unit, known as envelope, is added to form the PDU for exchange between MTAEs. The PDU for exchanging status reports contains only a header and no user information. The header fields are, however, different from those of the message PDU header.

The details of interpersonal messaging and message transfer services are spelt out in X.400 series specifications (see Table 11.2). They are classified as follows:

1. Basic facilities
2. Essential options
3. Additional options.

Basic facilities are an integral part of any implementation of X.400 messaging system. These features are exercised by the user every time he uses the X.400 system. Optional features are at the discretion of the user, either for a specific message or a set of messages. Essential optional features are to be implemented mandatorily by the service producer whether the user needs them or not. Implementation of additional optional features is not mandatory and is left to the discretion of the service provider. In the case of interpersonal messaging, some optional features are essential for either source or destination UAE, but not for both. Some examples of the features for both message transfer and interpersonal messaging are given in Table 11.2.

Table 11.2 Some Features of X.400 Services

Basic	Essential options	Additional options
I. Interpersonal Messaging Service Features		
Message identification	Originator indication	Autoforwarded indication (for sending)
	Autoforwarded indication (for reception)	
II. Message Transfer Service Features		
Submission/Delivery time stamp Message identification	Deferred delivery Delivery notification	Return of contents if undelivered Hold for delivery until UA is ready

11.2.3 Facsimile

Document exchange through facsimile systems is emerging as a major application of telecommunication systems. Unlike electronic mail, facsimile

is capable of transmitting and receiving printed matter which may include graphics, drawings, pictures, handwritten texts, etc. Two types of facsimile systems exist:

- Photographic facsimile
- Document facsimile.

In photographic facsimile, the gray level information is transmitted and printed in addition to black and white. Typically, there are 8 or 16 gray levels that can be recognised by the system. Document facsimile system handles only black and white levels, i.e. only two gray levels. Document facsimile system is more popular than the photographic system. The receiver/transmitter functions applicable to both the types of facsimile systems are shown in Fig. 11.6. A scanner converts the print material into a set of points

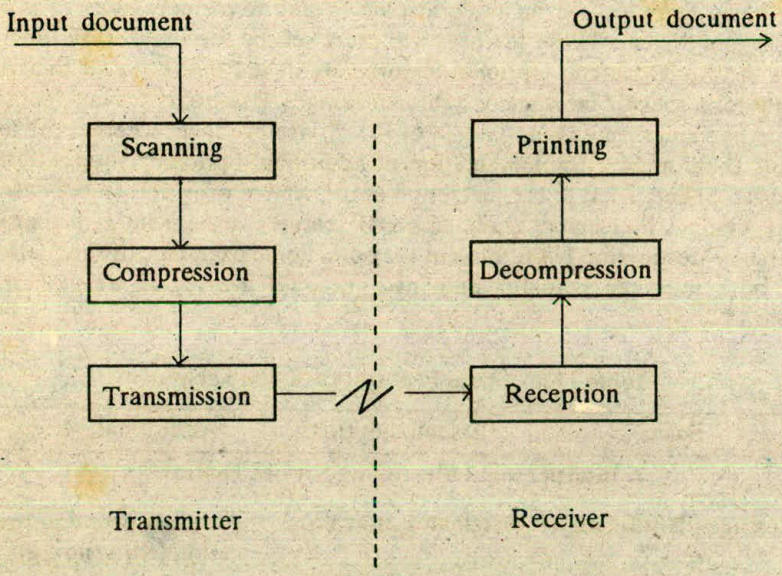


Fig. 11.6 Functions of a facsimile system.

represented by bit-mapped electrical signals. In document facsimile, every point is represented as a zero or one, whereas in photographic facsimile, each point is represented by 3 or 4 bits giving the gray level values. Scanning resolution is of the order of 4–16 dots per mm (100–400 dots per inch) horizontally across the page and 4 – 8 lines per mm (100–200 lines per inch) vertically down the page. An A4 size page (210 × 297 mm) generates about 3 million picture elements (pels). At a transmission rate of 9600 bps, the time required to transmit 3 million bits is of the order of 300 seconds or about 5 minutes. Facsimile information is transmitted using regular telephone lines at the same tariff as the telephone charges. If we were to transmit the

scanned page as it is, the cost of transmitting a page across two major metropolitan cities in India during working hours would work out to be about Rs.150. This is too expensive. Consequently, there is a need to compress the scanned data before it is transmitted. This is the second step in facsimile transmission. There are two types of compression techniques:

- Information preserving techniques
- Approximate techniques.

CCITT has standardised on two compression techniques, both belonging to the first category. These techniques reproduce an exact replica of the scanned image, whereas techniques belonging to the second category approximate the original in the output. The two techniques standardised by CCITT are:

- Modified Huffman (MH) technique
- Modified READ (MR) technique.

Before applying the Huffman or relative element address designate (READ) technique, the scanned information is coded using a basic technique known as **run-length coding.** In a document facsimile system, the pels are represented by zeros and ones. A quick observation of a printed document, taking into account the horizontal scan resolution of about 200 dots per inch, reveals that the scanned information would contain strings of zeros and ones. Longer strings of zeros and ones are more probable than short strings. Run length coding conveys a count of continuous ones and zeros instead of the pattern itself. Huffman coding is based on the principle that a more efficient coding scheme can be evolved by using short code words for frequently occurring symbols and long code words for sparingly occurring symbols, instead of using a uniform size for all symbols. In the case of document facsimile, longer strings of zeros or ones are coded with short code words as they are more probable and vice versa with the shorter strings. However, the string sizes vary over a very large range, theoretically one to a few millions per page. Consequently, the average size of the code word is large. To overcome this problem, the Huffman code is modified to view the run-lengths in two parts and code them independently, taking into account their probability of occurrence. The two parts of the run length are known as **made-up code part** and the **terminating code part.** The first part is used for coding run lengths that are less than 64 and the latter part is used for run lengths greater than or equal to 64. A detailed treatment of the coding scheme is beyond the scope of this text. Interested readers may refer to relevant CCITT standards or Further Reading [5].

Relative element address designate (READ) code is based on the principle that further coding efficiency may be gained by coding the relative position of changing elements. There is a strong correlation between the black-white patterns of two adjacent scan lines in a document. This fact is exploited in modified READ (MR) coding. A changing element is coded in

terms of its distance from a preceding changing element on the same or on the previous line. Which one of the two elements is chosen as reference depends on the relative position of the element to be coded. In general, the reference element is so chosen as to minimise the size of the code word required. A detailed treatment of this coding scheme is beyond the scope of this text.

Many studies have been conducted to assess the effectiveness of MH and MR coding schemes over a representative set of document pages. Both coding schemes achieve considerable compression. It has been observed that the MR coding is more effective than the MH coding. Compression ratios of 1:8 and 1:16 on an average have been reported for MH and MR coding schemes, respectively.

Since their introduction in the early 1970s, four models of facsimile machines, known as Group I, II, III and IV machines, have evolved. Group I and II machines are analog in nature, whereas Group III and IV use digital technology. Group I, II and III machines have been designed for operation with PSTN, whereas Group IV machines with ISDN. A comparison of the features of these machines is given in Table 11.3. Group IV machines have further models known as class 1, 2 and 3 machines. Class 2 and 3 machines have a horizontal and a vertical resolution of 11.8 pels/mm each. Class 2 machine is capable of receiving teletex messages, whereas class 3 machine can receive and transmit teletex messages.

Table 11.3 Features of Facsimile Machines

Feature	Group I	Group II	Group III	Group IV (class 1)
Resolution				
Horizontal (pels/mm)			7.8	7.8
Vertical (pels/mm)	3.85	3.85	7.8	7.8
Transmission time per page (s)	360	180	< 60	< 10
Compression scheme	None	Spectrum compression	MH (MR-optional)	MR

Another form of facsimile which is currently becoming available is the PC based facsimile. Here, a facsimile card is attached to a personal computer which can receive and send messages from/to other facsimile machines. This system is capable of sending messages, graphics, etc. prepared on a PC as a facsimile message. It is also capable of receiving facsimile messages and storing the same in the computer for viewing on the screen. The system is

catching up as a low cost alternative to standard facsimile machines, as the add-on card is less than one-tenth of the cost of a facsimile machine. It also uses ordinary paper for hardcopy printout as against the expensive thermal paper used by most facsimile machines. The lifetime of the thermal paper is about six months, after which the information in the paper starts fading. There is no such problem with computer based facsimile. Another interesting feature of the PC based solution is that it eliminates the cost of receiving 'junk fax'. Marketing companies are finding the large scale installation of facsimile machines at the user premises a boon for advertising. They 'fax' their product information to potential customers. While the telecommunication costs are borne by the marketing company, the reproduction cost is borne by the user. A user has no way of stopping an unwanted facsimile message and ends up paying for the 'junk fax'. This expenditure is avoided in a PC based facsimile system as the user may view the message on the screen and decide to delete the same if unwanted.

With the advent of ISDN high speed digital facsimile systems, transmitting at speeds of 384 kbps or above is a distinct possibility in the 1990s. These systems may be used for exchanging slow varying pictures or animated graphics.

11.2.4 Teletex

Teletex is an upgrade to the conventional telex service. The terminal-to-terminal communication service of telex will be turned into office-to-office document transmission system by teletex. Teletex envisages direct communication between electronic typewriters, word processors and personal computers. These units have storage for transmitting and receiving messages. The use of such equipments considerably enhances the character set available for document preparation. In addition to the standard character set, a rich set of graphic symbols and a comprehensive set of control characters are supported in teletex. The set of control characters helps in the preparation and reproduction of documents. In particular, they permit the positioning of the printing element, specification of page orientation, left and right margins, vertical spacing and the use of underlining. The page control feature allows standard A4 size papers to be used for receiving messages instead of the continuous stationery used in conventional telex systems.

A background/foreground operation is envisaged in teletex. Transmission/reception of messages should proceed in the background without affecting the work which the user may be carrying out in the foreground with the equipment. In other words, a user may be preparing a new document, while another document is being transmitted or is received. The teletex would also maintain compatibility with the present telex system and inter-operate with them. Teletex procedures call for the exchange of header information before the actual document transfer takes place. The header information consists of four parts:

Part 1 : Destination id
Part 2 : Originator id
Part 3 : Date and time stamp
Part 4 : Document reference.

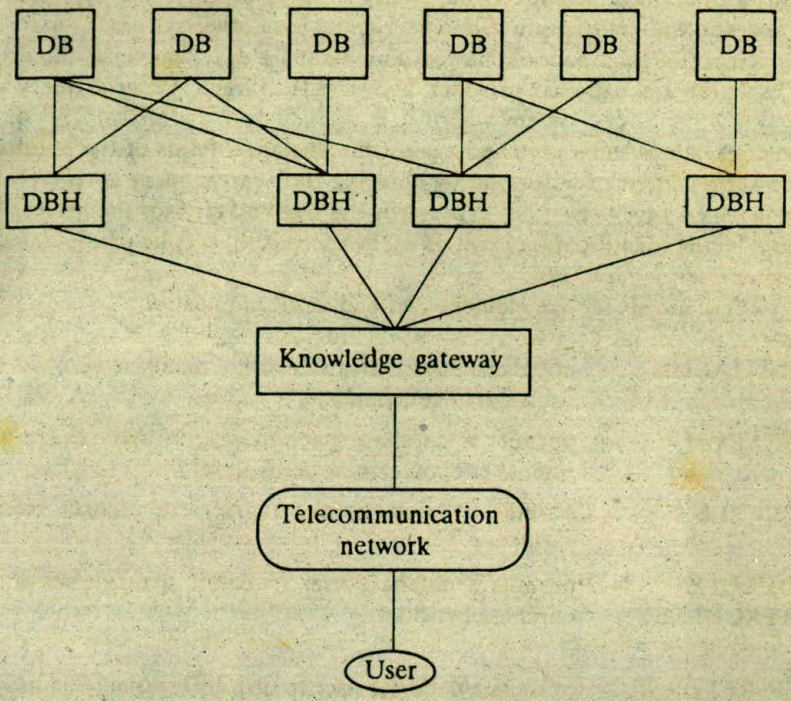
Twenty four characters are used for source/destination ids, 14 characters for date and time stamp, and 7 characters for document reference which also specifies the number of pages in the document. Destination/Source id consists of four fields:

Field 1 : Country/Network code
Field 2 : National subscriber number
Field 3 : Reserved for future use
Field 4 : Terminal/Owner code.

The number of characters allotted to each of the above fields is variable subject to a maximum for each field, the total being 24 characters.

11.2.5 Database Access

We are in an age of information explosion. There is so much of new information being generated/published periodically that it has become almost impossible to keep track of the new information in any field. It is in this context that electronic databases are becoming a boon. Here, a user can, by a suitable search query, obtain all the information generated in a particular topic. There are over 5000 databases in different parts of the world, covering a wide variety of subjects which include humanities and social sciences, science and technology, engineering and industry. These databases may be accessed online using the telephone network, modem and a personal computer. Accessing a database calls for an agreement to be signed with the database vendors and each vendor may have different terms and conditions. For an individual or an organisation to deal with a number of database vendors is cumbersome and difficult. In order to facilitate access to various databases through a single contact point, agents have sprung up and set up what are known as database 'hosts'. These host machines are connected to a number of databases. By entering into an agreement with one database host agent, the user can gain access to a variety of databases. While a database host simplifies the access procedure to databases, the user will still have to consult directories and other publications to decide on the right type of database to be accessed, involving considerable knowledge, expertise and effort on his part. Casual users and even frequent users would like a simpler way of accessing information of interest to them. In order to assist the users select the right type of database and to provide them transparent access to information, a new type of service is being introduced by setting up what are known as **knowledge gateways**. The scheme of access is shown in Fig. 11.7. The knowledge gateways use concepts from artificial intelligence and the techniques of expert systems. A user interacts with the knowledge



DB = database DBH = database host

Fig. 11.7 Electronic access of information.

gateways which provide him expert help to choose the right database pertaining to the subject of interest.

Interaction procedures with knowledge gateways, database host computers and databases are not standardised and vary considerably from one system to another and from one vendor to another. Some systems may be accessed over data networks whereas others accessed via PSTN only. Considering that database access will emerge as a major application in the 1990s, ISO has initiated standardisation efforts for **search and retrieval (SR)** of databases. In 1990, SR service definitions and the associated protocol specifications reached the second draft proposal stage and these are likely to be accepted as draft standards by 1992. The draft proposals are :

- DP 10162 – SR Service Definition
- DP 10163 – SR Protocol Specification.

The SR service is modelled as a pair of application processes, each of which is located in a separate open system. Within each application process, there are two types of functions: local processing functions, and communications related or OSI related functions. The former deals with activities

such as database manipulation, query evaluation etc, and the latter is concerned with interaction with a peer entity in another system.

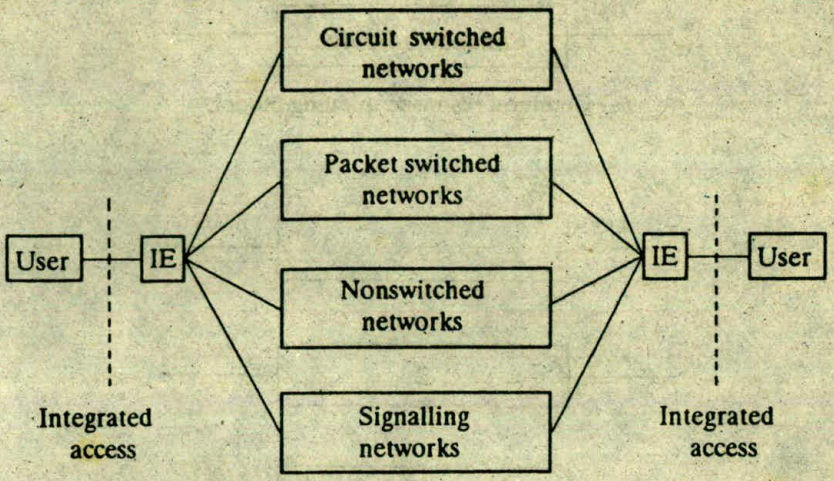
An application process may act as a database user, known as the origin or as a potential database provider, known as the target. A search query may be sent to a target system where it is applied to a database. The query specifies values to be matched against the specified fields of the database. The subset of records formed by applying the search query is termed the result set. A result set is not necessarily the selected records themselves but could be the identification pointers to the records. It may be manipulated in further query operations.

There are six service elements defined in SR application:

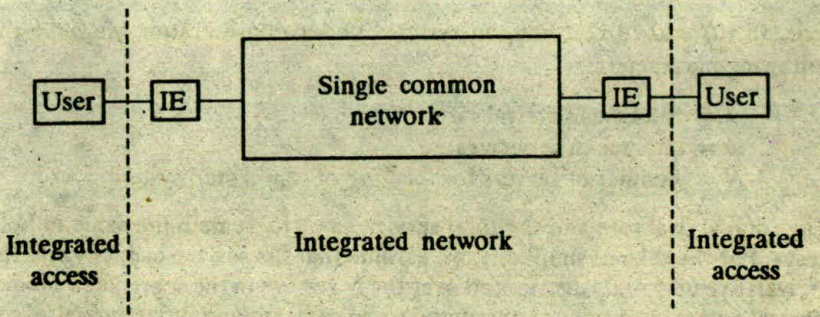
INITIALISE	–	initialises communication with a database provider for subsequent commands
SEARCH	–	permits a database user to search a database at a remote site; outcome is a result set
PRESENT	–	permits a database user to retrieve records from a result set
DELETE RESULT SET	–	permits a database user to delete a result set at the database provider
SR-RELEASE	–	permits a database user to orderly terminate an association
SR-ABORT	–	permits a database user or a database provider to request abrupt termination of the association.

11.3 Network and Protocol Architecture

Network architecture of ISDN is defined to follow an evolutionary path. It is natural that an evolutionary approach is taken with regard to ISDN. The wide range of telecommunication equipments and networks that exist today cannot be replaced overnight by ISDN. As a first step, the existing analog telephone networks should be converted to digital networks. These networks should then be operated along with other existing data and signalling networks. Such networks are termed Integrated Digital Networks (IDN). Thus, initially, ISDN will function as a collection of a few segregated networks to which an integrated access is provided as illustrated in Fig. 11.8(a). It is seen from the figure that four different types of networks form part of the segregated architecture. An ISDN exchange and a suitable user-network interface permit an integrated access to the network facilities. In the long run, as the economics and technology dictate, the segregated infrastructure will be replaced by an integrated infrastructure as shown in Fig. 11.8(b).



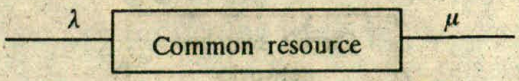
(a) Segregated architecture.



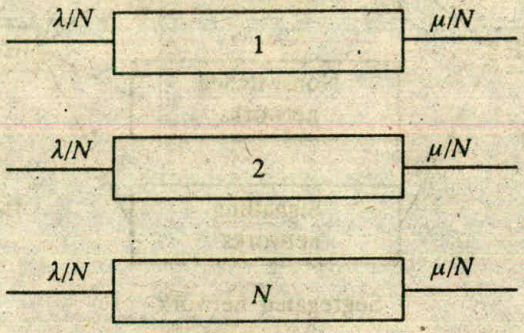
(b) Integrated architecture

Fig. 11.8 ISDN architecture.

The motivation for moving into an integrated infrastructure is based on factors like operational issues, management complexities and economies of realisation as discussed in Section 11.1. It is also based on the general belief that the pooling of resources would lead to an efficient utilisation of the resources and better performance of the system. This is true in most cases. We illustrate this point by modelling the segregated and the integrated networks as M/M/c and M/M/1 queuing systems respectively, and by analysing their performance. The models used are shown in Fig. 11.9. The dynamics of the queuing systems are governed by the steady state birth-death



(a) Common resource queuing model.



(b) *N*-independent resources queuing model.

Fig. 11.9 Different queuing configurations.

Eqs. (8.10) and (8.11). For the systems under consideration we have the following parameters:

λ = total mean arrival rate
μ = total mean service rate
N = number of servers for the case of segregated system

The total arrival rate and the total service capacity remain the same in both cases. For the sake of simplicity, we assume that the service capacity and the arrivals are uniformly distributed over the N servers in the segregated model. Arrival and service rates are assumed to be constant and independent of the state of the system. Under these conditions, Eqs.(8.10) and (8.11) reduce to

$$\lambda p_{k-1} + \mu p_{k+1} - (\lambda + \mu)p_k = 0 \quad \text{for } k \geq 1 \tag{11.1}$$

$$p_1 \mu = p_0 \lambda \quad \text{for } k = 0 \tag{11.2}$$

For different values of k, we get

$$\begin{aligned} p_1\mu &= p_0\lambda \quad \text{or} \quad p_1 = \rho p_0 \\ p_2\lambda &= p_1\lambda \quad \text{or} \quad p_2 = \rho^2 p_0 \\ p_3\lambda &= p_2\lambda \quad \text{or} \quad p_3 = \rho^3 p_0 \\ &\vdots \qquad\qquad \vdots \\ p_k\mu &= p_{k-1}\lambda \quad \text{or} \quad p_k = \rho^k p_0 \end{aligned} \tag{11.3}$$

where $\rho = \lambda/\mu$ is known as traffic intensity. To eliminate p_0 from Eq. (11.3), we use the fact that the probability sums to 1:

$$\sum_{k=0}^{\infty} \rho^k p_0 = 1 \tag{11.4}$$

$\sum_{k=0}^{\infty} \rho^k$ is a geometric series and the sum is given by $1/(1-\rho)$. Therefore,

$$p_0 = 1 - \rho, \qquad p_k = (1-\rho)\rho^k \tag{11.5}$$

Let us now use the mean waiting time of a customer T, as a measure of performance of the system. To determine T, we first determine the mean number of customers in the system M, by using the state probabilities

$$M = \sum_{k=0}^{\infty} k p_k = (1-\rho) \sum_{k=0}^{\infty} k\rho^k$$

Now, differentiating both sides of the equation

$$\sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho}$$

with respect to ρ, we get

$$\sum_{k=0}^{\infty} k\rho^{k-1} = \frac{1}{(1-\rho)^2}$$

Multiplying both sides by ρ, we obtain

$$\sum_{k=0}^{\infty} k\rho^k = \frac{\rho}{(1-\rho)^2}$$

Therefore,

$$M = \frac{\rho}{1-\rho} \tag{11.6}$$

We now use the well-known Little's result which states that the mean number of customers in any queuing system is equal to the product of the mean waiting time for a customer and the mean arrival rate. The result is valid for all queuing systems irrespective of the arrival time or service time distributions. Hence,

$$M = \lambda T \quad \text{or} \quad T = M/\lambda$$

For the integrated model, we have

$$T_I = \frac{\rho}{1-\rho} \times \frac{1}{\lambda}$$

Substituting $\rho = \lambda/\mu$, we get

$$T_I = \frac{\lambda}{\mu(1-\lambda/\mu)} \times \frac{1}{\lambda} = \frac{1}{\mu - \lambda} \tag{11.7}$$

For the segregated model, we have

$$T_s = \frac{1}{(\mu/N) - (\lambda/N)} = \frac{N}{\mu - \lambda} = NT_I \tag{11.8}$$

Equation (11.8) shows that the delay performance of the segregated system deteriorates by N times if the integrated system is broken into N segments with $(1/N)$-th of the capacity of the integrated system. It should be noted that this analysis is valid only under the following assumptions:

1. The total capacities of the integrated and segregated networks are identical.
2. Input services have identical requirements in both the cases.
3. Service rate distributions are identical.

These assumptions may not be valid in ISDN environment. Incoming traffic characteristics vary for different services like voice, data, facsimile etc. Similarly, the service time characteristics may also be different. Under the circumstances, it is possible that the integrated network performs worse than the segregated one. An analysis in this regard is beyond the scope of this text and the interested readers are referred to Further Reading [5].

The protocol architecture of ISDN is layered according to the OSI reference model; this is as shown in Fig. 11.10. However, the specific functionalities of the various layers are different from those of the OSI layer functions. The differences stem from the different fundamental characteristics of ISDN and data networks. While ISDN is largely a circuit switched network with provision for store and forward (S&F) communication, data networks are largely S&F networks with provision for virtual circuit switching. As a network, ISDN is largely concerned with the lower layers 1–3. Layers 4–7 are the concern of the service providers. Accordingly, the lower three layer functions are known as **bearer service functions**. Services offered by service providers are known as **teleservices**. Functionalities of layers 1–7 are required to offer teleservices which are built over bearer service functions. Layers 4–7 functionalities are not yet fully def ined. Some of the important functions performed in the lower three layers are:

Layer 1 (Physical Layer)

1. Encoding and decoding of signals
2. Transmission of B, D, and H channel data
3. Multiplexing to form basic or primary rates
4. Activation and deactivation of physical circuits.

Layer 2 (Data Link Layer)

1. Establishing and clearing datalinks
2. Error, flow and congestion control
3. Synchronisation

Layer 3 (Network Layer)

1. Addressing and routing
2. Establishing and clearing network level connections
3. User-to-user signalling
4. Network level multiplexing
5. Internetworking multiplexing.

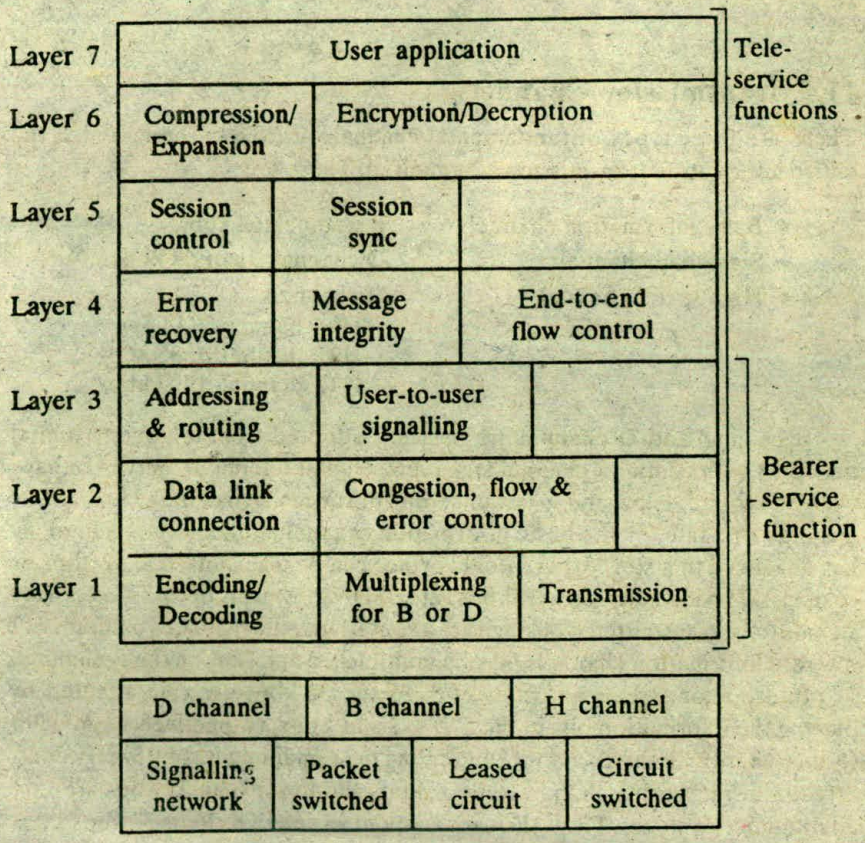


Fig. 11.10 ISDN protocol architecture.

Transmission channels B, D and H are discussed in Section 11.4. Channel level multiplexing is also dealt with in that section. Basic and primary rates are discussed in Section 11.5 which also deals with some layer 3 functionalities. User-to-user level signalling is covered in Section 11.6.

Addressing structure is discussed in Section 11.7, and Section 11.8 deals with interworking of networks. Further details of bearer and teleservices are presented in Section 11.8. Layer 1 protocol definitions apply uniformly to all transmission channels and services. But, from layer 2 upwards, the protocol structure differs for different channels and services. For example, in layer 2, a new protocol, LAP–D, has been defined for handling D Channel information. Level 2 of X.25 (see Section 10.4) is used for a packet switched connection over B channel. LAP–D protocol is modelled after LAP–B protocol of X.25, which has been discussed in Section 10.4. Other functions of layer 2 are similar to datalink layer functions of the OSI reference model, discussed in Section 10.4. Hence, ISDN layer 2 functionalities are not covered in this chapter. Interested readers may refer to Further Reading [2, 4, 6].

11.4 Transmission Channels

There are three types of fundamental channels in ISDN around which the entire information transmission is organised. These are:

- Basic information channel — B channel, 64 kbps
- Signalling channel — D channel, 16 or 64 kbps
- High speed channel. — H channels
  - H0 channel, 384 kbps
  - H11 channel, 1536 kbps
  - H12 channel, 1920 kbps

B channel and D channel are basically adopted from telephone digital networks with common channel signalling. Digital telephone networks have evolved around A-law and μ-law encoding of speech signal at 64 kbps. ISDN has the same rate for the basic information channel although the technology has advanced to a stage where quality voice can be transmitted at 32 kbps or even less. Having adopted a 64 kbps rate, ISDN, however, permits lower rate signals to be transmitted by using **rate adaptation**. Alternatively, a number of low rate information channels may be multiplexed and sent on the B channel.

Information streams at rates of 8, 16 and 32 kbps are rate adapted by placing the information bits in the first 1, 2 and 4 bits, respectively of an octet (8-bit quantity) in the B channel and filling the remaining bits by binary ones. Streams at rates other than 8, 16 or 32 kbps and lower than 32 kbps are rate adapted in two stages. First, they are adapted to one of 8, 16 or 32 kbps rates and then to 64 kbps by using the technique described above. Rate adaptation in the first stage is done by creating a frame similar to HDLC frame with the entire capacity filled in by the flag bits. The selection of frame size and frame rate are governed by the user information stream rate and the rate to which it is being adapted. Let

S = frame size
I = information stream rate in bps
R = rate in bps to which the information stream is being adapted; 8, 16 or 32 kbps
D = number of information bits in each frame
R_F = frame rate in frames per second

The frame size and the frame rate must be so chosen as to satisfy the equations

$$I = R_F \times D, \qquad R = R_F \times S$$

For information streams with rates between 32 and 64 kbps, a single-stage rate adaption is used by padding unused bits with binary ones in a 64 kbps stream.

B channel may be used efficiently by multiplexing two or more low rate signals. Multiplexing is limited only to 8, 16 and 32 kbps streams. All other data rates must be first rate adapted to one of these rates and then multiplexed. Two schemes of multiplexing are suggested:

- Fixed format multiplexing
- Flexible format multiplexing.

The former is less complex than the latter and is generally suitable when the mixture of subrate streams is known in advance. In this scheme, an 8 kbps stream may use any of the eight bit positions in a B channel octet as information bit; a 16 kbps stream may use any of the four twin bits (1,2), (3,4), (5,6), or (7,8), and a 32 kbps stream may use one of the two quadruplets (1, 2, 3, 4) or (5, 6, 7, 8).

Fixed-format multiplexing may not lead to efficient B channel utilisation if the mixture of subrate streams vary dynamically. Consider a situation where three 16 kbps streams, 1, 2 and 3 are multiplexed in bits (1,2), (3,4) and (5,6), respectively of the octet. Let the data from stream 2 end and a new 32 kbps stream come up. The new stream cannot be accommodated in the octet under the fixed-format scheme, even though there are four vacant bit positions in the octet. Flexible format multiplexing scheme permits accommodation of this stream in the octet. In this scheme, a new subrate stream may be added by inserting each successive bit of the new stream in the earlier available vacant bit positions. It may be noted that under the flexible format scheme, subrate streams can be always multiplexed to the maximum limit of 64 kbps. We have so far been considering the use of B channel in circuit switched mode. ISDN permits the use of B channel in packet switched, semi-permanent and permanent modes as well. In packet switched mode, the data exchange is done by using X.25 protocol. Semi-permanent and permanent modes correspond to short-term and long-term connections. It must be remembered that when the B channel is used in modes other than packet switched mode, all multiplexed channels must be to the same destination.

The D channel is primarily used for carrying signalling information to control and monitor ISDN services. It acts as the common signalling channel for all services. As an example, call establishment for a voice call using a B channel is entirely handled over the D channel without the involvement of the B channel. Information transfer on the D channels is governed by a layered protocol along the lines of ISO-OSI reference model. Spare capacity of D channel may be used to carry other packet switched data conforming to the layered model. Main applications envisaged on the D channel are user-to-user signalling, telemetry signals for energy management etc., and low speed videotex and computer terminal data.

H channels are used for high speed applications like high speed facsimile, video, high fidelity audio and high speed data. A user may also use an H channel as a high speed trunk on to which a number of low rate information streams are multiplexed. H11 channels at the 1526 kbps provide international compatibility, whereas H12 at 1920 kbps may be used in countries which are currently operating the 2048 kbps channel.

11.5 User-Network Interfaces

Comprehensive user-network interface definitions are key to ensuring worldwide ISDN compatibility. An excellent example of an interface standard that serves us so well and yet goes almost unnoticed is the electrical power user interface. We can purchase an electrical appliance almost anywhere in the world and plug it in our house socket and expect it to work. In fact, we are totally unmindful of the internal wiring, power distribution or generation scheme or the physical dimensions of the socket etc. as the final user interface conforms to a universal standard. By contrast, data communications suffer due to the presence of a variety of interface connectors, pin assignments and software interfaces.

In ISDN, user-network interfaces have been given careful consideration to avoid potential inconsistencies that may arise. ISDN caters to a variety of services such as voice, data, telemetry and image. In a situation like this, one encounters conflicting requirements. On the one hand, a number of custom designed interfaces may ideally suit each service but would lead to a proliferation of interfaces. On the other hand, one single multipurpose interface may turn out to be an 'overkill' for most of the services. One needs to adopt a via media. Keeping such factors in mind, two information rate access interfaces have been standardised for ISDN:

- Basic rate access
- Primary rate access.

In some sense, these rate accesses are analogous to 5 A and 15 A power sockets at home. The basic rate access caters to low bit rate services and the primary rate access caters to services that demand high bit rates in the same

way as the two types of power sockets cater to low-power and high-power electrical appliances.

The selection of access interfaces is somewhat influenced by the ISDN approach plan to achieve rapid and cost effective realisation. The plan is based on two factors:

- Use of much of the existing telephony networks; expensive replacement is not to be resorted to except for some special function or locations.
- Use of low cost VLSI technology to add new functionalities and to provide upgrade capabilities.

While the existing telephone networks provide extensive world wide connectivity with the associated infrastructure of buildings and equipment, their use imposes many constraints on the design of ISDN user-network interfaces. The existing digital switching and transmission systems are based on 64 kbps. Therefore, in ISDN the basic information channel is a B channel with 64 kbps rate. This rate is chosen to permit PCM encoded digitised voice to be transmitted. ISDN is also meant to support multiple services simultaneously. One single B channel cannot support multiple services. At the same time, the rate chosen for user interface must be carried by the existing telephony wire loops. Taking these factors into account, the basic rate access is defined as

Basic rate access : 2B + D

i.e. two B channels of 64 kbps each and a signalling channel of 16 kbps totalling to a rate of 144 kbps. Taking data encoding, header and other overhead bits into account, the actual rate on the line works out to be 192 kbps. This data rate can be supported over most existing two-wire loops. The basic access interface permits simultaneous use of voice, high speed data, facsimile and several text and telemetry applications. While one B channel may be used for voice at 64 kbps, the other B channel may support a wide variety of applications. It is also possible to transmit speech at 32 kbps using ADPCM, and the remaining 32 kbps in the B channel could be used for other applications. It is important to recognise that while a user sees logically two B channel sockets and a D channel socket, the entire information is physically carried over a single pair of wires to the ISDN exchange.

Primary rate access interfaces cater to the high speed requirements of the users. The higher bit rate may stem from a LAN or a PABX with a large number of terminals installed at the user premises or from high speed applications such as video and high speed facsimile. Keeping in mind the existing telephony transmission structure, two information rates have been chosen for primary rate access: 1544 kbps and 2048 kbps, the former being predominantly used in the USA, Canada and Japan, and the latter in Europe, India and other countries. The rate 1544 kbps corresponds to the T1

transmission standard of AT&T laboratories. A primary rate access interface may be defined around B, H0, H11, and H12 channels:

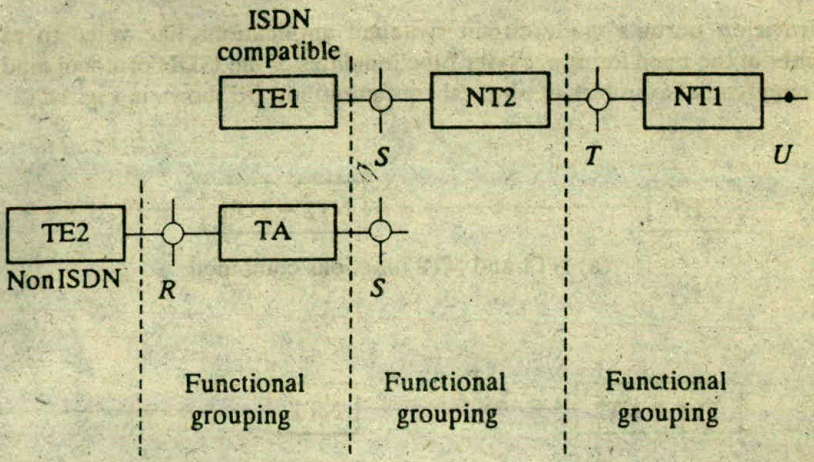
1544 kbps	*2048 kbps*
23B + D	30B + D
3H0 + D	5H0 + D
H11	H12 + D

All signalling channels (D channels) in the above are 64 kbps channels. A 1544 kbps around H11 channel implies that signalling is being carried in a D channel associated with another interface structure at 1544 kbps such as 3H0 + D or 23B + D. When there are multiple primary rate interfaces at a site, one D channel may suffice to carry the signalling information for all the interfaces. In this case, additional interfaces may contain all information channels. For example, a user may have a 23B + D and a 24B interfaces; or a 3H0 + D and a H11; or a 30B + D and a 31B. As mentioned in Section 11.4, H12 channel is not recognised for international compatibility.

In some cases, it is possible that a user may not want the full capacity of the standard interfaces. Hence, a general access capability of nB + D is permitted in both basic rate and primary rate interfaces. Value of n is 1 or 2 in the case of basic rate interface and varies from 1 to 23 for 1544 kbps and 1 to 31 for 2048 kbps primary rate interfaces.

11.5.1 Functional Grouping and Reference Points

In Section 11.3, we discussed the functions of ISDN layers 1–3. Different nations have different regulatory requirements that limit the functionalities that can be offered at the customer site by a network provider. Such regulations are used to restrain dominant carriers from monopolising the field and to provide business opportunities for small entrepreneurs. Taking into account these regulatory factors, ISDN user network interfaces are functionally grouped and the associated access points at different functional levels are known as **reference points**. Figure 11.11 shows the functional groupings and the points for the ISDN user network interfaces. There are three CCITT official reference points R, S and T. There is also a fourth reference point U which has come about on account of the current regulations in the USA, which prohibit some major carriers from providing even the minimal functionalities at the user site. The U reference point is not recognised by CCITT at present. The U interface is essentially the loop transmission system carrying digital signals at 192 kbps rates, which is the transmission rate for the basic access interface, 2B + D. In effect, at the U reference point, no functionalities are provided at the user site by the network operator. An equipment connected at the U interface will have to take care of the functions of all the ISDN layers 1–7. Network termination 1 (NT1) functional grouping includes the layer 1 functions of the ISDN protocol structure. Consequently, an equipment connected to the



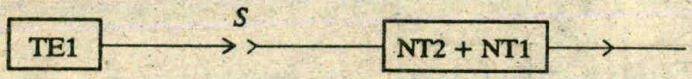
R, S, T = CCITT reference points U = reference point in the USA
NT1 = network termination 1 NT2 = network termination 2
TE1 = terminal equipment type 1 TE2 = terminal equipment type 2
TA = terminal adapter

Fig. 11.11 Functional grouping and ISDN reference points.

T reference point will have to concern itself only with the functions of layers 2–6. Network termination 2 (NT2) functional grouping includes the functions of ISDN layers 2 and 3 in addition to NT1 functions. Consequently, an equipment connected at the S reference point is concerned only with the function of higher level layers 4–7. Reference points S, T and U are ISDN compatible points and only ISDN compatible equipments can be connected to these points. ISDN compatible equipments are known as terminal equipment type 1 (TE1). Since ISDN is taking an evolutionary path, relying largely on the existing infrastructure, there is a need to have a functional grouping that would support the existing nonISDN equipments. This is achieved by using a terminal adapter (TA) between the S reference point and the nonISDN terminal. The access point for nonISDN equipment is the reference point R. The nonISDN terminals belong to the terminal equipment type 2 (TE2).

After having defined the reference points and the functional groupings, ISDN permits considerable flexibility in the physical realisation and implementation of functions. It is not essential that all reference points are accessible at a user site. For example, the functions NT1 and NT2 can be combined in a single equipment providing only the S access point to the user. Alternatively, NT2 functions can be implemented as part of the TE and connected to the reference point T. An important permisible case is the omission of NT2 functions when the reference points S and T coincide. Hence, TE1 directly connects to type S and T coincident points. This

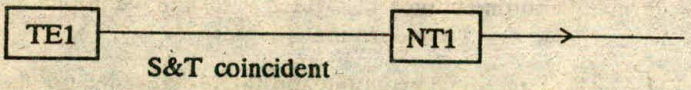
provision permits many circuit switched applications like voice to exist without the need for upper layer functionalities of the ISDN protocol model. Some typical examples of physical configurations are shown in Fig. 11.12.



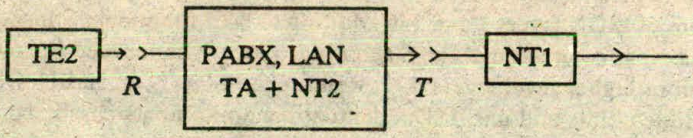
(a) NT1 and NT2 functions combined



(b) NT2 functions implemented as part of TE1



(c) NT2 functions omitted



(d) NonISDN terminal on LAN & PABX

Fig. 11.12 Some physical implementations of ISDN functional groupings.

11.6 Signalling

ISDN uses a common channel signalling scheme. The signalling is done over the D channel which acts as the common signalling channel for the B and H channels which carry the user information. D channel may also be used for carrying some user information, if there is spare capacity. In such cases also, the required signalling is done on the D channel. The concepts of common channel signalling and the CCITT's Signalling System 7 (SS7) have been discussed in Section 9.9. ISDN adopts SS7 for its interoffice and other network related signalling. It is the ISDN User Part (ISUP) of SS7 architecture (see Fig. 9.36) that deals with signalling facilities in ISDN. Unlike in other networks, signalling in ISDN falls into two distinct categories:

- User level signalling
- Network level signalling.

All user generated signalling and the signalling features that are open to the user are treated as user level signalling and are defined as part of the layer 3 user-network interface standards. The signalling facilities employed by the network to support user level signalling and to implement network control functions, not directly related to the user, are treated as network level signalling and are defined as part of the ISUP of SS7. In this section, we describe the user level and network level signalling features.

11.6.1 User Level Signalling

User level signalling in ISDN permits a user to

1. establish, control and terminate circuit switched connections in B channel,
2. carry out user-to-user signalling, and
3. establish, control and terminate packet switched connections in B or D channels.

User-to-user signalling is achieved by employing a symmetrical protocol for outgoing and incoming calls. User level signalling is of two types:

- Message based signalling
- Stimulus signalling.

Message based signalling is employed when the user end equipment is an intelligent terminal. In ISDN parlance, an intelligent terminal is known as **functional terminal**. It provides a user friendly interface for signalling and performs the functions of forming, sending, receiving and replying messages. The process of establishing, controlling and terminating a call is achieved by exchanging messages between the network and the terminal. The messages may be placed under four groups:

1. Call establishment messages
2. Call control messages
3. Call disconnect messages
4. Miscellaneous messages.

Call establishment group includes **set-up, call proceeding, alert, connect** and **connect acknowledge** messages. Alert signal corresponds to ring back signal and is used when a nonautomatic answering terminal is used at the receiving end. If the auto-answering facility is available, the terminal responds with connect signal directly and the alert signal is skipped. Call control group includes **suspend** and **resume** messages and also user-to-user messages. Call disconnect group includes **disconnect, release** and **release**

complete messages. The primary function of the miscellaneous messages is to negotiate network facilities to support additional service features like call forwarding, direct inward dialling, reverse charging etc.

All user level messages have a common message format. Three fields are mandatory for all messages:

1. Protocol discriminator
2. Call reference
3. Message type.

As the D channel may carry computer and telemetry data etc. in addition to signalling messages, it is necessary to have a mechanism for differentiating packets and their associated protocols. The protocol discriminator field is provided for this purpose. At present, only two message protocols are supported: the ISDN signalling message protocol and the X.25 level 3 packet protocol. The message format is shown in Fig. 11.13. The field has three

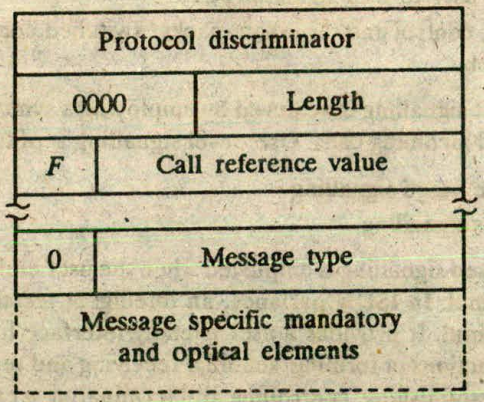


Fig. 11.13 User level signalling message structure.

subfields: length sub-field, flag and the reference value. The call reference field gives reference to the B, H or D channel information transfer activity to which a signalling packet pertains. Depending on the service and the channel used, the length of the call reference value may vary. At present, two lengths for identifying information transfer in the basic rate channel structure and the primary rate channel structure have been fixed as one and two octets respectively. Lengths of the reference value field for other services and channel structures have not yet been defined. The 1-bit flag is used to indicate which end of the connection initiated the call. Message type field identifies the messages in different categories as discussed above. Following the three common fields are the message specific information fields. Depending upon the type of the message, some of the information fields are mandatory and others optional.

Stimulus signalling is used when the user end equipments are dumb devices with no intelligence, e.g. digital telephone. As the devices do not have functional capabilities, stimulus signalling messages are generated as a direct result of actions by the terminal user. These signals just indicate events like handset off-hook or depression of a specific push button, which are all due to manual action by the user. Similarly, the signals sent by the network to an unintelligent device are in the nature of inducing specific events at the terminal end. Examples include connect or disconnect B channel, start alerting, etc. However, stimulus signalling procedures are defined as a compatible subset of the message based signalling procedures in order to facilitate functional expansion for simple terminals.

11.6.2 Network Level Signalling

Network level signalling in ISDN is concerned with interoffice signalling, signalling features accessible by the user to obtain enhanced services from the network and other network related signalling. Circuit suspension and call supervision messages are examples of network related signalling. The procedures for network level signalling are defined as the ISDN user part (ISUP) of the signalling system 7. One of the main aims in the context of ISUP has been to evolve a flexible design for the signalling system to accommodate new services and connection types that may come about in the future to be supported on ISDN. Flexibility is achieved by making the signalling entirely message based. For the transport of the signalling messages, ISUP relies on the services of the message transfer part or the network services part of SS7 (see Fig. 9.36).

About 40 network level messages have been standardised so far and these messages may be placed under nine broad categories:

1. Forward address
2. General setup
3. Backward setup
4. Call supervision
5. Circuit supervision
6. Circuit group supervision
7. In-call modification
8. End-to-end
9. User-to-user.

Messages belonging to 1–4 above are used to support the call setup process initiated by the user and start the accounting and charging functions. Circuit and circuit group supervision messages permit blocking and deblocking of circuits and circuit groups respectively. Other functions include connection release, temporary suspension and subsequent resumption of circuits. In-call modification messages are used to alter the characteristics or associated network facilities of an active call, e.g.

changeover from voice to data. End-to-end or node-to-node signalling is between the originating and the terminating ISDN exchanges and is accomplished in two ways. The **pass-along** service of the ISUP enables end-to-end signalling. As the name implies, the message is passed along the route of the transit exchanges without any processing of the information. The only action performed by the transit exchanges is to readdress the received pass-along message and forward the same to the next exchange in the chain. Another way of doing end-to-end signalling is to use the services of the signalling connection control part (SCCP) of the SS7 (see Fig. 9.36). This method is universally applicable and, unlike pass-along method, is independent of the presence or absence of a circuit connection between the message originating and terminating exchanges. In this case, the route taken by a message is determined by SCCP and may not relate to any user circuit.

A common format is used for all the messages defined in ISUP. The format is shown in Fig. 11.14. The message consists of six fields:

1. Routing label
2. Circuit identification code
3. Message type
4. Mandatory fixed part
5. Mandatory variable part
6. Optional part.

Routing label indicates the source and destination exchanges of a message and includes a link selection subfield. The field message type has been discussed earlier in this section. Other fields are self-explanatory.

11.7 Numbering and Addressing

In telephone and data networks, the end equipments are more often single units than multiple devices units like PABX or LAN. Historically, a telephone, a computer, or a terminal has been the predominant end equipment. The numbering systems for these networks have also evolved to identify single equipment end points. In ISDN, multiple devices at the end points are more of a norm than single units, in view of the multiple services environment. It then becomes necessary to identify a specific end equipment, e.g. facsimile or computer, to render the service. Identifying the specific equipment is a two-level process; first the end point is identified as in the case of telephone or data networks and then the equipment at the end point. ISDN addressing structure provides for this requirement. The component of the ISDN address which is used to identify the end point is known as the **ISDN number**, and the component for identifying the specific equipment at the end point is called the **ISDN subaddress**.

The numbering plan for ISDN is evolved using the following guidelines:

1. It is based on, and is an enhancement of, the telephone numbering plan. In particular, the country codes evolved for the telephone

numbering as defined in CCITT standard E.163 are adopted in toto for ISDN.

2. It is independent of the nature of the service (e.g. voice, facsimile or data) or the performance characteristics of the connection (e.g. 32 kbps voice or 64 kbps voice).
3. It is independent of routing, i.e. the numbering or addressing does not specify the intermediate exchanges through which the service is to be

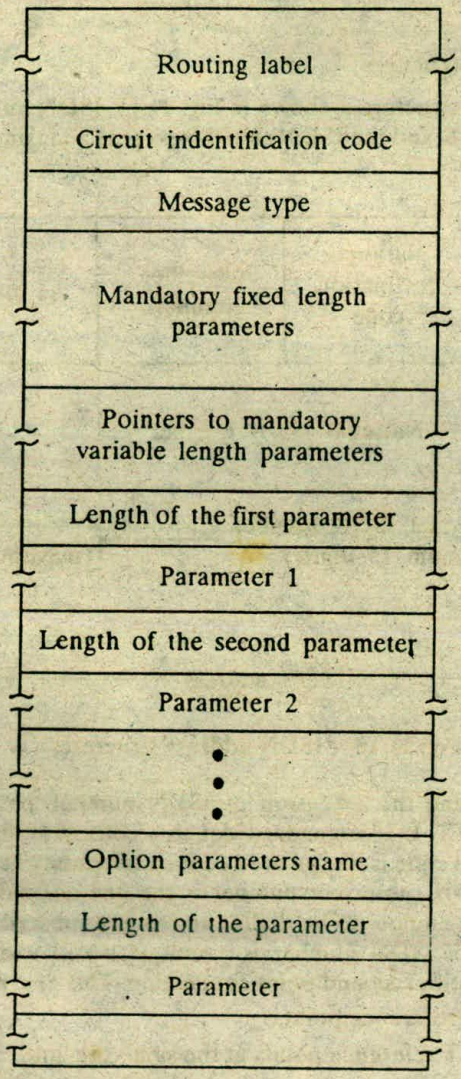


Fig. 11.14 Message format for ISDN user part.

put through. In contrast, some addressing schemes in data networks demand that the complete route from source to destination be specified as part of the address. UNIX based networks are typical examples of this.

4. It is a sequence of decimal digits. No alphabet or other characters are permitted as part of the address.
5. Its design is such that interworking between ISDNs requires only the use of ISDN number and no other additional digits or addressing signals.

11.7.1 Address Structure

The ISDN address structure is shown in Fig. 11.15. ISDN number part has a maximum of 15 digits and the ISDN subaddress part a maximum of 40 digits.

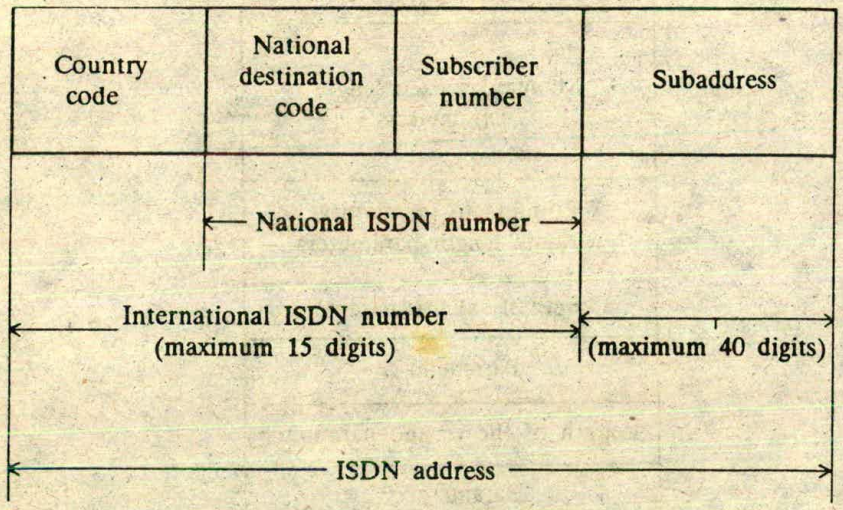


Fig. 11.15 ISDN address structure.

CCITT is considering the extension of ISDN number part to 16 or 17. According to CCITT, E.163 standard country code may have 1–3 digits. National destination code is like an area code in telephony network and is of variable length. ISDN subscriber number is the one normally listed in the directories. It is the number to be dialled to reach a subscriber in the same numbering area. An ISDN number is a unique world wide address and unambiguously identifies an end point connection. This end point may be

1. single S or T reference point,
2. one of many T reference points at the same site, and
3. one of many S reference points using direct inward dialling feature.

A single S or T reference point may also be addressed by multiple ISDN numbers. This feature is generally used in internetworking.

A subaddress, although a part of the ISDN address, is not considered as an integral part of the numbering scheme.

The subaddress is carried in a separate field in the user-network nterface message. It may or may not be present in a call setup message. The ubaddress is generally transparent to the network and it is the equipment at he destination which analyses the subaddress information for routing to the ppropriate terminal. A typical address using both ISDN number and the ubaddress is shown in Fig. 11.16. Here, an ISDN number identifies a T

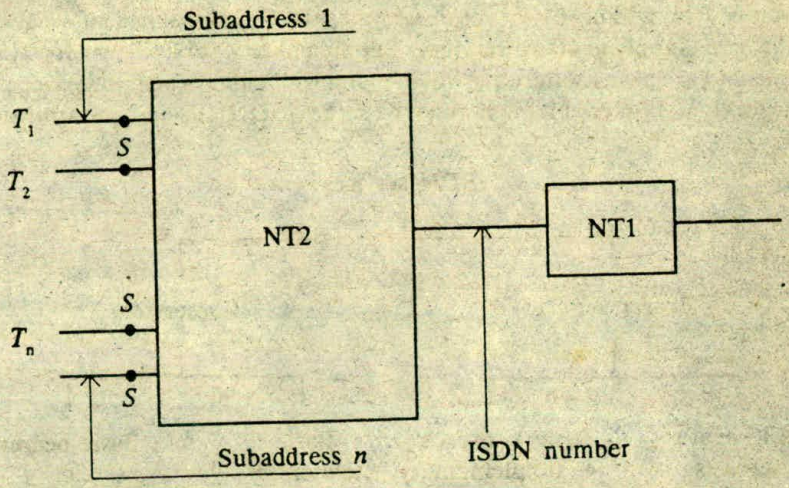


Fig. 11.16 Example of ISDN addressing.

reference point and a subaddress one of the many S reference points. Alternatively, a S reference point may also be addressed by using direct inward dialling (DID) feature.

11.8 Service Characterisation

In Section 11.2, we described some of the new services that may be supported on ISDNs. ISDN services are placed under two broad categories:

- Bearer Services
- Teleservices.

In Section 11.3 it was stated that the ISDN protocol architecture is layered according to the OSI reference model. Bearer services correspond to the functionalities offered by layers 1–3 (see Fig. 11.10). Provision of these three lower layer functionalities is generally the responsibility of the network provider (subject to the regulations in force in a country as mentioned in

Section 11.5). It is for this reason that the functions are termed bearer service functions, the bearer being the network provider. Bearer services are accessible at T and S reference points for ISDN compatible equipments (see Fig. 11.11). For nonISDN terminals the bearer services are accessible at R reference point of the user-network interface. The upper layer functions, viz. layers 4–7, support teleservices which employ the bearer services to transport information as per the requirements of the specific teleservices offered. A packet data transmission is a typical example of a bearer service, whereas telephony, teletex, videotex and facsimile fall in the class of teleservices.

Both the bearer service and the teleservice functionalities may be enhanced by adding to the basic service, the functionalities of what are known as **supplementary services**. Supplementary services cannot stand alone and are always offered in conjunction with either a bearer service or a teleservice. Figure 11.17 summarises the ISDN service possibilities.

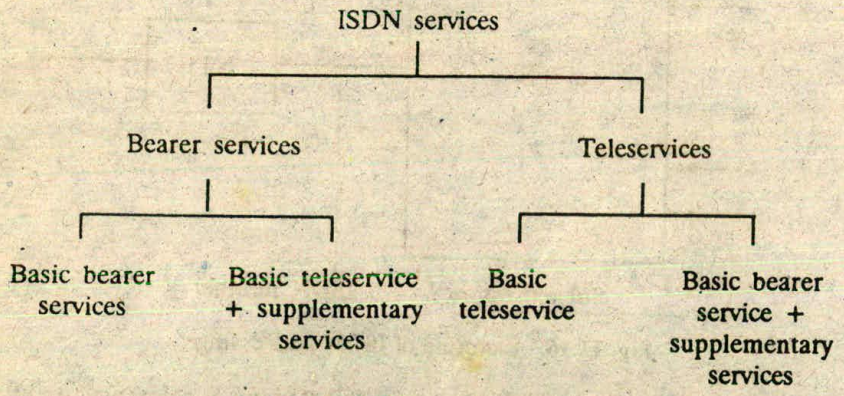


Fig. 11.17 ISDN service categories.

Supplementary services call for additional functionalities both in the lower layers and in the upper layers, depending on whether they supplement a basic bearer service or a basic teleservice.

Services are characterised by a set of attributes which may take on values from a domain of values permitted for each attribute. Bearer service or lower layer service attributes are grouped under three categories:

1. Information transfer attributes
2. Access attributes
3. General attributes.

Information transfer attributes describe the network capabilities for digital information transfer from S/T reference point of one customer to one or more S/T reference points of other customers. The attributes used for this purpose are:

1. Transfer mode
2. Transfer rate
3. Structure
4. Transfer capability
5. Establishment of communication
6. Establishment of connection
7. Communication configuration
8. Connection configuration
9. Symmetry.

Transfer mode may either be circuit switched mode or packet switched mode. Transfer rate may be one of the many standard rates discussed in Section 11.4. Structured transmission implies that the integrity of data is maintained by conveying timing information along with the data transmitted. For example, 8 kHz structured transmission ensures octet boundary of speech samples in 64 kbps transmission. Transfer capability specifies the ability of the network to carry structured or unstructured data, restricted or unrestricted data, speech, audio and so on. The term 'unrestricted' implies that the information is transferred as it is unaltered and as far as the user is concerned the channel is transparent. Consequently, a user may use an unrestricted service for any data transmission up to 64 kbps. Some typical examples of transfer capabilities are:

- 8 kHz structured, unrestricted
- 8 kHz structured, speech
- 8 kHz structured, 3.1 kHz audio.

The information transfer attributes 1–4 are referred to as **dominant attributes**, as they serve to identify specific bearer service categories.

The next four information transfer attributes, viz. attributes 5–8, deal with communication and connection aspects. It is important to recognise the difference between the terms **communication** and **connection** in this context. Communication deals with information transfer aspects and, in that sense, with end-to-end communication. Connection deals with network aspects, e.g. the type of circuit used for achieving a communication. The attribute **establishment of communication** can take on one of three values: demand, reserved, or permanent communication. In demand communication, a connection is set up and released on demand. In reserved mode, setup and release times are fixed in advance by the customer. In permanent communication, a connection is pre-established. The attribute **establishment of connection** may take one of three values: switched, semi-permanent, and permanent. Semi-permanent connection is also a switched connection but provided for an indefinite period. Permanent connection is a nonswitched connection, like a leased line, bypassing the ISDN exchange. The configuration for communication may be point-to-point, multipoint or broadcast. The attribute **connection configuration** has three subattributes:

1. Topology
2. Uniformity
3. Dynamics.

Topology may be simple with one connection element or tandem with two or more connection elements in series or parallel with a connection being formed with two or more elements put in parallel. **Uniformity** specifies the homogeneity of the elements involved in a connection. A connection configuration may be uniform or nonuniform. The **dynamics** of a connection configuration deals with the temporal aspects of establishing and releasing connections. Establishment or release of connection elements may be simultaneous or sequential. Elements may be added to or removed from a connection independently of others.

Symmetry deals with information flow characteristics. Information flow may be unidirectional, bidirectional symmetric or bidirectional asymmetric. The information transfer attributes 5–9 are referred to as **secondary attributes** as they identify services within the service categories defined by the dominant attributes.

Three access attributes have been defined by CCITT so far:

1. Access channel and bit rate
2. Signalling access protocol
3. Information access protocol.

The access channel is one of the standard channels discussed in Section 11.4, and the bit rate is the associated bit rate for the channel. Two signalling access protocols have been defined: one is the D channel protocol and the other defined on the B channel for packet mode equipments. Information access protocols have been defined only for voice transmission. A 64 kbps PCM coding and a 32 kbps coding in conformity with ADPCM have been standardised.

The third category of bearer service attributes, viz. **general attributes,** includes the following:

1. Supplementary services
2. Quality of service
3. Connection performance
4. Interworking
5. Commercial and operational attributes.

As mentioned earlier, supplementary services are not stand alone and are always associated with a bearer or a teleservice.

All supplementary services are formally defined using attributes so that they may be invoked in a uniform manner regardless of the bearer service or the teleservice that invokes them. Table 11.4 gives a partial list of the supplementary services. Most of them originated in the telephone network world. Some of them have already been discussed in Chapter 4.

Table 11.4 A List of Supplementary Services in ISDN

Abbreviated dialling	Closed user group
Do not disturb	Reverse charging
Call waiting display	Call forwarding
Call barring	Conference calls
Three-party service	Direct dialling
City-wide centrex	Credit card calling

A number of parameters (about a dozen) have been identified for characterising the quality of service:

Flexibility	–	Ability to choose different services or terminal equipments and to interwork services
Cost/Charging	–	Ability to give correct billing and appropriate billing/charging information as desired by the users
User friendliness	–	Provision of prompts and voice announcements
Ergonomicity	–	Aesthetic and ergonomic aspects of terminal equipments
Waiting time	–	Waiting time for a user to obtain a service or a terminal equipment
Enhanceability	–	Progressive introduction of new facilities and technology upgrades
Reliability/ Availability	–	Mean time between failures, mean time to repair, etc.
Promptness	–	Delay in establishment/release of connections
Fidelity	–	Quality of reproduction
Backup assistance	–	User training, maintenance, etc.
Transfer speed	–	End-to-end information transfer time or throughput
Universality	–	Ability to interwork other private or public networks.

Connection performance deals with attributes related to the performance of ISDN connections and is a subject for further study. Interworking is dealt with in Section 11.9. Commercial and operational aspects are not discussed in this text.

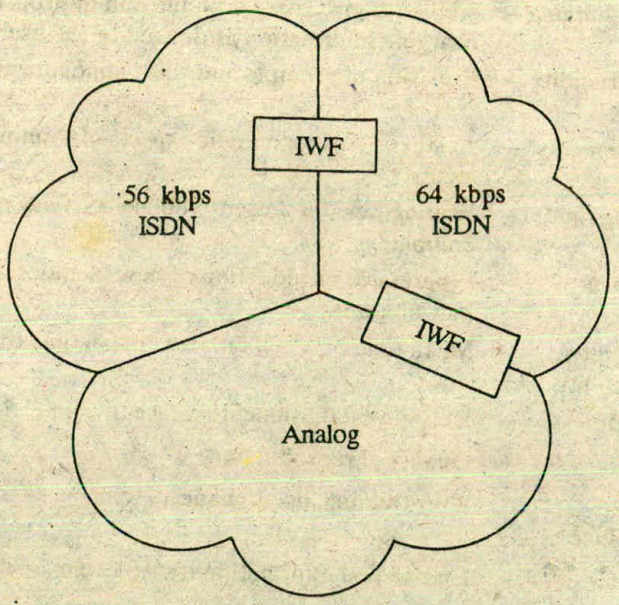
The attributes characterising teleservices are also placed under three broad categories:

1. Low layer attributes
2. High layer attributes
3. General attributes.

Low layer attributes are the same as the attributes for bearer services and in particular the information transfer and the access attributes. High layer attributes deal with different communication aspects such as voice, audio, text, facsimile and picture transmission. They are defined in the form of protocols conforming to the OSI model layers 4–7.

11.9 Interworking

ISDN is expected to evolve over a period of 20 years or so. During this period, it will have to coexist with other networks and this is essential for achieving market penetration and business success. Figure 11.18 depicts the situation.



IWF = Interworking functions

Fig. 11.18 Coexistence of ISDN with other networks.

At least three types of networks are involved: the analog networks, the 56 kbps digital networks for voice or data, and the ISDN with a 64 kbps clear channel. Interoperability is achieved through interworking functions (IWF). Interworking functions are different for different networks and so also the interface details. Taking this into account, separate reference points have been identified and their interface details have been defined to enable interworking with different networks:

Reference Point	*Interworking with*
K	Analog or IDN (voice or data) when the IWF is implemented in ISDN
L	Analog or IDN when the IWF is implemented as part of the existing network (analog or IDN)
M	Specialised service providers or value added networks (VANs)
N	Another ISDN
P	A specialised resource within ISDN.

Typical interworking functions to ensure interoperatibility between ISDN and other networks have the following objectives:

1. Determine if the network resources of the nonISDN network are adequate to meet the ISDN service demand.
2. In case the network resources are inadequate, take suitable action to cancel the call, reroute the call or renegotiate with ISDN to be consistent with the available resources.
3. Map signalling messages between the two (ISDN and nonISDN) signalling systems.
4. Ensure service and connection compatibility, such as compatibility of call process tones, call failure indication and signal processing modules like echo canceller.
5. Provide transmission structure conversion, including modulation technique and frame structure conversions.
6. Provide error and flow control.
7. Collect data for billing and charging.
8. Coordinate operation and maintenance procedures to be able to isolate faults.
9. Enable interworking of the numbering schemes.

The above interworking functions point to the need for parameter exchange among the existing nonISDN terminals, ISDN terminals, and the strategically located IWF modules. An illustrative example where parameter exchange is involved is when an ISDN user requests a 64 kbps bearer service to a destination user in a 56 kbps network. The service may be renegotiated with the ISDN user to operate at 56 kbps by exchanging parameters.

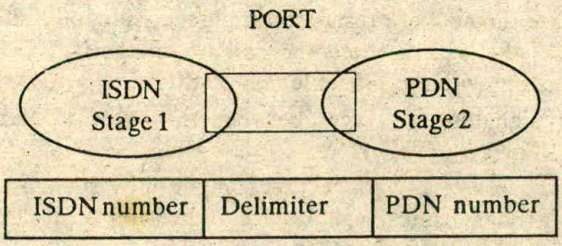
11.9.1 Numbering Interworking

The international PSTN numbering scheme used a 12 digit number with a two-part format of country code and national significant number as described in Section 9.5. ISDN uses a 15 digit number. Data networks have different schemes for access via different networks as described in

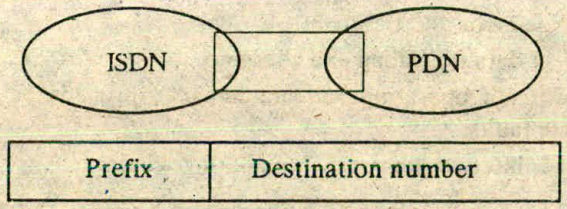
Section 10.3. ISO has defined its own numbering scheme in the context of OSI reference model as discussed in Section 10.3. The existence of many numbering schemes makes interworking of numbering plans a difficult proposition. However, two basic methods for interworking between the ISDN numbering plan and others have been proposed:

- Two-stage or port method
- Single-stage or integrated numbering method.

In the port method, the numbering plan of the network to which the called party is attached is treated as a port on the calling party's network and a two-stage dialling is used. First, the calling party dials a number in his own numbering scheme which takes him to the port to which the other network is attached. The user then dials a number in the destination network scheme. The two-stage interworking is illustrated in Fig. 11.19(a). The main



(a) Two-stage dialling



(b) Single-stage dialling

Fig. 11.19 Numbering interworking.

disadvantages of two-stage dialling are:

1. The user dials two sets of numbers.
2. Dialling is dependent on the relative numbering plans of the called and calling terminal.
3. A delimiter or a pause is necessary between the two stages, e.g. to obtain the dial tone of the second network.
4. The dialling becomes even more cumbersome if more than two networks are involved in establishing a connection.

An important advantage of the two-stage scheme is that it is simple to implement and is universally acceptable.

In single-stage or integrated numbering scheme, a prefix is used to identify a particular network and thereafter the numbering scheme of the destination network is dialled. This is illustrated in Fig. 11.19(b). As soon as the prefix is dialled, the originator local exchange routes the call to the appropriate interworking port. Thereafter, the dialled digits are processed by the destination network. ISDN numbering scheme is an example of integrated numbering. The information about the destination numbering plan is input by the user and is carried as part of the numbering identification field of the ISDN signalling message. For calls within the same numbering plan (e.g. ISDN), a null indication is used in this field.

11.10 ISDN Standards

Standardisation is an essential process in the introduction of any major and complex international service. The capability of providing true international connectivity and interoperability between networks is critically dependent on the availability of standards and the strict adherence to them. The importance of standards has been well recognised in the context of ISDN from the very early stages. CCITT has been playing a leading role and acting as a coordinating body by issuing ISDN related recommendations and thereby guiding the introduction of ISDN internationally.

The first definition of ISDN appeared in CCITT's recommendation, G.702 issued in 1972. Subsequent studies led to the emergence of the first ISDN standard, G.705 in 1980. The contents of the G.702 and G.705 have been presented in the beginning of this chapter. The conceptual principles on which ISDN should be based are laid down in G. 705. ISDN was declared as the major concern of CCITT during the study period 1981–84, and considerable efforts were put in to evolve standards for ISDN. These efforts culminated in a comprehensive set of recommendations in 1984 (although many were incomplete), known as I-Series recommendations. Figure 11.20 shows the general structure of I-Series recommendations.

The I.100 series serves as a general introduction to the concept of ISDN. Here an overall description and the expected evolution of ISDN are presented. This series gives the fundamental principles of ISDN and presents a vocabulary of terms specific to ISDN standards. General methods such as the attribute method to describe the services are described.

I.200 series specifies the classification and the description of ISDN services. Three CCITT services, telephony, telegraphy, and data are fully standardised. Four others, teletex, facsimile, videotex, and message handling are in the process of standardisation as of 1991. A CCITT service is said to be completely standardised only when

1. end-to-end compatibility is guaranteed;
2. terminals to provide the service are standardised;
3. procedures for obtaining the service are specified;
4. service subscribers are listed in an international directory;
5. testing and maintenance procedures are standardised; and
6. charging and accounting rules are spelt out.

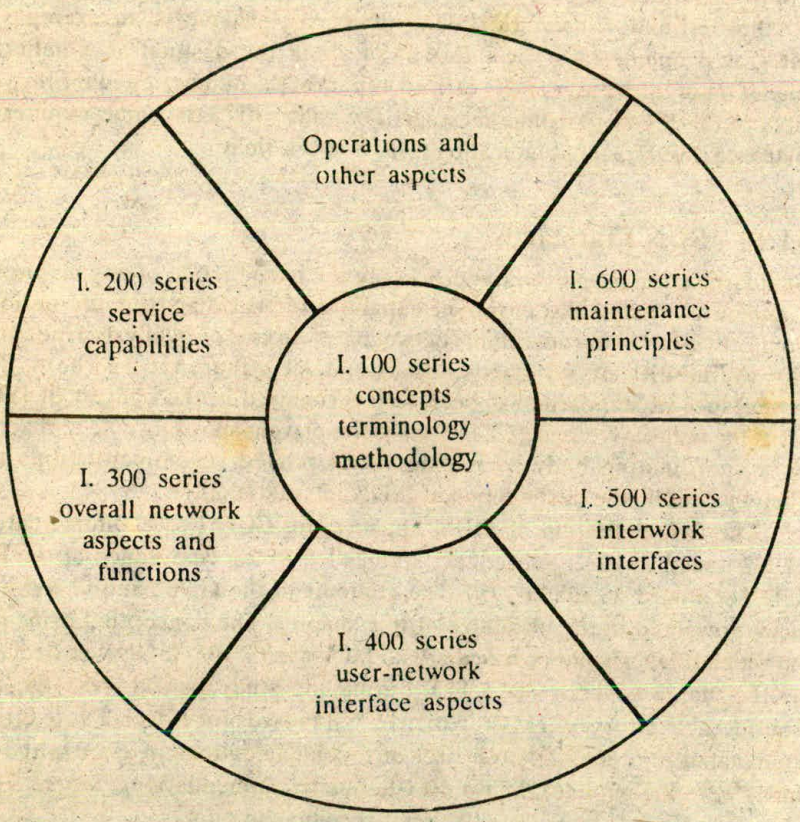


Fig. 11.20 Structure of ISDN recommendations.

The main aim of I.200 series is to provide an unambiguous description of service features but at the same time permit complete flexibility and freedom in the implementation of these services. In the past, definitions used in the context of different networks to mean broadly the same service have had subtle differences that often turned out to be important. I.200 series is an attempt to ensure that such problems do not arise in the future.

The I.300 series which deals with network capabilities discusses the protocol reference model, numbering and addressing principles, etc. The I.300 series maps the various services defined in I.200 series into elementary

network connections and functions. The protocol reference model is an extension of the open system interconnection model described in X.200 series of recommendations. There are three major extensions:

1. Separation of signalling and management operations from the flow of application information within a piece of equipment
2. Definition of communication contexts that operate independent of each other
3. Application of extensions 1 and 2 to internal network components.

Other extensions include multipoint and asymmetric connections, changeover/changeback of signalling links, and communications for maintenance and network management.

The I.400 series specifies the interface between the user equipments and the network. Apart from user-network interface definitions, the series also describes aspects such as electrical characteristics, permissible wiring configurations, connectors, powering arrangements and time-division multiplex structures. The basic rate interface may support up to eight user equipments bridged on the same bus. A contention resolution procedure is defined to resolve contentions, if multiple equipments have information to transmit simultaneously. Signalling procedures for establishing, modifying and terminating calls are also described in this series. The operation of terminal adaptor for nonISDN terminals is specified in this series.

The I.500 series describes the general principles of interworking and specifies ISDN-ISDN, ISDN-PSTN and ISDN-telex interworking principles.

The I.600 series enumerates the general maintenance principles and deals with the application of these principles to the maintenance of ISDN subscriber installation, basic rate access, primary rate access and multiplexed basic rate access interfaces. An important recommendation with regard to maintenance principles is the use of machine intelligence. It is expected that expert systems will play a significant role in ISDN maintenance in the years to come. Section 11.11 deals with the basics of expert systems. Table 11.5 gives a list of the 1988 I-Series recommendations of CCITT.

Table 11.5 CCITT ISDN Recommendations – 1988

I.100 Series – General Concepts, Terminology, and Methodology		
I.100	:	Preamble and general structure of the I-series recommendations
I.111	:	Relationship with other recommendations relevant to ISDN
I.112	:	Vocabulary of terms for ISDNs
I.113	:	Vocabulary of terms for broadband aspects of ISDN
I.120	:	Integrated services digital networks
I.121	:	Broadband aspects of ISDN

Table 11.5 CCITT ISDN Recommendations – 1988 (cont.)

I.122	:	Framework for providing additional packet mode bearer services
I.130	:	Method for the characterisation of telecommunication services supported by an ISDN and its network capabilities
I.140	:	Attribute technique for the characterisation of telecommunication services supported by an ISDN and its network capabilities
		I.200 Series – Service Capabilities
I.200	:	Guidance to the I.200 series of recommendations
I.210	:	Principles of telecommunication services supported by an ISDN and the means to describe them
I.220	:	Common dynamic description of basic telecommunication services
I.221	:	Common specific characteristics of services
I.230	:	Definition of bearer services categories
I.231	:	Circuit mode bearer services categories
I.232	:	Packet mode bearer services categories
I.240	:	Definition of teleservices
I.241	:	Teleservices supported by an ISDN
I.250	:	Definition of supplementary services
I.251	:	Number identification supplementary services
I.252	:	Call offering supplementary services
I.253	:	Call completion supplementary services
I.254	:	Multiparty supplementary services
I.255	:	Community of interest supplementary services
I.256	:	Charging supplementary services
I.257	:	Additional information supplementary services
		I.300 Series – Overall Network Aspects and Functions
I.310	:	ISDN network functional principles
I.320	:	ISDN protocol reference model
I.324	:	ISDN network architecture
I.325	:	Reference configurations for ISDN connection types
I.330	:	ISDN numbering principles
I.332	:	Numbering plan interworking
I.333	:	Terminal selection
I.334	:	Principles relating ISDN numbers/subaddress to the OSI reference model network layer addresses

Table 11.5 CCITT ISDN Recommendations – 1988 (cont.)

I.335	:	ISDN routing principles
I.340	:	ISDN connection types
		I.400 Series – User-network Interface Aspects
I.410	:	General aspects and principles relating to ISDN user-network interfaces
I.411	:	ISDN user-network interfaces – Reference configuration
I.412	:	ISDN user-network interfaces – Interface structures and recommendations on access capabilities
I.420	:	Basic user-network interface
I.421	:	Primary rate user-network interface
I.430	:	Basic user-network interface – Layer 1 specification
I.431	:	Primary rate user-network interface – Layer 1 specification
I.440	:	ISDN user-network interface – data link layer – general aspects
I.441	:	ISDN user-network interface data link layer specification
I.450	:	ISDN user-network interface – Layer 3 general aspects
I.451	:	ISDN user-network interface – Layer 3 specification
I.460	:	Multiplexing, rate adaptation and support of existing interfaces
I.461	:	Support of X.21 and X.21 *bis* based DTEs by an ISDN
I.462	:	Support of packet mode terminal equipment by an ISDN
I.463	:	Support of DTEs with V-series type interfaces by an ISDN
I.464	:	Multiplexing, rate adaptation and support of existing interfaces for restricted 64-kbps transfer capability
I.470	:	Relationship of terminal functions to ISDN
		I.500 Series – Internetwork Interfaces
I.500	:	General structure of ISDN interworking recommendations
I.510	:	General principles of ISDN interworking
I.511	:	ISDN-ISDN layer 1 internetwork interface
I.515	:	Parameter exchange for ISDN interworking
I.520	:	General arrangements for interworking between ISDNs
I.530	:	ISDN-PSTN interworking
I.560	:	Interworking with telex
		I.600 Series – Maintenance Principles
I.601	:	General maintenance principles of ISDN subscriber access and subscriber installation

Table 11.5 CCITT ISDN Recommendations – 1988 (cont.)

I.602	:	Application of maintenance principles to ISDN subscriber installation
I.603	:	Application of maintenance principles to ISDN basic rate accesses
I.604	:	Application of maintenance principles to ISDN primary rate accesses
I.605	:	Application of maintenance principles to static multiplexed ISDN basic rate accesses

11.11 Expert Systems in ISDN

ISDN is envisaged to be an intelligent network. CCITT recommendation G.705 (1980) states:

> The ISDN will contain intelligence for the purposes of providing service features, maintenance and network management functions...

While in the initial stages, intelligence in the ISDN may be limited to the use of processor based subsystems and terminals, in the future, concepts of artificial intelligence (AI) and expert systems will be applied to network functions. In particular, network maintenance and network management are the potential areas for the application of AI concepts and expert systems.

Artificial intelligence may be defined as the study of how to make computers do things at which human beings are better at the moment, or more generally, the study of how to produce artificial things that have intelligence. The ability to learn or understand from experience, the ability to acquire and retain knowledge, and the ability to respond quickly and successfully to new situations may be considered to be the symptoms of intelligence. Problem areas where AI techniques have been applied traditionally are:

1. Game playing
2. Theorem proving
3. General problem solving
4. Perception
5. Natural language understanding
6. Specialised domain problem solving.

Problem solving in specialised domains calls for expert knowledge in those domains and hence is done by experts. Computer systems or software packages that perform the tasks, which are normally carried out by experts in real life, are known as **expert systems**. The expert system is that branch of AI, which deals with expert reasoning rather than common sense reasoning. Expert reasoning demands knowledge of the field concerned. Therefore,

knowledge bases are an integral part of expert systems. The process of construction of an expert system is known as **knowledge engineering.**

A number of expert systems have been constructed over the last 15 years for a variety of applications. Table 11.6 presents a sample set of expert

Table 11.6 Some Expert Systems and their Application Areas

Expert system	Area of application
MYCIN	Antimicrobial therapy
CASNET	Glaucoma diagnosis
RI	Configuring computers
PROSPECTOR	Geological exploration
KBVLSI	Design aid for silicon chips
TAXMAN	Advice in corporate and tax laws
DART	Fault diagnosis of computer systems
WAVE	Telephone line fault diagnosis

systems and the applications for which they are developed. Construction of an intelligent system calls for a method of modelling intelligence and intelligent actions inside a computer. We also need techniques for representing knowledge and creating and maintaining knowledge bases. A modelling technique known as **physical symbol system** is almost universally adopted for representing intelligence and intelligent actions within a computer. The physical symbol system concept is only a hypothesis in the sense that it is felt that such a system has the necessary and sufficient means of representing intelligence and intelligent manipulations. This hypothesis has so far not been either proved or disproved. It has only been subjected to empirical validation. The bulk of the evidence says that the hypothesis is true.

A physical symbol system consists of a set of entities called symbols, which are physical patterns that can occur as components of another type of entity called an **expression** or **symbol structure**. At any instant of time, the system contains a collection of these symbol structures. Besides these structures, the system also has a collection of processes that operate on expressions to produce other expressions. Operations include creation, modification, reproduction and destruction of symbol structures. A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just these symbolic expressions themselves.

The significance of the physical symbol system is two fold. First, it is a significant theory of the nature of human intelligence. Secondly, it forms the basis of the belief that it is possible to build programs that can perform intelligent tasks as humans do. It is important to recognise that symbols can represent real world objects and the symbol structures the relationship among objects. Symbol manipulation should not be confused with string manipulation which is a nonintelligent computation.

Large knowledge base is critical for the effective functioning of an expert system. Hence, knowledge representation is an important consideration in the design of an expert system. Knowledge representation techniques are placed under two broad categories: **declarative** and **procedural**. In the former, knowledge is represented as a static collection of facts accompanied by a small set of general procedures for manipulating them. In the latter, bulk of the knowledge is represented as procedures, and there is only a small set of facts. Both the types have their distinct advantages and hence are well suited to different problem domains. It is important to recognise that both the types use static collection of facts as well as procedures. It is a question of where the bulk of the knowledge lies; whether in the procedures or in the set of facts. Major advantages of the declarative type are:

- Each fact is represented only once, irrespective of the number of different ways in which it is used.
- It is easy to add new facts to the system without having to change the representation of facts already stored or the procedures.

Widely used declarative techniques are:

1. Predicate logic
2. Frames and scripts
3. Semantic nets
4. Conceptual dependency.

Predicate logic is very useful for representing simple facts but is not suited for complex structures of the real world. An advantage of predicate logic representation is that simple but powerful inference mechanisms can be used for reasoning. Other declarative techniques are capable of representing complex knowledge, but each one is useful in a particular situation. Frames are useful in describing a collection of attributes that a given object, say a table, normally possesses. Scripts are used to describe common sequences of events, e.g. what happens when one goes to a theatre. Semantic nets are general enough to describe both the objects and events. Conceptual dependency is a specialised structure which facilitates drawing inferences from natural language sentences and the representation is independent of the language.

Main advantages of the procedural knowledge representation technique are that it is easy to represent

1. knowledge of how to do things,
2. unstructured knowledge like the one involved in rule-of-thumb, default and probabilistic reasoning, and
3. heuristics knowledge like short-cut methods of doing things efficiently.

Expert reasoning is often unstructured. Short-cut methods, rule-of-thumb decisions, empirical techniques etc are extensively used by experts

in solving problems in their specialised domains. Hence, the procedural type representation, i.e. more knowledge in the procedures and less in the static set of facts, is better suited for expert knowledge and is widely used in expert systems. In procedural representation, knowledge about the world is contained mainly in procedures that know how to do specific things. The most popular technique is the one where knowledge is represented in terms of modular procedures, known as **production rules**. A procedure is invoked by matching the system state with the preconditions needed for the procedure to be activated. A production rule is characterised by a format where the right hand side may be an action, a conclusion or a suggestion. An action is the end point of a reasoning process. For example, in an expert system based computer fault diagnosis, a statement like "replace the disk drive" is an action to be carried out at the end of a diagnostic process. A conclusion is an intermediate step in knowledge based interaction. An example here is a statement like "the fault is likely to be in the disk controller". A suggestion is an indication that if more information is provided, further investigation is possible. For example, "measure the loop resistance and enter the value" is a suggestion. The most popular form of production rule is: "if (condition), then (an action/a conclusion/a suggestion)". As an example:

> If the ringing tone does not reach the subscriber instrument, then the fault is in the ringing circuit of the exchange or the subscriber instrument.

Because of their modular representation and the easy expandability and modifiability, production rules are chosen for representing knowledge by most designers of expert systems.

11.11.1 Components of an Expert System

There are three essential components in an expert system as shown by the solid line boxes in Fig. 11.21. Knowledge base which stores the domain knowledge is at the heart of the system. When the domain knowledge is stored using production rules, the knowledge base is often called **rule base**. The knowledge base is developed with the help of an expert and the expert works with the knowledge base through the development engine. He adds new knowledge or deletes old knowledge or modifies the existing knowledge. A user interacts with the system through the inference engine which executes a set of strategies to solve a problem. The set includes:

1. Control level strategies
2. Reasoning strategies
3. Interpretation strategies
4. Action/Conclusion/Suggestion strategies
5. Explanation strategies.

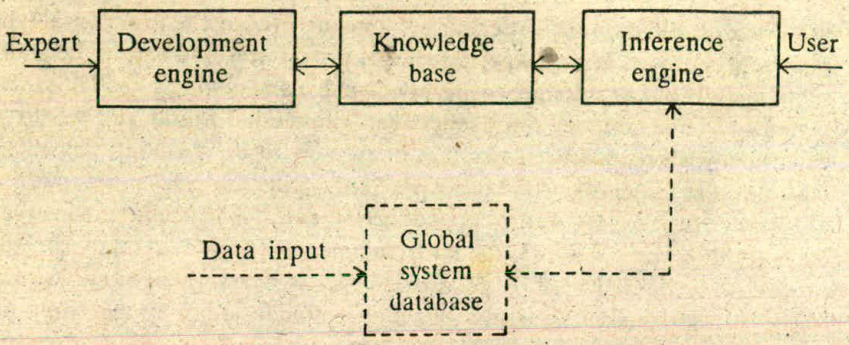


Fig. 11.21 Structure of an expert system.

In the context of the physical symbol system, a strategy in execution may be considered as a process. Control strategies are used to choose a problem-solving paradigm and to decide on a line of reasoning. Control strategies may be goal-driven or data-driven. Reasoning strategies may use rules of thumb, short cuts or heuristic inference as an expert would do. Interpretation strategies process the knowledge from the base. In addition to giving out the actions to be taken, conclusions, or suggestions, expert systems may also be required to explain the reasoning process as an expert would do. This is handled by building into the system a set of explanation strategies which keep track of the reasoning process and explain the same to the user.

Apart from the three main components, many expert systems also contain a global database, shown in dotted line box in Fig. 11.21, which keeps information about the state of the system. The application of rules changes the system status, enabling some rules and disabling some others. A global database contains information about such aspects.

Expert systems differ from conventional programs in many ways. Most importantly, in expert systems, there is a clear separation of knowledge about problems in a particular domain from the information (input data) pertaining to the current problem. Conventional programs deal only with current inputs. Table 11.7 summarises other differences between expert system software and conventional programs.

After having built an expert system, how do we know that it really behaves like an expert? This question is related to the more general question: "How do we know that an AI system is truly intelligent"? Alan Turing, the famous mathematician, has proposed a simple test to ascertain whether a given AI system is truly intelligent or not. The test has come to be known by his name as the **Turing test for intelligence**. In this test, which is experimental in nature, a human interrogator interacts with the system under test and another human being as shown in Fig. 11.22. The interrogator and the responders are in separate rooms. The interrogator is unaware as to which room the human or the machine responder is in. He interacts through a

Table 11.7 Expert Systems vs. Conventional Programs

Expert Systems	Conventional programs
Expert level processing	Machinistic level processing
Symbol manipulation	Arithmetic, logic and string manipulation
Deal with uncertainty	Always deterministic
Qualitative results	Quantitative results
Multiple solution paths	Only one solution path
Capable of explaining	Not capable of justifying
Dynamic environment	Static environment
Search intensive	Compute intensive
Self modifying	Not self modifying

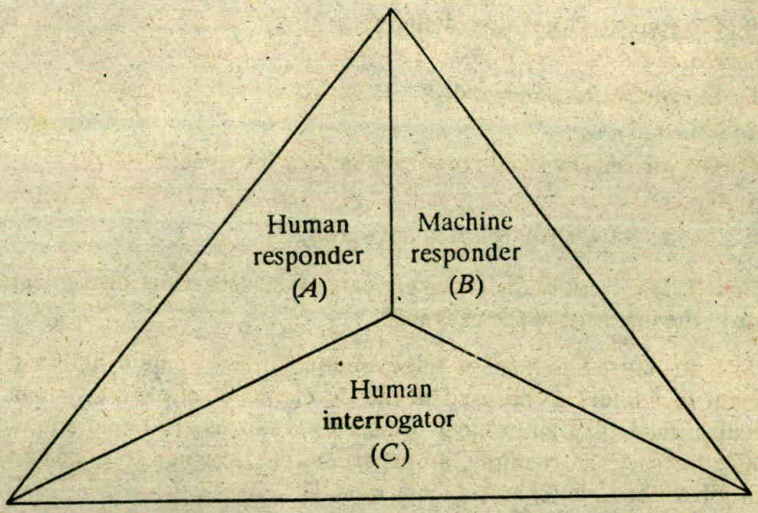


Fig. 11.22 Turing's test for intelligence.

keyboard with both the machine and the human responders for some time and draws one of three conclusions:

1. Unable to say which room contains the machine.
2. Human is in room *A* and the machine in room *B*.
3. Human is in room *B* and the machine in room *A*.

Conclusion 1 or 3 implies that the machine is intelligent. Conclusion 2 implies that the machine is not intelligent as the human interrogator is able to clearly differentiate the machine response from the human response. Although this test is rudimentary and subjective, it is indeed an effective way of judging an intelligent system.

11.11.2 Example of an Expert System

As an example of a rule based expert system, we present below a typical conversation that might take place between a telephone subscriber and an expert system at the time when the subscriber complains of a fault. Let *S* represent the subscriber and *E* the expert system. S dials and communicates:

S : Telephone No.414650 has a fault.
E : Are you complaining from your instrument or another instrument?
S : Some other instrument.
E : Do you get dial tone in your instrument?
S : Yes.
E : Are you able to make outgoing calls?
S : No.
E : Is your dial sticky?
S : No.
E : Do you get a tone after dialling?
S : No.
E : Do you get incoming calls?
S : No.
E : Do you see any visible breakage in your instrument?
S : No.
E : Thank you for your patient answering.

Conclusion Fault in the receiving part of the subscriber instrument or in the line card of the exchange.

The structured nature of telecommunications domain allows more powerful techniques to be used in the development of expert systems. In particular, modelling and object oriented paradigms are being applied in recent efforts. As an example, an expert system designed to allocate fault repair job to the right technician may model

1. the geographical area,
2. the equipments, and
3. the technician's skill and workload.

The inference engine may schedule a fault to a technician taking into account the following features:

1. The location of the fault and the technician
2. The priority of the fault
3. The skill and expertise of the technician.

Thus in the coming years, AI techniques and expert system will find increasing applications in telecommunication networks.

11.12 Broadband ISDN

Broadband ISDN (BISDN) is defined as a network capable of supporting data rates greater than the primary rate (1.544 or 2.048 Mbps) supported by ISDN. In the context of BISDN, the original ISDN concept is often termed narrowband ISDN (NISDN). The main aim of BISDN is to support video and image services. BISDN services are broadly classified as

- Interactive services
- Distribution services.

Interactive services may be classified as

1. Conversational services
2. Messaging services
3. Retrieval services.

Distribution services are classified as

- Broadcast services
- Cyclic services.

Conversational services support end-to-end information transfer on real time, bidirectional basis. There is a wide range of applications that may be supported using conversational services, the most important one being the video telephony or videophone. In this service, the telephone instrument has the capability to transmit, receive and display video signals. A dial-up connection brings about both video and audio transmission. Other applications include video conferencing and video surveillance. A number of data oriented conversational applications may also be supported. These include distributed databases, program downloading, interprocess communication and large volume, high speed data exchange as encountered in CAD/CAM, or graphics based applications.

Messaging services offer store and forward communication. Analogous to X.400 messaging services on NISDN, voice mail, video mail and document mail containing text, graphics etc. may become the important messaging services on BISDN.

Retrieval services in BISDN offer the capability to retrieve sound passages, high resolution images, graphics, short video scenes, animated pictures etc from centralised or distributed databases. In a sense, BISDN retrieval services are an enhancement of videotex services in NISDN. For example, stock market advisory information in NISDN may comprise only textual information, whereas in BISDN it may include graphic charts like histograms or pie charts.

Broadcast distribution services would provide support for broadcasting video, facsimile and graphical images to subscribers. Examples of such applications include television broadcasting over the network and electronic newspaper distribution. In broadcast services, the user has no control over what is being received on his/her screen.

Cyclic distribution services offer some control to the user in the presentation of information on the screen. An analogous NISDN service is the teletext service described in Section 11.2. Teletext normally uses only a portion of the analog TV channel, whereas the cyclic distribution services in BISDN may use a full digital broadband channel. The cyclic distribution services are an enhancement of the conventional teletext service. Unlike teletext where only textual information is transmitted, in cyclic distribution services, text, images, graphics and video and audio passages may be transmitted. A typical system may allow about 10,000 pages of frequently accessed information with a cycle time of one second. A user may randomly select and view pages.

The broad range of services envisaged to be supported on BISDN poses a challenge in designing a suitable architecture for the network. The difficulty stems from two widely varying characteristics of the services:

- Data rate required
- Average holding time.

At one end of the spectrum we have the conversational telephony service which needs a data rate of 64 kbps or less and lasts an average duration (holding time) of about 100 seconds. If we consider telemetering services, the data rates and holding times are even less. At the other end of the spectrum, we have the distributive high definition television (HDTV) service which demands a data rate of about 500 Mbps and with an average call duration of about two and a half hours. Compressed HDTV service requires 40 Mbps rate. Some of the new services like broadband information retrieval may demand as high as 1 Gbps data rate.

Another important parameter that influences network design is the duty ratio, i.e. the time for which the channel is actually busy transferring information divided by the holding time of the service. This parameter helps decide on the type of switching technique (circuit switching, packet switching or hybrid switching) to be used for a service.

Thus, the important requirements of BISDN may be listed as the ability to support

1. narrowband and broadband signals,
2. interactive and distributive services,
3. point-to-point, point-to-multipoint and broadcast connections,
4. different traffic patterns (e.g. for voice, data and video),
5. value-added services,
6. multirate switched and nonswitched connections, and
7. channel bandwidths up to 140 Mbps per service.

Some of the broadband channels proposed for BISDN are:

- H2 channel, 30–45 Mbps
- H3 channel, 60–70 Mbps
- H4 channel, 120–140 Mbps.

These channels are to be organised to provide a transmission structure at the user interface. As such, no standardisation has yet emerged in this regard.

11.13 Voice Data Integration

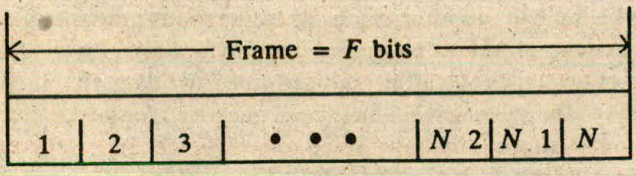
In Section 11.3, we pointed out that the present ISDN infrastructure is a segregated one and an integrated infrastructure is likely to evolve in the future. In the segregated infrastructure, we have a collection of different types of networks; packet switched, circuit switched, nonswitched and so on. Depending on the nature of the traffic of a service, the service is carried on the appropriate network. For example, voice traffic is carried on the circuit switched network, whereas data traffic on a packet switched network. In an integrated infrastructure, a single network carries traffic of all types. How do we integrate different types of traffic on a single network? Emergence of effective and efficient solutions to this problem would determine the adoption of a single network integrated infrastructure for future ISDNs. Many research investigations have been undertaken to study the issues related to this problem. The investigations are in preliminary stages. A lot remains to be done in this area which has tremendous potential for research in the yea s to come. Almost all studies done so far have only considered the integration of two types of traffic: digitised voice and data. In this section, we discuss some of the schemes that have been proposed for integration of voice and data on a single channel. Selection of voice and data traffic for integration studies is not accidental. The traffic characteristics and requirements of voice and data are so different that finding an efficient solution here may well turn out to be the key for solving the integration problem. Certain differences in the characteristics of voice and data traffic are given in Table 10.2. Other important parameters of voice and data traffic are summarised in Table 11.8.

Table 11.8 Parameters Relating to Digitised Voice and Data Traffic

Digitised voice	Data
Periodic bursty in nature	Aperiodic bursty in nature
Fixed length bursts	Variable length bursts
Small packet size	Large packet size
Packetisation time critical	Packetisation time is not critical
Hard bound on delay	Soft bound on delay
Hard bound on the variance of delay	Soft bound on the variance of delay
Loss of parts of speech acceptable	Loss of parts of data unacceptable
Low overhead as there is no error recovery	High overhead due to error detection and recovery

Data traffic is aperiodic whereas voice traffic is periodic. The traffic for digitised voice may be considered to be bursty, though periodic, as the transmission may be organised on sample by sample basis, i.e. one byte every 125 μs or on packet basis every few milliseconds by collecting many voice samples together. If packetised voice transmission is used, then the packet size is to be small to maintain the real time characteristics of the service. Large packet sizes would lead to broken speech transmission. For similar reasons of real time, the packetisation time is critical for voice packets. Voice packets must reach the destination within a specified time or else the speech would appear discontinuous. Hence, unlike in the case of data, the bound on the delay for voice traffic is to be strictly enforced. The variance of delay is also to be controlled. Otherwise, there will be fluctuations in speech signals leading to unnatural reproduction at the receiving end.

A basic model for integrating voice and data is to consider a TDM frame that is capable of carrying both circuit switched and packet switched data. A portion of the frame F_c bits is occupied by the circuit switched data and the remaining portion of F_p bits by packet switched data as shown in Fig. 11.23.



Circuit switched traffic = F_c bits
Packet switched traffic = F_p bits

$$F_c + F_p = F$$

Fig. 11.23 An integrated TDM frame.

A number of circuit switched sources may constitute the F_c bits of the frame. Each source may require different bandwidths and have different arrival and holding time characteristics. Similarly, F_p bits may contain data of a number of packet switched sources.

A TDM frame may be considered to contain N slots of b bits each, as a group of b bits is likely to be integral in any traffic type, e.g. a voice sample of 8 bits or a byte in data. We then have $F = Nb$, and one slot may be considered as the basic bandwidth unit (BBU). A source is assigned one or more BBUs, depending upon the bandwidth requirements. A circuit switched source is blocked if adequate number of BBUs are not available for assignment and a packet switched source is queued awaiting free BBUs. For voice data integration, we restrict the circuit switched traffic to voice, i.e. all sources have equal bandwidth requirement of one BBU and the same arrival time and holding time characteristics. Similarly, all the packet switched sources may be considered to have the same arrival and service time characteristics

for simplicity. Now, voice data integration schemes may be placed under two broad classes:

- Random allocation of BBUs
- Contiguous allocation of BBUs.

In random allocation, any available BBU is allotted to either voice or data using some criterion. The usual criteria include **first-come-first-serve (FCFS)** and **pre-emptive scheduling**. In the latter, a data source which has already been allotted BBUs may be scheduled out and placed in waiting in preference to a voice source that has just come up. In contiguous allocation schemes (also known as boundary schemes) attempts are made to gather together all BBUs allotted to voice and so also with BBUs allotted to data traffic. The contiguous allocation schemes can be classified as:

- Fixed boundary scheme
- Movable boundary scheme.

In fixed boundary scheme, the N slots in the frame are divided into two parts containing $N1$ and $N2 = N - N1$, slots respectively as shown in Fig. 11.24. One of the parts, say Part A, is permanently reserved for voice

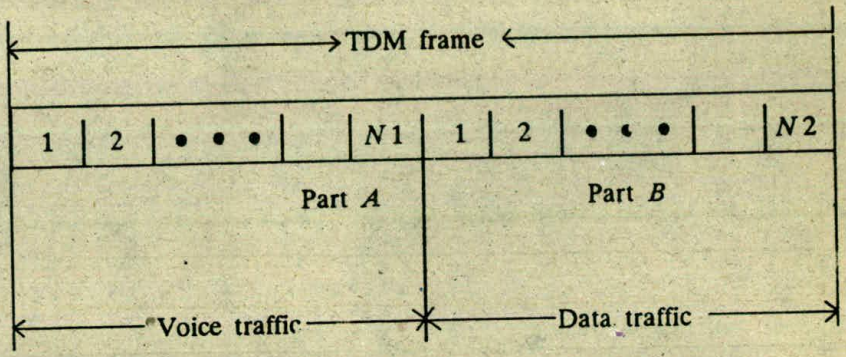
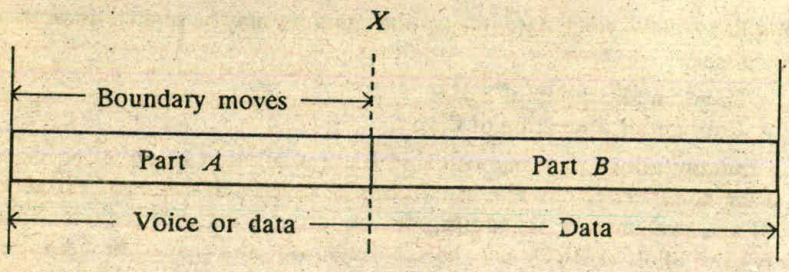


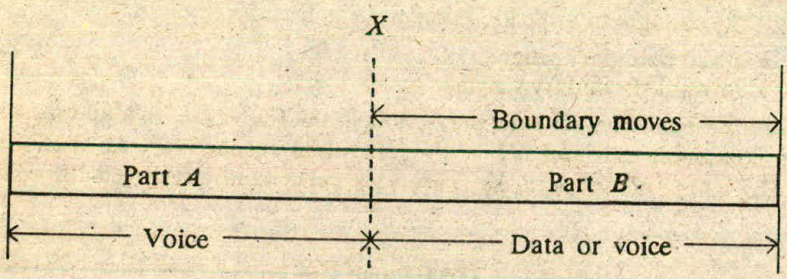
Fig. 11.24 Fixed boundary integration.

traffic and part B for data traffic. Voice traffic is always scheduled in part A, and if there are no free BBUs in that part, the call is blocked, even though free BBUs may be available in part B. Similarly, data traffic will be queued if there are no free BBUs in part B, even though free BBUs may be available in part A. The scheme performs like the segregated architecture although the traffic is physically carried on the same network. The performance is measured in terms of blocking probability for voice sources and mean delay for data sources. The fixed allocation of bandwidth to voice and data traffic logically reduces the network to two independent networks. This scheme of integration does not really exploit the advantage of a single common resource.

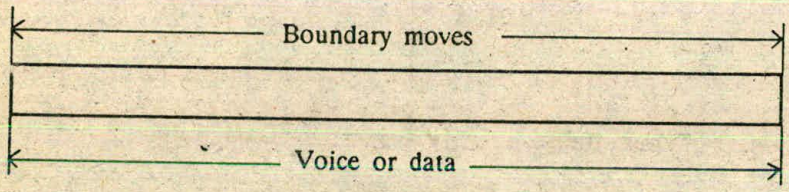
In movable boundary scheme, four strategies are possible as illustrated in Fig. 11.25. In strategies 1 and 2, there is a notional boundary shown at X.



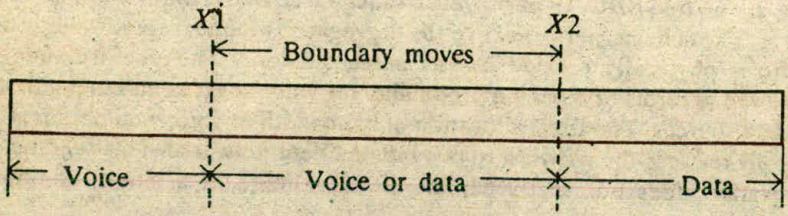
(a) Strategy 1



(b) Strategy 2



(c) Strategy 3



(d) Strategy 4

Fig. 11.25 Movable boundary scheme.

In the former, data traffic is allowed to use free BBUs, if any, in the voice traffic part. If a voice source arrives, then a data source occupying the voice area is pre-empted and the BBUs are allotted to the voice source. In this sense, the boundary moves back and forth to the left of X. If all the BBUs in part A are occupied by voice sources, further voice arrivals are blocked even if free BBUs are available in part B. This strategy gives the same blocking probability performance as the fixed boundary scheme for voice sources; but it provides a better delay performance for data sources.

In strategy 2, the boundary is allowed to move to the right of X into the data part area, thereby providing additional bandwidth to voice sources when available. Being a real-time source, once scheduled, a voice source cannot be pre-empted. Consequently, if the voice sources hold up BBUs in the data traffic part for a long time, the delay performance for the data traffic may be affected adversely. As one would expect, we get lower blocking probability for voice sources here than in the fixed boundary scheme.

In strategy 3, the boundary shifts dynamically throughout the frame, depending on the traffic load. Pre-empting may be permitted in this strategy for voice traffic. In all the three strategies discussed, there is a possibility that the entire TDM frame is occupied by only one type of traffic. In strategies 2 and 3, voice may completely dominate the channel, and hence some researchers have considered a strategy as shown in Fig. 11.25(d), where a minimum bandwidth allocation is always assured for both voice and data traffic or in some other cases for data traffic alone. Such a strategy is known as min-max scheduling as there is a minimum and a maximum bandwidth bound for either type of traffic. In all the discussions above, the boundary is assumed to be physical on the frame. This need not be the case. The allocation can be anywhere in the frame with only a bound on the bandwidth allocated. Contiguous allocation results in easy frame management. The random allocation with bounds on the bandwidth is left for further study.

FURTHER READING

1. Aleksander, Igor, *Designing Intelligent Systems: An Introduction,* Kogan Page Ltd., London, 1984.
2. *IEEE Journal on Selected Areas in Communications,* Vol. SAC– 4, No.3, May 1986. (This issue contains a series of articles covering different aspects of ISDN.)
3. Inose, H., *Telecommunication Technologies – Japan Annual Reviews in Electronics, Computers and Telecommunications,* Vol.23, North-Holland, Amsterdam, 1988.
4. Rich, Elaine, *Artificial Intelligence,* International Shindeli Edition, McGraw-Hill, New York, 1983.

5. Stallings, William, *ISDN – An Introduction*, Macmillan, New York, 1989.

6. Schwartz, Mischa, *Telecommunication Networks Protocols, Modeling and Analysis*, Addison-Wesley, Reading (Mass.), 1988.

7. Verma, Promode, K. (ed.), *ISDN Systems – Architecture, Technology and Applications*, Prentice Hall, Englewood Cliffs, New Jersey, 1990.

EXERCISES

1. Why is ISDN considered to be a well organised and planned area of development?

2. Discuss the CCITT 1984 definition of ISDN, giving the rationale behind each of the key points.

3. Give some examples of sociological needs in rural areas, which are best met by an ISDN.

4. In what way has technology contributed to the process of realising a single common network for all telecommunication services?

5. An open channel teletext broadcasts four magazines of 10 frames each. Each frame is displayed for 10 seconds and there is a gap of 20 seconds between the magazines. If a viewer is interested in seeing a particular frame, how long has he to wait on an average after he switches on the system?

6. The open channel teletext specified in Exercise 5 is converted to a pseudointeractive system by adding a decoder which is capable of storing four frames in sequence immediately following the frame selected by the user. What is the average wait time if a user wants to examine four frames at random?

7. A 100-frame teletext magazine is transmitted using four scan lines of the vertical blanking interval of a commercial TV system. Calculate the average access time for a given frame. If an access time of less than one second is desired, calculate the bandwidth to be used for transmitting the magazine.

8. Two schemes are being considered for handling multidestination messages in an Electronic mail system:

   (a) The originating MTA makes all the necessary copies and these are sent out independently.
   (b) Using mailbag approach in which the routes are analysed first and a single copy is sent out on the common portion of the route and copies are made only when routes diverge.

Compare the two schemes in terms of transmission load, processing requirements and routing considerations.

9. In a local area network, there are workstations and personal computers. Show the distribution of MTA and UA in this environment. How would the message exchange take place between two personal computers and between a personal computer and a workstation?

10. Determine the volume of data generated if an A4 size page is scanned with a resolution of 200 pels/inch both horizontally and vertically. Each pel is represented by 4 bits. If the page data is to be transmitted over 2400 bps line in 20 seconds, calculate the compression ratio required.

11. A word processor page is to be transmitted using teletex service to a personal computer located 500 km away. Estimate the time required to transmit the page if the operating speed of teletex is 2400 bps. Calculate the cost of transmitting the page in India given that teletex charging is the same as dial up line charges. Assume Re.1 as charge per unit and the transmission takes place at 8.00 p.m. on a Sunday.

12. An alphabet set has 20 symbols with the following probabilities of occurrence:

Number of symbols	*Probability of occurrence*
8	0.0125
6	0.03
4	0.08
2	0.2

Determine the theoretical minimum number of bits required to encode this source if variable length coding is used.

13. Compare the coding efficiencies when a well populated A4 size text page is encoded using teletex and Group III facsimile.

14. Compare the costs of accessing an online database system when the access is through a telephone network or a data network.

15. In an online database system, the centralised computer is capable of servicing 25 requests or enquiries per second. If a user generates on an average one query every 40 seconds, determine the maximum number of simultaneous users that may be supported on this system. Determine the total population that may be registered on the machine if 10% of the users are active during the peak hours and 5% of them may be blocked.

16. Show that the segregated architecture performs better than the integrated architecture for high traffic intensities and vice versa by comparing the performances at the traffic intensity values 0.1 and 0.9.

17. Study the details and list the differences between LAP-B and LAP-D protocols.

18. Explain how rate adaptation works in the following cases:
    (a) 600 bps to 8 kbps
    (b) 9600 bps to 16 kbps
    (c) 1200 bps to 64 kbps

    Use two-stage rate adaptation where required.

19. Flexible format multiplexing is used for an ISDN B-channel. Show the B-channel octet bit allocations for the following subchannel traffic pattern:

Source 1 arrives	8 kbps
Source 2 arrives	16 kbps
Source 3 arrives	32 kbps
Source 3 terminates	
Source 4 arrives	16 kbps
Source 1 terminates.	

20. What is the percentage overhead in basic and primary channel structures?

21. What is the rationale behind specifying different functional groupings and interfaces in ISDN? Would you think that U interface may be relevant to India? Why?

22. ISDN permits physical implementations where more than one device can be connected at a particular reference point either through a multidrop line or multiport interfaces units. Illustrate the implementations in the following cases:

    One TE1 and TE2 are connected to a 2-port interface
    Multiple TE1s to an NT2 using multidrop connection
    Multiple TE1s to a multiport NT2.

23. What makes ISDN signalling flexible? Why is it that a packet switched network is more suitable than a circuit switched network for ISDN signalling?

24. Distinguish between user level and network level signalling in ISDN. List some network level signalling messages under the following categories:

    Backward set up
    End-to-end
    Call supervision

25. What is the maximum number of unique international ISDN numbers possible in the ISDN structure? Compare this with the population of the world. Will everyone on this planet receive a unique number? If a number is not to be reassigned even after a person is dead, estimate the number of years in which the ISDN number space may be exhausted.

26. There is a suggestion that ISDN addressing be such that a person is identified irrespective of his location by a unique number assigned to him. For example, even if a person moves from USA to India, his ISDN number remains the same. Suggest a method by which this suggestion can be implemented.

27. If an ISDN is to interwork with a X.25 PDN, how many digits have to be dialled in the worst case, if two-stage dialling is used?

28. Comment on the structures of standards for PDN and ISDN.

29. A telephone company uses an expert system for fault registration. Describe a typical conversation between the subscriber and the expert system for the situation where the subscriber is able to dial out calls but unable to receive incoming calls.

Epilogue

This text started with what some may consider as an outdated system, viz. the Strowger switching system, and has ended with the most recent development in telecommunications, namely, the ISDN. The question that naturally arises then is : What does the telecommunications have in store for us in the years ahead?

Undoubtedly, full-fledged ISDN will be a reality within the next 20 years. Wideband ISDN would follow suit. Intelligent networks supporting real time machine translation etc. still appear to be a far away dream.

Cordless communications is likely to dominate the loal networks, with the result that we may find persons in street corners, parks, trains and other public places conversing with persons at distant places using a concealed pocket held instrument. Strangely, they may all appear as insane persons talking to themselves in a vacuum.

Slowly, underground cooper wires will give way to fibre optic cables. The use of light as a carrier for telecommunications would make available to us such enormous bandwidth that video phones, telemovies, telelibraries, etc. may become a reality. With such services, the travel needs of the society will be confined to social, pleasure and religious purposes only.

Satellites will play an increasing role in telecommunications in the coming years. Novel multisatellite configurations would emerge to support globalisation of telecommunications. For example, the use of a series of 20–30 near-earth orbit satellites for communications appears to be a definite possibility within the next few years.

The telecommunications of the future seems to have the tremendous potential of reducing pollution and energy consumption, and restoring the ecological balance of our dear planet. The realisation of the concepts of 'global village' or 'earth as one family' depends much on the developments in telecommunications.

Answers to Selected Exercises

Chapter 1

3. 80 GHz
4. 10 mW
5. 100 mW
6. 16 dB
7. 9.9 dBm
8. Signal power is one-half of the noise power.
9. Less than 0.1
14. 180, 380 or 220, 420
15. 50
16. 45
17. 15.5 km
18. Two-part design : one for the number range 400-699 and the other for the ranges 100-399 and 700-999. Number of operators during peak traffic is nine and during lean traffic two. Number of terminations is 2500.

Chapter 2

2. 16.44 seconds
5. A busy tone may occur due to the nonavailability of a switching path.
6. The ringing tone is not the usual one he hears and hence is originating from a different exchange.
7. There are at least five exchanges involved in setting up the call.
8. No. In a multilingual country, the segment may be identified by the language in which the announcement is made.
11. Level 6, contacts 1 and 10 respectively. Level 10, contact 5, and level 3, contact 5 respectively.
15. 20, 2
16. 24, 5
17. Number of line finders = 29, $S = 58$, $SC = 29$, $TC = 0.6$, $C = 116$, $CCI = 25$, $EUF = 1$

Chapter 3

4. (a) 6.12 seconds (b) 6.0 seconds (c) 13.2 seconds
7. 2 mW
8. 45 seconds
9. 10 days
10. 325,000 for full matrix; 199,750 for diagonal matrix.
11. IC crosspoint is 100 times more reliable.
13. 134,000

Chapter 4

11. (a) 7680 (b) 0.002
12. (a) 2.13 (b) 32.13
13. (a) 0.22 (b) 0.00032 (c) 8.6×10^{-8}
14. (a) 0.221 (b) 2.74×10^{-5} (c) 1.092×10^{-12}
15. (a) $s = 10$, $S = 5120$, $TC = 1$, $EUF = 0.075$, $CCI = 1.6$

Chapter 5

1. $10\ W/m^2$
2. Two times the residential level at the office and two times the office level at the airport.
4. (a) 20 kHz (b) 10 kHz
5. 697 Hz
6. 6 dB
7. 6 bits
8. 43.8 dB, 13.8 dB
9. 11 bits
10. 34 dB or the ratio of the largest to the smallest permissible value is 50.12.
11. (a) 10111110 (b) 01011111 or 01100000 (c) 11110111 or 11111000 (d) 11111100
12. (a) 812 mV (b) −29.54 mV
14. 10001110, 00110000, 01100001
15. 1.7 bits
17. DM bandwidth = 8.045 MHz, PCM bandwidth = 160 kHz
18. 20.1 kHz

19. 1920 bps
20. 0.243, 0.50, 0.195, 0.159
21. 0.318, 0.159, 0.106, 0.079, 93.75 μs
22. (a) Low frequency region 40–800 Hz, high frequency region $\geq$ 80 kHz
    (b) 3.17×10^7 m/s

Chapter 6

1. (a) 1.95 μs (b) 32 (c) 512 kHz
2. 500
3. (a) bus (b) $\frac{125-128\,tm}{32}$ μs per transfer (c) 7 bits (d) 1024 kHz
   (e) 16 channel (f) 128-to-16 and 16-to-128
4. The cost of the memory controlled switch is 1/4 of the output controlled switch.
5. (a) Data memory (b) 0.49 μs (c) 7 bits (d) 2.04 MHz (e) Single channel (f) 128-to-1 and 1-to-128
6. DM size = CM size = 1000 words,. CM width = 10 bits, DM width = 8 bits, 0.0625 μs
7. (a) 2.98×10^{-9} (b) 1.49×10^{-9}
11. No. of TSI switches = 41, cost of the switch = 26,880 units
12. $P_B = 9.7 \times 10^{-11}$, C_{TST} = 6144 units
14. 4096

Chapter 7

4. 1.386
5. 2.88×10^{-4} per μm
6. 1.732
7. 0.2934, 17.1°
8. 490
9. 5×10^{14} Hz, 3.3×10^{-19} J
10. 0.91, 4.55 ns
11. 55.58 cm^{-1}
12. 604 mA
13. 1.44 ns
14. 1.13 μm, 71.5%

15. 1.32 μm, 3.61 μW
16. 50%, 15.9 nA
17. 39.8 KΩ
18. 2.4 ns
19. 18 km

Chapter 8

1. (a) 40 E (b) 1440 CCS
2. 60,000, 60 ms
3. 30,000, 0.7
4. 19
5. 0.091
6. 0.091
9. (a) 0.036 (b) 0.037 (c) 1.73 , 0.072 E
10. 0.04, 0.96 E
11. 0.567, 3.9 minutes
13. 0.206
17. (a) 0.098 (b) 0.167 s (c) 2.43×10^{-4} (d) 0.84

Chapter 9

2. AWG 26
4. 17.46°
5. 0.983° K
6. 1.977 hrs, 4.94 hrs
7. 0.005 transponder, 0.01 transponder
8. 20 dB
9. 16
10. 0.37 μH, 0.67
11. Lower limit of high frequency region is 0.215 MHz.
12. (a) 0.85 km (b) 1.03 km (c) 98
13. The numbering space of Maldives can accommodate about 1000 times more than the present population of the country.
14. 1000, 100
16. 0.157
17. 16.8 s, 16 s
18. 1.39 signals per second per channel.

Chapter 10

3. 24 Mbps
4. 39 kbps
5. 123.3 dB
6. Amplitude shift keying
7. 2400 bps
8. 120.2%
10. Best = 1 hop, Worst = $(N-1)$ hops
11. $M < \frac{Rk\,(T_s - T_q)}{(pk + dk - p)}\,p$
12. Tcs = 18.33 s, Tps = 46.1 s
14. Datagram service is the cheapest.
15. 0.11%
16. 90.9%
17. Rs. 135
18. (a) True (b) False
20. 0.014%
21. 01101111101001111110011
22. 226 bits
23. 1.8%
24. 4.5%
25. 7, 3
26. (a) Presentation (b) Transport (c) Session (d) Transport (e) Network (f) Presentation (g) Data link
27. 1061, 2123
29. 2.3, 0.23
30. 200 bits, 66.7%
31. (a) 25.055 ms
32. 65 μs
33. 190 μs
34. 2.91%, 5.65%
35. 33 frames, 16 frames.

Chapter 11

5. 4 minutes
6. 6 minutes

10. 321.75
11. Rs. 1.67
12. 3.64 bits /symbol
15. 10,526
18. (a) One of the solutions is that the following bits are used as information bits 0,13, 26, 27, 39, 52, 65, 66.... and so on (b) Two stage rate adaptation is required. The information bits in the 16 kbps stream are: 0, 2, 4, 5, 7, 9, 11, 12, 14, 16, 18, 19 ... and so on. The information bits in 64 kbps stream are 0, 1, 8, 9, 16, 17 and so on. (c) Two stage rate adaptation.
19.

Event	Structure
Source 1 arrives	*b* 1 1 1 1 1 1 1
Source 2 arrives	*b* 1 *b b* 1 1 1 1
Source 3 arrives	*b* 1 *b b b b b* b
Source 3 terminates	*b* 1 *b b* 1 1 1 1
Source 4 arrives	*b* 1 *b b b b* 1 1
Source 1 terminates	1 1 *b b b b* 1 1

b = information bit 1 = pad bit