5

Speech Digitisation and Transmission

The human ear is capable of perceiving frequencies in the range of 16 Hz–20 kHz, known as the audio range. Speech produces a narrow band of frequencies 100 Hz–10 kHz in this audio range. If all the frequencies present in a speech waveform are transmitted, the received speech is most natural. However, the speech remains understandable or intelligible even if some upper and lower frequencies of the speech spectrum are omitted. A reduction in the bandwidth is desirable as it reduces the cost of the communication systems. An acceptable level of intelligibility of speech is obtained by transmitting frequencies in the range of 300–3400 Hz. Such a band-limited speech signal is often called 'toll' (telephone) quality speech. Most of the commercial speech communication systems are designed to have a bandwidth of 3.1 kHz. In this band-limited range of speech, the ear is most sensitive to frequencies that lie around 3 kHz. In the case of female voice, maximum energy is distributed around this frequency, whereas in the case of male voice, the maximum energy occurs at a much lower frequency. This is the reason why women are preferred as telephone operators and announcers.

As mentioned in Chapter 1, a channel in a communication system has a finite transmission loss and is subject to noise impairment. When the length of the transmission path increases, the signal-to-noise ratio at the receiving end decreases. Ultimately, the signal may be completely masked by the additive noise, particularly when the signal level is low. In analog voice transmission, the effect of noise and interference is most apparent during speech pauses when the signal amplitude is near zero. Even relatively low noise levels can be quite annoying to a listener during speech pauses. The same levels of noise may be unnoticeable when speech is present. Hence, it is the absolute noise level of an idle channel that determines the analog speech quality.

Digital speech transmission overcomes many of the problems faced in the analog systems. In a digital system, speech and speech pauses are encoded with data pattern and transmitted at a constant power level. Signal

regeneration at regular intervals bringing the signal to the original level virtually eliminates all noise due to the transmission medium. Thus, the idle channel noise is determined by the encoding process and not by the transmission link in a digital system. The ability of the digital transmission to reject crosstalk is superior to that of an analog system. First, low level crosstalks are eliminated because of the constant amplitude signals. Second, high amplitude crosstalks result in detection errors and as such are unintelligible. Intelligible crosstalks that occur in analog systems are particularly undesirable as someone's privacy is being violated. Other advantages of digital systems include the ability to support nonvoice services, and easy data encryption and performance monitoring. The signal structure in a digital system is independent of the nature of traffic and therefore the quality of the received signal can be ascertained with no knowledge of the traffic type.

Digital systems require greater bandwidth than analog systems and sometimes this can be a disadvantage. For example, transmission media like wire pairs cause greater attenuation when larger bandwidth signals are passed through them. Nevertheless, the advantages offered by the digital systems outweigh the bandwidth considerations.

5.1 Sampling

The first step in digitising speech, or for that matter any analog waveform, is to establish a set of discrete times at which the input waveform is sampled. The discrete sample instances may be spaced either at regular or irregular intervals. The prevalent digitisation techniques are based on the periodic and regularly spaced sample instances. The minimum sampling frequency required to reconstruct the original waveform from the sampled sequence is given by **Nyquist criterion** or **theorem** which can be stated as

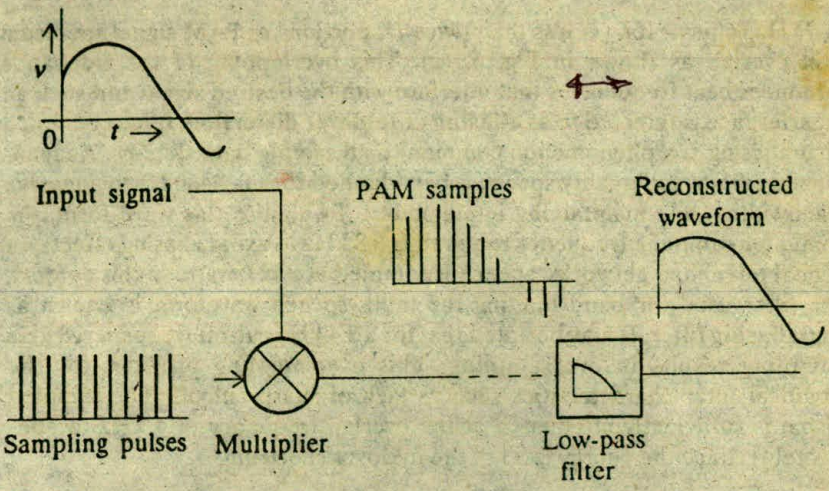
$$f_s \geq 2H \tag{5.1}$$

where

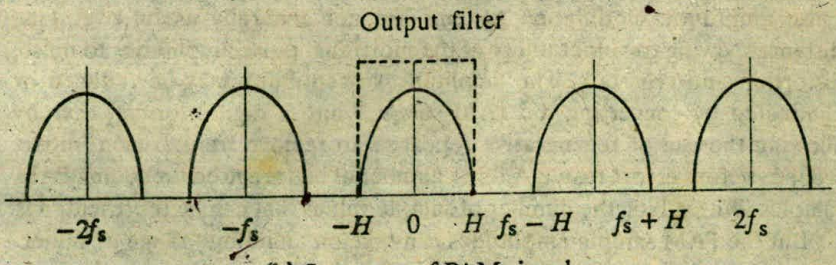
f_s = sampling frequency or the Nyquist rate

H = highest frequency component in the input analog waveform

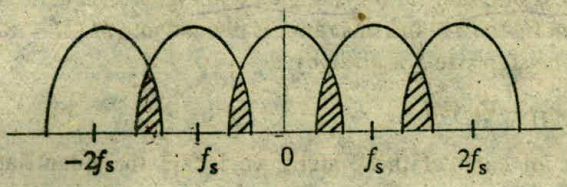
H is the bandwidth of the input waveform if it is not band limited with a lower cut-off frequency. In this case, the original waveform is reconstructed by passing the sampled values through a low pass filter which smoothens out or interpolates the signal between sampled values. The scheme is illustrated in Fig. 5.1(a). Sampling is a process of multiplying a constant amplitude impulse train with the input signal. In other words, it is an amplitude modulation process, where the pulse train acts as the carrier. Since the amplitude of the pulses is modulated, the scheme is called **pulse amplitude modulation** (PAM).



(a) Sampling and recovery of analog waveform



(b) Spectrum of PAM signal



(c) Spectrum overlap when $f_s < 2H$

Fig. 5.1 Digital transmission using PAM.

The frequency spectrum of an amplitude modulated signal, when the carrier is a sine wave, has frequencies ranging from $f_c - H$ to $f_c + H$, where f_c is the carrier frequency. If the carrier is a pulse train, as is the case in PAM, the output spectrum contains the fundamental as well as the harmonics of the fundamental shown in Fig. 5.1(b). If the pulse train is a square wave (50 per cent duty cycle), only the fundamental and odd harmonics are present. The low pass filter at the receiver end allows only the baseband component

0–H Hz to pass. If f_s is less than twice H, portions of PAM signal spectrum will overlap as shown in Fig. 5.1(c). This overlapping of the sidebands produces beat frequencies that interfere with the desired signal and such an interference is referred to as **aliasing** or **foldover distortion.** It may be noted that aliasing is a phenomenon common to all sample data systems. As mentioned earlier, toll quality speech is band limited to 300–3400 Hz to conserve bandwidth while maintaining intelligibility. To digitise this wave-form, the minimum sampling frequency required is 6.8 kHz to avoid aliasing effects. In digital telephone networks, speech is sampled at 8 kHz rate. In this context, the filter used for band limiting the input speech waveform is known as **antialiasing filter.** It may be noted that for a 3.4 kHz cut-off frequency, 8 kHz sampling results in oversampling. This oversampling provides for the nonideal filter characteristics such as lack of sharp cut off. The sampled signal is sufficiently attenuated at the overlap frequency of 4 kHz to adequately reduce the energy level of the foldover spectrum.

5.2 Quantisation and Binary Coding

Pulse amplitude modulation systems are not generally useful over long distances, owing to vulnerability of the individual pulse amplitudes to noise, distortion and crosstalk. The amplitude susceptibility may be reduced or eliminated by converting the PAM samples into a digital format, thereby allowing the use of regenerative repeaters to remove transmission imperfections before errors result. A finite number of bits are used for coding PAM samples. With n bits, the number of sample values that can be represented is 2^n. But the PAM sample amplitudes can take on an infinite range of values. It, therefore, becomes necessary to quantise the PAM sample amplitude to the nearest of a range of discrete amplitude levels.

The process of quantisation is depicted in Fig. 5.2. Signal V is confined to a range from V_L to V_H, and this range is divided into M ($M = 8$ in Fig. 5.2) equal steps. The step size S is given by

$$S = (V_H - V_L)/M \qquad (5.2)$$

In the centre of each of these steps we locate the quantisation levels $V_0, V_1, \ldots, V_{M-1}$. The quantised signal V_q takes on any one of the quantised level values. A signal V is quantised to its nearest quantisation level. The boundary values between the steps are equidistant from two quantisation levels and a convention may be adopted to quantise them to one of the levels. For example,

$$V_q = V_3 \quad \text{if } (V_3 - S/2) \leq V < (V_3 + S/2)$$
$$V_q = V_4 \quad \text{if } (V_4 - S/2) \leq V < (V_4 + S/2)$$

Thus, the signal V_q makes a quantum jump of step size S and at any instant of time the quantisation error $V - V_q$ has a magnitude which is equal to or less

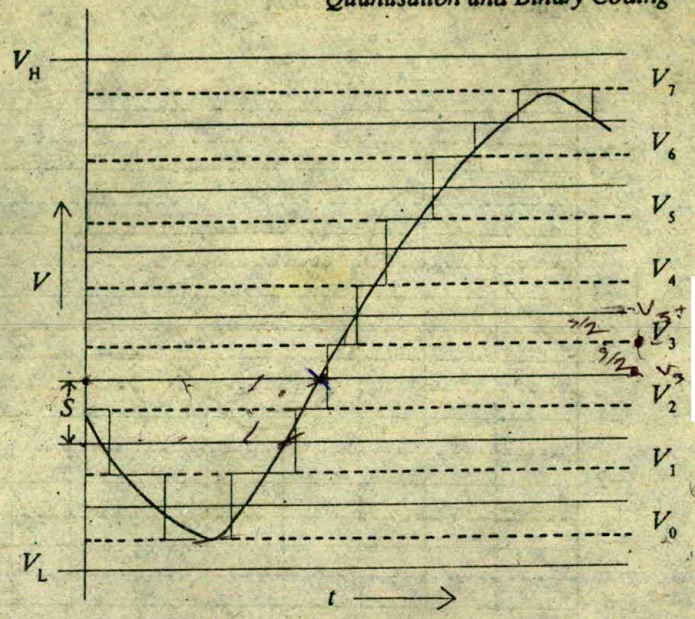


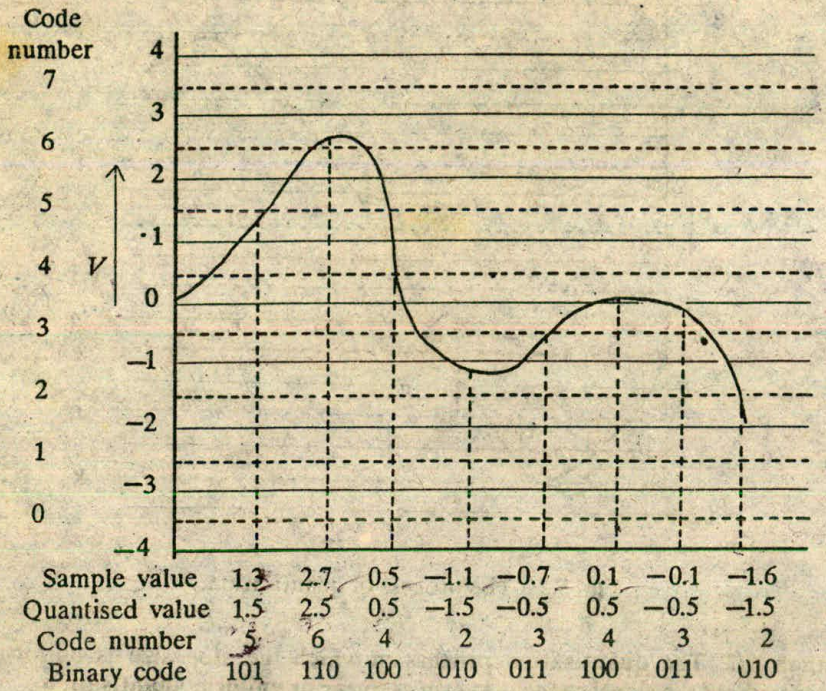
Fig. 5.2 The process of quantisation.

than $S/2$. The quantisation process in which the step size is uniform throughout the signal range is known as **linear** or **uniform quantisation.**

The process of quantisation itself brings about a certain amount of noise immunity to the signal. Consider the transmission of quantised sample amplitudes. We may construct repeaters which consist of quantisers and place them at intervals such that the instantaneous signal amplitude variation due to attenuation and channel noise in each interval is less than $\pm S/2$. Then the quantiser repeater reproduces the original quantised level with the noise removed. If, however, the instantaneous noise level is larger than $S/2$, an error occurs in the quantisation level.

The quantised signal is an approximation to the original signal. The quality of approximation may be improved by reducing the size of the steps and thereby increasing the number of allowable levels. However, reducing the step size makes the PAM signal more susceptible to noise in the sense that a smaller noise level can now cause an error in the quantisation level at the repeaters. The susceptibility to noise can be greatly minimised by resorting to digital coding of the PAM sample amplitudes. Each quantised level is represented by a code number which is transmitted instead of the quantised sample value itself. If binary arithmetic is used for coding, then the code number is transmitted as a series of pulses. Hence, such a system of transmission is called **pulse code modulation (PCM).**

The essential features of binary PCM are illustrated by means of an example shown in Fig. 5.3. The analog input signal is limited in its excursions



Sample value	1.3	2.7	0.5	−1.1	−0.7	0.1	−0.1	−1.6
Quantised value	1.5	2.5	0.5	−1.5	−0.5	0.5	−0.5	−1.5
Code number	5	6	4	2	3	4	3	2
Binary code	101	110	100	010	011	100	011	010

Fig. 5.3 Binary PCM – features.

to the range – 4 V to + 4 V. The step size is one volt. Eight quantisation levels are used and are located at – 3.5 V, – 2.5 V, ..., + 3.5 V. Code number zero is assigned to – 3.5 V, the code number one to – 2.5 V and so on. Each code number has its equivalent 3-bit representation. Considering the sample points shown in Fig. 5.3, if the sampled analog signals are transmitted, the voltage values 1.3, 2.7, 0.5 etc. would be transmitted. If the quantised values are transmitted, the voltage values 1.5, 2.5, 0.5 etc. would be transmitted. In binary PCM, the code patterns 101, 110, 100 are transmitted.

A diagram showing the functional blocks of a PCM system is given in Fig. 5.4. The analog input signal V is band limited to 3.4 kHz to prevent aliasing and sampled at 8 kHz. The samples are quantised to produce quantised PAM signals which are then applied to the encoder. The encoder generates a unique pulse pattern for each quantised sample level. The quantiser and the encoder together perform the **analog-to-digital (A–D) conversion.** The first operation at the receiver is the separation of signal from the noise. This is done by a quantiser which makes a simple decision of whether a pulse has been received or not or which of the two voltage levels of a pulse has occurred. It then regenerates an appropriate pulse or voltage level based on the decision. The regenerated pulse train is fed to a decoder which assembles the pulse pattern and generates a corresponding quantised

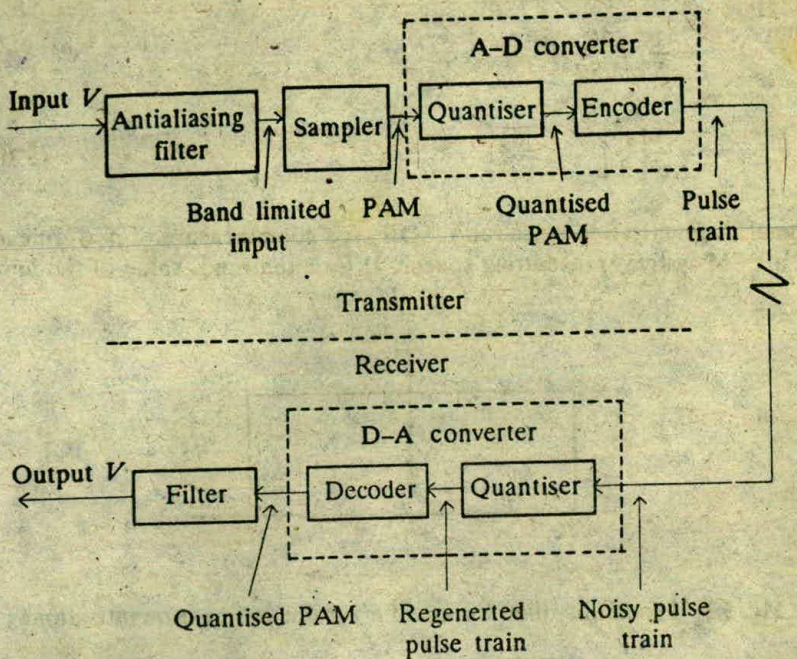


Fig. 5.4 A PCM system for speech communication.

voltage level. The quantiser and the decoder together perform **digital-to-analog (D–A) conversion**. The quantised PAM levels are then passed through a filter which rejects the frequency components lying outside the baseband and produces a reconstructed waveform of the original band limited signal.

5.3 Quantisation Noise

It was pointed out in Section 5.2 that the quantised signal is an approximation to the original signal and an error is introduced in the signal due to this approximation. The instantaneous error $e = V - V_q$ is randomly distributed within the range $(S/2)$ and is called the **quantisation error** or **noise**. We now calculate the mean square quantisation error S^2. For linear quantisation, we can assume that the probability distribution of the error is constant within the range $\pm(S/2)$ as shown in Fig. 5.5. The average quantisation noise output power is given by the variance

$$\sigma^2 = \int_{-\infty}^{\infty} (e - \mu)^2 p(e)\, de \tag{5.3a}$$

where μ = mean, which is zero for quantisation noise. The range of quantisation error $\pm$ $(S/2)$ determines the limits for integration. Therefore,

$$\sigma^2 = \int_{-S/2}^{S/2} (e-0)^2 \frac{1}{S}\, de$$

$$= \frac{1}{S} \int_{-S/2}^{S/2} e^2\, de$$

$$= \frac{1}{S}\left(\frac{e^3}{3}\right)_{-S/2}^{S/2} = \frac{S^2}{12} \qquad (5.3b)$$

Signal to quantisation noise ratio (SQR) is a good measure of performance of a PCM system transmitting speech. If V_r is the r.m.s. value of the input

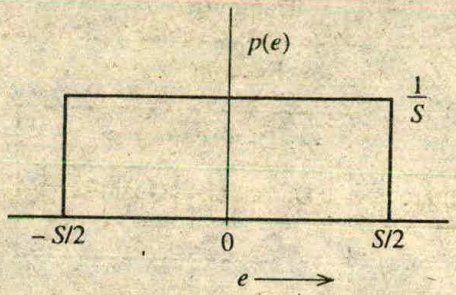


Fig. 5.5 Probability distribution of error due to linear quantisation.

signal and if we assume (for convenience) a resistance level of 1 ohm, then SQR is given by

$$\begin{aligned} \text{SQR} &= 10 \log [(V_r^2)/(S^2/12)] \text{ dB} \\ &= 10 \log (12) + 20 \log (V_r/S) \text{ dB} \\ &= 10.8 + 20 \log (V_r/S) \text{ dB} \end{aligned} \qquad (5.4)$$

If the input signal is a sinusoidal wave with V_m as the maximum amplitude, SQR may be calculated for the full range sine wave as

$$\begin{aligned} \text{SQR} &= 10 \log [(V_m/\sqrt{2})^2)/(S^2/12)] \text{ dB} \\ &= 10 \log (6) + 20 \log (V_m/S) \text{ dB} \\ &= 7.78 + 20 \log (V_m/S) \text{ dB} \end{aligned} \qquad (5.5)$$

Expressing S in terms of V_m and the number of steps, M, we have

$$\begin{aligned} \text{SQR} &= 10 \log \frac{(V_m^2/2)}{(4V_m^2/12M^2)} \text{ dB} \\ &= 10 \log (1.5M^2) \text{ dB} \\ &= 20 \log (1.225M) \text{ dB} \end{aligned} \qquad (5.6)$$

It may be noted that the quantity $1.225M$ represents the signal to quantisation noise voltage ratio for a full range sinusoidal input voltage. Now, $M = 2^n$, where n is the number of bits used to code a quantisation level. Therefore, Eq. (5.6) may be expressed in terms of n as

$$\begin{aligned} \text{SQR} &= 20\log(1.225) + 20n\log(2)\ \text{dB} \\ &= 1.76 + 6.02n\ \text{dB} \end{aligned} \tag{5.7}$$

Table 5.1 gives the values of SQR for different binary code word sizes for sinusoidal input systems. It may be seen that every additional code bit gives an improvement of 6 dB in SQR.

Table 5.1 Binary Code Word Size vs. Quantisation Noise

n (bits)	M (levels)	Voltage ratio ($1.225M$)	SQR (dB)
2	4	4.9	13.8
4	16	19.6	25.8
6	64	78.4	37.8
7	128	156.8	43.8
8	256	313.6	49.8
9	512	627.2	55.8
10	1024	1254.4	61.8

5.4 Companding

In linear or uniform quantisation, the magnitude of quantisation noise is absolute for a particular system and is independent of the input signal amplitude. Therefore, comparatively, the weak and low-level signals suffer worse from quantisation noise than the loud and strong signals. This is illustrated in Fig. 5.6 which gives a plot of input signal amplitude versus quantisation error in percentage for $n = 6$. The quantisation error as a fraction of the input signal amplitude may be expressed as

$$e_f = \frac{S/2}{|V|}$$

For sinusoidal input we have $S = 2V_m/M$, and hence

$$\begin{aligned} e_f &= \frac{V_m}{M}\ \frac{1}{|V|} \\ &= \frac{V_m}{M}\ \frac{100}{|V|}\ \% \end{aligned} \tag{5.8}$$

The very high percentage error at low input signal levels actually represents idle channel noise. The effect of this is particularly bothersome

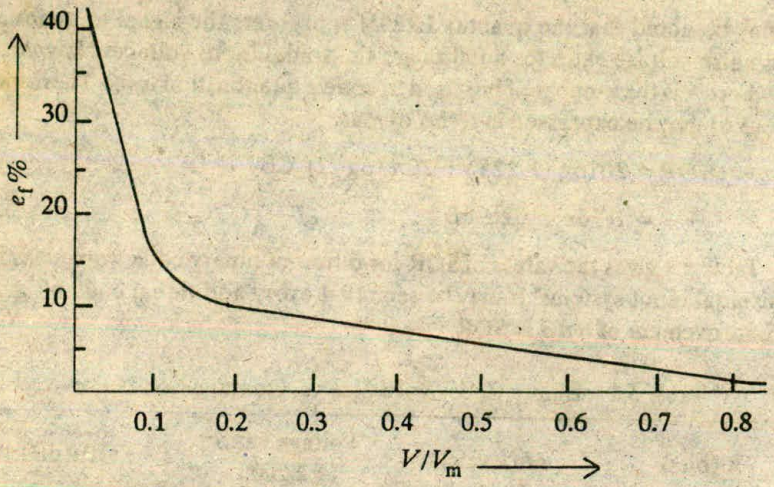
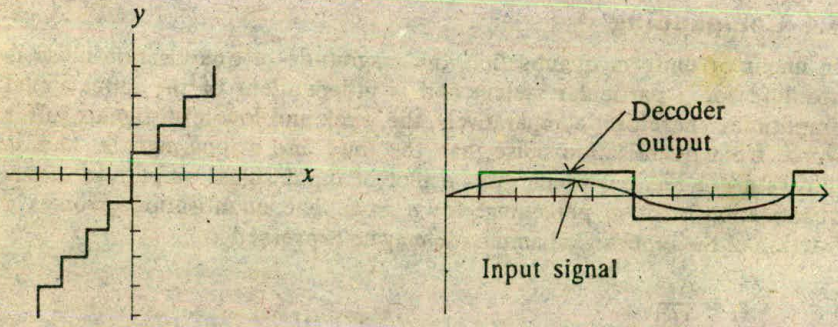


Fig. 5.6 Percentage quantisation error.

during speech pauses and can be minimised by choosing zero volt level as a quantisation level and avoiding the mid points of the first intervals on either side of the zero level as quantisation levels. The scheme which uses the two first mid points is known as **mid-riser** scheme and the other as **mid-tread** scheme. The two schemes and their effects on the idle channel noise are shown in Figs. 5.7 and 5.8. Obviously, the mid-tread scheme uses odd number of quantisation levels. In this case $M = 2^n - 1$. In the mid-tread scheme, very

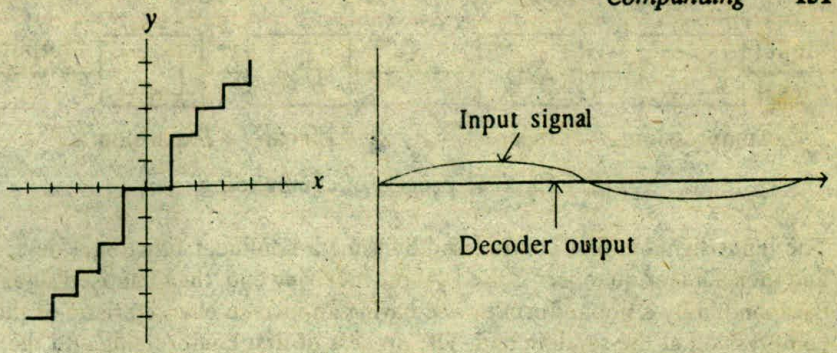


(a) Mid-riser scheme (b) Idle channel output

Fig. 5.7 Mid-riser quantisation and the idle channel noise.

low signals are decoded into a constant, zero-level output. However, if a d.c. bias exists in the encoder, idle channel noise is still a problem with mid-tread quantisation.

A more efficient method of minimising large variations in the percentage quantisation error over the signal range is to use nonlinear or nonuniform



(a) Mid-tread scheme (b) Idle channel output

Fig. 5.8 Mid-tread quantisation and the idle channel noise.

quantisation. Here, the quantisation intervals are not uniform and are allowed to increase with the signal magnitude. With this technique it is possible to maintain a constant SQR for all signal levels and a fewer bits are required to maintain the chosen SQR. It is interesting to note that uniform quantisation intervals result in nonuniform SQR over the signal range and nonuniform intervals result in uniform SQR. The effect of permitting larger quantisation intervals at higher signal amplitudes is to compress the input signal to achieve a uniform quantisation level as shown in Fig. 5.9. It may be

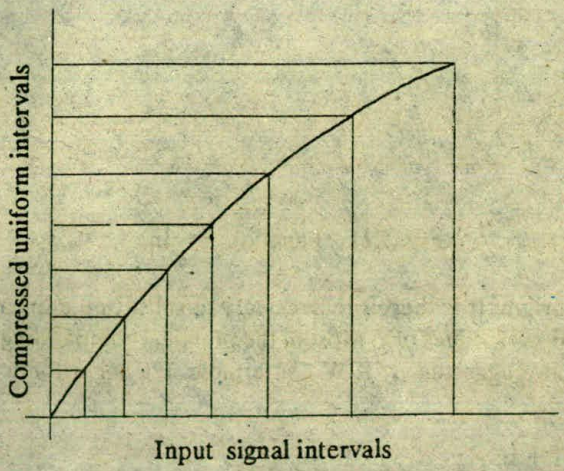


Fig. 5.9 Compressing input signal intervals to a uniform interval.

seen that successively larger input signal intervals are compressed into constant length quantisation intervals. Thus, the larger the sample value, the more it is compressed before encoding. The process is illustrated in Fig. 5.10.

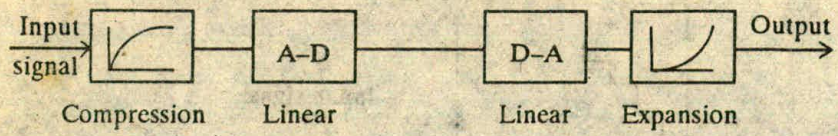


Fig. 5.10 The process of companding.

The input signal is first compressed by using a nonlinear functional device and then a linear quantiser is used. At the receiving end, the quantised signal is expanded by a nonuniform device having an inverse characteristic of the compression at the sending end. The process of first compressing and then expanding is referred to as **companding**.

A variety of nonlinear compression-expansion functions can be chosen to implement a compandor. The obvious one is a logarithmic law. Unfortunately, as is seen from Fig. 5.11, the function $y = \ln x$ does not pass

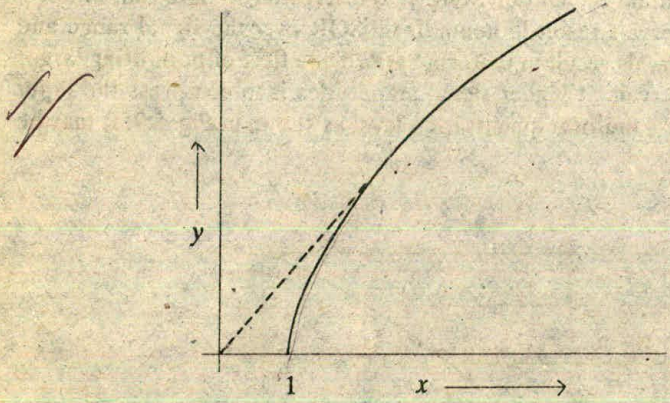


Fig. 5.11 Function $y = \ln x$.

through the origin. It is therefore necessary to substitute a linear portion to the curve for lower values of x. Most of the practical companding systems are based on a law suggested by K.W. Cattermole. For logarithmic section the law is

$$y = \frac{1 + \ln Ax}{1 + \ln A} \qquad \text{for } \frac{1}{A} \leq x \leq 1 \tag{5.9}$$

and for linear section,

$$y = \frac{Ax}{1 + \ln A} \qquad \text{for } 0 \leq x \leq \frac{1}{A} \tag{5.10}$$

where, for $x = 1/A$, the values of Y in Eqs. (5.9) and (5.10) are the same and the function is continuous at $x = 1/A$. The quantity A is called the compression coefficient. A set of curves can be obtained by choosing different values for A. The expansion function which is the inverse of compression function is given by

$$x = \frac{e^{y(1 + \ln A) - 1}}{A} \quad \text{for } \frac{1}{1 + \ln A} \leq y \leq 1 \tag{5.11}$$

$$x = \frac{y(1 + \ln A)}{A} \quad \text{for } 0 \leq y \leq \frac{1}{1 + \ln A} \tag{5.12}$$

Equations (5.9)–(5.12) are valid for positive values of x and y. Appropriate signs and magnitude quantities must be introduced for negative values. The companding characteristics represented by Eqs.(5.9)–(5.12) are collectively known as A-law. These characteristics are recommended by CCITT and are adopted by India and other European countries. The United States and Japan follow a variation of A-law known as μ-law.

The implementation of these laws requires nonlinear signal processing followed by linear quantisation as shown in Fig. 5.10. In practice, nonlinear signal processing at the transmitting end together with the inverse nonlinear processing at the receiving end presents considerable problems. Therefore, a piecewise linear segment approximation is used in practice. A-law companding consists of eight linear segments for each polarity. The slope halves for each segment except the lowest two segments which have the same slope. The piecewise linear approximation is shown in Fig. 5.12. The lowest two segments of positive and negative polarities coalesce into one straight line segment. As a result, there are 13 effective segments in the curve and the law is sometimes referred to as 13-segment companding law. (In μ-law, the slope halves in the lowest two segments also, giving rise to 15 effective segments.) Each segment is divided into 16 linear steps. Eight bits are required to represent each sample value: 1-bit sign, 3-bit segment number and a 4-bit linear step number. There are in all 256 defined signal levels. The piecewise linear companding rules are easily implementable within the analog-to-digital conversion process. As the signal value passes up from one segment to another, the step size of the A–D converter is doubled by simply discarding the least significant bit of the converter output.

The companding may be done in the digital domain too, i.e. after A–D conversion at the transmitting end and before D–A conversion at the receiving end. The process is depicted in Fig. 5.13. Here, we use an A–D converter that has a higher resolution than the one used in the process shown in Fig. 5.10. Since the companding laws use 8-bit words, A–D and D–A converters in Fig. 5.10 have a resolution of 8 bits. To illustrate companding in digital domain, let us consider a 12-bit A–D converter and the A-law companding. In Fig. 5.13, the output lines of the A–D converter are connected to the address pins of a read only memory (ROM) which, in our

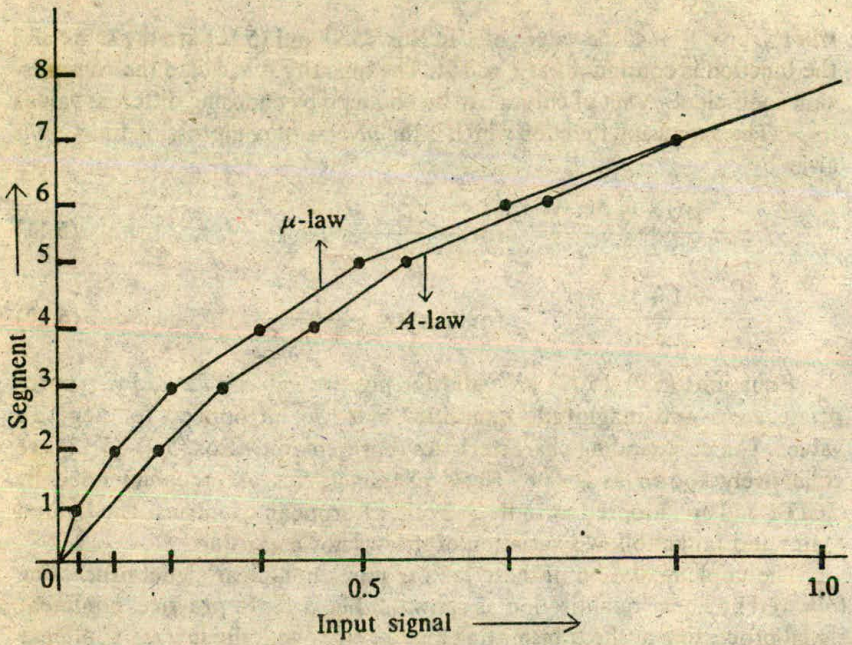
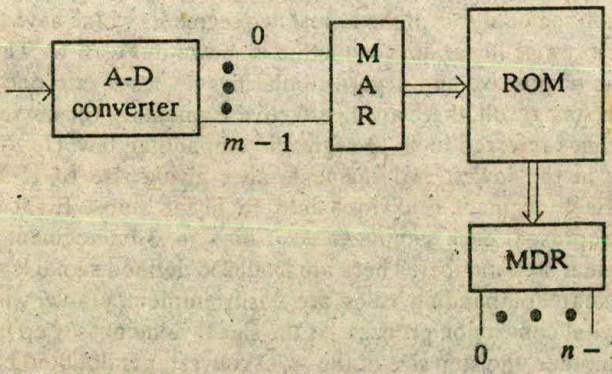


Fig. 5.12 Piecewise linear companding.



MAR = memory address register MDR = memory data register
ROM = read only memory

Fig. 5.13 Digital companding.

example, has 4096 locations of eight bits each. Every location of the ROM contains a code word value. Each A–D converter output addresses a unique location in the ROM and the code word value contained in that location is output to the communication channel. By suitably replicating the code word values in more than one location, compression is achieved.

As mentioned earlier, each segment in the *A*-law is quantised linearly with 16 steps. The first two segments have the same slope and the subsequent ones have one-half of the slope of the previous segment. The slope is halved by doubling the input signal range for the same number of quantisation steps. In our example of digital companding, let us consider only the positive signal. The entire positive input signal range is represented by 2048 discrete output values of the A–D converter. The lowest 16 output values (000–00F) correspond to the 16 quantisation levels of the first segment. The next 16 (010–01F) values correspond to the second segment. The slope of each subsequent segment is halved by assigning double the number of output values of the previous segment. Thus, 32 discrete output values (020–03F) constitute the third segment, 64 values the fourth segment and so on as shown

Table 5.2 Digital Companding Coding Details

Segment number	Range of ROM addresses in hexadecimal	No. of ROM locations	Range of code word values in hexadecimal	No. of locations per code word (Replication level)
1	000 – 00F	16	00 – 0F	1
2	010 – 01F	16	10 – 1F	1
3	020 – 03F	32	10 – 2F	2
4	040 – 07F	64	30 – 3F	4
5	080 – 0FF	128	40 – 4F	8
6	100 – 1FF	256	50 – 5F	16
7	200 – 3FF	512	60 – 6F	32
8	400 – 7FF	1024	70 – 7F	64

in Table 5.2. Every location corresponding to the first two segments contains a distinct code word. For segment three, every two adjacent code locations contain the same code word. Hence the transmitted code word changes after a change of two steps at the input, thus doubling the step size or halving the slope. Similar reasoning holds for segments 4–8.

5.5 Differential Coding

Pulse code modulation discussed in Section 5.4 is a generic coding scheme that can be used to code any arbitrary random waveform whose maximum frequency component does not exceed one-half of the sampling rate. The scheme is not specifically designed for digitising speech waveforms. If it were the case, the properties of speech signals can be exploited profitably to arrive at a more efficient scheme than the PCM.

Speech waveforms exhibit considerable redundancy which can be usefully exploited in designing coding schemes. The following characteristics of speech signals contribute to the redundancy:

- Nonuniform amplitude distributions
- Sample-to-sample correlations
- Periodicity or cycle-to-cycle correlations
- Pitch interval-to-pitch interval correlations
- Speech pauses or inactivity factors.

In speech signal, lower amplitude values are more common than the higher amplitude values. The power levels of active speech signals tend to occur at the lower end of the signal range. There is a strong correlation between adjacent 8 kHz samples and the correlation coefficient is estimated to be 0.85 or higher. Although telephone quality speech signal requires the entire 300–3400 Hz bandwidth, at any instant of time only a few frequencies exist in a sound and the waveform tends to exhibit strong correlations over several samples corresponding to cycles of an oscillation. A sizeable fraction of the human speech sounds is produced by the flow of puffs of air from the lungs into the vocal tract. The interval between these puffs of air is known as the pitch interval. Speech waveforms display repetitive patterns corresponding to the duration of a pitch interval in a given sound. There may be as many as 20 to 40 pitch intervals in a single sound. Analysis of telephone conversations have indicated that a party is typically active for about 40% of a call duration largely on account of listening to the other party's talk. In addition, there are pauses in one's speech. All these properties of speech are useful in designing coding schemes that reduce the bandwidth requirements of the channel. In the digital domain significant reductions in bit rates can be achieved. A number of coding schemes exploiting different redundancy properties have been evolved and we discuss some of these in this section and in Section 5.6.

Delta or differential coding systems are designed to take advantage of the sample-to-sample redundancies in speech waveforms. Because of the strong correlation between adjacent speech samples, large abrupt changes in levels do not occur frequently in speech waveforms. In such situations, it is more efficient to transmit or encode and transmit only the signal changes instead of the absolute value of the samples. **Delta modulation (DM)** is a scheme that transmits only the signal changes and **differential pulse code modulation (DPCM)** encodes the differences and transmits them.

A delta modulator may be implemented by simply comparing each new signal sample with the previous sample and transmitting the resulting difference signal. At the receiver end, the difference signals are added up to construct the absolute signal by using an integrator. However, such a system, being open loop, suffers from the possibility of the receiver output diverging from the transmitter input due to system errors or inaccuracies. The system can be converted into a closed loop system by setting up a feedback path with an integrator at the transmitting end as shown in Fig. 5.14. The integrator must be identical to the one at the receiving end. The difference signal $e(b)$ has only two possible levels. A positive level indicates that the new sample

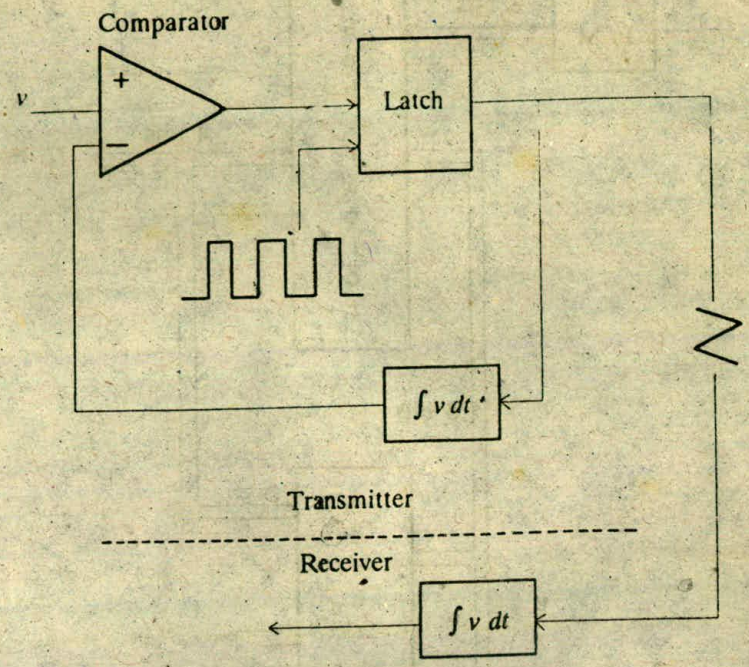


Fig. 5.14 Delta modulator scheme.

value is greater than the previous output of the integrator and the negative level the vice versa conditions. A latch provides a delay of one sample time, so that the input is always compared with the previous output. When the input is constant, the output of the transmitter is an alternating positive and negative pulse train. This constitutes the quantisation noise in delta modulators and is also known as granular noise. If the transmitter input signal changes too rapidly, the receiver output is unable to keep up and this phenomenon is known as **slope overload**. This problem may be overcome by using a variable slope integrator whose output slope is increased or decreased, depending on the rate of change of the input signal. Manufacturers have developed continuously variable slope delta (CVSD) modulators and demodulators on ICs.

In differential pulse code modulation (DPCM), the difference values are coded before transmission. Fewer bits are required to encode the difference values than are required to encode the absolute values. Instead of computing the difference values in analog domain and encoding them, they can be computed in digital domain using a subtractor as shown in Fig. 5.15. The signal is first PCM coded and the A–D converter output is fed to a subtractor. The other input to the subtractor comes from an adder that acts as the integrator. The difference value is represented in 4-bit quantities that are transmitted on the line. At the receiver end another adder sums up the

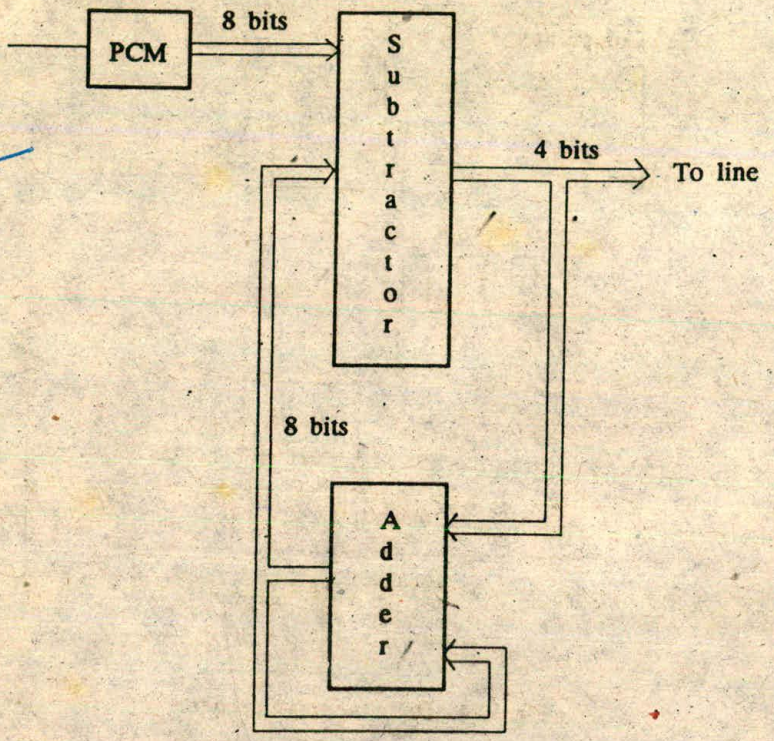


Fig. 5.15 Digital DPCM encoder.

differences to produce the absolute values. Obviously, the bit rate required in this case is 32 kbps as against the 64 kbps required in PCM.

Just as in the case of variable slope delta modulator, the DPCM concept may be extended to let the integrator respond better to larger variations in the input signal by considering several past samples to determine the rate of change of the input. Such an adaptive differential PCM (ADPCM) has been standardised by CCITT for speech digitisation at 32 kbps rate.

5.6 Vocoders

Differential coding schemes described in Section 5.5 use the property of adjacent sample correlations to reduce the bit rates. They are still usable for signals other than speech signals, e.g. waveforms originating from modems. By considering some of the properties that are more or less unique to speech, such as pitch interval and cycle correlations, further significant reductions can be achieved in bit rates. But such coding schemes then become applicable to speech signals only and cannot be used to encode other signals even if they lie in the same frequency band. Coding systems that are so specifically

designed for voice signal are known as **voice coders** or vocoders for short and they operate typically at bit rates in the range 1.2–2.4 kbps.

Vocoders take into account the physiology of the vocal cords, the larynx, the throat, the mouth, the nasal passages and the ear in their design. The basic purpose of the vocoders is to encode only the perceptually important aspects of speech and thereby reduce the bit rate significantly. As a result, the reproduced voice is synthetic sounding and unnatural with artificial quality. The main applications of vocoders include recorded message announcements and encrypted voice transmission (usually in military communications). Voice mail appears to be a potential area for the application of vocoders.

Modelling of speech is at the core of the design of vocoders. Human speech can be considered as being generated in two basic ways. The first category of sounds arises as a result of vibrations in the vocal cords and is known as **voiced sounds**. The second category known as **unvoiced sounds** comprises sounds that are formed by expelling air through lips and teeth as happens when a speaker pronounces letters such as 's', 'p', 't' and 'f'. Human speech can now be modelled as a sequence of voiced and unvoiced sounds passed through a filter which represents the effect of mouth, throat, etc. on the generated sounds. The model that underlies the design of vocoders is depicted in Fig. 5.16. Vocoders use frequency domain techniques to analyse and process redundancies in speech signal. The time domain redundancies described in Section 5.5 have corresponding frequency domain characteristics. For example, the high correlation between adjacent speech samples reflects as nonuniform spectral density in the frequency domain. It is often easier and more efficient to process signals in the frequency domain and most vocoders resort to this.

There are three basic types of vocoders:

1. Channel vocoders
2. Formant vocoders
3. Linear predictive coders.

The speech spectrum exhibits sound-specific structures with energy peaks at some frequencies and energy valleys at others over short periods. Channel vocoders attempt to determine these short term signal spectrums as a function of time and take advantage of them. A bank of bandpass filters is used to separate the speech energy into sub-bands that are full wave rectified and filtered to determine relative power levels. The individual power levels are encoded and transmitted to the destination. In addition to measuring signal spectrum, channel vocoders also determine the nature of speech excitations, i.e. voiced or unvoiced, and the pitch intervals. The excitation information is used at the receiver end to synthesise speech by switching the appropriate signal source, viz. impulse generator or noise source (see

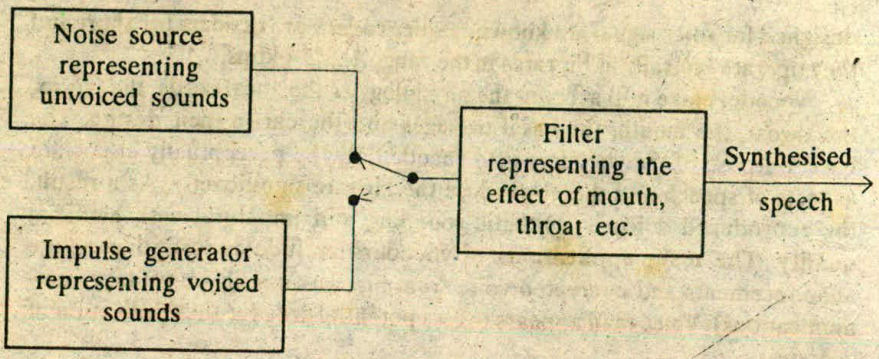


Fig. 5.16 Speech model used in vocoder design.

Fig. 5.16) for the required duration. The filter at the receiver end in Fig. 5.16 implements a vocal tract transfer function. It is constructed by using a bank of bandpass filters whose input power levels are the respective sub-band power levels at the transmitting end. Superposing the individual band outputs along with the appropriate switching of the signal source, generates, in a spectral sense, the original signal. Recent advances in digital signal processing (DSP) permit the use of Fourier transform algorithms to determine the input spectrum in lieu of the bank of analog filters.

The three or four energy peaks in the short-term spectral density of speech are known as formants. A formant vocoder determines the location and amplitude of these spectral peaks and transmits this information instead of the entire spectrum envelope.

In addition to determining the nature of excitation (voiced or unvoiced) and the pitch interval or period, the linear predictive coders (LPC) also predict parameters for vocal tract. The vocal tract filter (see Fig. 5.16) is an adjustable one whose parameters are varied based on the prediction. The LPC functions as a feedback system similar to adaptive DM and ADPCM. As a result, quality of speech synthesised by LPC is superior to that of the speech synthesised by the channel or formant vocoders.

5.7 Pulse Transmission

In the earlier sections of this chapter, we have been considering how analog signal waveforms are represented in the form of binary digital signals or pulses. In this section, we consider aspects relating to the communication of such pulse trains in practical transmission channels.

Any reasonably behaved periodic function $g(t)$ with period T can be represented in the form of a Fourier series:

$$g(t) = C/2 + \sum_{n=1}^{\infty} a_n \sin(2\pi nft) + \sum_{n=1}^{\infty} b_n \cos(2\pi nft) \qquad (5.13)$$

where $f = 1/T$ is the fundamental frequency and a_n and b_n are the sine and cosine amplitudes of the n-th harmonics. Equation (5.13) indicates that if the period T of a waveform and its harmonic amplitudes a_n and b_n are known, the original function of time can be found from the Fourier series. The harmonic amplitudes may be computed in the following way:

1. a_n can be determined by multiplying both sides of Eq. (5.13) by $\cos(k\omega t)$ and then integrating with respect to time over the interval 0 to T. (Note : $\omega = 2\pi f$.)
2. b_n can be determined by multiplying both sides of Eq. (5.13) by $\sin(k\omega t)$ and then integrating over the interval 0 to T.
3. C can be determined by integrating both sides of Eq. (5.13) over the interval 0 to T.

We note that:

(a) $\int_0^T \sin(k\omega t)\cos(n\omega t)\,dt = 0$

(b) $\int_0^T \sin(k\omega t)\sin(n\omega t)\,dt = \begin{cases} T/2 & \text{for } k = n \\ 0 & \text{for } k \neq n \end{cases}$

(c) $\int_0^T \cos(k\omega t)\cos(n\omega t)\,dt = \begin{cases} T & \text{for } k = n = 0 \\ T/2 & \text{for } k = n \neq 0 \\ 0 & \text{for } k \neq n \end{cases}$

(d) $\int_0^T \cos(n\omega t)\,dt = 0$

(e) $\int_0^T \sin(n\omega t)\,dt = 0$

Performing the operations listed in 1–3 above and applying the results given in (a)–(e), we get

$$a_n = \frac{2}{T}\int_0^T g(t)\cos(n\omega t)\,dt \tag{5.14a}$$

$$b_n = \frac{2}{T}\int_0^T g(t)\sin(n\omega t)\,dt \tag{5.14b}$$

$$C = \frac{2}{T}\int_0^T g(t)\,dt \tag{5.14c}$$

The readers are urged to actually derive Eqs. (5.14a)– (5.14c). The root mean square amplitudes of the harmonic components are given by

$$r_n = \sqrt{a_n^2 + b_n^2} \tag{5.15}$$

The r_n values are of interest because their squares are proportional to the energy transmitted at the corresponding frequencies.

A pulse or data train can be considered repetitive and analysed using the Fourier series. Consider a square wave defined by

$$g(t) = \begin{cases} -1 & \text{for } T/2 \le t \le T \\ 1 & \text{for } 0 \le t < T/2 \end{cases}$$

as shown in Fig. 5.17(a). The harmonic amplitudes have the following values:

$$c = \frac{2}{T}\int_0^{T/2} dt + \frac{2}{T}\int_{T/2}^{T} (-1)\,dt = \frac{2}{T}(T/2 - T + T/2) = 0$$

$$a_1 = \frac{2}{T}\int_0^{T/2} \cos \omega t\, dt + \frac{2}{T}\int_{T/2}^{T} -\cos \omega t\, dt$$

$$= \frac{2}{T}\left[\frac{\sin \omega t}{\omega}\right]_0^{T/2} - \frac{2}{T}\left[\frac{\sin \omega t}{\omega}\right]_{T/2}^{T} = 0$$

Similarly, $a_2 = 0$ or, in general, $a_n = 0$.

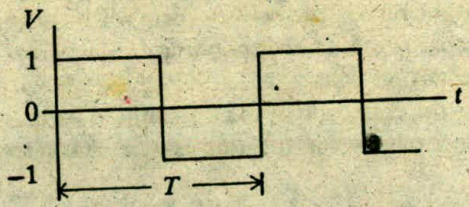
$$b_1 = \frac{2}{T}\int_0^{T/2} \sin \omega t\, dt + \frac{2}{T}\int_{T/2}^{T} -\sin \omega t\, dt$$

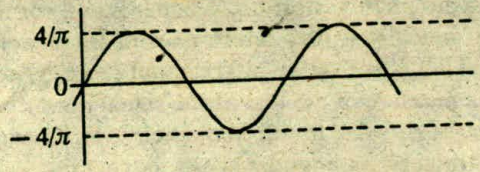
$$= -\frac{2}{T}\left[\frac{\cos \omega t}{\omega}\right]_0^{T/2} + \frac{2}{T}\left[\frac{\cos \omega t}{\omega}\right]_{T/2}^{T} = \frac{2}{T}\frac{T}{2\pi}(4) = \frac{4}{\pi}$$

Similarly, $b_2 = 0$, $b_3 = 4/3\pi$ $b_4 = 0$, $b_5 = 4/5\pi$ or, in general $b_n = 4/n\pi$ if n is odd; $b_n = 0$ if n is even.

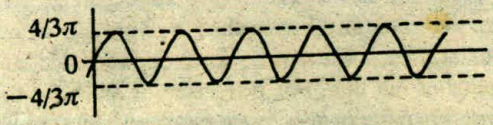
Thus, the square wave spectrum consists of only odd harmonics. Figure 5.17(a) shows the fundamental and the third harmonic components and the sum of the two harmonics using the correct relative amplitudes and phase. While the sum is beginning to show a rectangular shape, perfect square wave can be reconstructed only if all the harmonic components are added together. Hence, transmitting a square wave or a pulse through a transmission medium requires transmission of all the frequency components of the wave for a perfect reconstruction. This implies that a channel must have infinite bandwidth. However, Fig. 5.17(b) shows that the amplitudes decrease exponentially. As a result, it is adequate that the channel has a capability to transmit the fundamental and a few more harmonic components that would permit unambiguous determination of the pulse levels at the receiving end. Most of the pulse trains are not square waves and have a d.c. component. Hence the transmission channel must be capable of transmitting d.c. components as well. Alternatively, techniques may be adopted to remove d.c. components from the waveforms before transmission.

Apart from the adequate bandwidth, the channel must offer **equal attenuation** and **equal delay** for all the frequencies within the bandwidth.

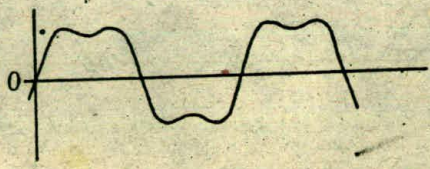




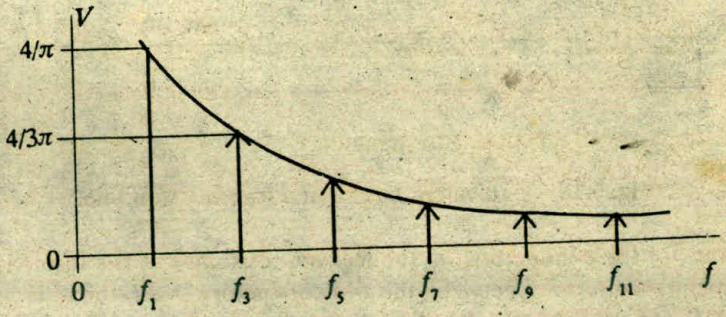
Fundamental $f_1 = \frac{1}{T}$ Hz



Third harmonic $f_3 = \frac{3}{T}$ Hz



(a) Harmonic components and their sum



$f_i = i$-th harmonic

(b) Peak amplitudes of harmonics

Fig. 5.17 Harmonic contents of a square wave.

Otherwise, the harmonics will not combine in proper amplitude and phase at the receiver with the result that the reconstructed waveform will be distorted. Measurable transmission line parameters corresponding to these two preceding requirements are **attenuation constant** A_c and **phase velocity** V_p. Another source of distortion is the reflection of the transmitted wave by the transmission line. The related parameter for this is the **characteristic impedance** Z_0 of the line.

The phase velocity V_p is a measure of the speed of wave propagation along the transmission line in meters per second. Ideally, it should be constant for all frequencies to avoid phase distortion. When it is not, certain frequencies may be delayed so much that they interfere with frequencies corresponding to later pulses. This is known as **intersymbol interference**. The attenuation constant A_c is a measure of the loss of the signal power per meter of the line length. A constant A_c value for all frequencies would avoid amplitude distortion. The characteristic impedance Z_0 is the inherent impedance that is presented to the signal by the transmission line. If the source and load impedances are equal to the characteristic impedance, there is no reflection of the wave from the load, and hence there is no distortion.

In a transmission line, resistance, inductance and capacitance are distributed along the entire length of the line. For purposes of analysis, the distributed parameters are lumped usually over a distance of one meter and an equivalent circuit formed as shown in Fig. 5.18. R_s represents the series

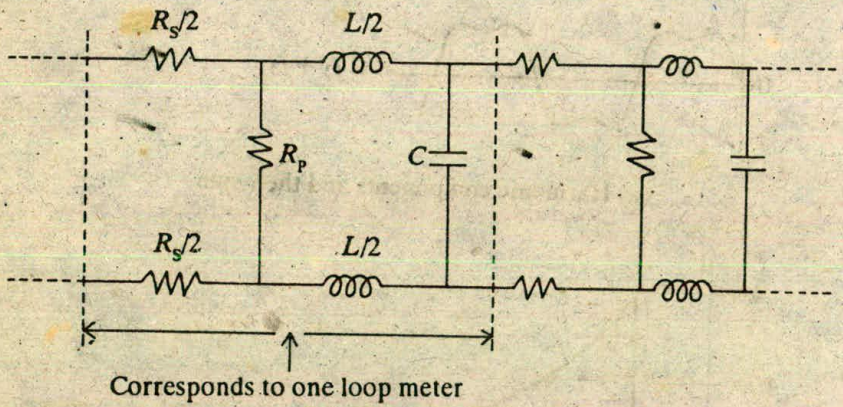


Fig. 5.18 Equivalent circuit of a transmission line.

resistance of the conductors, R_p the leakage resistance of the insulation, C the capacitance formed between the two conductors and L the inductance. R_s and L are values corresponding to one loop-meter and R_p and C are values corresponding to one running meter of the transmission line. Based on the equivalent circuit shown in Fig. 5.18, the general expression for V_p, A_c and Z_0 may be developed. But these are rather complex functions of R_s, R_p, C, L and frequency. However, for practical purposes, we may consider three

separate frequency bands, viz. d.c., low frequencies and high frequencies, that greatly simplify the equivalent circuit and the corresponding expressions for V_p, A_c and Z_0. At d.c. level the inductive reactance is zero and the capacitive reactance is infinite. At low frequencies the series resistive component and the parallel capacitive component dominate and at high frequencies inductive and capacitive reactances play a dominant role. The simplified equivalent circuits are shown in Fig. 5.19. For the purposes of our analysis, we may define the three frequency ranges in terms of specific relative values of reactances as follows:

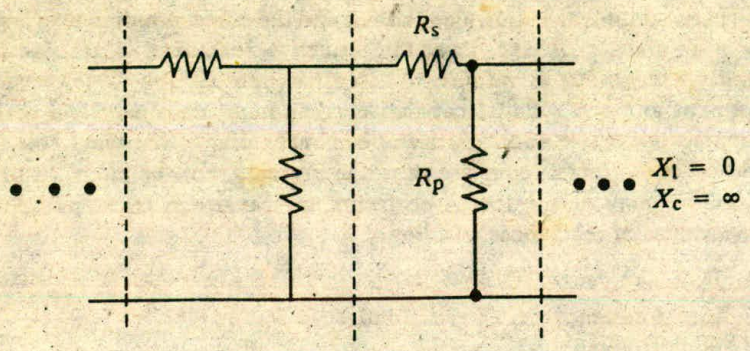
1. D.C. : $X_L = 0, \quad X_c = \infty$
2. Low frequency : $X_L < R_s/10$ and $X_c < R_p/10$
3. High frequency : $X_L > 10\,R_s$ and $X_c \ll R_p$

For these relative values, the expressions for Z_0, A_c and V_p are shown in Table 5.3. It is seen from Table 5.3 that all the three parameters have values

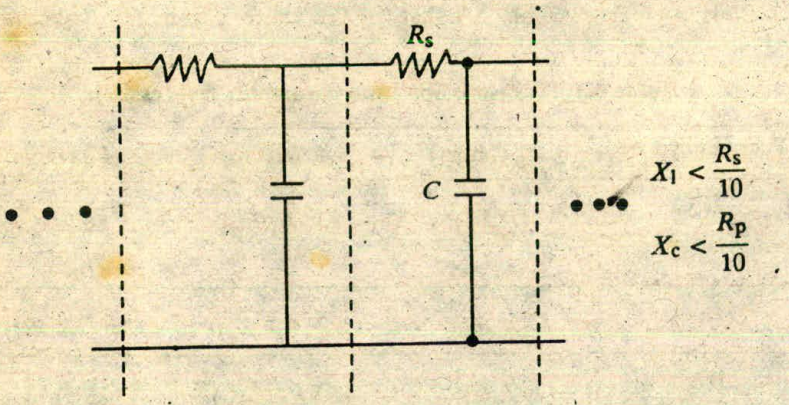
Table 5.3 Transmission Line Parameter Expressions

Frequency band	Z_0 (ohms)	A_c (dB/m)	V_p (m/s)
D.C.	$(R_p R_s)^{\frac{1}{2}}$	$8.7\,(R_s/R_p)^{\frac{1}{2}}$	$2\frac{R_s/R_p}{(L/R_p)^{\frac{1}{2}}} + R_s C$
Low	$(X_c R_s)^{\frac{1}{2}}$	$8.7\,(R_s/2X_c)^{\frac{1}{2}}$	$2\pi f\,(2X_c/R_s)^{\frac{1}{2}}$
High	$(L/C)^{\frac{1}{2}}$	$4.35\,R_s\,(C/L)^{\frac{1}{2}}$	$(LC)^{-\frac{1}{2}}$

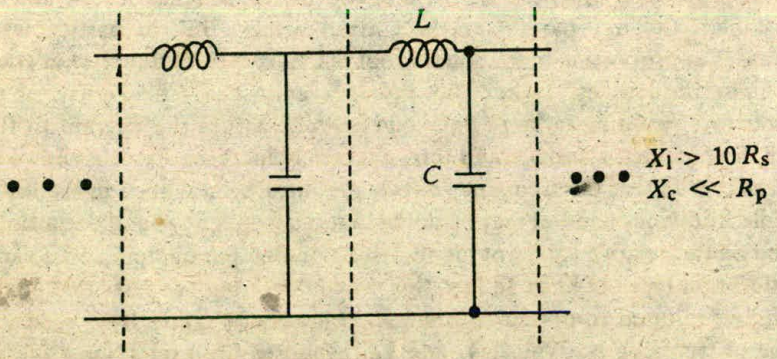
independent of frequencies at the d.c. and high frequency levels, and are constant if the line parameters have fixed values. But, in particular, R_s increases in proportion to the square root of the frequency due to **skin effect**. At high frequencies, higher flux linkages exist near the centre of the conductors resulting in large back e.m.f. which forces the current to flow near the conductor surface and thus effectively increase R_s. It is, however, generally true that the frequency effects are most pronounced in the low to middle frequency band. It may be noted that the frequency band definition is based on the relative values of the transmission line parameters and it varies for different types of lines. In fact, this is an interesting property that can be usefully exploited to obtain constant A_c, V_p and Z_0 values over a desired range of frequencies. Typical A_c and V_p responses for a telephone line are shown in Fig. 5.3. Now, it is possible to obtain uniform A_c and V_p values throughout the voice frequency range by deliberately introducing additional inductance in the network and satisfying the high frequency band condition,



(a) D.C.



(b) Low frequency region



(c) High frequency region

Fig. 5.19 Simplified equivalent circuits of a transmission line.

$X_L > 10\ R_s$, for the entire frequency range. Such additional lumped inductances are known as **loading coils** which are placed at regular intervals. Typical inductance value for the coil is 88 mH and the typical spacing interval is 1.8 km.

Apart from the channel related properties that have been discussed so far, mismatches between the timing pulses at the transmitter and receiver can also cause distortion and intersymbol interference. There are two basic transmission techniques for establishing a time reference between the transmitter and the receiver; **asynchronous** and **synchronous transmissions.**

In asynchronous transmission, two independent clocks, one at the transmitter and the other at the receiver, which have the same nominal frequency depending upon the bit rate of transmission, are used. Since the clocks are independent and are matched only at the nominal frequency values within a certain tolerance, instantaneous rates are likely to be different. When a long stream of bits is transmitted, the difference in frequency and phase of the two clocks may result in erroneous reception. If the clocks are synchronised at the beginning of a transmission, one may ensure correct reception for short streams of bits, before the clock pulses drift far apart to cause erroneous reception. This is the principle adopted in asynchronous transmission which is illustrated in Fig. 5.20. The start bit at the beginning of the transmission synchronises the phase of the receive clock with that of the transmit clock. The incoming data bits are sensed correctly as long as the difference in the phase of T_x and R_x clock frequencies remain within a pulse period. The

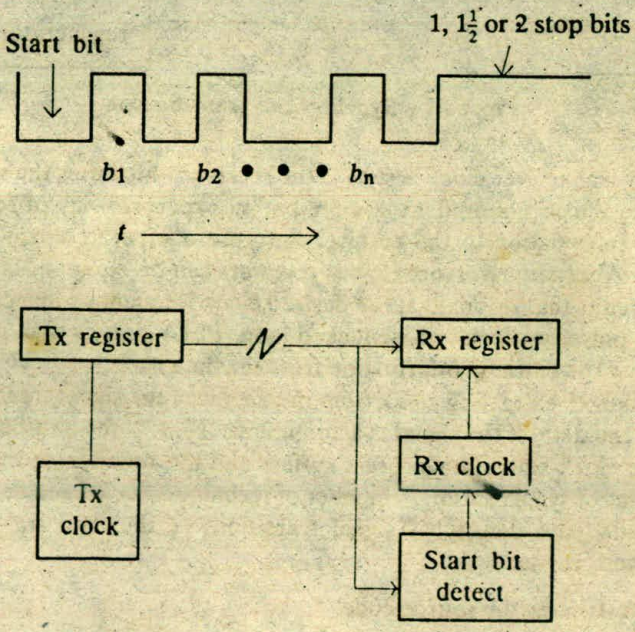


Fig. 5.20 Asynchronous serial transmission.

number of bits b_n that can be transmitted reliably depends on the frequency tolerances of the clocks. The end of transmission is signalled by 1, $1\frac{1}{2}$ or 2 stop bits. The number of stop bits depends on the receive side equipment. The stop bits provide the minimum buffer period required by the receive equipment between two successive transmissions. Electromechanical devices require longer buffer periods than the electronic ones. In asynchronous transmission, which is also known as **start-stop transmission** for obvious reasons, the bit stream length is usually eight bits corresponding to a character. Hence, this transmission is often termed as character mode transmission. Two additional bits, one start and one stop bit for every eight bits represents a 25% transmission overhead. Therefore, it is not economical to transmit large volume data using asynchronous mode.

Synchronous transmission cuts down the transmission overhead by using a single clock for both transmission and reception as illustrated in Fig. 5.21.

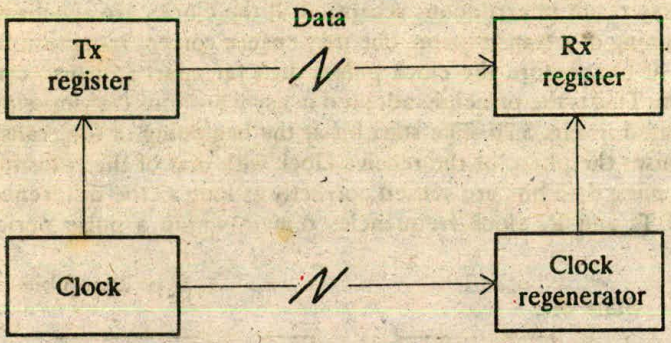


Fig. 5.21 Synchronous transmission.

Here, the transmitter clock is sent to the receiver which uses the same clock to sample the incoming data. Simple but an expensive way of sending the clocking information to the receiver is to use a separate channel for this purpose. Alternatively, synchronous receivers can be equipped with special tuned circuits that resonate at the desired clock frequency when excited by a received pulse or special phase locked loop (PLL) circuits that are capable of deriving the clocking information from the data itself.

The key to extracting clock from the data is in the signal transitions that occur in the data. If the signal is a continuous '1' or '0' for a sufficiently long time, the clock extraction process suffers and the receiver and transmitter may go out of synchronism. A number of techniques have been developed to ensure adequate density of signal transitions in the data transmitted in synchronous mode:

1. Restricting the source code
2. Use of dedicated timing bits

3. Bit stuffing
4. Forcing bit errors
5. Data scrambling
6. Line coding.

Of the above techniques, line coding has some other objectives too and as such is discussed in Section 5.8. We describe briefly the other five techniques in this section.

An n-bit binary code set has 2^n code words. If we now restrict the valid code set to m words, where m is less than 2^n, by eliminating the code words that do not have adequate signal transitions, we can ensure that the receive clock remains in synchronism with the transmit clock. The main drawback of the technique is that new applications wher the source does not exclude the unwanted data patterns cannot be supported on the network.

Just as the start bit in asynchronous transmission, transition bearing timing bits can be introduced at regular intervals in synchronous transmission. The interval is so chosen as to ensure adequate number of signal transitions per second required for reliable clock synchronisation. It may, however, be noted that these timing bits add to the transmission overhead.

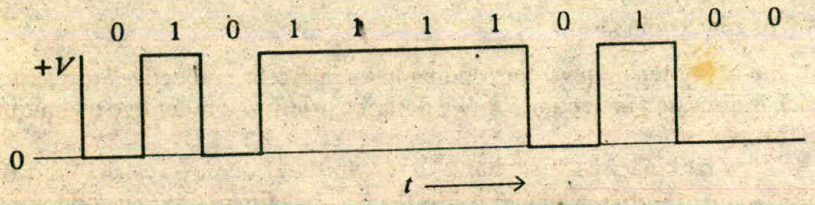
Bit stuffing is a technique where an extraneous bit is introduced in the data stream whenever there are no signal transitions in the data for a continuous period of given duration. For example, a signal transition bit may be introduced in the data whenever a series of seven ones or zeros is observed on the data line. Bit stuffing breaks the continuous flow of data and delays the data stream by one bit, every time a bit is inserted. In real-time services like voice, this causes jitter, thereby affecting the quality of service. In such services, the quality is less affected by errors than delays. Considering this aspect, another technique introduces deliberate errors in the data streams to bring about signal transitions.

Data scramblers randomize the data patterns and prevent the transmission of repetitive data patterns on the line continuously. Although designed to reduce the strong interference that the repetitive data patterns are capable of, data scramblers do produce adequate timing information in the process which can be usefully exploited for synchronisation.

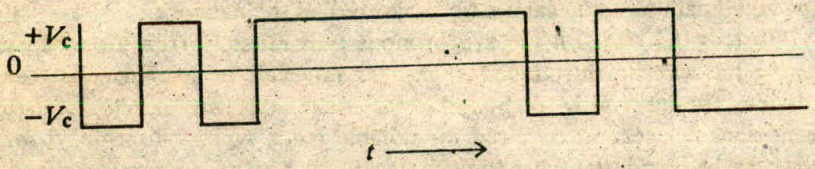
5.8 Line Coding

Binary digital information is usually represented by either unipolar or bipolar codes using either positive or negative logic systems. In unipolar codes, a nonzero voltage or current level which may either be positive or negative, and a zero level are used to represent the binary ones and zeros. In bipolar codes, nonzero negative and positive levels are used to represent zeros and ones. If the positive logic system is adopted, the relatively positive level in the code is used to represent the binary zero and vice versa in the

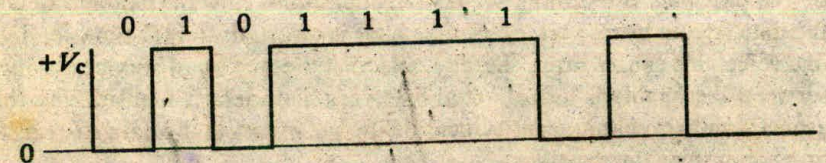
negative logic system. Positive logic unipolar and bipolar codes for an arbitrary binary sequence are shown in Fig. 5.22. The unipolar and bipolar



(a) Unipolar NRZ code



(b) Bipolar NRZ code



(c) Unipolar NRZ and RZ code

Fig. 5.22 NRZ and RZ codes.

codes are sometimes known as unbalanced and balanced codes respectively. Bipolar codes are more efficient from the point of view of power consumption. As an illustration, the average power required for equally likely +5 V and 0 V unipolar coding at 1 Ω resistance is 12.5 W, whereas for the same level distance a bipolar code with 2.5 V requires only 6.25 W.

In the unipolar and bipolar codes shown in Fig. 5.22(a) and (b), the voltage levels are maintained for the entire duration of the signal. Hence, these codes are known as nonreturn-to-zero (NRZ) codes. In return-to-zero (RZ) codes, there is a voltage transition for every '1' state of the signal as shown in Fig. 5.22(c). Therefore, the RZ codes are somewhat better suited than NRZ codes for clock synchronisation. However, long strings of zeros

have no transitions in RZ coding and one of the techniques described in Section 5.7 can be used to bring about transitions when continuous zeros are encountered. The codes shown in Fig. 5.22 have a nonzero d.c. or average level which varies depending on the ratio of ones and zeros in the data pattern. Most transmission lines do not pass d.c. signals as they are a.c. coupled using transformers or capacitors to eliminate d.c. ground loops. Some systems purposely remove d.c. components to allow line powering of repeaters or to facilitate single sideband transmission. The elimination of the low frequency components by the transmission system results in gradual decay of amplitudes of NRZ waveforms as shown in Fig. 5.23. This phenomenon, known as **d.c. wander**, causes amplitude reference to be lost for optimally discriminating between a one level and a zero level. For example, in Fig. 5.23, '1' levels are likely to be interpreted as '0' levels after the d.c. wander. It is possible to nullify the effect of d.c. wander by using techniques that restore the d.c. value to zero level after every pulse. Alternatively, line codes can be designed such that they do not contain any d.c. energy.

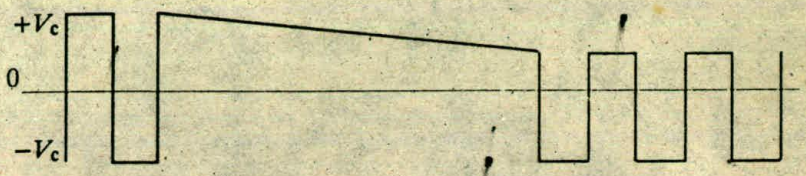
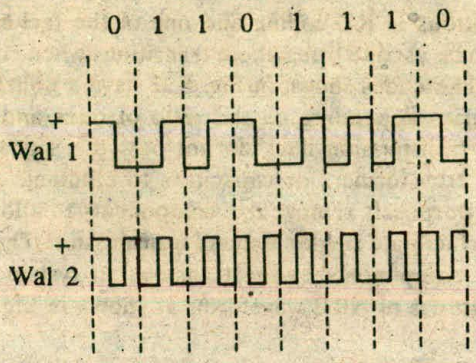


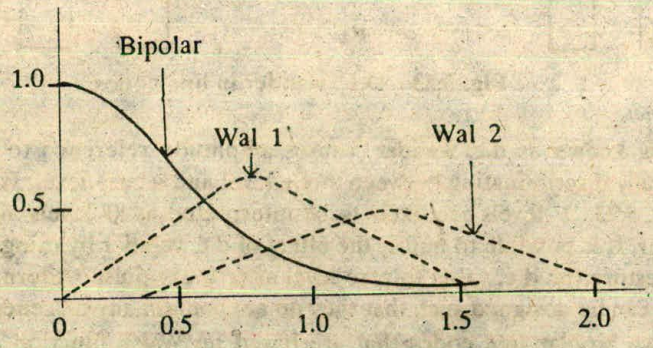
Fig. 5.23 D.C. wander in line codes.

Two bipolar line codes that are based on Walsh function and do not contain any d.c. component are shown in Fig. 5.24(a). The codes contain a large number of transitions from which timing information can be recovered easily. The power spectra of the two codes are shown in Fig. 5.24(b) where they are compared with the spectrum for the simple bipolar binary code. While there is a significant d.c. component in the simple code, both the Walsh codes have no d.c. component at all. The low frequency contents are also very small. The Walsh Code 2 occupies a slightly larger bandwidth than the Code 1 but is better with regard to low frequency components. The Walsh Code 1 is also known as **Manchester coding**.

Another class of line codes are **ternary codes** which operate on three signal levels, usually positive, zero and negative. These codes are sometimes referred to as pseudo-ternary because of the fact that the coding is effectively binary although three levels are used. The best known of the pseudo-ternary codes is the **alternate mark inversion (AMI)** code. In AMI code, a logical zero (space) is encoded with zero voltage while a logical one (mark) is alternatively encoded with positive and negative voltages as shown in Fig. 5.25(a). The first useful property of such coding is that there are no d.c.



(a) Coded pattern

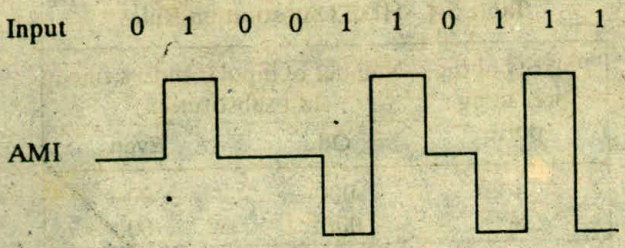


Frequency × Period

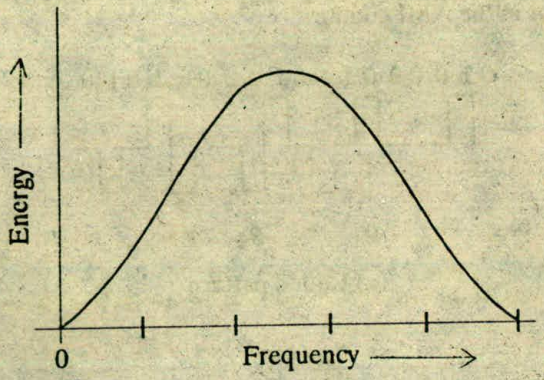
(b) Line spectra

Fig. 5.24 Walsh codes.

components in the spectrum. The low frequency component is also insignificant as seen from Fig. 5.25(b). The code has some inherent error detection features. Errors result in AMI rule violation, which can be recognised at the receiving end. Deliberate violation of AMI rule can be made to send some control information from the transmitter to the receiver. AMI coding resolves two important problems encountered by many other bipolar codes. Firstly, there is no d.c. component in the transmitted signal and hence the phenomenon of d.c. wander is absent. Secondly, excellent timing information is available for continuous ones along with error detection capability. However, the code fails to provide timing information when a long series of zeros are transmitted.



(a) AMI coding



(b) AMI spectrum

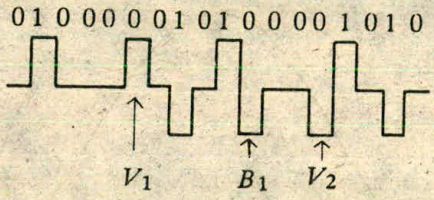
Fig. 5.25 AMI coding and spectrum.

A modification to AMI coding known as **binary *N* zero substitution** breaks up long runs of zeros (*N* zeros) by substituting a group of pulses that violate the AMI alternating pulse polarity rule. A widely used code in this category is the HDB3 (high density binary) code recommended by CCITT. In HDB3, each string of four zeros in the source data is encoded with either 000*V* or *B*00*V* pattern, where *V* stands for a pulse that violates AMI rule and *B* stands for a pulse that respects AMI rule. There is a violation pulse at the last zero digit in both the patterns, which enables identification of the encoding at the receiving end. In order to discriminate between a violation due to channel error and a violation due to HDB3 coding, the number of bipolar transitions is kept odd between two violations. The violation pulses alternate in polarity to avoid d.c. wander. One of the two patterns, 000*V* or *B*00V, is chosen to ensure odd number of bipolar pulses between two purposeful violations. This is illustrated in Table 5.4. Figure 5.26 shows HDB3 coding for a given data pattern and the spectrum of the code. It is seen from Fig. 5.26(a) that the first group of four zeros is coded with 000*V* pattern and the second group with *B*00*V* pattern to ensure odd number of bipolar

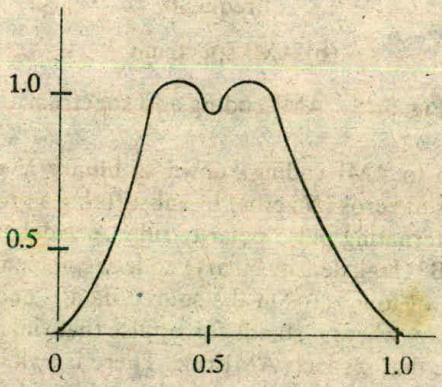
Table 5.4 HDB3 Substitution Rules

Polarity of the preceding pulse	Number of bipolar pulses since last substitution	
	Odd	Even
−	000−	+00+
+	000+	−00−

pulses between the two violation pulses. It may be noted that the bipolar pulse, i.e. *B* pulse in the *B00V* pattern, is considered as a mark pulse and as a result the mark pulses subsequent to the violation pulse V_2 have a polarity opposite to that in the AMI coding.



(a) Coded pattern



(b) Line spectrum

Fig. 5.26 HDB3 code.

Another class of ternary coding is known as alphabetic coding. In this coding, a signal element is formed by taking n binary digits together and coded into an m-digit ternary character. Any binary signal element may be considered as an alphabet from a set of 2^n characters and a ternary signal

element as an alphabet from a set of 3^m characters. Obviously, 3^m should be greater than 2^n. Such codes are normally described as nBmT (n binary and m ternary) codes.

The simplest of nBmT codes occurs when $n = m = 2$, and is known by a special name as pair selected ternary (PST). The message signal elements and the corresponding ternary codings are given in Table 5.5. Since there are nine 2-digit ternary codes but only four 2-digit binary codes, there is

Table 5.5 Pair Selected Ternary Coding

Message signal	Mode *A*	Mode *B*
00	− +	− +
01	0 +	0 −
10	+ 0	− 0
11	+ −	+ −

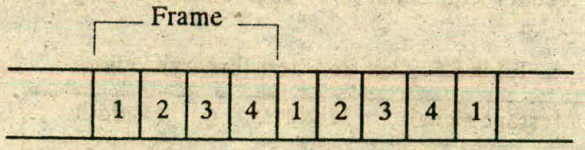
considerable flexibility available in selecting the codes, and hence two modes of coding are defined. A balance between positive and negative pulses and a strong timing component are maintained by switching between the modes appropriately. As the ternary digits are transmitted, a modulo–1 sum of +ve and – ve pulses is kept. If the sum is zero, the mode remains unchanged. If the sum is +ve, i.e. if a single positive pulse has been transmitted, mode *B* is selected, and if the sum is – ve, mode *A* is selected.

The efficiency of transmission can be improved by the use of alphabetic codes, where $m < n$. A widely used code of this class is 4B3T, where 4-bit binary words are coded into 3-digit ternary words. Since there are only 16 binary combinations but 27 ternary combinations, there is still considerable flexibility to define two modes and alternate between modes. Readers are advised to form a table of code translation for 4B3T coding similar to Table 5.5 given for PST coding.

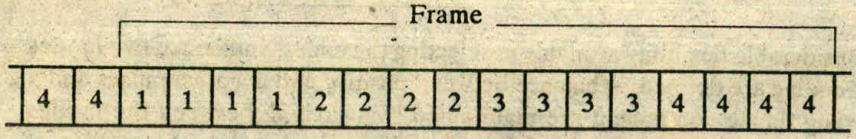
5.9 Time Division Multiplexing

Time division multiplexing (TDM) is by far the most commonly used means of subdividing the capacity of a digital transmission facility among a number of sources. There are two basic modes of operation for TDM: one that repeatedly assigns a portion of the transmission capacity to each source and the other that assigns capacity as and when needed. The former mode is known as synchronous TDM (STDM) or simply TDM and the latter as asynchronous TDM (ATDM) or statistical TDM or statistical multiplexing (Stat-Mux).

In TDM, capacity allocation may be done either bitwise or wordwise. In bitwise allocation, each source is assigned a time slot corresponding to a single bit and in wordwise allocation, a time slot corresponds to some larger number of bits (often 8 bits) referred to as a word. The two allocation techniques give rise to two different TDM frame structures as shown in Fig. 5.27. In one case, the frame size is equal to the number of channels



(a) Bit interleaving



(b) 4-bit word interleaving

Fig. 5.27 Four-channel TDM frame structures.

multiplexed and in the other it is equal to the product of the word size and the number of channels. Bitwise interleaving is natural if delta modulator coders/decoders (codecs) are used and wordwise interleaving is natural if PCM codecs are used for digitisation. It may be recalled that in delta modulation, the coder produces a bit-by-bit output representing the slope variation whereas in PCM, sampled value is output as an 8-bit quantity. The bit rate of a TDM stream should be equal to or greater than the sum of the bit rates of the individual channels.

Having formed a frame by bit interleaving or word interleaving, it is necessary to have a mechanism to synchronise the frames at the transmitting and the receiving end, i.e. the start of the frame must unambiguously be identified at the receiving end. Frame synchronisation may be done in a number of ways. In all the ways, there are one or more framing bits with an identifiable data sequence. Frame bits may be added additionally to data bits or some of the data bits may be used as frame bits. For example, in Bell Systems T1 channel frame, an additional bit is introduced for every frame which alternates in value. The T1 channel structure multiplexes 24 channels and hence the frame has a length of 193 bits ($24 \times 8 + 1$). In one of the early

TDM channel structures, British Telecom (then the British Post Office) used 7-bit PCM samples with one bit per octet for framing and signalling. Thus the data rate for 24-channel TDM works out to be $24 \times 64 \times 10^3 = 1.536$ Mbps including framing and signalling bits.

At the receiving end, frame synchronisation is achieved by searching for the frame synchronisation pattern. In the case of T1 channels, an alternating bit pattern occurring after every 193 bits is searched for. Framing is established by examining bit after bit successively until a sufficiently long framing pattern is detected. Average frame acquisition time F_t is an important factor in designing frame synchronisation schemes. We now estimate F_t for the scheme which uses one framing bit per frame with alternating ones and zeros. F_t has two components, F_t' and F_t'' defined as follows:

F_t' = average time required examine the required number of bits before a framing bit is hit upon,

F_t'' = average time required test and ensure that chosen bit is a frame bit, and

$$F_t = F_t' \times F_t'' \tag{5.16}$$

Equation (5.16) implies that the bits are chosen one by one and tested one after another sequentially. It is possible to examine bits in parallel. We, however, confine our discussion to sequential search. Let

p = probability of finding a 1 in a bit position

q = probability of finding a 0 in a bit position

We then have

$$p = 1 - q$$

If the first bit observed is a 1, then the probability that a mismatch occurs at the end of

one frame = p

two frames = $q \cdot q = q^2$

three frames = $q \cdot p \cdot p = qp^2$

four frames = $q \cdot p \cdot q \cdot q = pq^3$

five frames = $q \cdot p \cdot q \cdot p \cdot p = p^3q^2$

six frames = $q \cdot p \cdot q \cdot p \cdot q \cdot q = p^2q^4$

and so on. Therefore, the average time required to ensure that a chosen bit is not an information bit is given by the following in terms of frame times:

$$F_t'' = p + 2q^2 + 3qp^2 + 4pq^3 + 5p^3q^4 + 6p^2q^4 \tag{5.17}$$

In order to simplify the arithmetic, let us assume $p = q$. Then Eq. (5.17) becomes

$$F_t'' = p + 2p^2 + 3p^3 + 4p^4 + \ldots$$

$$= p[(1 + p + p^2 + \ldots) + (p + 2p^2 + 3p^3 + \ldots)]$$

$$= \frac{p}{1-p} + p^2[(1 + p + p^2 + \ldots) + (p + 2p^2 + 3p^3 + \ldots)]$$

$$= \frac{p}{1-p} + \frac{p^2}{1-p} + \frac{p^3}{1-p} + \ldots = \frac{p}{1-p}(1 + p + p^2 + \ldots)$$

$$= \frac{p}{(1-p)^2} \quad \text{frame times} \tag{5.18}$$

Setting $p = \frac{1}{2}$, we have

$$F_t'' = 2 \text{ frame times} = 2N \text{ bit times} \tag{5.19}$$

where N = the number of bits per frame.

If the search is started randomly, the average number of bits that must be tested before the framing bit is encountered is $N/2$, i.e. $F_t' = (N/2)$. Therefore,

$$F_t = (2N)(N/2) = N^2 \text{ bit times} \tag{5.20}$$

Adding extra bits to the data stream may interfere with periodic sampling process by introducing aperiodicities in the process, which in turn may result in sample time jitter. On the other hand, use of data bits for framing does not interfere with the periodic transmission rate, but amounts to tampering with data quality. More recent techniques use dedicated separate channels for framing which affect neither the periodicity of sampling nor the data quality. Such techniques are known as **added-channel framing**. Here, framing digits are added in a group such that an extra channel is formed. The scheme adds considerable performance and flexibility to the framing process. An example of added-channel framing is the CCITT multiplexing standard which recommends 32 channels per frame with one channel carrying framing information, one channel signalling information and the remaining 30 channels data. The average frame acquisition time F_t for multibit frame code grouped as a separate channel can be shown to be

$$F_t = \frac{N^2}{2(2^L - 1)} + \frac{N}{2} \text{ bits} \tag{5.21}$$

where

N = length of the frame
L = length of the frame code.

Equation (5.21) assumes that 1's and 0's occur with equal probability in the frame.

Table 5.6 CCITT TDM Hierarchy

Level	Transmission rate (Mbps)	Number of PCM voice channels
1	2.048	30
2	8.448	120
3	34.368	480
4	139.264*	1920
5	565.0	7680

* Nominally 140 Mbps.

More details on frame synchronisation such as digit assignment etc. are presented in Section 9.8.

Channel multiplexing as in the case of 32-channel CCITT frame, is only the first level of multiplexing. Groups of channels may also be multiplexed so that high quality digital links may be used over long distances to carry data at much higher rates. Accordingly, a hierarchy of digital multiplexing is used by various agencies for carrying digital information. Different standards exist for the hierarchical levels and the CCITT standard is presented in Table 5.6. At each level of the hierarchy, the number of channels multiplexed is multiplied by four. Each level of multiplexing requires additional digits to indicate framing, control and signalling information. As a result, the data rate increases by a factor greater than four. The rates are chosen to be multiples of 64 kbps, so that the standard octet structure can be maintained throughout the hierarchy.

FURTHER READING

1. Bates, P., *Practical Digital and Data Communications*, Reston, Virginia, 1987.
2. Bellamy, J. C., *Digital Telephony*, 2nd ed., John Wiley & Sons, New York, 1991.
3. Brewster, R .L., *Telecommunications Technology*, Affiliated East-West Press, New Delhi, 1988.
4. Jayant, N. S., Digital coding of speech waveforms: PCM, DPCM, and DM quantisers: *Proceedings of IEEE*, 1974, pp 611–632.
5. Taub, H. and Schilling, D. L., *Principles of Communication Systems*, McGraw-Hill, New York, 1986.

EXERCISES

1. The threshold of hearing or the reference level for the minimum discernible sound is internationally agreed upon as 10^{-12} W/m^2. The threshold of pain is 130 dB above this threshold of hearing. What is the actual power level at the t. reshold of pain?

2. The human ear has logarithmic response characteristics. How would a person perceive the change in loudness if he moves from a quiet residence which has an intensity level of 30 dB to an office with a level of 60 dB and an airport with a level of 120 dB?

3. List and discuss at least five advantages of digital transmission of speech over analog transmission.

4. What is the minimum sampling frequency required for a signal with a frequency range of d.c. –15 kHz if it is (a) band limited between 1 kHz and 10 kHz, (b) passed through a low pass filter which has a cut off frequency of 5 kHz.

5. The tone generated by push button 3 of a DTMF telephone is sampled at 1.6 kHz and passed through a PAM decoder which has a 'cut off frequency of 800 Hz. What frequencies will be present at the output of the PAM decoder?

6. How much does the SQR of a uniform PCM encoder improve when one bit is added to the code word?

7. If a minimum SQR of 33 dB is desired, how many bits per code word are required in a linearly quantised PCM system?

8. A 7-bit uniform PCM system has a bit rate of 56 kbps. Calculate the SQR when the input is a sine wave covering the full dynamic range of the system. Calculate the dynamic range of the sine wave input if the SQR is to be at least 30 dB.

9. If a linear PCM system is to have a minimum SQR as in the first segment of A-law companded PCM, calculate the number of bits required per sample. Assume a dynamic range equal to that obtained between the first and the eighth A-law segments.

10. What is the dynamic range provided by a uniform PCM encoder with 12-bit code words and a minimum SQR of 40 dB?

11. A PCM encoder using A-law has a dynamic range of ±2 V. Determine the code word when the input signal is (a) –120 mV, (b) 0.5 V, (c) –1.5 V, (d) 1.8 V.

12. For Problem 11, determine the input signal value if the code word is (a) 0110 1001, and (b) 1001 1011.

13. A piecewise linear PCM system is to be designed for positive inputs with four segments and four equally spaced intervals in each segment. The slope halves in second, third and fourth segments when compared to first, second and third segments respectively. Construct a coding table for this design.

14. Given two A-law code words 11001100 and 01100111 determine their linear representations, add them and convert back to compressed representation.

15. Determine the number of bits per sample that can be saved in a linear DPCM system compared to a linear PCM system when a 400-Hz signal is coded in both schemes.

16. Explain the terms granular noise and slope overload in the context of differential coding schemes.

17. A delta modulator is designed to avoid slope overload for an input signal voltage range of ± 5 V and a maximum signal frequency of 20 kHz. Compare the required bandwidth with that of an 8-bit PCM system for the same signal.

18. A 4-bit DPCM system is used to transmit signals of amplitude range ± 1V and maximum frequency component 4 kHz. If slope overload is to be avoided for intersample changes of less than 0.2 V, what must be the sampling frequency and the transmission line bandwidth?

19. A channel vocoder uses a bank of 16 low pass filters of 20 Hz bandwidth along with the fullwave rectifiers to determine the relative power levels of the sub-bands of speech energy. If the outputs of the filters are sampled at the Nyquist rate and encoded as 3-bit samples, determine the bit rate of speech transmission.

20. For the ASCII character 'b', determine the root mean square amplitudes of the first four harmonics.

21. A series of rectangular pulses of 100 μs duration are transmitted at intervals of 1 ms. Determine the relative amplitudes of the first four harmonic components. What pulse duration would be required to make the sine and cosine components of the fourth harmonic equal?

22. A 22-gauge twisted pair wire has the following line parameters:

    $L = 2\ \mu$H/m
    $C = 50$ pF/m
    $R_s = 0.1\ \Omega$/ m
    $R_p = 800\ \text{M}\Omega$ /m

    (a) Determine the boundaries for low and high frequency regions as defined in Section 5.7.
    (b) Determine the phase velocity at 400 Hz.

23. What is the characteristic impedance of the 22-gauge wire specified in Exercise 22, at 100 Hz and 1 MHz?

24. Sketch the serial asynchronous link waveform for the ASCII character 'C'. Assume even parity and two stop bits.

25. For the binary data pattern 001101111011, sketch the coded waveform when the following coding schemes are used:

    (a) Conventional bipolar
    (b) Manchester Coding
    (c) AMI Coding
    (d) Pair selected binary
    (e) 4B3T.

26. A TDM system operating at 2 Mbps is to have a maximum frame synchronisation time of 20 ms. What is the maximum possible frame length if one synchronisation bit per frame is used with alternating values?

6

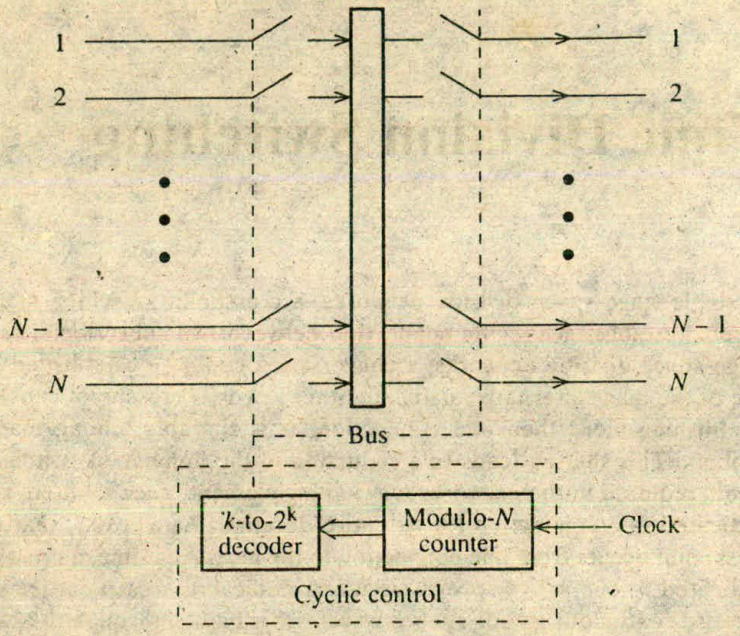
Time Division Switching

In single-stage space division networks, a crosspoint switching element is used to establish a specific connection between two subscribers. In multistage space division networks, a conversation could be established via any one of the many alternative paths. In other words, a crosspoint is usable for establishing more than one connection. It is sharable among many subscribers. This sharing leads to a reduction in the number of switching elements required in the network. A switching element, once allotted, remains dedicated to a connection for its entire duration. As a result, sharing of a crosspoint occurs from one connection to the next. At a time, a crosspoint is dedicated to one active speech circuit. A dedicated crosspoint is required because continuous analog speech waveform is passed through the switch in space division switching.

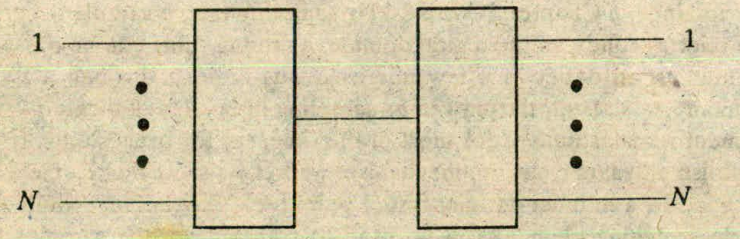
In digital transmission, sampled values of speech are sent as pulse amplitude modulated (PAM) values or pulse code modulated (PCM) binary words as explained in Chapter 5. With 8-kHz sampling rate, a sample occurs every 125 microseconds. In the digital domain, a sample value can be passed from an inlet to an outlet in a few microseconds or less, through a switching element. As a result, during 125-μs sampling interval, a dedicated switching element remains unused for most of the time, say for over 120 μs. If we can establish a dynamic control mechanism whereby a switching element can be assigned to a number of inlet-outlet pairs for a few microseconds each, a single switching element can be used to transmit speech samples from a number of inlets to the corresponding outlets. In other words, a switching element can be shared by a number of simultaneously active speech circuits. This is the principle of time division switching. Obviously, with the way the switching elements are shared in time division switching, much greater savings can be achieved in the number of switching elements when compared to multistage space division switching.

6.1 Basic Time Division Space Switching

A simple $N \times N$ time division space switch is shown in Fig. 6.1(a). The switch can be represented in an equivalent form as a two-stage network with $N \times 1$



(a) Switching structure



(b) Two-stage equivalent.

Fig. 6.1 Simple PAM time division switching.

and $1 \times N$ switching matrices for the first and second stages respectively as shown in Fig. 6.1(b). The network has one link interconnecting the two stages. Each inlet/outlet is a single speech circuit corresponding to a subscriber line. The speech is carried as PAM analog samples or PCM digital samples, occurring at 125-μs intervals. When PAM samples are switched in a time division manner, the switching is known as **analog time division switching**. If PCM binary samples are switched, then the switching is known as **digital time division switching**. In Fig. 6.1(a), the interconnecting link is shown as a bus to which a chosen inlet-outlet pair can be connected by a suitable control mechanism and the speech sample transferred from the inlet

to the outlet. When PAM samples are switched, the bus is analog in nature and if the binary samples are switched, a digital bus is used. The number of simultaneous conversations, SC, that can be supported on a network of this kind is inversely proportional to the time required to select and connect an inlet-outlet pair and transfer the speech sample from the inlet to the outlet and is given by

$$SC = \frac{125}{t_s} \tag{6.1}$$

where t_s is the time in microseconds to set up a connection and transfer the sample value.

The selection of inlet/outlet is controlled dynamically. The simplest form of control is to select the inlets and outlets in a cyclic manner in synchronism. In this case, there is a fixed one-to-one correspondence between the inlets and the outlets. Suppose a cycle starts at inlet/outlet 1 and proceeds as 2, 3, 4, ..., N, the next cycle as 1, 2, 3,..., N and so on, then inlet i is always connected to outlet i. In effect, there is no switching in the true sense except that a single switching element is being shared by all the connections. The cyclic control is organised using a modulo-N counter and a k-to-2^k decoder as shown in Fig. 6.1(a). N and k are related by the equation

$$\lceil \log_2 N \rceil = k \tag{6.1a}$$

where the symbol $\lceil\ \rceil$ is a ceiling function that gives the lowest integer equal to or higher than the quantity inside the symbol. Since all the inlets/outlets are scanned within 125 μs, the switching capacity, SC, of the network is the same as the number of inlets or outlets in the system and the switch is nonblocking. But it lacks full availability as it is not possible to connect any inlet to any outlet.

If we make one of the controls, say on the output side, memory based, then full availability can be obtained. Such a scheme is shown in Fig. 6.2. The input side is cyclically switched. There is a control memory on the output side which contains the addresses of the outlets stored in contiguous locations in the order in which they are to be connected to the inlets. For example, an address sequence 4–7–1–5 stored in locations 1, 2, 3 and 4 of the control memory (see Fig. 6.2) implies that inlet 1 is connected to outlet 4, inlet 2 to outlet 7, inlet 3 to outlet 1 and so on in that order. This switch is said to be **input-controlled** or **input-driven** as the outlet is chosen depending on the inlet that is being scanned at any instant.

The modulo-N counter of the cyclic control also acts as the memory address register (MAR) of the control memory. Control memory has N words corresponding to N inlets and has a width of $\lceil \log_2 N \rceil$ bits which are used to address the N outlets. Cyclic control at the input implies that all the subscriber lines are scanned irrespective of whether they are active or not. For an active inlet i, the corresponding outlet address is contained in the i-th location of the control memory. It is read out and passed to the address

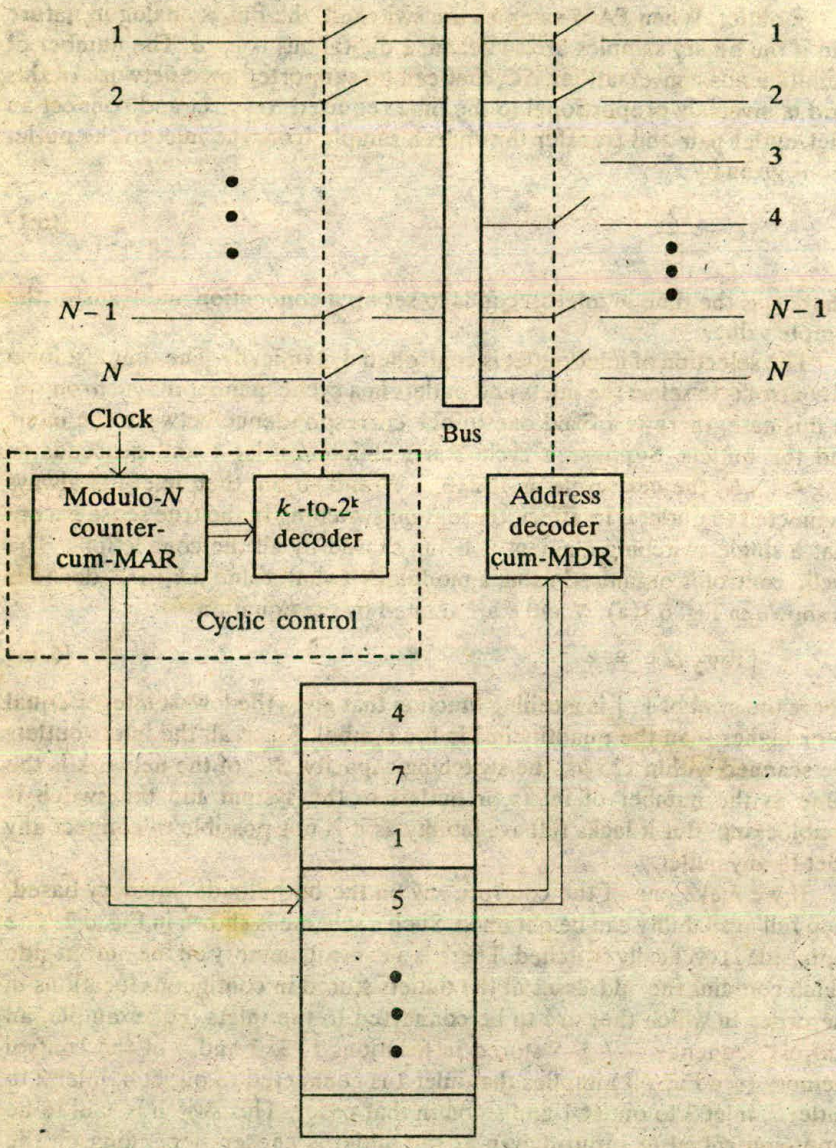


Fig. 6.2 Input-controlled time division space switch.

decoder which also acts as the memory data register (MDR) of the control memory. The decoder output enables the proper outlet to be connected to the bus. The sample value is then transferred from the inlet to the outlet. Thus, any inlet i can be connected to any outlet k ensuring full availability. In case an inlet is not active, the corresponding location in the control memory contains a null value which forbids the address decoder from enabling any of

the output lines. Since a single switching element, the bus, is being time shared by N connections, all of which can be active simultaneously, and a physical connection is established between the inlet and the outlet for the duration of the sample transfer, the switching technique is known as **time division space switching.**

It is possible to organise a fully-available time division space switch with cyclic control for the outlets and control memory based selection for the inlets as shown in Fig. 6.3. In such a case, the switch is said to be **output-**

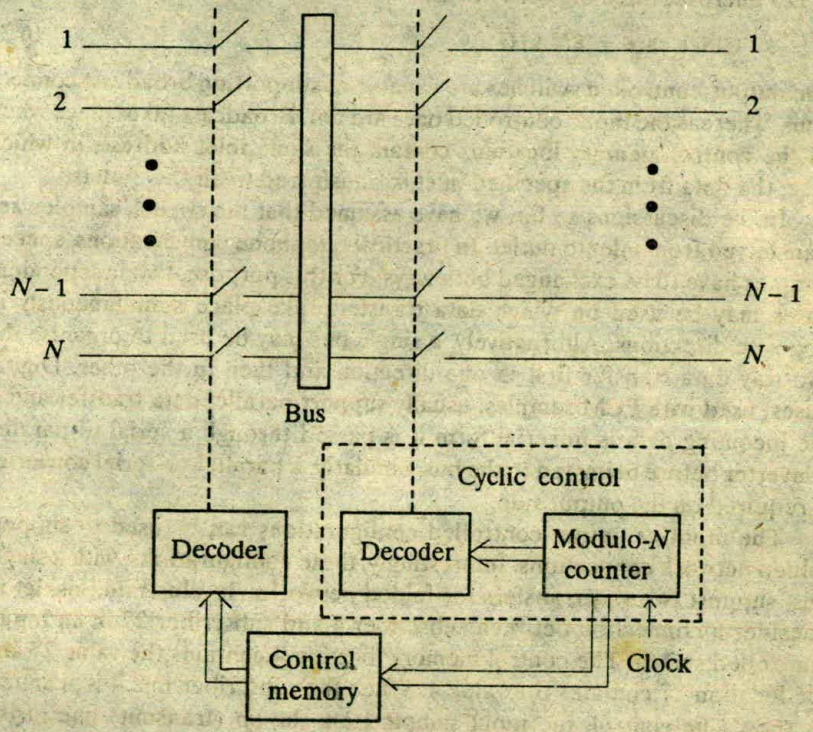


Fig. 6.3 Output-controlled time division space switch.

controlled because each location of the control memory is rigidly associated with a given outlet. For both input and output-controlled configurations, the number of inlets or outlets N, which is equal to the switching capacity, is given by

$$N = SC = \frac{125}{t_i + t_m + t_d + t_t} \tag{6.2}$$

where

t_i = time to increment the modulo-N counter

t_m = time to read the control memory

t_d = time to decode address and select the inlet or outlet as the case may be

t_t = time to transfer the sample value from inlet to outlet

All time values are expressed in microseconds. Equations (6.1) and (6.2) are valid for a 8-kHz sampling rate and a nonfolded network. The clock input to the modulo-N counter has a rate such that N sample transfers are organised in 125 microseconds. In other words,

$$\text{Clock rate} = 8N \text{ kHz} \tag{6.3}$$

The output-controlled switches are capable of supporting broadcast connections, whereas the input-controlled ones are not. Broadcast takes place when all the control memory locations contain the same inlet address, in which case, the data from the specified inlet is transferred to all the outlets.

In the discussions so far, we have assumed that the speech samples are transferred from inlet to outlet. In practical telephone conversations, speech samples have to be exchanged both ways. For this purpose, two independent buses may be used on which data transfers take place simultaneously in opposite directions. Alternatively, a single bus may be used to organise the two-way data transfer first in one direction and then in the other. Digital buses, used with PCM samples, usually support parallel data transfer and if the incoming data is in serial form it is passed through a serial-to-parallel converter before being fed to the bus. Similarly, a parallel-to-serial converter is required on the output side.

The input or output-controlled configurations can be used to support folded network connections. Interestingly, these configurations, with a single bus, support two way transfers for folded networks. To illustrate this, let us consider a connection between subscriber 4 and subscriber 27 in an input-controlled switch. The control memory location 4 contains the value 27 and the location 27 contains the value 4. When the subscriber line 4 is scanned, by the cyclic control, the input sample from the *tip* (transmit) line of the subscriber 4 is transferred to the *ring* (receive) line of the subscriber 27. When the subscriber line 27 is scanned, the input from the line 27 is transferred to the *ring* line of the subscriber 4. Thus, data samples are exchanged both ways in one 125-μs cycle. When the switches are operated in this fashion, the Eqs. (6.1)–(6.3) apply to folded networks as well.

The fact that only $N/2$ simultaneous conversations are possible in a folded network can be used to arrive at another switch configuration with only $N/2$ control memory locations. In this case, no cyclic control is possible and both the input and output are memory controlled. But before discussing this configuration, we take a look at the design parameters of a time division space switch and compare the same with that of a space division switch. The definitions of the design parameters of a switch have been given in

Section 2.5. We only calculate the values here. Let us assume same unit cost for each switching element and for each word of the control memory. The of the stored program control (SPC) system may be ignored, as it is of the same order irrespective of whether the switching network uses space division or time division techniques. For nonfolded networks, we have

No. of switching elements:
- on the input side $= N$
- on the output side $= N$
- total $= 2N$

Switching capacity, $SC = N$

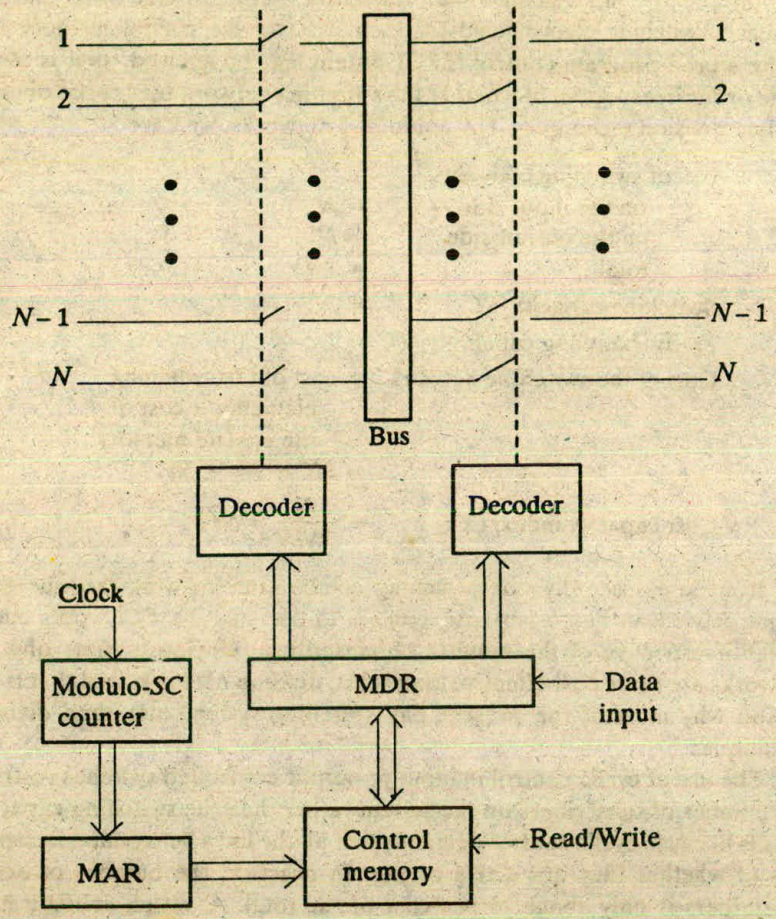
Traffic handling capability, $TC = 1$

Cost of the switching network = cost of the switching elements + cost of the control memory
$= 2N + N = 3N$

Cost capacity index, $CCI = \frac{N}{3N/N} = N/3$

If we use space division switching on the same network (see the equivalent network in Fig. 6.1(b)), *SC* reduces to one, and the *CCI* works out to be 0.5 irrespective of the number of subscribers. Obviously, time division networks are more cost effective than space division networks and that is the reason why most of the present day switching systems use time division techniques.

The use of cyclic control in input or output-controlled switches restricts the number of subscribers on the system rather than the switching capacity. This is because cyclic control demands that all the lines be scanned irrespective of whether they are active or not. In practice, the number of active subscribers is only about 20 per cent of the total. A switch configuration based on the use of control memory for controlling both inlets and outlets would permit a much larger number of subscribers than the switching capacity of the network and can be made to function either as a folded or a nonfolded network. Such a configuration is shown in Fig. 6.4 and we call this **memory-controlled time division space switch**. This structure is more general and versatile than the input or output-controlled structures. Here, each word of the control memory has two addresses: an inlet address and an outlet address. The control memory width is $2\lceil \log_2 N \rceil$. The operation of the switch proceeds as follows. The modulo-*SC* counter is updated at the clock rate. The control memory words are read out one after another. The inlet address is used to connect the corresponding inlet to the bus and so also the outlet address to connect the outlet. The sample is then transferred from the inlet to the outlet. Next, the clock updates the counter and the cycle of operations is repeated. When a connection is to be set up between inlet k and



MDR = memory data address MAR = memory address register

Fig. 6.4 Generalised time division space switch.

outlet *j*, the two addresses are entered into one of the free locations of the control memory via the data input facility and the location is marked busy. When the conversation terminates, the two addresses are replaced by null values and the location is marked free. The free or busy information may be maintained in a bit vector, where one bit corresponds to one location in the control memory. If a bit is set, the corresponding location is busy, otherwise it is free. The busy/free vector and the data input are managed by the SPC processor. Managing the entries in the control memory is further discussed in Section 6.2. It may be noted that there is no reservation of the control memory locations for particular inlets or outlets or inlet-outlet pairs unlike in the case of input or output-controlled configurations.

The number of words in the control memory and the size of the modulo counter are now equal to the switching capacity of the network and have no relation to the number of subscribers connected to the network. The switching capacity SC and the clock rate are given by

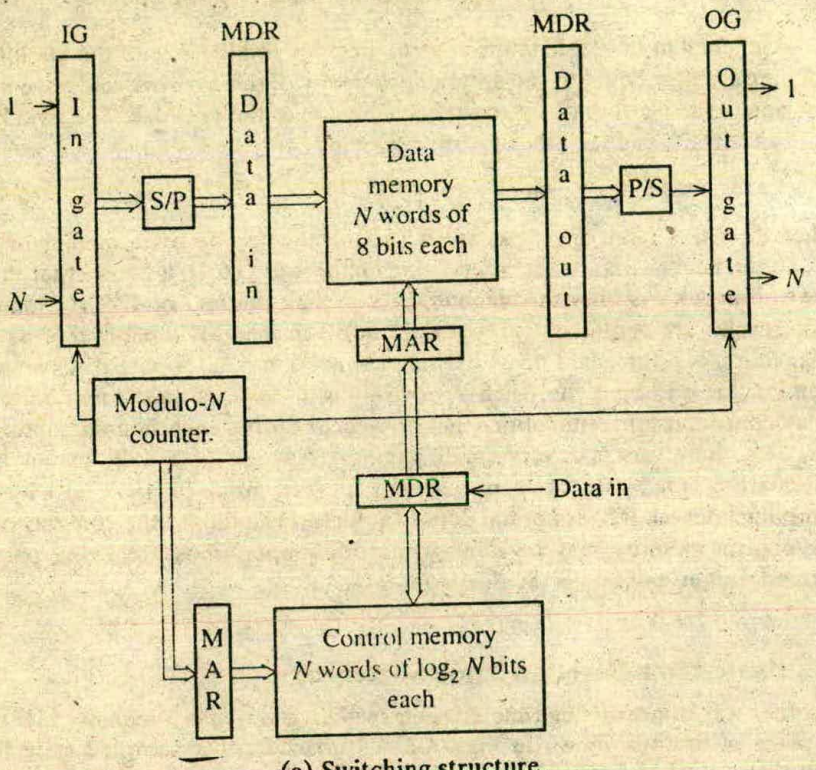
$$SC = 125/t_s, \qquad \text{clock rate} = 8SC \text{ kHz} \tag{6.4}$$

where $t_s = t_i + t_m + t_d + t_t$ as in Eq. (6.2). If the time to read the memory, i.e. component t_m of t_s, is the dominating factor in Eq.(6.4), it means that the control memory is busy throughout the sampling interval of 125 μs. Some write cycles are required to input data into the memory whenever a new connection is established or an existing connection is terminated. A provision for this is to be made. Such a provision is dependent on the number of new connections or terminations that may occur during each sampling interval. This, however, is a very small number (fraction of a call) and it is adequate if one write cycle is reserved for the input purpose in every sampling interval. If t_m is not the dominant factor, changes to the contents of the control memory may be done while other operations are being performed. But, usually, t_m is the dominant factor.

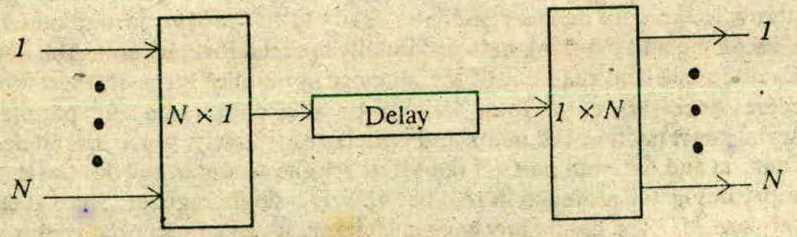
6.2 Basic Time Division Time Switching

Another way of organising time division switching is to use a memory block in place of the bus shown in Figs. 6.2–6.4, provided the sampled speech values are PCM and not PAM samples. The functional blocks of a memory based time division switch is shown in Fig. 6.5(a) and its equivalent circuit in Fig. 6.5(b). In this organisation, the data coming in through the inlets are written into the data memory and later read out to the appropriate outlets. The incoming and outgoing data are usually in serial form whereas the data are written into and read out of the memory in parallel form. It, therefore, becomes necessary to perform serial-to-parallel conversion and parallel-to-serial conversion at the inlets and outlets respectively. For convenience, the data-in and data-out parts of the MDR are shown separately for the data memory in Fig. 6.5 although in reality, MDR is a single register. Since there is only one MDR, a gating mechanism is necessary to connect the required inlet/outlet to MDR. This is done by the in-gate and out-gate units. Unlike the time division space switches discussed in Section 6.1, there is no physical connection, even momentarily, between inlets and outlets in the case of data memory based operation. Further, the information is not transferred in real time; it is first stored in the memory and later transferred to the outlet. There is a time delay between the acquisition of a sample from an inlet and its delivery to the corresponding outlet. Signifying these aspects, the data memory based scheme is known as **time division time switching.**

Recall that a time division space switch can be input, output or memory-controlled. In a similar fashion, a time division time switch may be controlled in any of the following three ways:



(a) Switching structure.



(b) Equivalent circuit.

Fig. 6.5 Basic time division time switching.

1. Sequential write/random read
2. Random write/sequential read
3. Random input/random output.

In the first two methods of control, the sequential/random read/write operations refer to the read/write operations associated with the data memory. In

both these cases, the inlets and outlets are scanned sequentially. In the last case, the inlets and outlets are scanned randomly, and the data memory is accessed sequentially.

There are two modes in which a time division time switch may be operated under any of the three forms of control already mentioned. We term these modes as:

- Phased operation
- Slotted operation.

The phased operation of the time switches proceeds in two phases. When the method of control is sequential write/random read, the inlets are scanned in the first phase one after another and the data is stored in the data memory sequentially. There is a one-to-one correspondence between the inlets and the locations of the data memory. The data memory location i corresponds to inlet i. The control memory locations, 1, 2, 3, ..., N contain the addresses of the inlets corresponding to the outlets 1, 2, 3, ..., N. The inlet addresses are read out from the control memory in the second phase, the corresponding locations in the data memory are accessed and the data transferred to the outlets in sequence. Since any inlet may be connected to any of the outlets, the inlet addresses are randomly distributed in the control memory. Consequently, the read access to the data memory is random. Since the write access to the data memory in the first phase proceeds sequentially and the read access in the second phase randomly, the nomenclature sequential write/random read is used to describe this form of control. It may, however, be noted that the inlets, outlets and the control memory are accessed sequentially.

In the first phase, one memory write is involved per inlet whereas in the second phase two memory reads, one at the control memory and the other at the data memory, are involved per outlet. The time taken for the two-phase operation t_s is given by

$$t_s = Nt_d + N(t_d + t_c) \qquad (6.5)$$

where

t_d = read/write time for the data memory
t_c = read/write time for the control memory

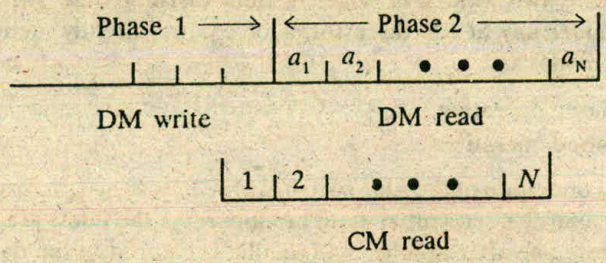
If $t_d = t_c = t_m$, we have

$$t_s = 3Nt_m \qquad (6.6)$$

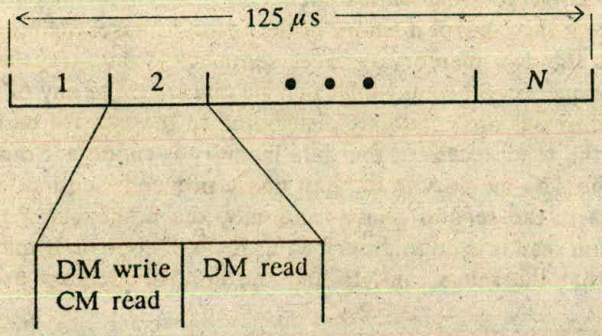
Since the entire operation is to be completed within 125 μs, we have the expression for the number of subscribers as

$$N = \frac{125}{3t_m} \qquad (6.7)$$

where t_m is expressed in microseconds. The number of subscribers can be increased by overlapping the read cycles of the data memory and the control memory in the second phase of operation as shown in Fig. 6.6(a). During the



(a) Overlapping of data and control memory reads



(b) Simultaneous DM write and CM read

Fig. 6.6 Two modes of operation of time switches.

last write cycle of phase 1, the first location of the control memory which contains an inlet address, say a_1, is read. During the first cycle of the second phase, the location a_1 of the data memory and the second location of the control memory are read out simultaneously. Thus Eq. (6.7) may be modified for overlapped operation as

$$N = \frac{125}{2t_m} \tag{6.8}$$

In the slotted operation the 125-μs period is divided into N subperiods of duration $125/N$. In each subperiod i, the following operations are performed:

1. Read inlet i and store the data in data memory location i.

2. Read the location i of the control memory which contains the value, say j.
3. Read the data memory location j and transfer the data to outlet i.

The first two operations can be carried out simultaneously such that Eq.(6.8) holds valid for this mode of operation which is depicted in Fig. 6.6(b).

Choosing between the two modes of operation of time switches is a matter of design choice. For example, in the phased mode of operation, the modulo-N counter goes through two cycles for the operation as against only one cycle in the slotted mode. The data transfer from inlets to the data memory can be organised using direct memory access facility in the phased mode whereas this is not possible in the slotted mode of operation. More importantly, in the phased mode of operation, samples transferred to the outlets belong to the same 125-μs period in which they were acquired whereas in the second mode, the sample values may belong to the previous 125-μs period, thus introducing one sample delay between the inlet and the outlet. This aspect is further studied in Section 6.4.

In random write/sequential read form of control, the control memory contains the addresses of outlets corresponding to the inlets. In the first phase, the control memory is read and the inlet data is written into the data memory location specified by the contents of the control memory. The inlets are scanned sequentially but the data are written into the data memory randomly. In the second phase of operation, the data memory is read out sequentially and the data sent to the outlets sequentially. There is no correspondence between the inlets and the data memory locations. But, there is one-to-one correspondence between the outlets and the data memory locations. It may, however, be noted that there is a one-to-one relationship between the inlets and the control memory locations. Observe the similarity to the input-controlled space switch discussed in Section 6.1.

Equations (6.7) and (6.8) are applicable to random write/sequential read form of control as well. The overlapped operation corresponding to Eq. (6.8) implies that the first read cycle of the control memory will be overlapped with the last read cycle of the data memory in the previous 125-μs cycle. Therefore, the writing into the data memory proceeds simultaneously with the reading of the control memory locations. While the contents of the i-th location of the control memory is being used as address for the data memory to write data into, the $(i + 1)$-th location of the control memory is read out. This is illustrated in Fig. 6.7.

Both sequential write/random read and random write/sequential read controls impose a serious restriction on the number of subscribers that can be connected to the system. Since all the inlets and the outlets are scanned irrespective of whether they are active or not, the total number of subscribers is limited to the number of read/write operations that can be performed in a 125-μs interval. In other words, the number of subscribers is not more than the switching capacity, SC, of the system. The switches are obviously

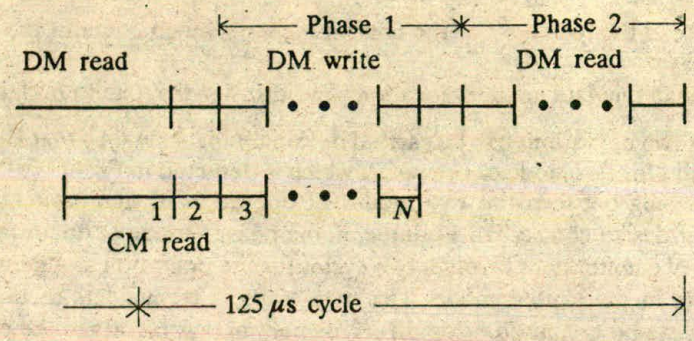


Fig. 6.7 Overlapping of data memory write and control memory read operations.

nonblocking. If we have a scheme whereby only the active subscribers are scanned, then the total number of subscribers connected to the system can be increased significantly.

Random input/random output form of control permits a larger number of subscribers than the switching capacity of the system. The switch in this case is, however, blocking in nature. The random input/random output scheme is illustrated in Fig. 6.8. Functionally, there are two control memory

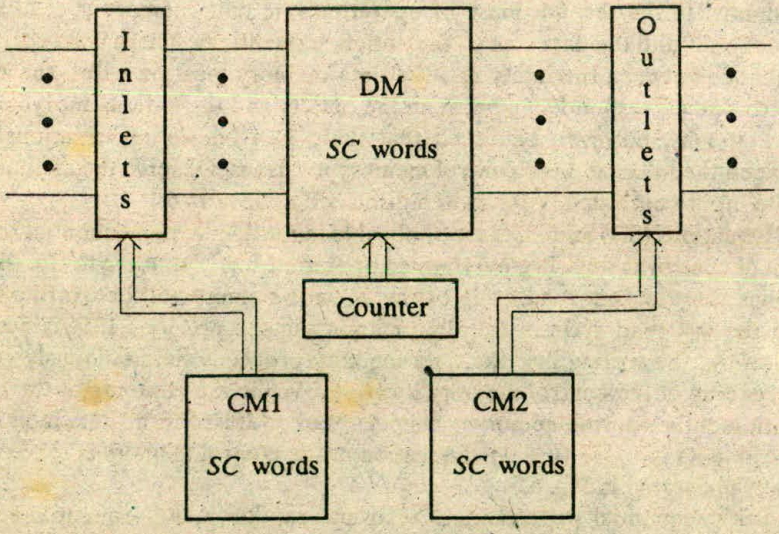


Fig. 6.8 Random input/random output time switch.

modules, CM1 and CM2, which hold the addresses of the active inlets and outlets respectively. There is a one-to-one correspondence between the locations of the two control memories. If the address of an active inlet is

placed in location x of CM1, the address of the outlet to which this inlet is connected, is placed in the location x of CM2. The phased operation of this configuration also proceeds in two phases as in the case of the two time switches discussed earlier. In the first phase, the addresses of the active inlets are read out of CM1 one by one and the data transferred from the respective inlets to the data memory starting from its first location. In the second phase, the addresses of the outlets are read out of CM2, and the data from the data memory is transferred to outlets specified by these addresses. There are two memory read/write operations involved in each phase. As a result, the switching capacity of the system may be expressed as

$$SC = \frac{125}{4t_m} \tag{6.9}$$

By overlapping the control memory and data memory operations, the switching capacity can be doubled and Eq. (6.9) changes to

$$SC = \frac{125}{2t_m} \tag{6.10}$$

With the advent of integrated semiconductor memories, dual port memory chips, where two accesses can be made simultaneously, are commonly available. By using such memory chips the input and output operations can be done simultaneously, doubling the switching capacity given by Eq. (6.10). However, one sample delay occurs in the process, as the sample value sent to the outlet corresponds to the previous 125-μs cycle. In this case, the two control memories are read out simultaneously. Alternatively, only one control memory may be used with each location containing both the inlet and the outlet addresses. By using dual port memories and overlapped operations, the switching capacity becomes

$$SC = \frac{125}{t_m} \tag{6.11}$$

Entries in the control memory modules specify the active inlet/outlet pairs. Whenever a call is set up, the corresponding inlet and outlet addresses are entered in CM1 and CM2 respectively at the same location addresses. When a call is terminated, the corresponding entries are set to null values. The entries in the control memory may be managed in any of the following three ways:

1. By maintaining a free list
2. By compacting the entries every time a call terminates
3. By maintaining free and occupied lists.

The free list is a linked list maintaining information about the free locations in the control memory. This list may be organised using a

first-come-first-served (queue) discipline or a last-come-first-served (stack) discipline. Whenever a call is to be set up, a free location is obtained from this list, and whenever a call is terminated, the freed location is added to the list. The free locations are distributed throughout the control memory interspersed with the busy locations. When a free list alone is used for managing the entries in the control memories, all locations of these memories have to be scanned in the respective phases of operation. Whenever a free location is encountered with a null value, no data transfer takes place between the inlet and data memory or between the data memory and the outlet. Thus the time required to complete the two phases of operation is given by

$$T = 4s_1t_m + 2(SC - s_1)t_m \ \mu s \tag{6.12}$$

where s_1 is the number of active subscriber pairs. Equation (6.12) corresponds to the case where single port memory chips are used and there is no overlapped operation. The remaining time $(125 - T)$ in the 125-μs cycle may be used for maintenance and diagnostic purposes as explained in Section 4.5.

Accessing of all the locations of the control memory may be avoided by maintaining an occupied list in addition to the free list. In this case, only those locations in the occupied list are accessed and the Eq. (6.12) becomes

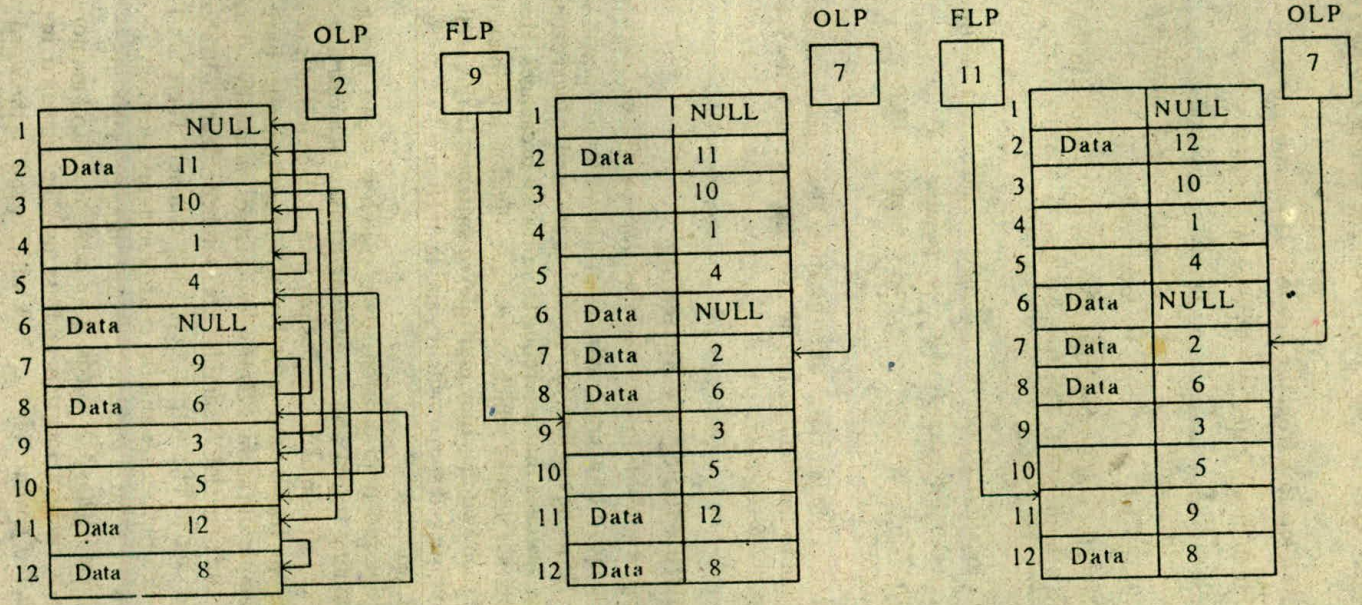
$$T = 4s_1t_m \tag{6.13}$$

Maintenance of the lists may be done away with if compaction is resorted to every time a call terminates. Compaction ensures that all entries corresponding to active inlet/outlet pairs are gathered at the beginning of the control memory. In such a case, the control memory access is terminated as soon as a location with null value is encountered. The time taken to complete the two phases of operation is now given by adding a term to Eq. (6.13) as

$$T = 4s_1t_m + 2t_m \tag{6.14}$$

The free list (or the occupied list) is usually managed by using pointers. Every entry in the list now consists of two parts: data and a pointer to the next entry in the list. For a free list, data value has no significance. The beginning of the list is identified by a variable which contains the address of the first free (or occupied) location which, in turn, points to the next free (or occupied) location. A link is thus established to all the free (or occupied) locations. The last free (or occupied) location contains a null pointer. This is illustrated in Fig. 6.9(a). At the head of the free list is location 7 as indicated by the free list pointer (FLP) and the free list comprises locations 7, 9, 3, 10, 5, 4 and 1. The first location of the occupied list is 2 as indicated by the occupied list pointer (OLP) and the occupied list is 2–11–12–8–6.

Associated with any list, there are two fundamental list management operations: *add an item* to the list and *remove an item* from the list. In the case of free and occupied lists pertaining to the same set of memory locations, as is the case here, a location is added to the free list when it is



FLP = free list pointer OLP = occupied list pointer

(a) A sample list (b) Location 7 allocated (c) Location 11 freed

Fig. 6.9 List management.

taken out of the occupied list and vice versa. The actions involved in these operations are:

1. Allocate a location from the free list to the occupied list

Location to be allocated, F	=	contents of FLP
Contents of FLP	=	pointer in location F
Pointer in location F	=	contents of OLP location F
Contents of OLP	=	location F

Figure 6.9(b) shows the status of the lists after the first location in the free list is moved to the occupied list. The free list now is 9–3–10–5–4–1 and the occupied list is 7–2–11–12–8–6.

2. Free the location X from the occupied list and add to the free list

Pointer in the predecessor of X	=	pointer in X
Pointer in X	=	contents of FLP
Contents of FLP	=	X

Figure 6.9(c) depicts the list status after location 11 is freed from the occupied list and added to the free list. The new free list is 11–9–3–10–5–4–1 and the occupied list is 7–2–12–8–6.

6.3 Time Multiplexed Space Switching

In Sections 6.1 and 6.2, we dealt with time division switches where an inlet or an outlet corresponded to a single subscriber line with one speech sample appearing every 125 μs on the line. Such switches are used in local exchanges. We now consider switches that are required in transit exchanges. Here, the inlets and outlets are trunks which carry time division multiplexed data streams. We call such switches **time multiplexed switches**. In this section, we consider time multiplexed space switches and in Section 6.4 we discuss time multiplexed time switches.

A time multiplexed time division space switch is shown in Fig. 6.10. There are N incoming trunks and N outgoing trunks, each carrying a time division multiplexed stream of M samples per frame. Each frame is of 125-μs time duration. In one frame time, a total of MN speech samples have to be switched. One sample duration, $125/M$ microseconds, is usually referred to as a **time slot**. In one time slot, N samples are switched. Figure 6.10 shows an output-controlled switch. The outputs cyclically scanned. There is a 1-to-M relationship between the outlets and the control memory locations, i.e. there are M locations in the control memory corresponding to each outlet.

The control memory has MN words. If we view the control memory as M blocks of N words each, a location address may be specified in a two dimensional form, (i, j), where i is the block address and j is the word within the block. We have $1 \leq i \leq M$ and $1 \leq j \leq N$. The block address i corres-

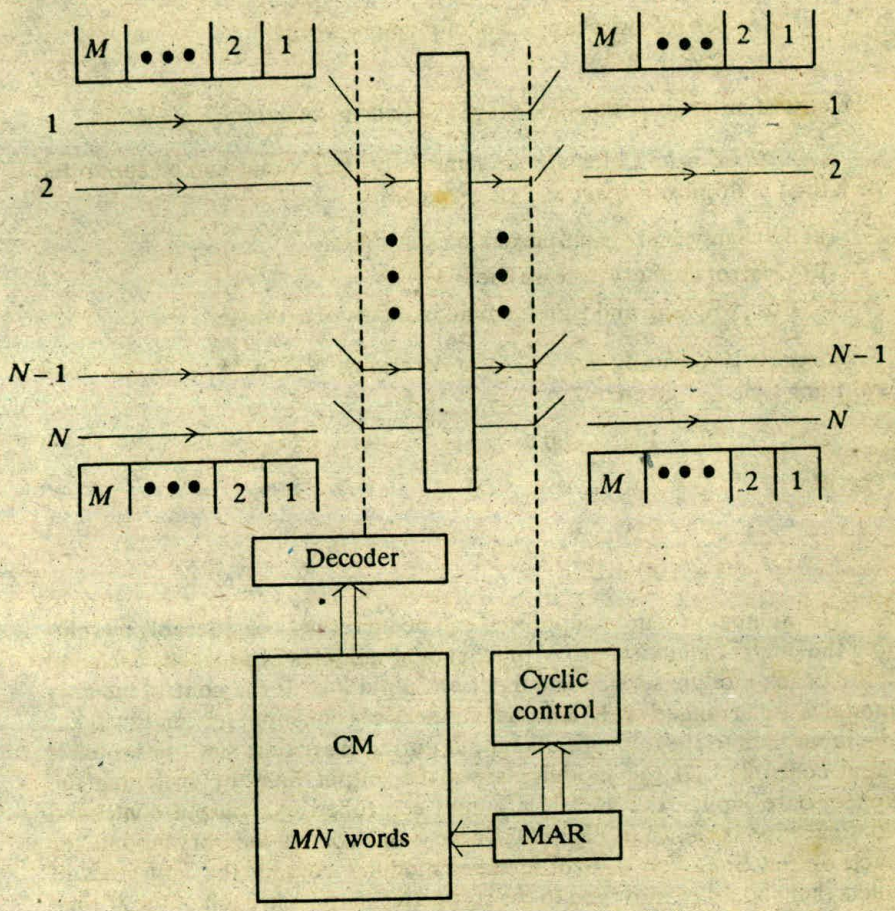


Fig. 6.10 Time multiplexed space switch.

ponds to the time slot i and the word address j to the outlet j. The first N locations of the control memory correspond to the first time slot, the next N locations, i.e. locations $N + 1$ to $2N$ when addressed linearly, or locations $(2, 1)$ to $(2, N)$ when addressed in a two dimensional form, correspond to the time slot 2 and so on. Therefore, if the location (i, j) contains an inlet address k, it implies that inlet k is connected to the outlet j during the time slot i. The number of trunks that can be supported on this switch is given by

$$N = \frac{125}{Mt_s} \tag{6.15}$$

where t_s is the switching time including memory access time per inlet-outlet pair. The operation of the switch is very similar to the one shown in Fig. 6.3 and described in Section 6.1. The cost of the switch is estimated as

$$C = \text{No. of switches} + \text{No. of memory words}$$
$$= 2N + MN \tag{6.15a}$$

The cost of an equivalent single-stage space division network is $(MN)^2$.

EXAMPLE 6.1 Calculate the number of trunks that can be supported on a time multiplexed space switch, given that

(a) 32 channels are multiplexed in each stream

(b) Control memory access time is 100 ns

(c) Bus switching and transfer time is 100 ns per transfer.

Solution We apply Eq. (6.15) to determine N. For $M = 32$, the total switching time t_s is given by

$$t_s = 100 + 100 = 200 \text{ ns}.$$

Therefore,

$$N = \frac{125}{32 \times 200 \times 10^{-3}} \approx 20$$

The number of trunks supported can be increased considerably by splitting the control memory into N independent modules of M words each and letting each module service one input or output line. If the control memory modules are arranged such that each one services one input line and switches the input data to the appropriate output during every time slot, the switch is input-controlled. If the modules serve the output lines by switching the appropriate inputs, the switch is output-controlled. An output-controlled configuration is shown in Fig. 6.11. There is one control memory module for each output line. The control memory modules contain the addresses of inlets that should be switched to the respective outputs in each of the M time slots. Location 1 of every module corresponds to the time slot 1, location 2 to time slot 2 and so on. All locations corresponding to a particular time slot are read out in parallel. As a result, there is no constraint due to time on the number of trunks that can be supported by the switch. The only time constraint that needs to be satisfied is

$$M = \frac{125}{t_m} \tag{6.16}$$

where t_m is the control memory access time. In writing Eq. (6.16), we assume that the time required for decoding addresses from the control memory and switching the gates are negligible when compared to the memory access time. If this is not the case, t_m may be substituted by t_s which includes these times as given in Eq. (6.4).

We now estimate the cost of the switch. The switch has N switching matrices of size $N \times 1$ each at the inputs. There are NM memory words. Therefore, the cost is

$$C = N(N \times 1) + NM = N^2 + NM \tag{6.16a}$$

The cost of the switch is lower than that of the equivalent single-stage space division switch but is much more expensive than that of the one shown in

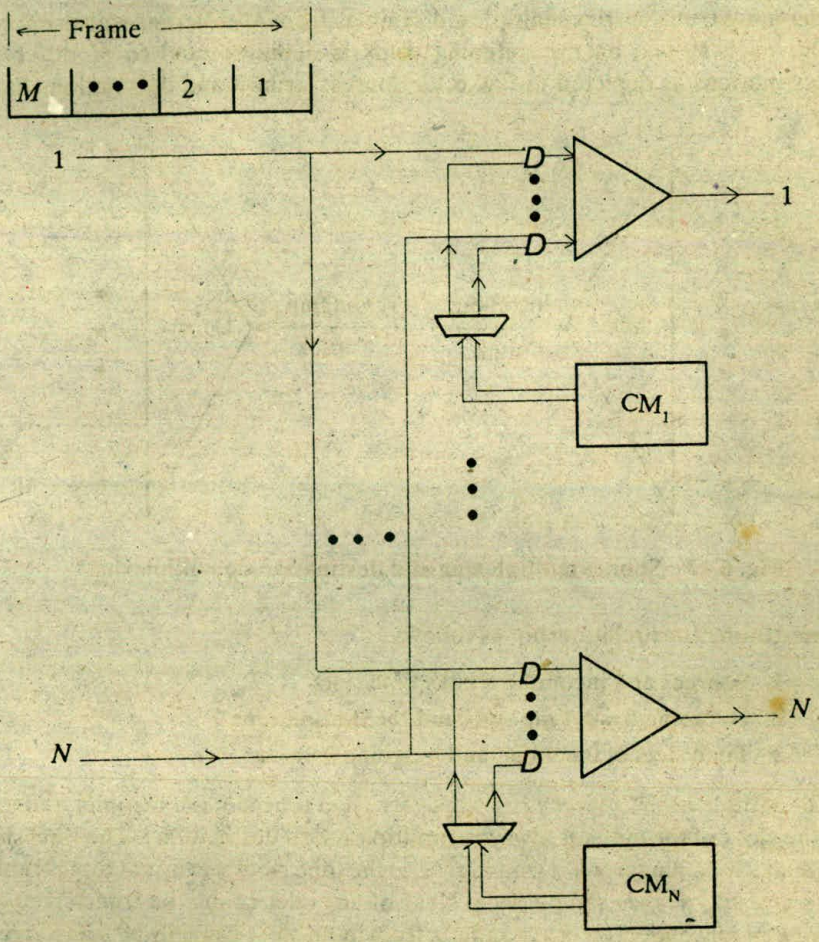


Fig. 6.11 Time multiplexed space switch with N control memory modules.

Fig. 6.10. It may be observed that there is no cost increase on account of the memory words. The cost has gone up due to the number of space division switching matrices used for gating the inlets.

As noted in Section 6.1, output-controlled configurations permit broadcast or one-to-many connections. During a given time slot, any particular input may be enabled by more than one control memory module leading to a

one-to-many connection. Such a connection is not possible in an input-controlled switch. In view of this, output-controlled configurations are preferred to the input-controlled switches. It may be noted that the width of the control memory is $\log_2 N$.

Time multiplexed space switches do not provide full availability. Each incoming trunk carries multiplexed samples from M different voice sources and each stream on the outgoing trunk is demultiplexed to M different destinations as depicted in Fig. 6.12. Sources, trunks and destinations have

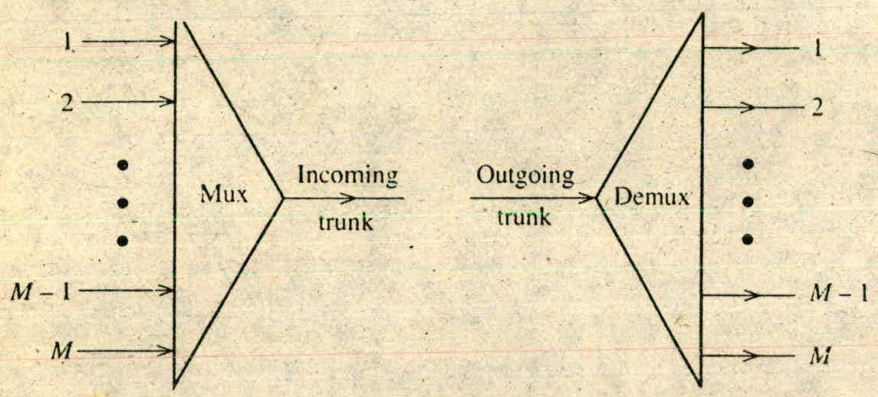


Fig. 6.12 Source multiplexing and destination demultiplexing.

one-to-one time relationship as follows:

1. Sources and incoming trunks time slots
2. Outgoing trunk time slots and the destinations
3. Time slots of incoming and outgoing trunks.

The sample of the source i is always carried in time slot i of the inlet and the time slot j of the outlet is always demultiplexed to destination j. The time slot k of any incoming trunk is transferred to the time slot k of any outgoing trunk. As a result, a voice sample from slot i of any inlet cannot be transferred to slot k of any outlet ($i \neq k$). A sample from input slot i can only be transferred to destination i of one or more outlets. In other words, interchange of samples among different time slots is not possible. Thus, the switch does not support full availability. For every input, there are $N(M-1)$ outputs that cannot be reached.

With the advent of large scale integration in the semiconductor technology, integrated circuit space arrays that are suitable for digital space switching are becoming available. A digital space array of size $N \times N$ is similar to a $N \times N$ crossbar switching matrix, except that bulky expensive, and power consuming electromagnets are replaced by digital gates.

A time multiplexed time division space switch can be configured around a space array as shown in Fig. 6.13. The array consists of M input horizontals

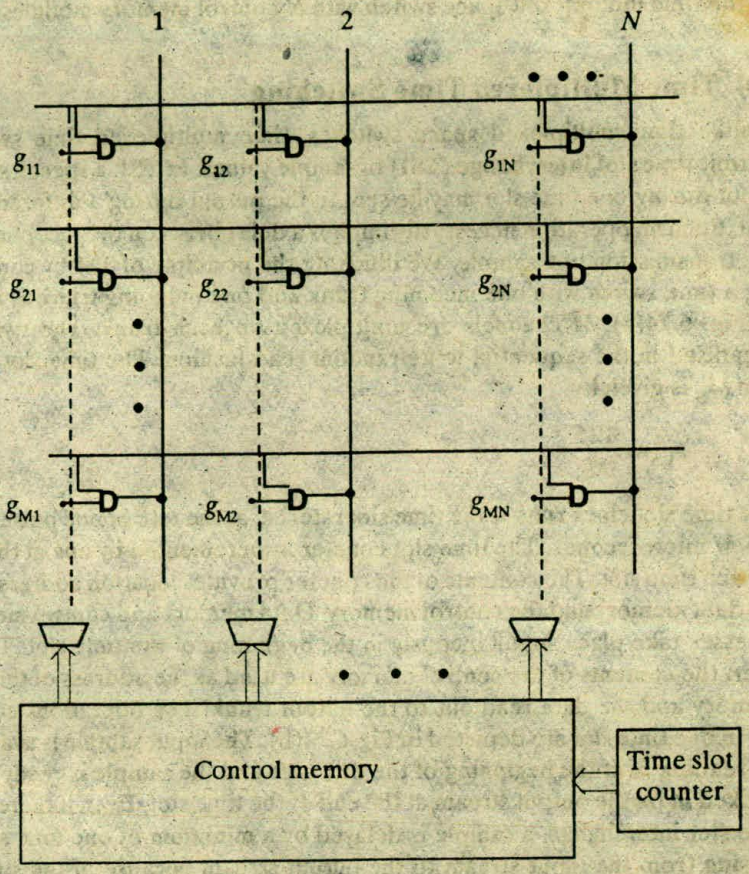


Fig. 6.13 Time division space switch using space array.

and N output verticals. If $M = N$, the switch is nonblocking. If $M > N$, the switch is concentrating, and if $M < N$, the switch is expanding. In every time slot, one logic gate per vertical if $M > N$, or one logic gate per horizontal if $M < N$ is enabled for one-to-one connections. It may be observed that the space array is capable of establishing one-to-many and broadcast connections. In any time slot, only one gate for vertical and up to N gates per horizontal can be enabled for these connections. An enabled logic gate passes the input digital sample to the corresponding output vertical.

In every time slot, upto N or M samples are switched simultaneously. The control store has N addresses corresponding to N vertical outputs with each address selecting one gate in each vertical output. The size of the control

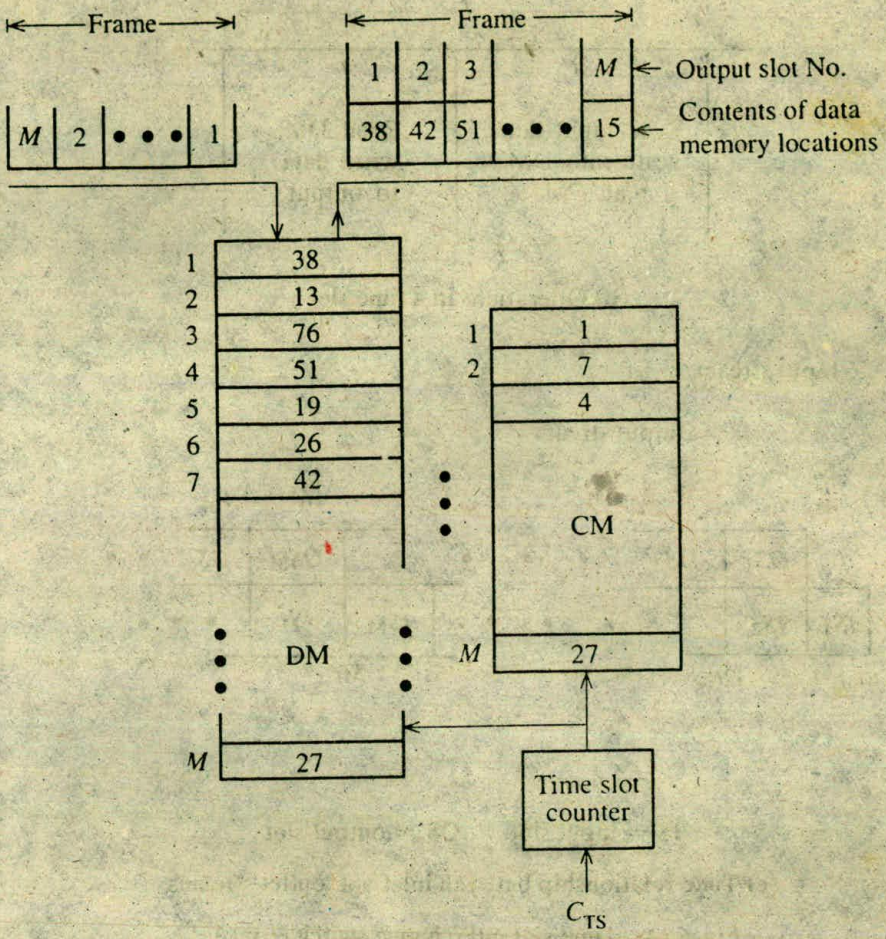
memory is N and its width $\lceil \log_2 M \rceil$. Because of the parallel transfer of N or M data samples in each time slot, a large number of channels can be multiplexed per input line. The performance of this switch is similar to that of the time multiplexed space switch with N control memory modules.

6.4 Time Multiplexed Time Switching

Unlike time multiplexed space switches, time multiplexed time switches permit **time slot interchange (TSI)** of sample values. In TSI, a speech sample input during one time slot may be sent to the output during a different time slot. Such an operation necessarily implies a delay between the reception and the transmission of a sample. We illustrate the principle of TSI by considering a time switch with one incoming trunk and one outgoing trunk as shown in Fig. 6.14(a). M channels are multiplexed on each trunk. The switch is organised in the sequential write/random read fashion. The time slot duration t_{TS} is given by

$$t_{TS} = \frac{125}{M} \tag{6.17}$$

The time slot clock runs at the time slot rate, i.e. at the rate of one pulse every $125/M$ microseconds. The time slot counter is incremented by one at the end of each time slot. The contents of the counter provides location addresses for the data memory and the control memory. Data memory and control memory accesses take place simultaneously in the beginning of the time slot. Thereafter, the contents of the control memory are used as the address of the data memory and the data read out to the output trunk. The operations carried out in one time slot are depicted in Fig. 6.14(b). The input sample is available for reading in at the beginning of the time slot and the sample is ready to be clocked in on the output stream at the end of the time slot. Even if there is no time slot interchange, a sample is delayed by a minimum of one time slot in passing from the input stream to the output stream because of the storage action. In other words, a time slot switch may be considered to have an inherent time delay of one time slot. In effect, the output stream is delayed by t_{TS} microseconds when compared to the incoming data stream as shown in Fig. 6.14(c). Depending on the output time slot to which an input slot contents are switched, the sample experiences a delay in the range of t_{TS} to Mt_{TS} microseconds. In the example entries, shown in the control memory of Fig. 6.14(a), the first location contains the value 1. This implies that the contents of input time slot 1 is switched to output time slot 1. The sample, in this case, experiences a delay of t_{TS} microseconds. The second location of the control memory contains the value 7 and, therefore, the input time slot 7 is switched to output time slot 2. This sample experiences a delay of $((M-7)+2+1)t_{TS}$ or $(M-4)t_{TS}$ microseconds. Output time slot 3 carries the contents of input time slot 4 and the delay experienced by the sample is



CM = control memory DTS = time slot clock DM = data memory

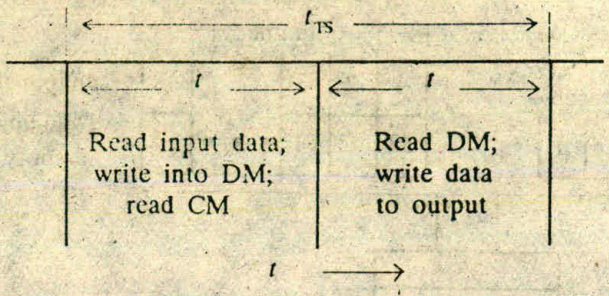
(a) Principle of time slot interchange

Fig. 6.14 Time slot interchange switch.

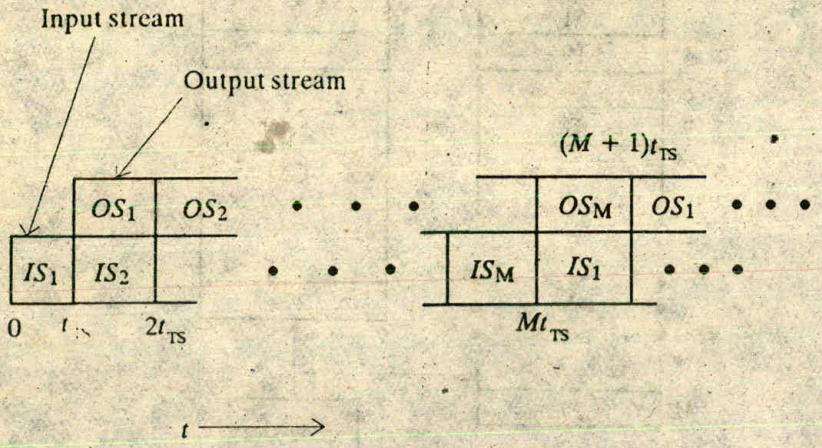
$(M-(4-3)+1)t_{TS}$ or Mt_{TS}, i.e. 125 microseconds. There are two sequential memory accesses per time slot and hence the time constraint may be stated as

$$t_{TS} = 2t_m, \qquad 125 = 2Mt_m \tag{6.18}$$

where t_m is the access time of the memory modules in microseconds. When there is a two-way traffic and the network is nonfolded, another set of data and control memories is used. In the second control memory, the locations 1, 7 and 4 contain the values 1, 2 and 3 respectively, corresponding to the



(b) Operations in a time slot



IS = input slot OS = output slot

(c) Time relationship between inlet and outlet streams

Fig. 6.14 Time slot interchange switch (cont.).

sample entries shown in Fig. 6.14(a). When the 125-μs cycle is complete, the values in the input time slots 1, 7 and 4 are interchanged with the output time slots 1, 2 and 3 respectively. When the network is folded, there is only one set of data and control memories even for two-way traffic. For the example shown in Fig. 6.14(a), control memory locations of 7 and 4 contain the values 2 and 3 respectively. When the 125-μs cycle is complete, the values in the time slots 7 and 2 and the time slots 3 and 4 are interchanged. For a folded network, transferring the data between the same input and output slots, e.g. from input slot 1 to output slot 1, is not relevant.

Since there are no switching elements in this configuration, the cost is equal to the number of memory locations. There are M locations each in the control and in the data memory. Therefore, the cost is given by

$$C = 2M \text{ units} \qquad (6.18a)$$

EXAMPLE 6.2 Calculate the maximum access time that can be permitted for the data and control memories in a TSI switch with a single input and single output trunk multiplexing 2500 channels. Also, estimate the cost of the switch and compare it with that of a single stage space division switch.

Solution Applying Eq. (6.17), we have

$$t_m = \frac{125 \times 10^3}{2500 \times 2} = 25 \text{ ns}$$

$$C = 2 \times 2500 = 5000 \text{ units}$$

This switch is nonblocking and supports full availability. An equivalent single stage space division switch uses a matrix of 2500 × 2500. Hence, the cost of such a switch is 6.25 million units.

$$\text{Cost advantage of the time switch} = \frac{6.25 \times 10^6}{5000} = 1250$$

A TSI switch may be designed to be expanding or concentrating. In such switches, the number of time slots (samples) per frame in the input stream and in the output stream are different. If we represent these numbers as M_1 and M_2 respectively, the switch is expanding when $M_2 > M_1$. The bit rates of the input and output streams are also different. For an expanding switch, the output bit rate is higher and for a concentrating switch, the input bit rate is higher.

An expanding or a concentrating time switch can be realised by delinking the read and write operations of the data memory. They are carried out independently and in an asynchronous fashion as shown in Fig. 6.15 for

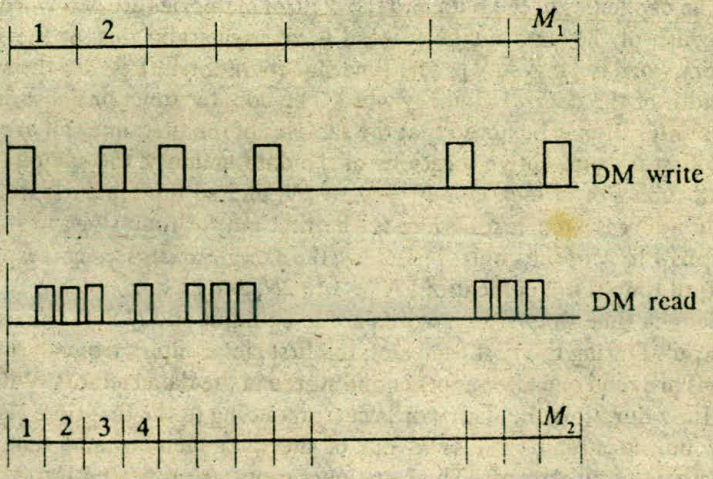


Fig. 6.15 Asynchronous read/write operations.

an expanding switch. The time constraint for the switch may be stated as in an expanding TSI switch

$$125 = (M_1 + M_2)t_m \tag{6.19}$$

Buffering may be required at the input and output of the data memory to ensure uniform data rates on the lines. An application of the expanding and concentrating time switches is seen in the context of combination switching in Section 6.5 and 6.6. The control memory read is assumed to be overlapped with the data memory read. There are M_1 data memory write operations and M_2 data memory read operations in every 125-μs interval.

After having illustrated the principle of operation of a TSI switch, we now consider practical configurations of time multiplexed time switches. In such configurations, there are N time multiplexed input streams each multiplexing M subscribers, and there are N time multiplexed output streams each carrying M subscribers. The problem is to handle NM subscribers in the time duration of 125 μs. This can be done in four different ways:

1. Serial-in/serial-out
2. Parallel-in/serial-out
3. Serial-in/parallel-out
4. Parallel-in/parallel-out.

The serial-in/serial-out configuration is similar to the one shown in Fig. 6.14(a) except that there are N input and N output streams instead of one each. The multiple input and output streams call for gating circuits on the input and output sides similar to the ones shown in Fig. 6.5(a). The gating circuits are required to route data from the inlets to the data memory and from the data memory to outlets. They are equivalent to $N \times 1$ and $1 \times N$ switching matrices.

The capacities of the data and the control memories are NM words each. The width of the data memory word is eight bits and that of the control memory word is $\lceil \log_2 MN \rceil$. The time slot counter width is also the same as the width of the control memory word. The counter functions as a modulo-MN counter. It may be noted that the samples of the first time slot are stored in the first N consecutive locations of the data memory, the samples of the second time slot in locations $N + 1$ to $2N$, and so on. Similarly, the input sample addresses corresponding to the first output time slot are stored in locations 1 to N of the control memory, the sample addresses corresponding to output time slot 2 in locations $N + 1$ to $2N$, and so on.

In each time slot, N data words are to be input and N data words are to be output. During the first time slot, the first slot samples from all the input streams are read one after another and stored in the data memory. Within the same time duration, the data words corresponding to the first time slot of all the N output streams are read out of the data memory and sent to the respective output streams. The same operations are repeated for every one

of the M time slots. The data memory write and the control memory read operations are carried out simultaneously followed by data memory read operation. The time constraints may be stated as

$$t_{TS} = 2Nt_m, \quad 125 = 2NMt_m \tag{6.20}$$

In parallel-in/serial-out configuration, the data memory is organised as N modules of M words each as shown in Fig. 6.16(a). Each module is associated with one input line. The time slot duration t_{TS} is divided into $N + 1$ subslots $t_0, t_1, ..., t_N$. During the subslot t_0, data from all the input lines are read into the respective data memory modules simultaneously. A two dimensional addressing structure comprising module address and word address, is used to access the data memory words. The word address corresponds to the time slot number and the module address to the inlet number. The word address, word 1 corresponds to time slot 1, word 2 to time slot 2 and so on. A special control signal is used to enable all the data memory modules for write access during the subslot t_0. During the subslots t_1 to t_N, the control memory locations are accessed one after another, the contents of the locations are used to address the data memory for read access, and the data words read out are transferred to the corresponding output lines. During the subslot t_1 of the time slot i, a data word read from the data memory is transferred to the outgoing trunk 1, for insertion in the time slot i of this stream. Similar transfers take place during subslots t_2 to t_N to outgoing trunks 2 to N.

The control memory locations contain the two-part address of the data memory. The higher-order address bits select the module and the lower-order bits the word within a module. The width of the control memory word is $\lceil \log_2 N \rceil + \lceil \log_2 M \rceil$. The control memory clock, C_{cm}, runs at the rate of one pulse per $t_{TS}/(N + 1)$ microseconds except that one pulse is inhibited during the subslot t_0. The control memory read may be overlapped with data memory as indicated in the timing diagram in Fig. 6.16(b). In this case, C_{cm} is inhibited during subslot t_N. The time constraints may now be stated as:

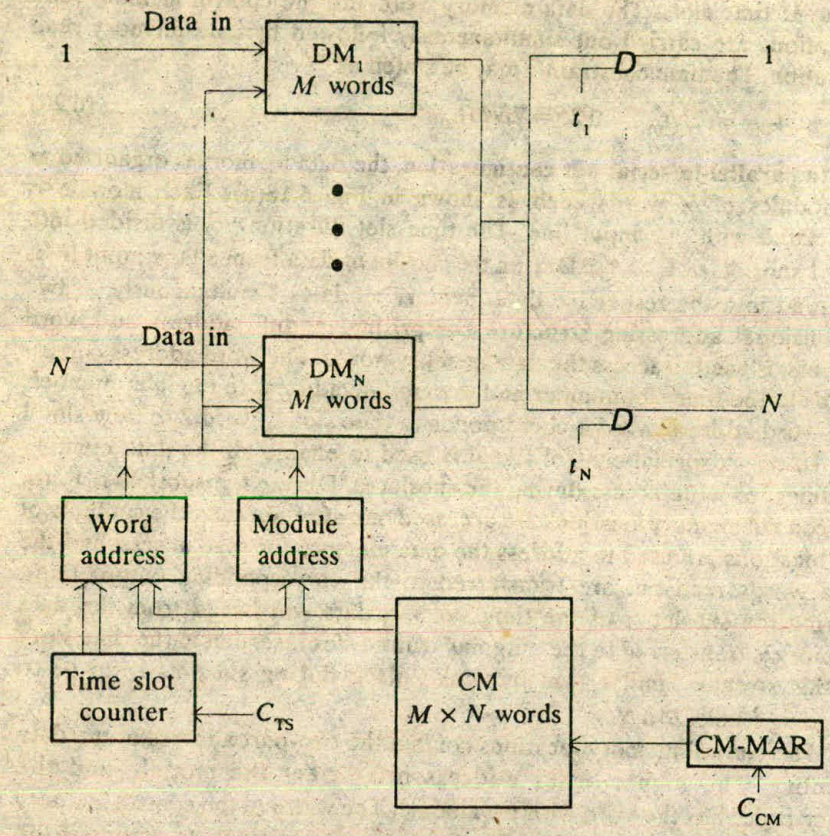
For nonoverlapped operation,

$$t_{TS} = (2N + 1)t_m, \quad 125 = M(2N + 1)t_m \tag{6.21}$$

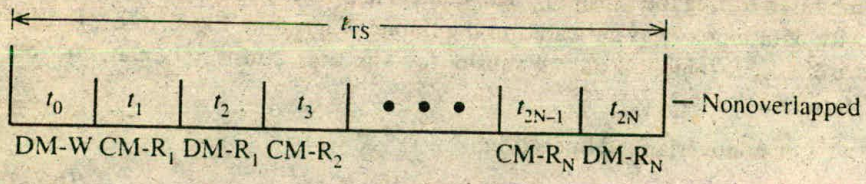
For overlapped operation,

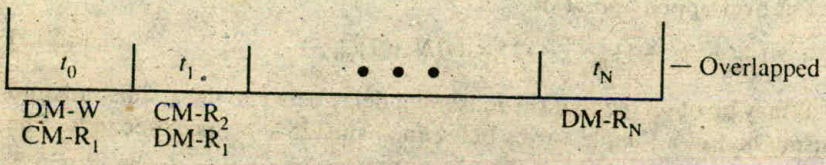
$$t_{TS} = (N + 1)t_m, \quad 125 = M(N + 1)t_m \tag{6.21a}$$

It may be observed that the maximum delay between the acquisition and transmission of a sample varies between t_0 and $125 - t_N$ microseconds. The parallel-in/serial-out switch functions like an output-controlled switch,



(a) Configuration





(b) Timing details

Fig. 6.16 Parallel-in/serial-out configuration.

because each storage location of the control memory is rigidly associated with a given output time slot and a specific outlet.

The serial-in/parallel-out switch is the dual of the parallel-in/serial-out switch. Here, the data memory modules are associated with the output lines instead of the input lines. The data input is serial and is carried out during the subslots t_1 to t_N. The subslot t_0 is used for parallel read-out from all the data memory modules to the corresponding output lines. With this arrangement, it appears as though the output precedes input. Since the entire operation is cyclic, it is immaterial whether the output operation is carried out in subslot t_0 or t_N. It is convenient to have the output operation done during the subslot t_0. The serial-in/parallel-out configuration is called input-controlled for obvious reasons. It may be noted that it is easier to organise broadcast connections in an parallel-in/serial-out switch rather than serial-in/parallel-out switches.

EXAMPLE 6.3 Calculate the access time of the memory modules in parallel-in/serial-out time switch using 64 input and 64 output streams with each stream multiplexing 32 channels.

Solution Substituting the values in Eq. (6.21) and assuming overlapped operation, we have

$$125 = 32 \times 65 \times t_m$$

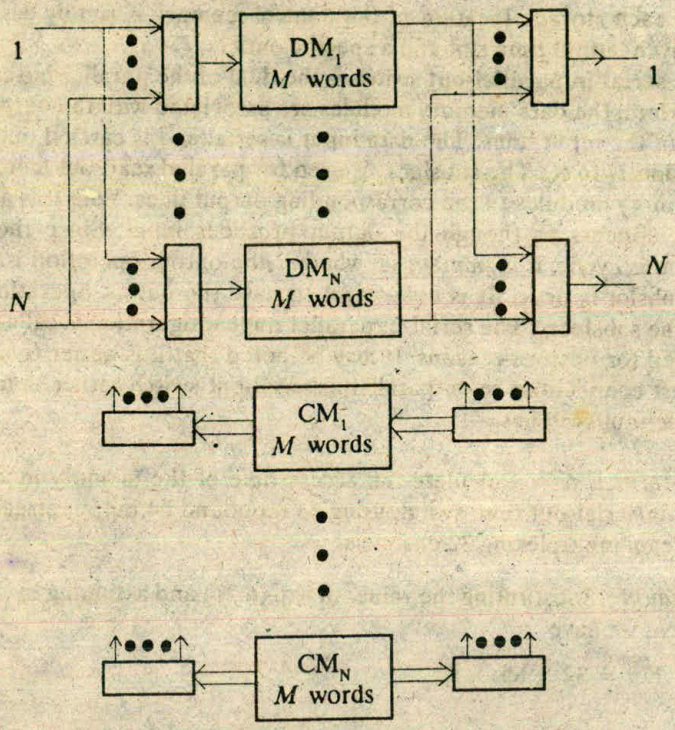
or

$$t_m = \frac{125}{32 \times 65} = 0.06 \ \mu s$$

Present day memory chips have access times in the range of 20–50 ns which implies that a time division time switch with either input or output operation performed in parallel and the other in serial fashion can, at best, support about 6000 subscribers.

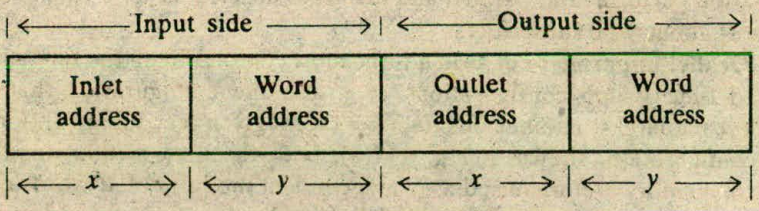
Parallel-in/parallel-out switch is the most complex of all the time multiplexed time switch configurations. It is also the configuration that can support the largest number of subscribers for a given technology option. In the configuration, shown in Fig. 6.17(a), both the data memory and the control memory are organised as N independent modules of M words each. There is one-to-one correspondence between the data memory and control memory modules. Address outputs from the control memory module i are always used to access words in the data memory module i. The correspondence is limited to the module level and does not extend to word level.

Each location of the control memory module contains information about both the input side and the output side as shown in Fig. 6.17(b). The input side contains an inlet address part and a word address part and the output side an outlet address part and a word address part. Inlet/outlet address selects one of the N inlet/outlet streams. Word address is used to access one of the M words in the data memory module corresponding to the control



CM = control memory DM = data memory

(a) Configuration



$x = \lceil \log_2 N \rceil \quad y = \lceil \log_2 M \rceil$

(b) Contents of control memory locations

Fig. 6.17 Parallel-in/parallel-out time switch.

memory module from which the word address is read. There is one-to-one correspondence between the control memory words and the time slots. The word i in each module contains information pertaining to time slot i. There is no such correspondence with data memory words.

The operation of the switch takes place in three phases in each time slot:

1. Control memory read
2. Data memory write
3. Data memory read.

Let us consider the operation of the switch corresponding to time slot 1. In the first phase, the first location of all the control memory modules are read. The inlet addresses are decoded and the respective inlets are gated to the MDR of the data memory modules. The word addresses of the input side are transferred to the respective MARs of the data memory modules. In the second phase, a *write* operation is performed in all the data memory modules simultaneously, the data contained in the MDRs are transferred to the respective locations. In the third phase, the output side word addresses are transferred to respective MARs and a *read* operation is performed in all the data memory modules. The MDRs now contain the data to be transferred to the required outlets. The outlet addresses from the control memory word are decoded to select the respective outlets and the data from the MDRs transferred. The time constraint may be expressed as

$$t_{TS} = 3t_m \tag{6.22}$$

It may be noted that Eq. (6.22) is independent of N and thus permitting any number of trunks to be supported theoretically. In practice, however, engineering considerations would limit the number of trunks. It is possible to organise the control memory in two sets of modules, one set for the input side and the other for the output side. The output side set may be read out during the second phase of operation described above simultaneously with the write operation of the data memory modules.

6.5 Combination Switching

We have seen that time multiplexed time division space switches do not provide full availability as they are not capable of performing time slot interchange. Time slot interchange switches are not capable of switching sample values across the trunks without the help of some space switching matrices. Therefore, a combination of the time and space switches leads to configurations that achieve both time slot interchange and sample switching across trunks. These structures also permit a large number of simultaneous connections to be supported for a given technology. A combination switch can be built by using a number of stages of time and space switches. We consider two-stage switches in this section, three-stage switches in Section 6.6 and n-stage switches in Section 6.7. A two-stage combination switch may be organised with time switch as the first stage and the space switch as the second stage or vice versa. Accordingly, the two switch configurations are known by the nomenclature **time-space (TS)** or **space-time (ST) switches** respectively.

A time-space switch is shown in Fig. 6.18. The first stage consists of one time slot interchanger per inlet and the second stage a $N \times N$ space switch.

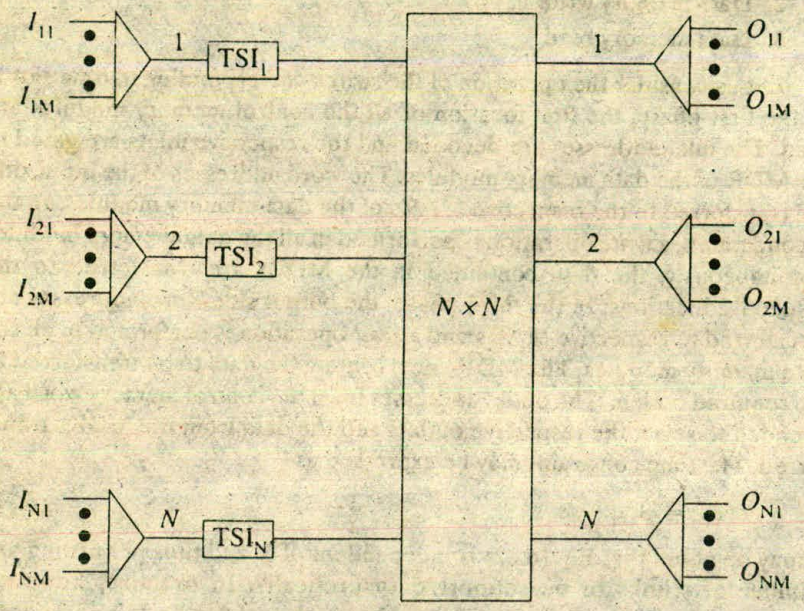


Fig. 6.18 Two-stage TS switch.

The control memories for the TSI and space switch are not shown in Fig. 6.18. Each time multiplexed inlet/outlet stream carries M channels. A subscriber on the input side is assigned to one of the inlets and a time slot in that inlet. An input subscriber assigned to line 4 at time slot 7 is identified by the label I_{47}. Similarly, a subscriber connected to the outlet 5 and time slot 6 is identified by O_{56}. The corresponding time slots are identified as IS_{47} and OS_{56}. Suppose that a communication is to be established between these two subscribers. The input sample from IS_{47} is first moved to IS_{46} at the output of TSI switch. During the time slot 6, a connection is established between the inlet 4 and the outlet 5 at the space switch. While this switch configuration ensures full availability, it is not nonblocking. Consider the two connections to be established between I_{47} and O_{29}, and I_{43} and O_{69}. Both the samples originate from the same inlet and are destined to the same time slot in different outlets. Both input samples require to be switched to time slot 9, which is not possible. Only one of them can be switched to slot 9. In general, blocking occurs if two inputs I_{ij} and I_{ik} are destined to outputs O_{pq} and O_{rq}. In other words, it is only possible to set up a single connection between any of the subscribers on an inlet and any of the subscribers connected to the same time slot on the output side.

Let the traffic intensity in erlangs per inlet trunk $= \rho$ E

Probability that a subscriber X is active $= \frac{\rho}{M}$

Probability that any other subscriber is active on the same inlet $= \frac{(M-1)}{M}\rho$

Probability that a particular outlet subscriber is chosen by X $= \frac{1}{MN}$

Probability that the same time slot on a different outlet is chosen by the other subscribers on the same inlet $= \frac{(M-1)\rho(N-1)}{M(MN-1)}$

Therefore, the blocking probability is given by

$$P_B = \frac{\rho}{M \times MN} \frac{(M-1)(N-1)}{M(MN-1)} \tag{6.23}$$

If $M >> 1$ and $N >> 1$, then Eq. (6.23) becomes

$$P_B = \frac{\rho}{M^3 N} \tag{6.23a}$$

Theoretically, the TS switch can be made nonblocking by using an expanding time switch and a concentrating space switch. In the worst case, all the subscribers in one line may want to establish connection to the same output time slot. In order that they can be accommodated without blocking, we need $(M-1)$ additional slots for each time slot. The space switch must be able to establish M connections for each input time slot. This implies M^2 time slots on the output side of TSI. A time switch providing expansion from M inlet time slots to M^2 outlet time slots is very expensive. Moreover, the space switch at the second stage has to concentrate the time slots on the output side. Design of such a space switch is complex. Instead, the space switch may be symmetric with regard to time slots and a concentrating TSI switch may be added as another stage in the network. This takes us on to three-stage networks which we discuss in Section 6.6.

A space-time (ST) switch configuration is shown in Fig. 6.19. Consider a connection between I_{47} and O_{69}. During the time slot 7, the input sample is switched from inlet 4 to outlet 6 by the space switch. It is then switched to time slot 9 by the time switch. Similar to a TS switch, this switch is also blocking. This happens when the input samples originate from two different inlets during the same time slot and are destined to the same outlet though to different time slots. As an example, consider I_{75} and I_{85} to be switched to O_{96} and O_{92} respectively. Only one of the inputs can be space switched

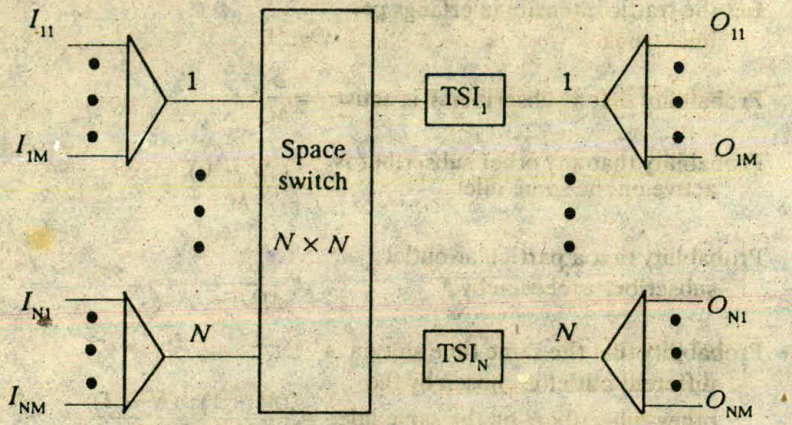


Fig. 6.19 A space-time switch.

during time slot 5 to the output line 9 and the other is blocked. The blocking probability calculations can be done in a manner similar to the one given for TS switch. The performance of ST and TS switches are identical. Since two switches are in series in these configurations, the time constraint is the sum of the two constraints specified by Eqs. (6.15) and (6.20) and is given by

$$125 = 2NMt_m + NMt_m = 3NMt_m \tag{6.24}$$

In writing Eq. (6.24), it is assumed that $t_s = t_m$ in Eq. (6.15).

6.6 Three-Stage Combination Switching

Three-stage time and space combination switches are more flexible than their two-stage counterparts. The most common three-stage configurations are:

- Those which place time stages on either side of a space stage giving rise to TST configuration
- Those which place space stages on either side of a time stage giving rise to STS configuration.

A TST network is shown in Fig. 6.20. The two time stage exchange information between external channels and the internal space array channels. The first flexibility that becomes obvious in this arrangement is that there is no need to have a fixed space stage time slot for a given input or output time slot. For example, IS_{47} may be moved to OS_{96} via an intermediate time slot 5. An incoming channel time slot may be connected to an outgoing channel time slot using any possible space array time slot. There are many alternative paths between a prescribed input and output unlike a two-stage network which has only one fixed path. This factor reduces the value of the blocking probability of a three-stage combination switch.

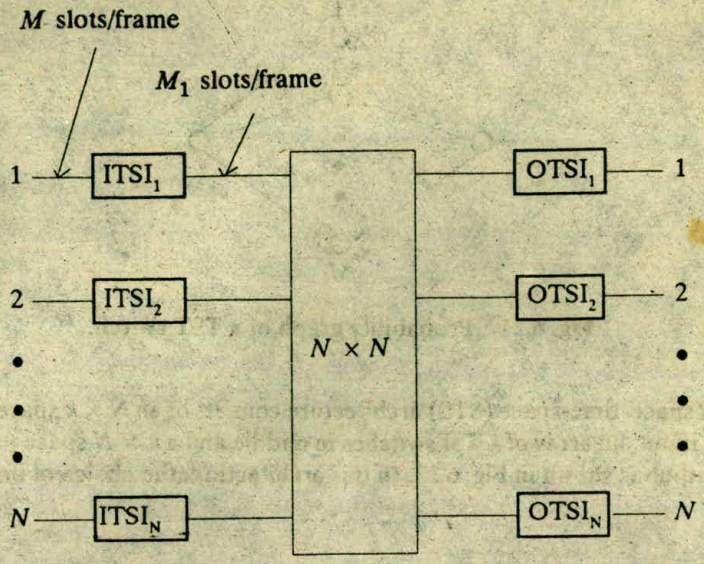


Fig. 6.20 A time-space-time switch.

Three-stage combination switches are still blocking. Consider a situation when $(M-1)$ slots in an inlet I_j are all busy. Let traffic arrive in the M-th slot destined to a time slot outlet O_k. It is possible that during the time slot M, the outlet O_k is busy receiving some other output. As a result, blocking occurs. The worst case happens when $(M-1)$ slots of O_k are busy and the outlet is occupied for this purpose during $(M-1)$ slots when the inlet I_j is free. If this happens, we would need one more additional time slot to establish a connection between the free time slots of I_j and O_k. This means that we need a total of $(M-1)+(M-1)+1$ time slots, i.e. $2M-1$ time slots in the intermediate space stage. Readers are advised to note the similarity of consideration in the case of three-stage space division networks in Section 4.8.

In a functional sense, a TST switch is identical to a three-stage space division network. An equivalent Lee's probability graph of a TST switch is shown in Fig. 6.21. The general expression for the blocking probability of a TST switch is similar to Eq. (4.20) and is given by

$$P_B = \left(1-(1-\alpha/L)^2\right)^{M_1} \tag{6.25}$$

where

M_1 = number of time slots on the output side of the TSI switch

L = expansion or concentration factor = M_1/M

α = traffic intensity on an inlet

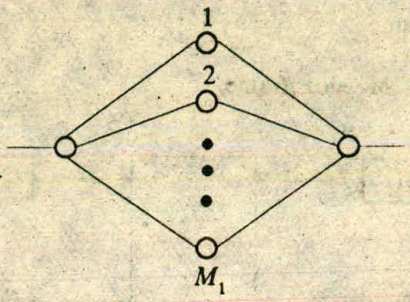


Fig. 6.21 Probability graph of a TST switch.

A space-time-space (STS) architecture consists of an $N \times k$ space matrix at the input, an array of k TSI switches in middle and a $k \times N$ space matrix at the output as shown in Fig. 6.22. In this architecture, the choice of input and

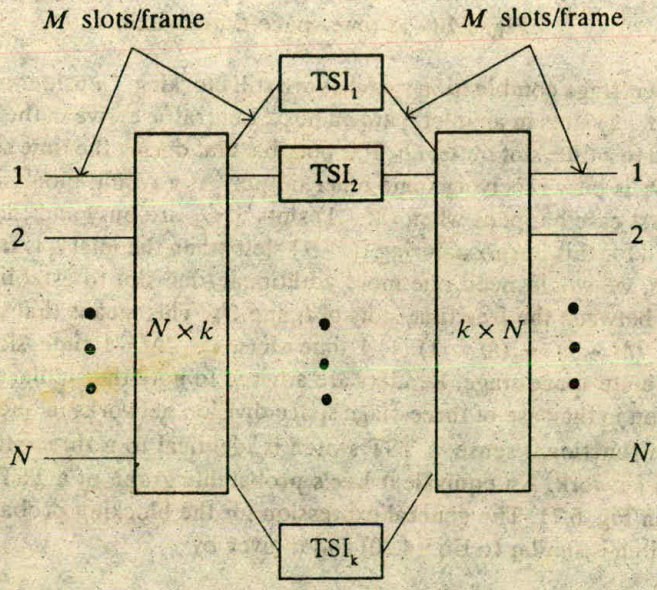


Fig. 6.22 A STS switch.

output time slots is fixed for a given connection. But the flexibility is provided by the ability to utilise any free TSI switch by space switching on the input and the output side. There are as many alternative paths for a given connection as there are TSI switches. For example, I_{79} may be connected to O_{84} by passing through TSI_3. The sequence is as follows:

$$I_{79} - IS_{79} - IS'_{39} - OS'_{34} - OS_{84} - O_{84}$$

where IS' and OS' represent the input and output time slots of the TST switch. The expansion and concentration take place at the space switch level and not at the time slot level. The time slots are symmetrical throughout the switch. Based on a reasoning similar to the one used for TST switch, a STS switch is nonblocking if $k = (2N - 1)$. The blocking probability is governed by an expression similar to the one given in Eq. (6.25).

The cost comparison of a STS switch and a TST switch shows up some interesting points. Equation (6.15a) gives the cost of a space switch that uses $2(N \times 1)$ space matrices. Modifying this for a space switch that uses a $N \times k$ space matrix, we have the cost as

$$C_s = Nk + MN \tag{6.26}$$

Therefore, the cost of the STS switch is

$$C_{STS} = 2 \times (Nk + MN) + 2MN$$

$$= 2Nk + 4MN \tag{6.26a}$$

From Eqs. (6.18a) and (6.26), we have the cost of a TST switch as

$$C_{TST} = 2N(2M) + (N^2 + MN) = 5MN + N^2 \tag{6.27}$$

Please note that a TST switch uses N TSI switches. We now compare Eqs. (6.26a) and (6.27) for different values of M and k, keeping N fixed. For concentrating switches we have $M_1 < M$ for TST and $k < N$ for STS switch. For $M = k = N/2$, we have

$$C_{STS} = 3N^2, \quad C_{TST} = 3.5N^2$$

For symmetric switches, $M = k = N$; then

$$C_{STS} = 6N^2, \quad C_{TST} = 6N^2$$

For expanding switches, $M_1 > M$ and $k > N$, say $M_1 = k = 2N$; then

$$C_{STS} = 12N^2, \quad C_{TST} = 11N^2$$

The above calculations bring out that for small switches STS architecture is to be favoured and for large switches, TST configurations are better.

Switches are designed to be concentrating when the utilisation of the input links is low. As the input traffic intensity increases, less and less concentration is acceptable. When the input loading becomes sufficiently high, space expansion in the STS switch and time expansion in the TST switch are required to maintain low blocking probabilities. Calculations shown above indicate that time expansion is cheaper than space expansion. Hence, TST architecture is more cost effective than STS architecture for higher loads. There are, of course, other factors like modularity, testability and maintainability which must also be taken into account before deciding on a particular architecture.

6.7 *n*-Stage Combination Switching

Very large time division switches, supporting 40,000 lines or more, can be economically designed by using more than three stages of time and space combination switching. We call this *N*-stage switching where *N* is greater than three. *N*-stage combination networks may also be designed to provide better flexibility and less blockage. Basically, two approaches are seen in designing *N*-stage switches:

- Expanding a TST switch with additional space stages
- Modular design using a number of time-space (TS) modules.

Typical configurations include TSST, TSSSST and TSTSTSTS. Examples of practical combination switches with their salient features are presented in Table 6.1.

Table 6.1 Examples of Time-Space Combination Switches

System	Configuration	Traffic capacity (erlangs)	Maximum number of trunks
E10 B (France)	TST	1600	3600
No.4 ESS (USA)	TSSSST	47,000	107,520
C-DOT MAX-XL (India)	TST	16,000	40,000
DMS 100 (Folded) (Canada)	TSTS	39,000	61,000
System 12 (USA)	TSTSTSTSTSTSTS	25,000	60,000
NEAX 61 (Japan)	TSST	22,000	60,000

When the space stage of a TST switch is large enough to justify additional control complexity, multiple space stages can be used to reduce the total crosspoint count. As an example of this approach, we consider the No.4 ESS switch which has a TSSSST architecture. The No.4 ESS was developed by Bell Laboratories and was the first digital switch in the U.S. telephone network when it became operational in 1976. Even today, it is one of the largest capacity switches developed in the world (see Table 6.1). It is capable of supporting a maximum of 107,520 voice channels with a blocking probability of less than 0.005 at a traffic intensity of 0.7 E per channel. The No.4 ESS is designed to be a transit or tandem exchange providing significant

cost savings by being capable of replacing several of the largest space division switches available previously, in particular No.4A crossbar systems.

A block diagram of the different time and space stages in No.4 ESS is shown in Fig. 6.23. The interface between the No.4 ESS and the transmission systems is based on analog 12 or 24 group trunks which carry 120 time slots each (12 × 10 or 24 × 5). Internally, the No.4 ESS has design parameter values that are powers of two to take advantage of the efficiencies allowed by binary digital systems. The time stage uses 128-time slot TSI switches. The eight extra slots available internally are used for diagnostic and maintenance purposes.

The time stages at the input and output consist of 128 time modules each. A time module comprises eight of 128 time slots each. Thus we get TSI switches giving a total capacity of 128 × 8 × 128 = 131,072 voice channels. However, the capacity is limited by the incoming trunks, each of which carries only 120 channels. There are seven trunks input to each time module giving a maximum capacity of 7 × 120 × 128 = 107,520 terminations. In order to ensure low blocking, a **decorrelator** is used, which performs an expansion function by spreading the traffic from seven 120-time slot streams to eight 128-time slot streams. The effect of this is that only 105 time slots are active in the 128-time slot streams. The distribution/expansion function is performed by using a fixed wiring pattern. There is a corresponding **recorrelator** at the end of the network to concentrate traffic from eight lines to seven lines.

The first and the fourth space stages consist of 128, 8 × 8 space switches each. The eight TSI switch outputs are connected to the eight inputs of the 8 × 8 space switch in the second stage of the network. The third and fourth stages of the network, i.e. second and third space stages, contain four, 256 × 256 space matrices each. These 256 × 256 space matrices are built using 16, 16 × 16 space arrays. With 128 time slots in one 125-μs frame, each time slot has a duration of 976 ns. During every time slot the four-stage space switch is clocked 16 times giving 16 different possible paths to each of the time slot sample. The clock duration works out to be 61 ns requiring a basic clock frequency of 16.384 MHz. With 128 possible time slots and 16 space matrix paths for each sample, there are 2048 alternative paths available for each connection in the network, ensuring a low blocking probability even at high loads.

In order to achieve high system availability, reliability and maintainability, hot standby is provided for the entire switching network. Comparators are provided in a number of appropriate points between the duplicated units to detect faults and aid diagnostics. If any path fails, system control can switch to the hot standby unit which maintains the identical path configuration details. Faulty unit can be repaired while the services are maintained by the functional unit. All memory modules are provided with parity checking feature for both address and data. Two additional bits are

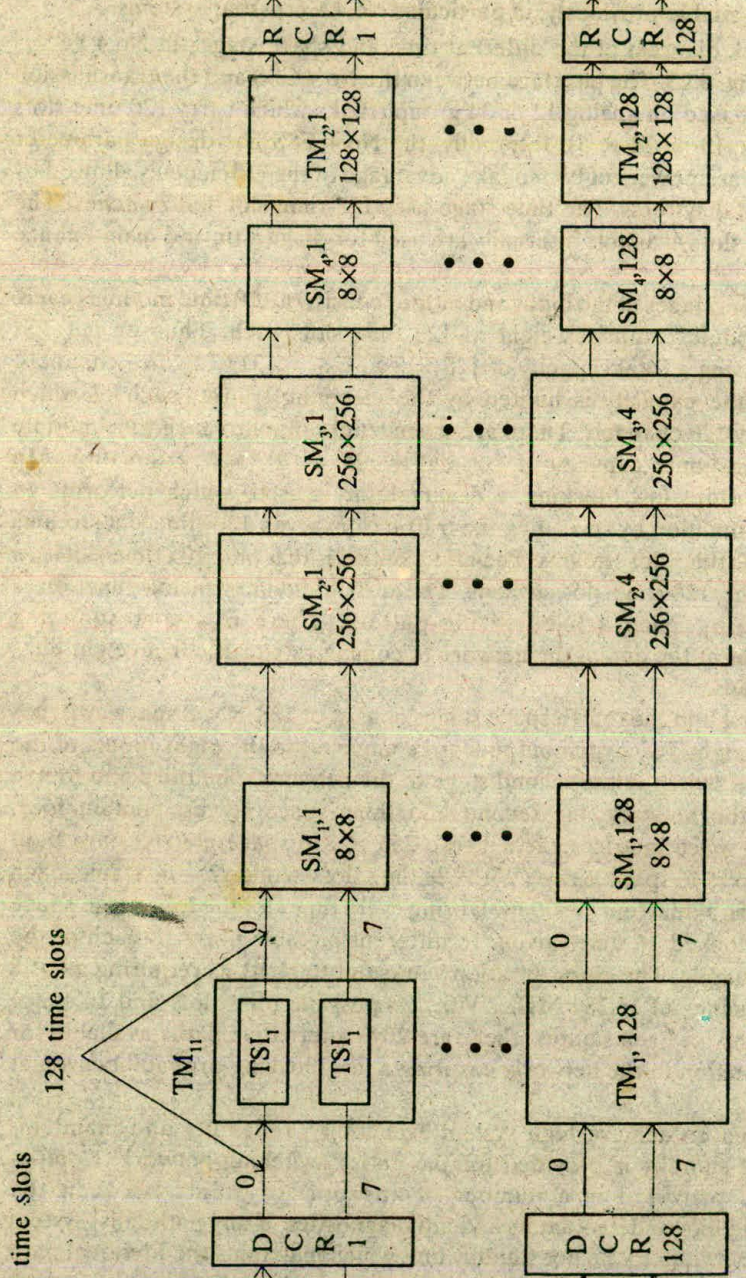


Fig. 6.23 No.4 ESS time and space stage.

used with 8-bit PCM samples, one for enabling space matrices to be switched between time slots and the other as parity bit.

In designing *n*-stage combination systems, a modular approach has been adopted by some manufacturers. The interest here has been to use one type of module containing both time and space switches for all stages. Main advantages of the modular approach include the ability to implement a wide range of switch sizes, simplified manufacturing and maintenance and reduced redundancy and spare parts inventory costs. A major disadvantage is the introduction of potentially long delays due to too many stages in the switch. Propagation delays of the order of a few milliseconds are not as such detrimental to voice traffic but can cause 'singing' on local connections. Singing is discussed in Section 9.3. An example of the modular design is the System 12 shown in Table 6.1.

System 12 ESS is a new generation of digital switching equipment designed by ITT, USA to fulfill the needs of a futuristic network concept known as Network-2000. The futuristic network is expected to meet the requirements of ISDN incorporating new device technology with wideband media like optical fibres. Under System 12 four models, viz. ITT-1210, ITT-1220, ITT-1230 and ITT-1240 have been developed. The model ITT-1210 is designed to provide end-to-end (subscriber end) digital connectivity. The model ITT-1210 is a toll exchange with medium to large capacity. The ITT-1240 is the most versatile digital exchange in the System 12 series. It exploits the modular structure to a great extent and can grow up to a capacity of 100,000 lines. The number of switching stages in the network varies with the number of terminals to be connected. A switching network stage is realised using a standard T-S switching module in as many numbers as required. Each T-S module consists of 16 identical LSI switch ports, eight input and eight output ports, which function as TSI switches. These TSI switches are interconnected by a 39-line parallel TDM bus which enables information to be transferred from any of the input channels to any of the output channels. Thus, the LSI switch ports and the TDM has together act as aT-S switch module. The T-S modules are laid out in the form of a plane and multiple planes as required are used to make up a switching network.

The design of ITT-1240 uses a fully distributed control with considerable intelligence at the end points of the network. The line and trunk interface units are microprocessor controlled and share many of the functions of the SPC processor. The operation, administration and maintenance (OA&M) facilities of the exchange are software controlled. The OA&M functions provide for day-to-day supervision, automatic fault location and recovery, and exchange extension and reconfiguration. A wide range of measurement facilities are incorporated to evaluate performance and help in network management. A centralised OA&M centre, ITT-1290 is also available for remote OA&M of a group of System 12 exchanges.

FURTHER READING

1. Andre, G. *et al.*, Digital switching in France, *IEEE Transactions on Communications,* Vol. Com-27, 1979, pp1047-1055.
2. Briley, B. E., *Introduction to Telephone Switching,* Addison-Wesley, Reading (Mass.), 1983.
3. Joel, A. E. (ed.)., *Electronic Switching : Digital Central Office Systems of the World,* IEEE Press, New York, 1981.
4. McDonald, J. C. (ed.), *Fundamentals of Digital Switching,* 2nd ed., Plenum Press, New York, 1989.
5. Special Issue on ITT-1240 Digital Exchange, *Electrical Communication*, Vol. 56, No. 2/3, 1981.
6. Special Issue on No.4 ESS, *Bell System Technical Journal*, Vol.56, No.7, Sept. 1977.
7. Special Issue on System 12 Digital Exchanges, *Electrical Communication,* Vol. 59, No.1/2, 1985.
8. Sueyoshi, H., *et al.*, System design of digital telephone switching system – NEAX 61, *IEEE Transactions on Communications,* Vol. Com-27, 1979, pp 993-1000.

EXERCISES

1. An analog nonfolded time division space switch is to be designed to support 32 inlets and outlets each. Assuming bidirectional traffic,
   (a) What is the time duration that is available for exchanging samples for one connection?
   (b) What is the size of the control memory if the switch is output controlled?
   (c) What is the clock rate?
2. How many subscribers can be supported in bidirectional PAM switching bus, if the pulse width of the PAM sample is 125 ns?
3. A PABX is designed using an output-controlled digital time division space switch for supporting 128 subscribers. The transmission between the subscribers and the exchange is analog. On an average, twenty-five per cent of the subscribers are active simultaneously. Determine the switch parameters in terms of
   (a) Space matrix size
   (b) Operating speed of the space matrix
   (c) Control memory configuration and size
   (d) Clock speed

(e) A–D and D–A converter configuration
(f) Multiplexer/demultiplexer sizes.

4. For the specifications given in Problem 3 for PABX, compare the cost of output-controlled and memory-controlled configurations.

5. Work out the different parameter values if a time division time switch were to be used in Problem 3.

6. A 1000-inlet and 1000-outlet digital switch is to be built using TSI. Determine the size of the control and data memories, and the speeds with which the memories have to be accessed.

7. In a TS switch, $M = 128$, $N = 16$ and the number of subscribers connected to the system is $0.1\,MN$. Determine the blocking probability of the switch if
   (a) all the subscribers are active at the same time,
   (b) only 50% of the subscribers are active simultaneously.

8. Four 32-channel TDM streams are multiplexed to form one input stream of an 8-stream TS switch. Determine the cost of the configuration including that of the multiplexers.

9. Draw the Lee's probability graph for ST switch. What is the blocking probability if the input channel loading is 0.2 erlang (E) and space and time stages are symmetrical with $N = 8$ and $M = 32$.

10. Determine the maximum and minimum delay that may be experienced by an input sample in a TTTT switch if the operating mode for all the switches is
    (a) phased mode
    (b) time slotted mode.

11. Design a STS switch for supporting 128 TDM trunks each carrying a primary CCITT channel. The blocking probability should be less than 0.002. The loading is 0.2 E per channel. What is the cost of the switch?

12. A TST switch supports 32 trunks of 32 channels each. A time expansion/concentration factor of 2 and a single-stage space switch are used. What is the blocking probability of the switch if the channel loading is 0.9 E per channel? Determine the cost of the switch.

13. Derive an expression for the blocking probability of a TSTS switch if each stage is individually nonblocking.

14. How many alternative paths are available in a TSST switch if the space stages are capable of eight switchings per input time slot and the first and second time stages provide an expansion and a concentration factor of 2 respectively? Assume 32 time slots.

15. A TSSTSST switch supports 1024 trunks each carrying 128 channels. Determine the basic clock rate of the switch if the input channels were to have 4096 alternative paths. The T stages are symmetrical. Is any useful purpose served by the middle T stage? Explain.

16. In n-stage combination switching a trade-off between blocking probability and time delay is possible. Explain.

17. Is TS nework nonblocking? Explain.

7

Optical Fibre Systems

Fibre optic transmission medium is fast emerging as an alternative and strong competitor to coaxial cables in telecommunication networks. Optical fibres have several advantages over the electrical cables, but are not without problems. The merits of the optical fibres stem from the fact that the basic material used in their construction is nonmetallic and electrically non-conductive. The problems are mostly due to their delicate structure.

The conductive metallic paths of electrical cables act as antenna and pick up electromagnetic radiation from radio frequency sources, power line currents and other industrial machinery. They also give out electromagnetic radiation when carrying high speed signals, and thus produce cross-talk in cables running adjacently. Long distance data transmission in electrical cables suffers from ground loop problems. Electrostatic discharges also find a path in the metallic conductors, thereby destroying voltage sensitive circuit devices connected to these conductors. Lastly, the skin effect associated with electrical cables considerably reduces the bandwidth of the signals that can be carried over them.

In contrast, the nonmetallic and totally dielectric fibre optical cables are immune to radio frequency and other electromagnetic interferences. Ground loop and common mode voltage problems do not exist and data can be transmitted between points of vastly different potential. In optical cables the information is transmitted by packets of photons which have no charge and hence do not lead to electrostatic discharges. There is no possibility of sparks or short circuit when a fibre is cut. The bandwidth-distance product of a fibre optic cable is very high compared to that of the electrical cables. The standard RG-58 coaxial cable has a bandwidth-distance product of 300 kHz-km. On the other hand, the bandwidth -distance product of a typical optical fibre is about 100 MHz-km. This factor lends itself to long distance high speed transmission without repeaters. Signal transmission in fibres is essentially a waveguide communication with the optical flux contained totally within the fibre waveguide. Therefore, no electromagnetic radiation is possible from the optical fibres.

Fibre cables are about the thickness of a human hair. Any dirt obstructing the optical port causes poor transmission. The thin dimension

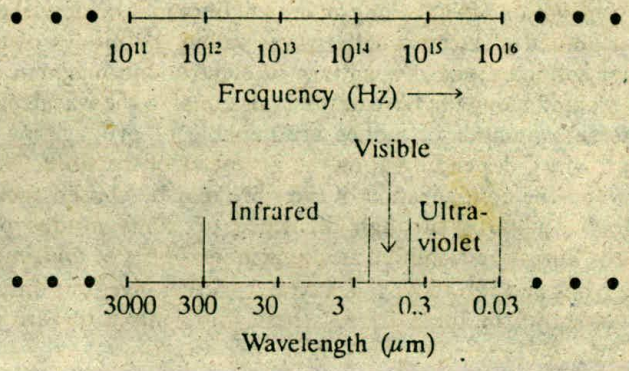
results in a low weight for a given length when compared to electrical cables. However, being thin and somewhat brittle in nature, fibres tend to break easily if bent beyond a certain limit. While the fibre cables are safe from electrical discharge and shock, a direct viewing into the optical point can be harmful to the eyes.

7.1 Types of Optical Fibres

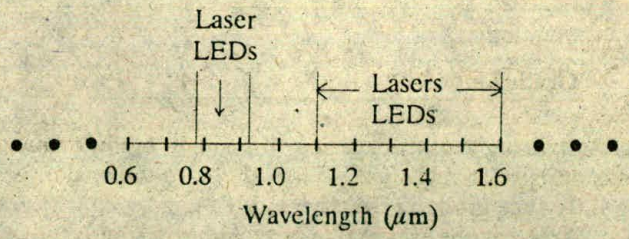
Functionally, an optical fibre is a dielectric waveguide capable of carrying light energy. A waveguide confines electromagnetic energy to within its surfaces and guides the energy in a direction parallel to its axis. An optical waveguide operates at optical frequencies which lie in the range 10^{12}–10^{16} Hz covering infrared, visible and ultraviolet energy spectrum as shown in Fig. 7.1(a). In this entire optical energy spectrum, there is a region where the present day optical fibres function efficiently. The transmission characteristics of optical fibres are dictated by their structural features and material properties. For example, impurities in fibre material may result in extensively large losses of more than 1000 dB/km. In fact, this was the case with glass fibres until 1970, when Kapron, Keck and Maurer of Corning Glass Works, UK fabricated a fibre with high purity glass, giving an attenuation of 20 dB/km. It was with this achievement that a whole new era of optical fibre communications was launched. By reducing the concentrations of hydroxyl ions and metallic ion impurities in the fibre material, manufacturers are now able to produce optical fibres with very low losses in the wavelength region of 600–1600 nm. The low loss region is depicted in Fig. 7.1(b).

Physically, an optical fibre is composed of three layers: a core, a cladding and a protective coating as shown in Fig. 7.2. In low and medium loss fibres, the core material is generally glass and the cladding is either glass or plastic. Fibres which use plastic for core have higher losses than the ones that use glass. Being economical, the fibres with plastic core and cladding are used in many applications. In principle, cladding is not necessary for light to propagate along the core. But it reduces scattering losses due to discontinuities at the surface of the core and protects the core from absorbing surface contaminants. The protective coating is an elastic, abrasion resistant plastic material. The protective encapsulation gives mechanical strength to the fibre and prevents random microscopic bends that can arise in the core due to geometric irregularities or roughness of the supporting structures. Microscopic bends and other perturbations in the core cause scattering losses.

Although different configurations are possible for the core, the most widely accepted structure is the single solid dielectric cylinder. The refractive index of the cladding is less than that of the core. Depending on the material used, the refractive index of core may be uniform or it may gradually decrease from the centre towards the surface of the core. If the index is uniform in the core, there occurs an abrupt change in the refractive



(a) Optical frequency spectrum



(b) Low loss region

Fig. 7.1 Energy spectrum for optical fibres.

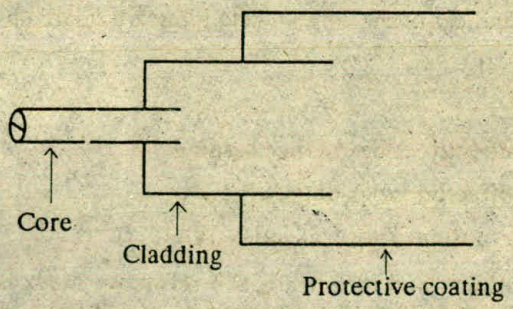


Fig. 7.2 Structure of an optical fibre.

index value when we move from the core to cladding or vice versa, i.e. at the core cladding interface. Signifying this abrupt change, these fibres are known as **step index fibres. When the refractive index changes gradually, the fibres are known as graded index fibres.** The cladding has generally a uniform refractive index.

The propagation of light along an optical fibre may be described in terms of a set of guided electromagnetic waves called the modes of the optical fibres. Each guided mode is a pattern of electric and magnetic field lines which is repeated along the fibre at intervals equal to the wavelength. Only a certain discrete number of modes are capable of propagating along the waveguide (fibre), depending upon the geometry of the fibre. The physical dimension and the configuration of the fibre may be chosen such that only one mode propagates through the fibre or many modes propagate. Accordingly, fibres are classified as **single mode fibres** or **multimode fibres**.

Mode types and refractive index types together give four different types of fibres, as shown in Table 7.1. Of these four, only three are practically

Table 7.1 Fibre Types

Fibre	Single mode	Multimode
Step index	√	√
Graded index	X	√

realisable. Restricting the propagation to a single mode demands that the core dimension is relatively small. When the core dimension is small, it is difficult to introduce graded refractive index. As a result, the graded index single mode fibres are not practicable as technology stands today.

The physical dimensions and the values of the refractive indices for commonly used telecommunication fibres are shown in Fig. 7.3. The core diameter used for single mode fibres lies in the range of 8–12 μm, and the diameter of the cladding lies in the range of 100–125 μm. A multimode fibre has a core diameter of 50–200 μm and a cladding diameter of 125–400 μm. The refractive indices of the core and cladding are related by the expression

$$n_2 = n_1(1 - \Delta) \tag{7.1}$$

where

n_2 = refractive index of the cladding

n_1 = refractive index of the core

Δ = a constant known as index difference

Beyond the cladding and the sheath, the refractive index is that of the air which has a value 1.0. From Eq. (7.1) the core-cladding index difference is given as

$$\Delta = \frac{n_1 - n_2}{n_1} \tag{7.2}$$

In the case of a graded index fibre, the most commonly used law for varying the refractive index from the centre to the surface of the core is a power law which is stated as

$$n(r) = \begin{cases} n_1[1 - 2\Delta(r/a)^{\alpha}]^{\frac{1}{2}} & \text{for } 0 \le r \le a \\ n_1(1 - \Delta) = n_2 & \text{for } r > a \end{cases} \tag{7.3}$$

where

r = radial distance from the fibre axis

a = radius of the core

α = a dimensionless quantity which defines the shape of the index profile

When $\alpha = \infty$, Eq. (7.3) reduces to the step index profile of $n(r) = n_1$ for $0 \le r \le a$.

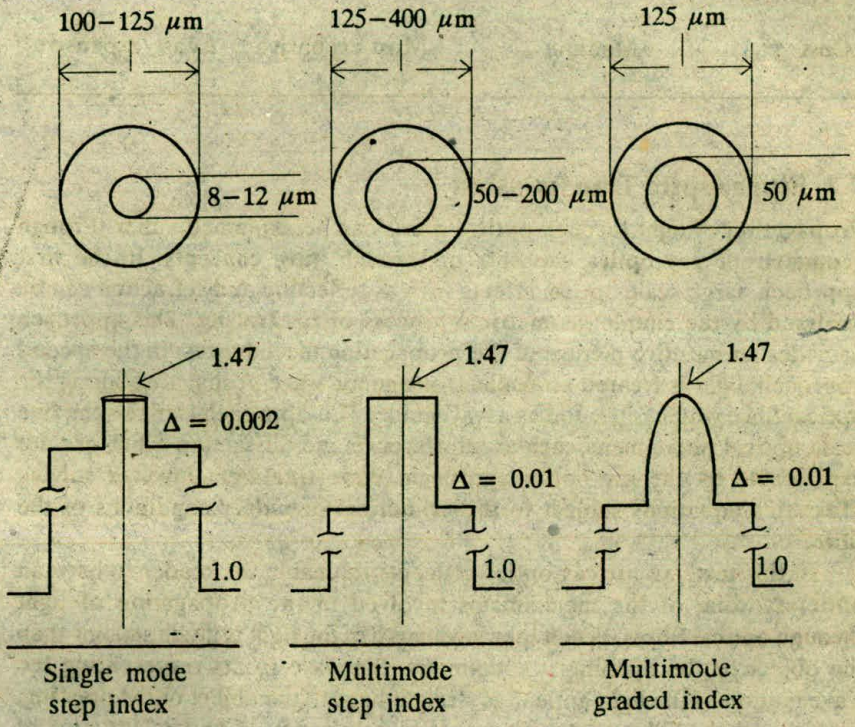


Fig. 7.3 Fibre configurations and dimensions.

Table 7.2 enumerates some of the characteristics, advantages, constraints and applications of the three types of fibres. A single mode fibre offers much higher bandwidth distance product than a multimode fibre. But, the small core dimension poses problems in launching the power into the core and in splicing. Only sharp beam sources, like lasers, can launch adequate power into single mode fibres. Such sources are more expensive than, say, LEDs. Multimode step index fibres are the least expensive ones and are widely used in many applications.

Table 7.2 Characteristics of Fibre Types

	Single mode	Multimode graded index	Multimode step index
Source	Laser	Laser or LED	Laser or LED
Bandwidth-distance product	> 3 GHz-km	200 MHz-km to 3 GHz-km	< 200 MHz-km
Splicing	Very difficult	Difficult	Relatively easy
Application	Under sea cables	Intercity trunks	Data link
Cost	Medium	Most expensive	Least expensive

7.2 Fibre Optic Transmission

Propagation of light through optical fibres can be explained either through **geometric or ray optics concepts** or through **wave concepts**. In the first approach, large scale optical effects such as reflection and refraction can be analysed by the simple geometrical process of ray tracing. This approach provides an intuitive picture of the propagation mechanisms. In the second approach, light is treated as an electromagnetic wave propagating along the optical fibre which is treated as a waveguide. This approach explains the fine scale optical phenomena such as interference and diffraction which are not explainable by the ray optics approach. Wave treatment involves solving Maxwell's equations subject to the cylindrical boundary conditions of the fibre.

In this text, we use ray optics concepts to enable the reader to have an understanding of the mechanisms involved in the propagation of light through optical fibres. When the wavelength of the light is much smaller than the object or the opening it encounters, the wave fronts representing the wave motion of the light appear as straight lines to the object or the opening. In this case, the light can be represented by a plane wave and its direction of travel by a light ray drawn perpendicular to the wave front. In fibre optic transmission, the wavelength of light is around 1 μm and the core diameter around 10 μm for single mode fibres and 50 μm for multimode fibres. Consequently, ray optics treatment of the propagation of light through optical fibres provides fairly accurate results.

As we know, the speed of light c in free space is related to the frequency f and the wavelength λ by the equation

$$c = f\lambda \tag{7.4}$$

where c has a value of 3×10^8 m/s. In a dielectric or a nonconducting medium, the light travels at a lower speed v than c. The speeds v and c are related through the refractive index n as

$$n = \frac{c}{v} \tag{7.5}$$

Obviously, for air, $n = 1$ as $c = v$. Typical refractive index values for some dielectric materials are shown in Table 7.3.

Table 7.3 Typical Refractive Indices

Material	Refractive index n
Air	1.0
Water	1.33
Glass	1.5–1.9
Diamond	2–2.42
Silicon	3.4
Gallium Arsenide	3.6

When a light ray encounters a boundary of two materials with different refractive indices, part of the ray is reflected back into the first medium and the rest passes through the second material refracted, i.e. with the direction of travel changed. This is illustrated in Fig. 7.4. The angle the ray makes with

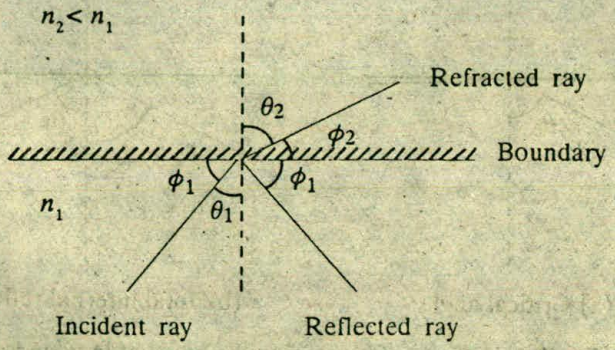


Fig. 7.4 Behaviour of light ray encountering material boundary.

the normal to the plane of incidence, i.e. θ_1, is known as the **angle of incidence**. Similarly, angle θ_2 is known as the **angle of refraction**. Although angles of incidence and refraction are conventionally defined in this fashion, some authors define ϕ_1 and ϕ_2 as angles of incidence and refraction respectively. These angles are related by Snell's law which defines the relationship as

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \tag{7.6}$$

or, equivalently,

$$n_1 \cos \phi_1 = n_2 \cos \phi_2 \tag{7.7}$$

The reflected ray is such that the angle of reflection is equal to the angle of incidence. Figure 7.4 shows the behaviour when light passes from a denser medium to a rarer medium, i.e. $n_2 < n_1$. It should be noted that angle θ_1 is smaller than angle θ_2.

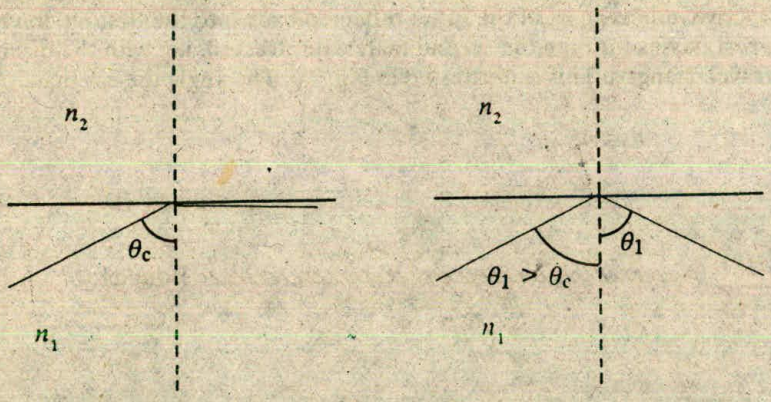
EXAMPLE 7.1 Determine the angle of refraction when a light ray passes from glass to air and the angle of incidence is 30°.

Solution From Eq. (7.6),

$$\sin \theta_2 = (n_1/n_2) \sin 30° = 1.5 \times 0.5 = 0.75$$

Therefore, $\theta_2 = 48.6°$.

As the angle of incidence θ_1 becomes larger, the refracted angle θ_2 approaches 90°. When this happens, the angle of incidence is known as **critical angle** θ_c. At this point, the light ray in the rarer medium is parallel to the surface of the denser medium, as shown in Fig. 7.5(a). We may now write



(a) Critical angle (b) Total internal reflection

Fig. 7.5 Effects of increasing the angle of incidence.

the expression for the critical angle θ_c in terms of Snell's law as

$$n_1 \sin \theta_c = n_2 \sin 90°, \text{ i.e. } n_1 \sin \theta_c = n_2$$

Therefore,

$$\theta_c = \text{arc sin } (n_2/n_1) \tag{7.8}$$

If the angle of incidence is further increased beyond θ_c, no refraction is possible and the light rays become totally internally reflected as shown in Fig. 7.5(b). In practice, there is always some tunnelling of optical energy

through the interface. Such a fine-scale phenomenon can only be explained in terms of the wave theory which is beyond the scope of this text. For our study, we assume that the entire light energy is reflected back into the denser medium.

EXAMPLE 7.2 Determine the critical angle for the glass-air interface.

Solution From Eq. (7.8), we have

θ_c = arc sin (0.67) = 62.5°

Any light in the glass incident on the interface at an angle greater than 62.5° is totally reflected back into the glass.

Let us now see, how ray optics can be qualitatively used to represent propagation in a multimode fibre. A guided mode in a fibre travelling along the axis is composed of a family of plane waves which, when superimposed, form the particular mode. Since any plane wave can be represented by a light ray, the family of plane waves can be represented by a **ray congruence**. Each ray of a ray congruence travels in the fibre at the same angle with respect to the fibre axis. In a multimode fibre, there is a finite number M of discrete guided modes. Correspondingly, there are M ray congruences.

A simple ray picture suggests that any ray which has an angle of incidence greater than the critical angle can propagate through the fibre. In other words, there can be infinite modes or ray congruences through a fibre. However, the modes are limited to a finite number on account of the phase differences and wave cancellation experienced by the guided waves as they travel along the fibre. Consider the light ray shown in Fig. 7.6. Phase changes occur on two counts:

- Reflection at the boundary points
- Travel through the fibre.

Without going into the derivations, we state the expressions for the phase change due to reflections, δ_r, and the phase change due to travel along the

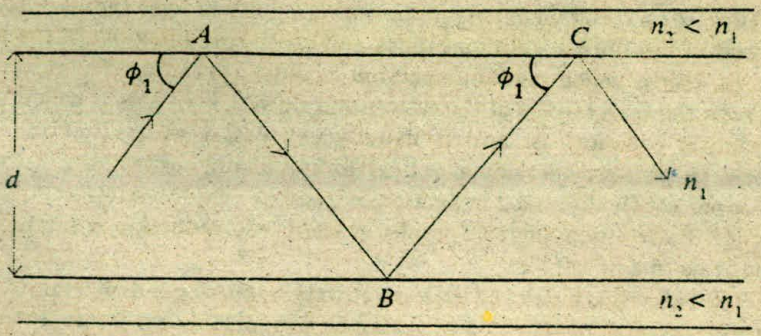


Fig. 7.6 Ray propagation in the fibre.

fibre δ_t. For simplicity, we consider only plane waves that are polarised normal to the reflecting surface. Phase change due to reflections is given by

$$\delta_r = 2 \arctan \frac{n^2(\cos^2\phi_1 - 1)^{\frac{1}{2}}}{n \sin \phi_1} \tag{7.9}$$

where $n = n_1/n_2$.

Phase change due to travel along the fibre is given by

$$\delta_t = n_1 ks \tag{7.9a}$$

where

k = propagation constant in free space

s = distance travelled by the wave along the ray

In order for a plane wave to propagate along the fibre, the phase of the twice reflected wave must be the same as that of the incident wave. In Fig. 7.6, the phase must be the same at points A and C. If this is the case, the wave superimposes constructively with itself. Otherwise, the wave interferes with itself and dies out. Therefore, the phase shift between the points A and C must be an integral multiple of 2π for proper propagation through the fibre. There are two refraction points, A and B, and a travel of two distances, AB and BC, between the points A and C. The distance $AB = BC = d/\sin\phi_1$. Therefore, from Eqs. (7.9) and (7.9a) the total phase shift is obtained as

$$2\pi M = 2\delta_r + \frac{2n_1 kd}{\sin \phi_1} \tag{7.10}$$

where M is an integer that determines the allowed ray angles for wave-guiding.

In multimode fibres, light rays belonging to different modes take different paths for travel and hence take different times to reach the destination end of the fibre. A light ray that propagates straight down the axis of the fibre takes the least amount of time, and a light that strikes the core-cladding interface at the critical angle takes the longest time to reach the other end of the fibre. As a result, light rays representing a pulse of light energy reach the far end of the fibre at different times and cause spreading out of the pulse energy. This is called **modal dispersion** or **pulse spreading**. The difference between the travel times of the fastest and slowest light rays is called **pulse spreading constant** Δt and is usually expressed in nanoseconds per kilometre. Modal dispersion in optical fibres is analogous to phase distortion in copper cables discussed in Section 5.7, and can lead to intersymbol interference. Pulse spreading also results in amplitude reduction at the far end, giving rise to a power loss.

The refractive index of a material is wavelength dependent. If some light energy consisting of different wavelengths is launched into a fibre, the different wavelength components will travel through the fibre at different speeds,

resulting in signal distortion. This is known as **wavelength dispersion** or **chromatic distortion**. This distortion can be eliminated by using a monochromatic light source such as an injection laser diode discussed in Section 7.3.

The intersymbol interference caused by pulse spreading and the consequent errors in digital transmission can be minimised by using suitable line coding schemes for transmission. Various line coding schemes have been discussed in Section 5.8. Unipolar return-to-zero coding (UPRZ) can tolerate twice as much pulse spread as unipolar nonreturn-to-zero coding (UPNRZ). Consequently, the maximum bit rate for UPRZ is twice the maximum bit rate for UPNRZ. The maximum bit rates are given by

$$B_{RZ} = (\Delta t \times L) \text{ bps}, \qquad B_{NRZ} = \frac{\Delta t \times L}{2} \text{ bps} \tag{7.11}$$

where L is the fibre length.

7.3 Optical Sources

Two principal light sources are used in fibre optic communication systems: **light emitting diodes (LEDs)** and **injection laser diodes (ILDs)** or, simply, **laser diodes**. Another less known source is the **super radiant diode (SRD)**. All the three devices are *p–n* junction devices. The LED is a low-injection level device whose light output has a wide emission angle and a broad spectral width. The ILD is a high-injection level device where **light amplification by stimulated emission of radiation (laser)** takes place. The light output from ILDs has a narrow spectral width. SRDs perform somewhere between the LEDs and ILDs. In this text, we discuss only the LEDs and ILDs. These devices are suitable for fibre optic transmission systems for the following reasons:

1. Physical dimension of these devices are compatible with those of the optical fibres.
2. Their optical power output can be directly modulated by varying the input current..
3. They emit optical power at wavelengths that fall in the low-loss regions of optical fibres.
4. Their output power is large enough to support a wide range of applications.
5. Their operating and fibre coupling efficiencies are high.

A fundamental difference between LEDs and ILDs is that LEDs are incoherent sources, whereas ILDs are coherent sources. In a coherent source, the optical output is highly monochromatic and the output beam is very directional. In an incoherent source, the output power has a large spectral width and the beam is extremely divergent. The spatially directed output from an ILD can be launched into either single mode or multimode

fibres easily, whereas an LED output can be launched only into large-dimension (core size 50 μm or more) multimode fibres.

The coherence properties govern the amount of optical power that can be coupled into a fibre and the effect of chromatic dispersion on the bandwidth of the fibre medium. Although both LEDs and ILDs have similar output power levels, only a small fraction of this power can be launched into the fibre from LEDs, whereas a larger fraction can be launched from ILDs. **Brightness** or **radiance** is another factor determining the power levels that can be launched into fibres. Radiance is a measure of the optical power radiated into a unit solid angle per unit area of the emitting surface.

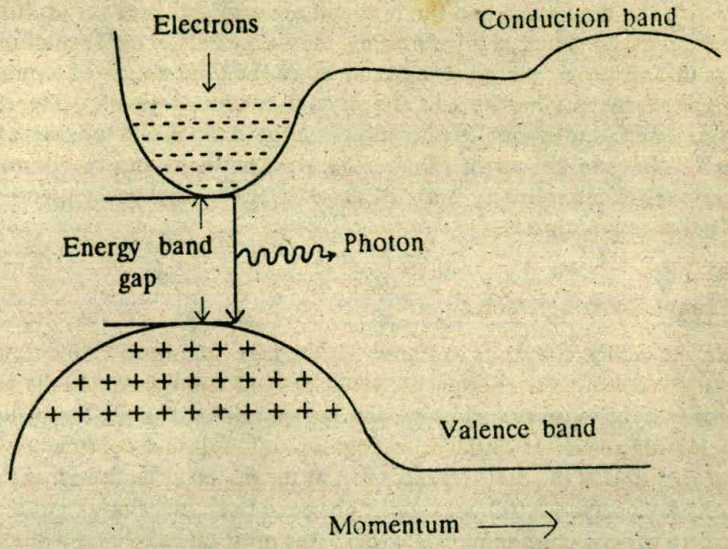
The light emitting region of both LEDs and ILDs is a semiconductor *p-n* junction known as **active** or **recombination region**. When this junction is forward biased, electrons and holes are injected into the *p*- and *n*-regions respectively. These injected minority carriers combine with majority carriers radiatively or nonradiatively. In the former case, a photon of energy is emitted, and in the latter case, the energy is dissipated in the form of heat. The fraction of the injected electron-hole pairs that combine radiatively is known as **quantum efficiency**. The injection of minority carriers in the recombination region takes place on account of stimulated emission in the case of ILDs. The stimulated emission in ILDs results in faster rise and fall times of the optical output when compared to those of LEDs. **Emission response time** is a parameter that affects the bandwidth of a light source. It is defined as the time delay between the application of an input current to the diode and the onset of optical emission.

In the case of LEDs, there is a near linear relationship between input drive current and the output light. On the other hand, ILDs are threshold devices and the *lasing* action takes place only above the threshold. The threshold current is not a constant but is a function of the device temperature and age. Therefore, laser diodes require thermal or optical stabilisation circuits whereas LEDs do not. As a result, the LED drive circuitry is less complex and LEDs can be fabricated less expensively with higher yields. Table 7.4 gives a comparison of the characteristics of LEDs and ILDs.

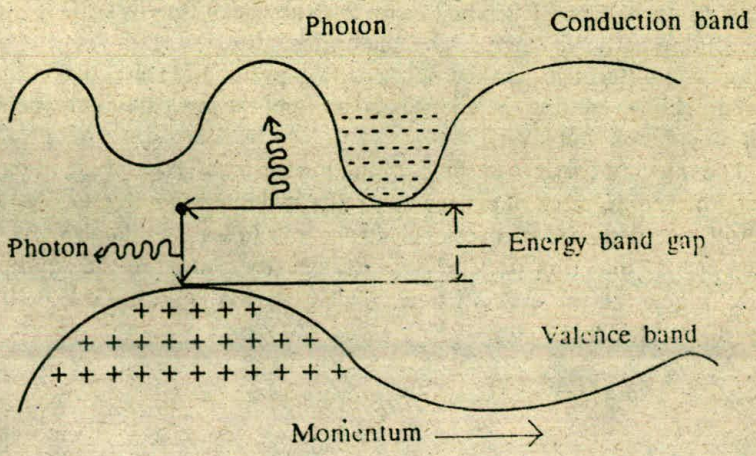
Table 7.4 Characteristics of LEDs and ILDs

Characteristic	LEDs	ILDs
Output power (mW)	1–10	1–10
Power launched into fibre (mW)	0.03–0.3	0.5–5
Spectral width (nm)	15–60	2–4
Radiance (W/cm^2/sr)	10–10^3	10^5
Rise time (ns)	2–20	$\leq$ 1
Frequency response (MHz)	< 200	> 500
Voltage drop (V)	1.5–2.5	1.5–2
Forward current (mA)	50–300	10–300

As explained earlier, optical radiation is generated by a recombination process of minority carriers with majority carriers in the active region of a *p-n* junction. Recombination takes place when an electron experiences a transition from the conduction band to the valence band as shown in the energy band diagram (Fig. 7.7a). Recombination occurs between the minimum energy level of the conduction band and the maximum energy level



(a) Direct band gap material



(b) Indirect band gap material

Fig. 7.7 Momentum vs. energy band diagrams.

of the valence band. The recombination process must conserve both energy and momentum. The energy band diagrams shown in Fig. 7.7 are plotted against momentum. In Fig. 7.7(a), the minimum and the maximum energy levels of the conduction and valence bands respectively have the same momentum value. Hence, there is a direct recombination between electrons and holes. Semiconductor materials which exhibit energy level characteristics as shown in Fig. 7.7(a) are called **direct band gap semiconductors.** In Fig. 7.7(b), the maximum and the minimum energy levels occur at different momentum values. Here, a third particle must take part in the recombination process in order to conserve momentum since the momentum of a photon is very small. Lattice vibrations in the crystal known as **phonons** serve this purpose. Semiconductors that exhibit energy-momentum characteristics similar to the one shown in Fig. 7.7(b) are known as **indirect band gap materials.** Optical sources, both LEDs and ILDs, are constructed using direct band gap materials.

7.3.1 Light Emitting Diodes

In order to achieve a high radiance and a high quantum efficiency, it is necessary to confine the charge carriers (electrons and holes) and the optical emission to the active region of the *p-n* junction. Carrier confinement achieves a high level of radiative recombination. Optical confinement prevents absorption of the emitted radiation by the material surrounding the *p-n* junction. A variety of LED configurations have been investigated to achieve carrier and optical confinement. Of these, the most effective and widely used structure known as **double heterostructure** is shown in Fig. 7.8. A double heterojunction configuration may be fabricated to emit light through a surface or an edge of the structure. The surface-emit type shown in Fig. 7.8 is called **Burrus type LED.** The sandwich structure of Burrus LED is capable of confining both the carriers and the optical emission to the central active region. The differences in the energy band gaps and the refractive indices of adjoining layers confine the charge carriers and the optical field respectively to the recombination layer.

The injected carriers in the *p-n* junction form and recombine in pairs to satisfy the requirement of charge neutrality in the semiconductor crystal. As a result, the densities of excess electrons and holes are equal, i.e. $n = p$. When carrier injection stops, the carrier density returns to the equilibrium value. In general, the excess carrier density decays exponentially with time according to the relation

$$n = n_0 \exp(-t/\tau) \tag{7.12}$$

where

n_0 = initial injected excess electron density

τ = carrier lifetime

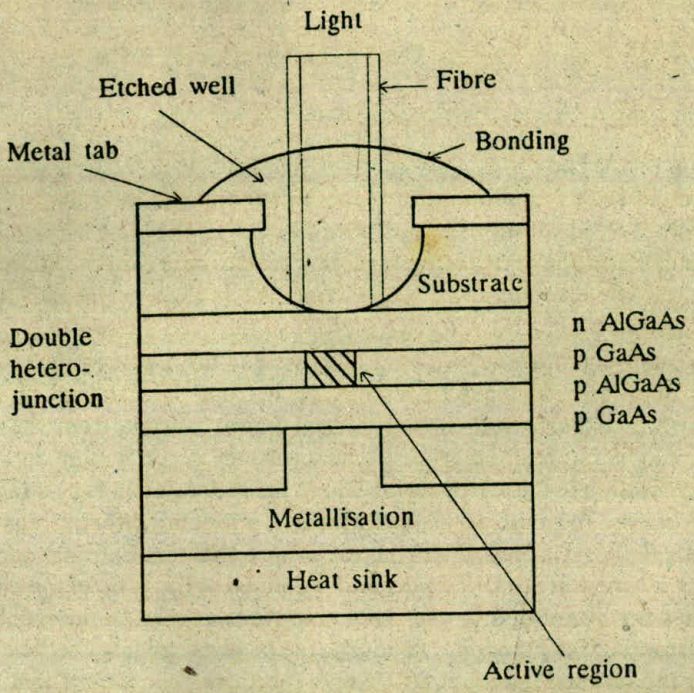


Fig. 7.8 Burrus type LED with fibre pigtail.

The carrier lifetime is one of the most important operating parameters of an electro-optic device. Its value ranges from a fraction of a nanosecond to a few milliseconds, depending on the material and manufacturing process. The quantum efficiency in the active region is given by

$$\eta = \frac{R_r}{R_r + R_{nr}} \tag{7.13}$$

where

R_r = radiative recombination rate per unit volume

R_{nr} = nonradiative recombination rate per unit volume

For exponential decay of excess carriers, we have the radiative and non-radiative recombination times as

$$\tau_r = \Delta n/R_r, \qquad \tau_{nr} = \Delta n/R_{nr} \tag{7.14}$$

Since the two recombination processes happen in parallel, the **bulk recombination lifetime** τ is governed by the equation

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{7.15}$$

We then have

$$\eta = \frac{\tau}{\tau_r} \tag{7.16}$$

Equations (7.15) and (7.16) are somewhat simplistic as there are a number of other factors such as nonradiative recombination at the interfaces of different layers, self-absorption of optical emission in the active region, and doping concentration of the active layer that tend to decrease the lifetime. As a result, the effective lifetime of the carriers is less than τ. However, in this text, we assume τ to be the effective lifetime of the excess carriers.

The quantum efficiency we have discussed so far is often referred to as **internal quantum efficiency** to distinguish it from the **external quantum efficiency** which deals with useful optical power output from the semiconductor. The internal quantum efficiency is concerned with main optical energy generation in the active region where the generated energy is radiated in all directions. To be useful, an optical source must emit effectively as much main optical energy as possible from the semiconductor. External quantum efficiency is a measure of this aspect and is defined as the ratio of the number of photons finally emitted by the semiconductor material to the number of carriers crossing the junction. Obviously, the external quantum efficiency is less than the internal quantum efficiency. There are four factors that affect the external quantum efficiency:

1. Amount of generated light emitted in the direction of semiconductor-air surface
2. Light reaching the emitting surface at an angle of incidence less than the critical angle
3. Reflection at the semiconductor-air surface
4. Absorption between the point of generation and the emitting surface.

These factors are illustrated in Fig. 7.9.

The frequency response of an LED is limited by two factors:

- Diffusion capacitance
- Parasitic diode space charge capacitance.

The first factor delays the storage of injected carriers in the active region and the second delays the injection process itself. It is the parasitic space charge capacitance that determines the emission response time. By applying a small constant forward bias, the emission response time can be made negligible, in which case, the frequency response is entirely determined by the diffusion capacitance. The relationship between the input signal frequency ω and the optical output $I(\omega)$ is given by

$$I(\omega) = I_0[1 + (\omega\tau)^2]^{-\frac{1}{2}} \tag{7.17}$$

where

I_0 = d.c. drive current when signal frequency is zero

τ = effective carrier lifetime

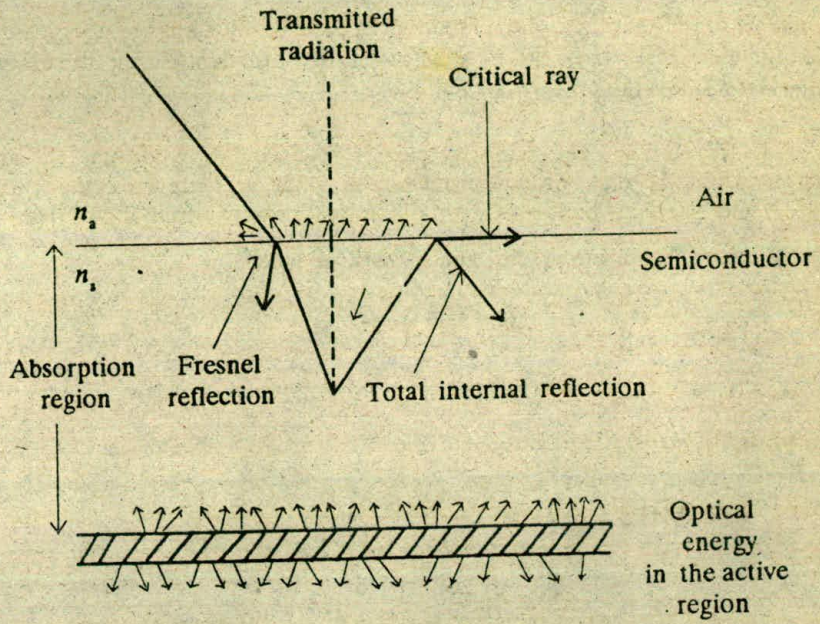


Fig. 7.9 Factors affecting external quantum efficiency.

The optical output power of an ideal LED is directly proportional to the current flowing through the device. Hence, the optical power relationship of the LED may be stated as

$$P(\omega) = P_0[1 + (\omega\tau)^2]^{-\frac{1}{2}} \tag{7.17a}$$

where P_0 is the d.c. optical power output corresponding to the drive current I_0.

It is now possible to define the modulation bandwidth of an LED in either electrical or optical terms. The **electrical bandwidth** corresponds to the frequency at which the electrical output power is reduced by 3 dB. This is obtained by setting $I^2(\omega) = I_0^2/2$ since the electrical power is directly proportional to the square of the current. Hence, we get a 3 dB electrical bandwidth which is of the form

$$I_0^2/2 = I_0^2/(1 + (\omega_e\tau)^2)$$

i.e. $1 + (\omega_e\tau)^2 = 2$ or

$$\omega_e = 1/\tau \tag{7.18}$$

The **optical bandwidth** is obtained by setting $P(\omega) = P_0/2$. That is,

$$P_0/2 = P_0[1 + (\omega_0\tau)^2]^{-\frac{1}{2}}$$

i.e. $1 + (\omega_0\tau)^2 = 4$, or

$$\omega_0 = \sqrt{3}/\tau \tag{7.18a}$$

It may be noted that the optical modulation bandwidth is apparently greater than the electrical bandwidth by a factor $\sqrt{3}$. In a practical circuit, the lower value determines the performance limit.

EXAMPLE 7.3 The minority carrier recombination lifetime for an LED is 5 ns. The d.c. optical output power is 320 W. Determine the 3 dB optical and electrical bandwidth and the optical power output at 40 MHz.

Solution From Eqs. (7.18) and (7.18a), we have

$$\omega_e = \frac{1}{5 \times 10^{-9}} = 200 \text{ MHz, and}$$

$$\omega_0 = \sqrt{3}\,\omega_e = 350 \text{ MHz.}$$

From Eq. (7.17a),

$$P(40\text{MHz}) = \frac{320 \times 10^{-6}}{[1 + (2\pi \times 40 \times 10^6 \times 5 \times 10^{-9})^2]^{\frac{1}{2}}}$$

$$= \frac{320 \times 10^{-6}}{(2.57)^{\frac{1}{2}}} = 200\ \mu\text{W}$$

The transient response of the LED is limited by the carrier lifetime, junction space charge capacitance and the diffusion capacitance. If the amplitude of the step current used to determine the transient response is large, the effects of capacitances become relatively small as compared to the effect due to carrier lifetime. In this case, the rise and fall times can be shown as:

$$t = \tau(\ln 9) \tag{7.19}$$

In an LED, the emitted power and the bandwidth of operation are inversely related. For a given injection level, the **power-bandwidth product** remains constant. An attempt to increase the bandwidth results in reduced power emission. Hence, the power-bandwidth product, *PBW*, is an important factor and can be shown to be:

$$PBW = \frac{J}{\tau}\frac{hf}{q}\eta \tag{7.20}$$

where

J = current density

η = quantum efficiency
τ = carrier lifetime
f = frequency of the signal
q = charge of the electron
h = Plank's constant $= 6.63 \times 10^{-34}$ J-s

7.3.2 Laser Diodes

There are three key processes that take place in a semiconductor laser diode:

1. Photon absorption
2. Spontaneous emission
3. Stimulated emission.

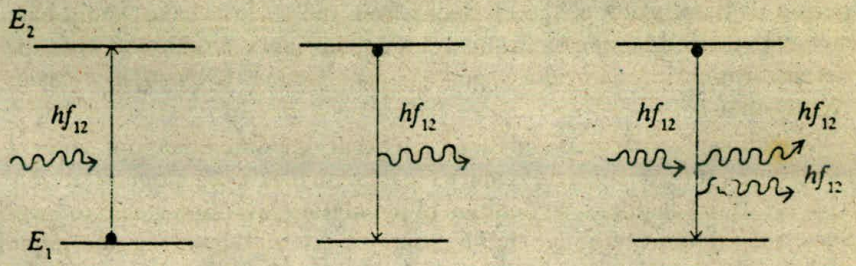
These processes are illustrated in the energy band diagrams shown in Fig. 7.10. E_1 represents the stable ground state energy level and E_2 the unstable excited state energy level. The region between the two levels is the band gap. According to Planck's law, a transition between the two energy states involves either absorption or emission of a photon of energy given by the equation

$$hf_{21} = E_2 - E_1 \tag{7.20a}$$

where

h = Planck's constant
f_{21} = frequency of emission or absorption.

Normally, the device is in the stable ground state. When a photon of energy with proper wavelength (λ_{12}) impinges on the device, an electron in state E_1 may absorb the energy and transfer itself to the upper excited state as depicted in Fig. 7.10(a). This is the photon absorption process. The excited state being unstable, the electron returns to the ground state after staying in the excited state for a short duration. It emits a photon of energy while



(a) Absorption (b) Spontaneous emission (c) Stimulated emission

Fig. 7.10 Absorption and emission process in laser diodes.

returning to the ground state as illustrated in Fig. 7.10(b). This is called **spontaneous emission**. These emissions are isotropic and random in nature and thus appear as a narrowband Gaussian spectrum. The electrons can also be induced to make a downward transition from the excited state to the ground state by an external stimulation. When a photon of energy with proper wavelength impinges on the device, an electron in the excited state may be stimulated to drop to the ground state, immediately giving out a photon of energy as shown in Fig. 7.10(c). The emitted photon and the incident photon are in phase with each other and the resultant emission is known as **stimulated emission**.

In laser diodes, the emitted radiation is almost entirely produced by stimulated emission. This endows them with three important properties:

1. The emitted radiation is confined to a narrow band of wavelengths.
2. The radiation is very directional, thereby improving the external quantum efficiency significantly.
3. The modulation bandwidth of the devices is larger compared to that of LEDs.

Under thermal equilibrium conditions, the number of atoms and, therefore, the number of electrons per unit volume, n_1 and n_2 at the energy levels E_1 and E_2, are related by the Boltzmann equation

$$\frac{n_1}{n_2} = \exp(E_2 - E_1)/kT \tag{7.21}$$

where

k = Boltzmann's constant which has value of 1.38×10^{-23} J/K
T = absolute temperature of the active region

Equation (7.21) implies that under normal circumstances the value n_1 is greater than n_2. Under thermal equilibrium the density of excited electrons is also very small. As a result, most photons incident on the device are absorbed rather than being instrumental in stimulating emissions. After a period of time, many of the electrons from the valence band would have moved to the conduction band and a new equilibrium is established when the net absorption is balanced by spontaneous emission. Even so, n_1 remains greater than n_2.

Significant stimulated emissions can occur only when the population of electrons in the excited state is much higher than that in the ground state, i.e. $n_2 > n_1$. If this happens, a condition of **population inversion** is said to exist. Such a condition can be created by using an external 'pumping' technique such as injecting electrons into the material at the device contacts to fill the lower energy states of the conduction band. A direct consequence of population inversion is that the rate of stimulated emission exceeds the rate

of absorption. The injected photons are being supplemented as they pass through the medium and an optical amplification takes place.

The energy level structure inside a semiconductor is rather a great deal more complex than just the two-level energy system we have considered so far. A detailed discussion of lasing action inside a semiconductor laser is beyond the scope of this book. However, we present here some of the results pertaining to the two-level energy system as they provide a reasonable insight into the phenomenon.

The gain coefficient pertaining to optical amplification in a two-level system is given by

$$g = \frac{c^2}{8\pi} \frac{(n_2 - n_1)P(f_{21})}{f_{21}\tau_{sp}} \tag{7.22}$$

where

$P(f_{21})$ = power spectral density of the external pumping source
τ_{sp} = mean carrier lifetime for spontaneous emission
c = velocity of light

Creating a population inversion and producing optical gain represent two necessary but not sufficient conditions to obtain laser action. In order to achieve sustained amplification at a resonant frequency, an optical cavity is required to be formed. This is achieved by providing some positive feedback to the amplifier to make it act as a resonating oscillator. The optical radiation then occurs at the resonating frequencies rather than being influenced by the spectral properties of emission from the two energy levels. Positive feedback is introduced by using a pair of mirrors that reflect the amplified light back and forth through the active medium, thus forming the cavity. Once the oscillations are set in, the optical power density at any of the cavity resonant frequencies comes into equilibrium such that the optical power loss during each transit of external radiation through the medium just balances the optical power gain. Self-oscillation cannot start before the gain exceeds the losses. This sets a **threshold gain** for population inversion which has to be exceeded before laser action can begin. Some of the light generated by stimulated emission is scattered out of the active medium during transit. This can be represented by a loss coefficient α. The threshold condition may then be stated as

$$g > \alpha + \frac{1}{l} \ln (1/R) \tag{7.23}$$

or

$$g_{th} = \alpha + \frac{1}{l} \ln (1/R) \tag{7.23a}$$

where

l = effective length of the active medium
R = reflection coefficient of the mirrors
g_{th} = threshold gain

The typical relationship between the diode drive current and the optical output power of a laser diode is shown in Fig. 7.11. At low drive currents, the

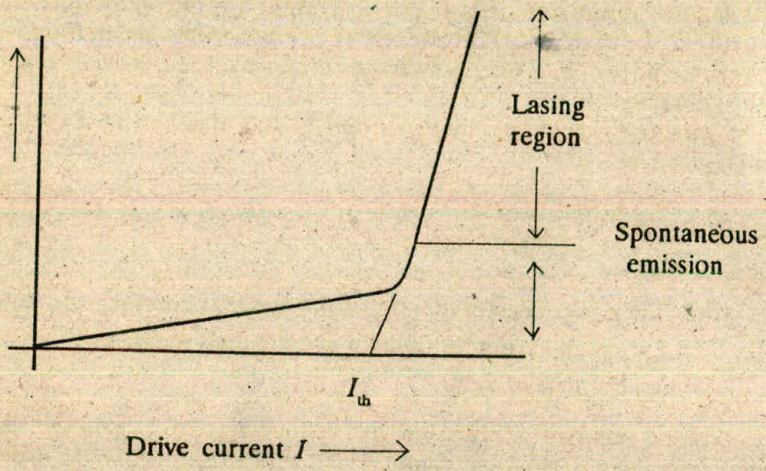


Fig. 7.11 Drive current vs. optical power in laser diodes.

emission characteristics are similar to that of the LEDs. Only spontaneous emission is present and the spectral range and the lateral beam width of the emission are broad. A sharp increase in the power output occurs at the lasing threshold. The beam width and the spectral range narrow down to about 8° and 1 nm respectively, as the drive current increases beyond the threshold value. The **threshold current** I_{th} is defined by extrapolating lasing region curve to intersect the X-axis.

The slope of the drive current, viz. power characteristic in the spontaneous emission region, corresponds to the external quantum efficiency discussed in Section 7.3.1 with regard to LEDs. The slope in the lasing region corresponds to the **external differential quantum efficiency** which is defined as the number of photons emitted per radiative electron-hole pair recombination in the lasing region. The efficiency is given by

$$\eta = \frac{\eta_i(g_{th} - \alpha)}{t_{th}} \tag{7.24}$$

where

η_i = internal quantum efficiency

α = loss coefficient or effective absorption coefficient

In Eq. (7.24), it is assumed that the gain coefficient remains at a constant value of g_{th}, the threshold.

The frequency response of the laser diode depends on a variety of factors such as the following:

1. Operating region
2. Analog or pulse modulation
3. Spontaneous carrier lifetime, τ_{sp}
4. Stimulated carrier lifetime, τ_{st}
5. Photon lifetime, τ_{ph}

If the operating region includes the spontaneous emission region, the frequency response or the modulation rate is limited by the lifetime of the spontaneous carriers, τ_{sp}. This may happen if the sinusoidal signal excursions are large enough to traverse the spontaneous emission region or if the pulse modulation turns off the diode completely after each pulse. τ_{sp} is a function of the semiconductor band structure and carrier concentration and is governed by Eqs. (7.12) and (7.15).

When the laser is biased beyond the threshold, a sinusoidal modulation is used, and the operating region is confined to stimulated emission region, and the modulated optical power goes through a resonance with a natural frequency which is a function of both τ_{st} and τ_{ph}. The stimulated carrier lifetime τ_{st} depends on the optical density in the lasing cavity. The photon lifetime τ_{ph} is the average time that a photon resides in the lasing cavity before being lost either by absorption or emission through the facets. τ_{ph} is given by

$$\frac{1}{\tau_{ph}} = \frac{c}{n} g_{th} \tag{7.25}$$

where

c = velocity of light

n = refractive index of the lasing material

The photon lifetime sets the upper limit to the modulation capability of a laser diode. If pulse modulation is carried out in the stimulated emission region, then modulation rates are limited by the stimulated carrier lifetime. The photon lifetime is typically of the order of a few picoseconds, stimulated carrier lifetime around 10 picoseconds, and the spontaneous carrier lifetime around 10 nanoseconds.

7.4 Optical Detectors

An optical detector or a photodetector is the first element of the receiver of an optical communication system. It senses the luminescent power incident upon it and converts the same into a corresponding electronic current. To be useful in a fibre optic transmission system, a photodetector must have

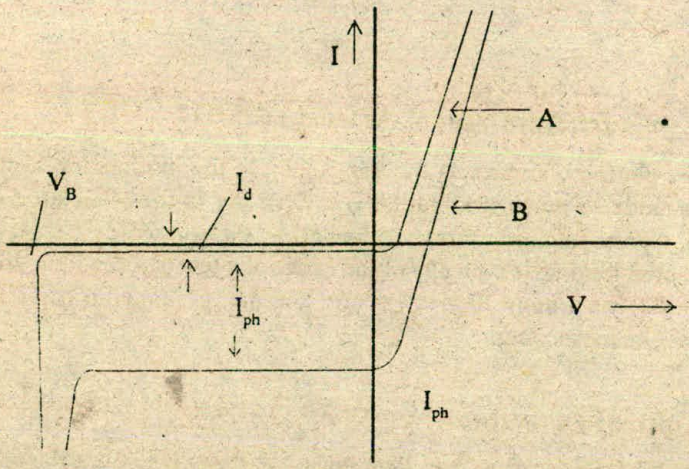
1. physical dimensions which are compatible with those of the optical fibres,
2. high sensitivity in the emission wavelength range of the optical sources, and
3. a fast response and sufficient bandwidth to handle the desired data rates.

A variety of photodetectors are commonly used. These are:

1. Photomultipliers
2. Pyroelectric detectors
3. Semiconductor photoconductors
4. Phototransistors
5. Photodiodes.

Of these, the photodetectors 1–4 are not suitable for fibre optic systems for one reason or the other. Therefore, semiconductor photodiodes are almost used on an exclusive basis as photodetectors in fibre optic systems. They have a small size compatible with optical fibres, high sensitivity and fast response time.

There are two principal types of photodiodes: *p-i-n* **photodiode** and **avalanche photodiode (APD)**. Both the types of detectors are reverse biased *p-n* junction devices. When not illuminated, their voltage-current characteristics are similar to those of *p-n* junction diode as indicated by the curve *A* in Fig. 7.12. When light of wavelength less than some threshold wavelength λ_{th} is shone into the junction, the *V-I* characteristics shift by a



I_d = dark current I_{ph} = photocurrent V_B = breakdown voltage

Fig. 7.12 *V-I* characteristics of a *p-n* junction photodiode.

constant current I_{ph}, as indicated by curve *B* in Fig. 7.12. This shifting is a result of the quantum interaction between the optical radiation and the valence band electrons in which the electrons are excited to the conduction band and contribute to the higher current as shown in Fig. 7.13. The

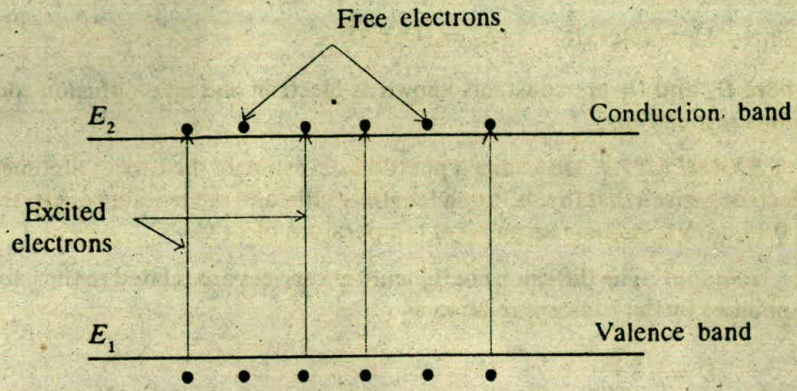


Fig. 7.13 Quantum interaction in photodiodes.

threshold wavelength is related to the band gap energy by

$$\lambda_{th} = \frac{hc}{E_2 - E_1} = \frac{1.24}{E_g}\ \mu\text{m} \tag{7.26}$$

where E_g is expressed in electron volts. The threshold or cutoff wavelength of silicon is about 1.06 μm and that of germanium is about 1.6 μm. For longer wavelengths, the photon energy is not sufficient to excite an electron from the valence band into the conduction band.

The current due to illumination, I_{ph}, known as **photocurrent**, increases linearly with the incident optical power, P_0, over a range of many orders of magnitude. The ratio of the photocurrent to the incident optical power, known as **responsivity** *R*, is an important parameter of a photodiode and is given by

$$R = \frac{I_{ph}}{P_0} \tag{7.27}$$

Typical responsivity values are 0.65 μA/μW for silicon at 800 nm wavelength and 0.45 μA/μW for germanium at 1.3 μm.

The **quantum efficiency** in the case of photodiodes is defined as the ratio of the number of electron-hole pairs generated per incident photon energy and is given by

$$\eta = \left(\frac{I_{ph}}{q}\right) \Big/ \left(\frac{P_0}{E_g}\right) = \frac{I_{ph}}{P_0}\,\frac{hf}{q} = \frac{hf}{q}R \tag{7.28}$$

where q is the electron charge.

As the charge carriers flow through the semiconductor material, some electron-hole pairs recombine and get absorbed. On the average, the charge carriers travel a certain distance known as **diffusion length**, before they recombine. The diffusion lengths, L_n for electrons and L_p for holes, are related to the respective carrier lifetimes by the expressions

$$L_n = (D_n \tau_n)^{\frac{1}{2}}, \qquad L_p = (D_p \tau_p)^{\frac{1}{2}} \tag{7.29}$$

where D_n and D_p are constants known as electron and hole diffusion coefficients respectively.

EXAMPLE 7.4 In a silicon photodiode, estimate the carrier lifetimes of electrons, given that the diffusion length is 1 μm and the mobility of electrons is 0.15 m^2/Vs. Assume an ambient temperature of 27°C.

Solution The diffusion coefficients of carriers are related to the carrier mobilities by the Einstein relation as

$$D = \mu \frac{kT}{q}$$

where μ is the mobility of carrier, k is the Boltzmann constant and T is the temperature in degrees Kelvin. Therefore,

$$D_n = \frac{0.15 \times 1.38 \times 10^{-23} \times 300}{1.6 \times 10^{-19}} = 4.14 \times 10^{-3}\ \text{m}^2/\text{s}$$

From Eq. (7.29),

$$\tau_n = \frac{L_n^2}{D_n} = \frac{10^{-12}}{4.14 \times 10^{-3}} = 0.24\ \text{ns}$$

Optical radiation is absorbed in the semiconductor material according to the exponential power law

$$P(x) = P_0 e^{-\alpha x} \tag{7.30}$$

$$P'(x) = P_0 - P(x) = P_0 - (1 - e^{-\alpha})^x \tag{7.31}$$

where

α = absorption coefficient at a given wavelength

P_0 = incident optical power at the starting point

Equation (7.30) gives the radiation power density at a distance x and Eq. (7.31) provides the radiation power absorbed in the distance interval x. The absorption coefficient for a given semiconductor material is a function of wavelength and rises sharply for lower wavelengths. Equation (7.26) gives an upper cutoff value for the wavelength. At wavelengths much lower than λ_{th}, the material absorption may become significant and limit the useful

photocurrent. In particular, absorption within the depletion region is of importance as any recombination here is a loss to the external circuitry. If w is the width of the depletion region, the power absorbed in this distance is given by

$$P'(w) = P_0(1 - e^{-\alpha w}) \tag{7.32}$$

The useful power at the end of the depletion layer is

$$P(w) = P_0 e^{-\alpha w} \tag{7.33}$$

If we take into account the reflectivity R_f at the entrance face of the photodiode, then the power entering the semiconductor is given by

$$P_0' = (1 - R_f)P_0 \tag{7.34}$$

The quantum efficiency may now be expressed as

$$\eta = (1 - R_f)P_0 e^{-\alpha w}/P_0 = (1 - R_f)e^{-\alpha w} \tag{7.35}$$

The primary photocurrent lost due to absorption is given by

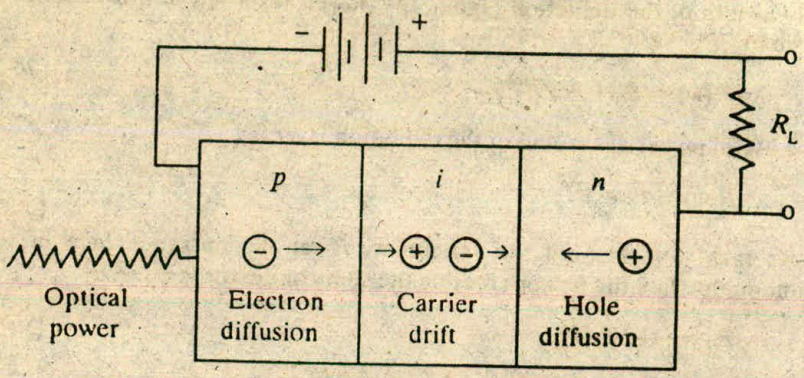
$$I_{ph}' = \frac{q}{h_f} P_0(1 - e^{-\alpha w})(1 - R_f) \tag{7.36}$$

The useful primary current is given by

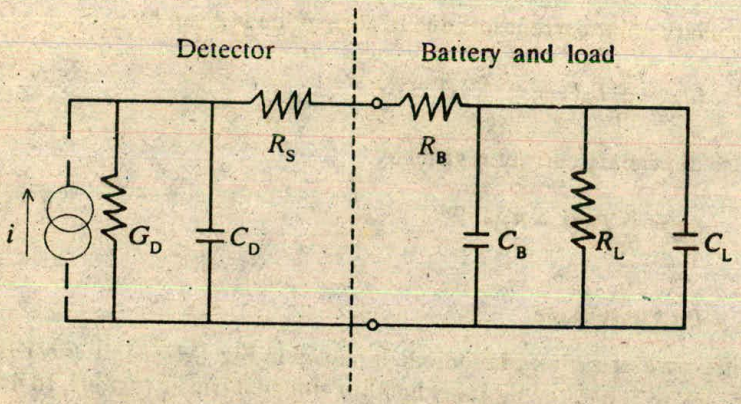
$$I_{ph} = \frac{q}{h_f} P_0 (1 - R_f)e^{-\alpha w} \tag{7.37}$$

7.4.1 *p-i-n* Photodiode

The structure of a *p-i-n* photodiode is shown in Fig. 7.14(a). It consists of *p* and *n* regions separated by a very lightly *n*-doped intrinsic region *i*. In normal operation, the intrinsic region is kept fully depleted of carriers by means of a sufficiently large reverse-bias voltage applied across the device. A *p-i-n* photodiode is normally designed so that the photo carriers are generated mainly in the intrinsic region where most of the incident light is absorbed. The high electric field produced by the large reverse-bias voltage causes the carriers to separate and be collected in the *p*- and *n*-regions. This gives rise to current flow in the external circuit. Under reverse-biased conditions, the small signal behaviour of the photodiode can be represented by the equivalent circuit shown in Fig. 7.14(b). The circuit consists of an ideal current generator with a shunt conductance G_D, series resistance R_s, and a capacitance C_D. G_D represents the slope of the reverse bias characteristics, R_s the resistance offered by the bulk semiconductor and the contacts, and C_D the capacitance of the *p-n* junction, the leads and the package. R_B and C_B are the battery parameters and R_L and C_L are the load parameters.



(a) Structure



(b) Equivalent circuit

Fig. 7.14 Structure and the equivalent circuit of a *p-i-n* photodiode.

The response time of a *p-i-n* photodiode depends mainly on three factors:

1. Carrier transit time in the *i*-region
2. Carrier diffusion times in the *p*- and *n*-regions
3. Time constant of the photodiode and its load.

Carriers generated in an electric field acquire an average drift velocity v_d and travel at that speed across the field. At low field strengths, the drift velocity is proportional to the field but at sufficiently high field strengths, the carriers drift at the saturation velocity, v_s, due to the large reverse-bias voltage. We, therefore, have the carrier drift time as

$$t_d = \frac{w}{v_s} \tag{7.38}$$

where w is the width of the i-region. If the carriers are generated in the p-region, the transit time is determined by electron saturation velocity, whereas the hole saturation velocity determines the transit time if the carriers are generated in the n-region.

Diffusion processes are slower than drift processes and hence diffusion times are greater than drift times for a given length. Diffusion times can be calculated from Eq. (7.29) for electrons and holes. The effect of slow diffusion rates on the response time of the photodiode can be minimised by ensuring that most of the carrier generation takes place at the edges of the p-i or n-i boundaries and the carriers are absorbed within the depletion region.

EXAMPLE 7.5 In a silicon p-i-n photodiode, the optical energy is incident on the p-region. If the saturation drift velocities of the electrons and holes in the silicon are 8×10^6 and 4×10^6 cm/s respectively, and the depletion region width is 10 μm, calculate the response time of the photodiode. Assume that p- and n-layers are very thin.

Solution Since the photons are incident on the p-region, the drift velocities of the electrons determine the response time. Since the p-layer is very thin, we may assume that all the carriers are generated at the p-i edge. Therefore,

$$\text{Response time} = \frac{10 \times 10^{-4}}{8 \times 10^6} = 0.125\,\text{ns}$$

If all the resistive and the capacitive components are lumped together, the equivalent circuit shown in Fig. 7.14(b) reduces to a low pass filter with equivalent R and C components given by

$$\frac{1}{R} = G_D + \frac{1}{R_s} + \frac{1}{R_B} + \frac{1}{R_L} \tag{7.39}$$

$$C = C_D + C_B + C_L \tag{7.40}$$

The bandwidth of the photodiode circuit is then determined by

$$B = \frac{1}{2\pi RC} \tag{7.41}$$

The junction capacitance of the photodiode, C_j, which is a component of C_D, is given by

$$C_j = \frac{\epsilon_s A}{w} \tag{7.42}$$

where

ε_s = permittivity of the semiconductor

A = cross-sectional area of the depletion layer

w = width of the depletion region

It may be noted that if the depletion layer is very thin, the junction capacitance is excessive.

EXAMPLE 7.6 In a *p-i-n* photodiode the depletion region has a width of 10 μm and a cross-sectional area of 10 μm^2. The diode is connected to a resistive load of 1 kΩ. If the load resistance and the junction capacitance are the dominating resistive and capacitive values respectively, determine the bandwidth of the diode circuit. Assume that the dielectric constant of the semiconductor is 1.1.

Solution

$$\varepsilon_s = \varepsilon_0 k$$

where

ε_0 = permittivity of free space = 8.85×10^{-12} F/m

k = dielectric constant of the semiconductor

Therefore,

$$\varepsilon_s = 9.74 \times 10^{-6} \text{ pF}/\mu\text{m}$$

$$C_j = \frac{9.74 \times 10^{-6} \times 10}{10 \times 10^{-6}} = 9.74 \text{ pF}$$

$$B = \frac{1}{2 \times 3.14 \times 10^3 \times 9.74 \times 10^{-12}} = 16.35 \text{ MHz}$$

7.4.2 Avalanche Photodiode

Signal-to-noise ratio is an important parameter in any receiver system. Noise is introduced into a receiver system from external sources as well as by the components of the receiver system. For example, the dark current of a photodiode contributes to the noise in the system. In an optical receiver system with a *p-i-n* photodetector, the electronic noise introduced by the load resistor and the amplifier circuits following the photodetector forms a dominant component of the total noise. It would, therefore, be beneficial to multiply the signal level within the photodetector itself. **Carrier multiplication** is a fundamental process in **avalanche photodiodes (APDs)**. A *p-i-n* photodiode converts one photon to one electron and has a conversion efficiency of less than unity. In contrast, one photon of energy results in multiple carriers in the case of APD. Obviously, there is a gain G associated with an APD.

In order that carrier multiplication takes place, the photon-generated carriers must impart adequate energy to the bound electrons in the valence band and excite them to the conduction band. This they do by acquiring additional energy in the presence of a very high electric field and colliding with the valence band electrons. Thus, the existence of a very high electric field to accelerate the photon-generated carriers is a basic requirement for carrier multiplication. The process of carrier multiplication by collision is known as **impact ionisation**. The carriers generated by impact ionisation are also accelerated by the electric field and they, in turn, cause further ionisation leading to an **avalanche effect**. The process is illustrated in Fig. 7.15.

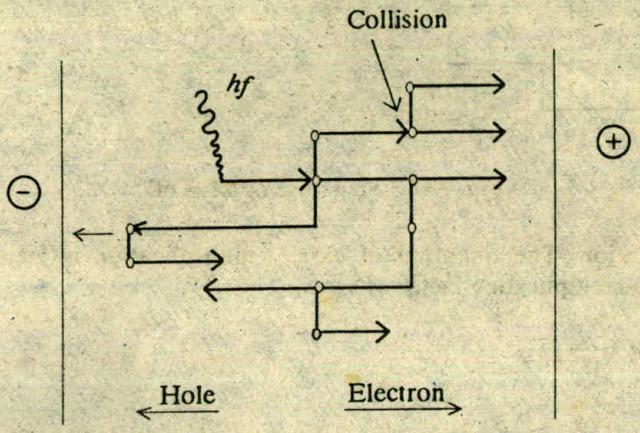


Fig. 7.15 Carrier multiplication process.

A commonly used structure for APDs is known as **reach-through structure (RAPD)** and is shown in Fig. 7.16. It is composed of a high resistivity p-type material, known as π-**layer**, deposited as an epitaxial layer on a heavily doped p-type (p^+) substrate. A p-type diffusion or ion implant is then made in the high-resistivity material followed by the formation of a heavily doped n-type (n^+) layer. This structure is referred to as a p^+-π-p-n^+ reach through structure. The π-layer is basically an intrinsic region which has some p-doping on account of imperfect purification.

During normal operation, an APD is reverse biased with a very high reverse-bias voltage (100–400 V). When the reverse-bias voltage is increased from a low value, the depletion region widens until it 'reaches through' to the intrinsic region. Most of the potential drop is across the p-n^+ junction and a high electric field is built across this junction. Light enters the device through the p^+ region and is absorbed in the π-region. The electron-hole pairs created in the region are separated and the electrons and holes move towards n^+ and p^+ regions respectively. The very high electric field that exists across the p-n^+ junction accelerates the electrons, leading to carrier

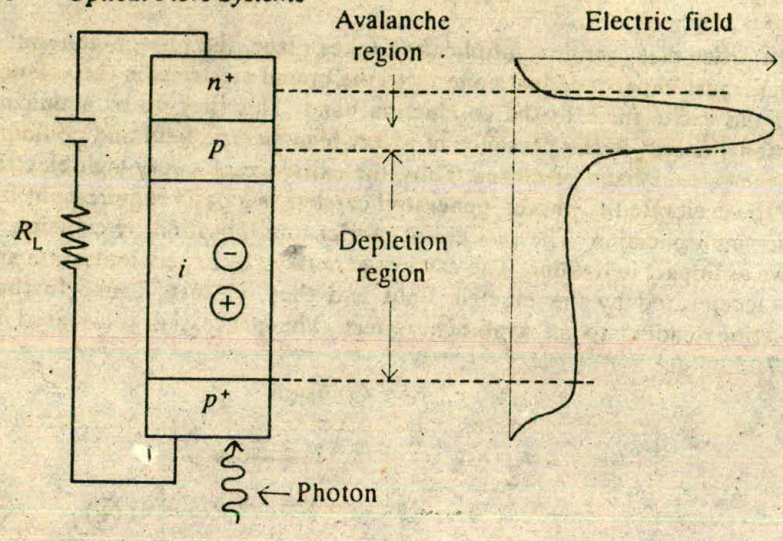


Fig. 7.16 Structure of a RAPD.

multiplication. The quantum of carrier multiplication is measured by defining a multiplication factor M as

$$M = \frac{I}{I_p} \tag{7.43}$$

where

I_p = primary unmultiplied photocurrent

I = average output current after carrier multiplication

EXAMPLE 7.7 The quantum efficiency of a particular silicon APD operating at a wavelength of 0.8 μm is 90%. The incident optical power is 0.5 μW. The output current is 13 μA. Determine the multiplication factor of the photodiode.

Solution From Eq. (7.28), the responsivity of the photodiode is

$$R = \frac{q\eta}{hf} = \frac{0.9 \times 1.6 \times 10^{-19} \times 0.8 \times 10^{-6}}{6.63 \times 10^{-34} \times 3 \times 10^{-8}} = 0.576\,\text{AW}^{-1}$$

From Eq. (7.27) the photocurrent is

$$I_p = P_0 R = 0.5 \times 10^{-6} \times 0.576 = 0.288 \times 10^{-6}\,\text{A}$$

From Eq. (7.43),

$$M = \frac{I}{I_p} = \frac{13 \times 10^{-6}}{0.288 \times 10^{-6}} = 45.2$$

Response times and bandwidth of APDs are limited by four factors:

1. Transit time of the electrons across the depletion region
2. Time constant of the photodiode and its load
3. Transit time of the holes produced in the avalanche region to travel back across the depletion region
4. Time required for the avalanche to develop.

Factors 1 and 2 are the same as in the case of *p-i-n* photodiodes. Factors 3 and 4 are specific to avalanche photodiodes. In APDs, the holes produced in the avalanche region have to drift back across the depletion region. This movement affects the fall time of the output waveform. As a result, an asymmetric pulse shape is obtained from the APD with relatively fast rise time as the electrons are collected and a fall time dictated by the transit time of the holes.

The avalanche delay time, t_A, is a function of the ratio of ionisation coefficients or rates, α_e and α_h for electrons and holes respectively. Ionisation rate is the number of electron hole pairs generated by an electron or a hole per unit distance travelled. The ratio is defined as

$$K = \alpha_h/\alpha_e \tag{7.44}$$

The factor K lies in the range 0.01–100, implying that in some materials $\alpha_e > \alpha_h$, in some $\alpha_h > \alpha_e$, and in others, they are the same. If $\alpha_h = 0$, the avalanche would develop in a rather simple manner within the normal electron transit time across the avalanche region, as given by

$$t_e = w_A/V_{se} \tag{7.45}$$

where

w_A = width of the avalanche region

V_{se} = saturation velocity of electrons

When $K > 0$, the avalanche develops in multiple passes across the avalanche region; at high levels of multiplication for $0 < K < 1$, t_A is given by

$$t_A = MKw_A/V_{se}$$

The hole travel time is given by

$$t_h = (w_A + w)/V_{sh} \tag{7.46}$$

where

V_{sh} = hole saturation velocity

w = width of depletion region

Therefore,

$$\text{Total transit time} = \frac{w_A + MKw_A}{V_{se}} + \frac{w + w_A}{V_{sh}} \tag{7.47}$$

The carrier multiplication process is not itself free of noise. While on an average, each photon-generated carrier leads to M carriers at the end of the multiplication process, some carriers may produce more than M carriers and others less than M. This statistical nature of the process introduces further noise in the system. The result is that while the signal current is multiplied by the factor M, the r.m.s. noise level increases by a factor MF, where F is the noise factor. This noise factor is always greater than unity and an increasing function of M given by

$$F \approx M^x \tag{7.48}$$

where the index x lies between 0.2 and 1.0, depending upon the material and the type of carrier initiating the avalanche.

To conclude, APDs have a distinct advantage over *p-i-n* photodiodes for the detection of very low light levels due to their inherent carrier multiplication property. But they have several drawbacks too:

1. Their structure is more complex than that of *p-i-n* photodiodes.
2. The fabrication is difficult and hence they are more expensive.
3. Carrier multiplication is a random process and hence introduces additional noise into the system.
4. High reverse-bias voltages required are wavelength dependent.
5. Multiplication factor is temperature dependent. Hence temperature compensation circuits are required for stabilised operation.

7.5 Power Budget Analysis

System performance and cost constraints are important factors in any communication system. In Chapter 1, we briefly described the structure of an optical communication system. In the simplest form, a fibre optic communication system consists of a source with its associated modulator circuits, an optical fibre, and a detector with its associated receiver circuits. Performance of the system is evaluated by analysing the **link power budget** of the system and the cost is kept minimum by carefully selecting the system components from a variety of available choices. Various sources, detectors and optical fibres have already been discussed in the earlier sections of this chapter. In this section, we consider link power budget calculations.

In a fibre optic system, optical power loss occurs due to

1. source-to-fibre coupling losses,
2. connector loss,
3. splice loss, and
4. fibre attenuation.

The link power budget should provide for these losses and an additional power margin to allow for component aging and temperature fluctuations.

Generally, a system margin of 6 to 8-dB is provided. The power budget equation may be written as

$$P_D = P_s - P_r = (1 - \eta_c)P_s + m_1 l_c + m_2 l_s + \alpha_f L + SM \qquad (7.49)$$

where

P_s = output power from the source
P_r = receiver sensitivity
η_c = coupling efficiency
l_c = connector loss in dB/connector
m_1 = number of connectors in the system
l_s = splice loss in dB/splice
m_2 = number of splices in the system
α_f = fibre attenuation in dB/km
L = transmission distance in km
SM = system margin

Source-to-fibre coupling efficiency is defined as

$$\eta_c = \frac{P_f}{P_s} \qquad (7.50)$$

where P_f is the power launched into the fibre. Two factors affect the coupling efficiency:

- Unintercepted illumination loss
- Numerical aperture loss.

Unintercepted illumination loss occurs when the emitting area of the source is greater than the fibre core area or when the source and fibre axes are misaligned. However, unintercepted illumination loss may be eliminated by properly matching the source emitting area to the fibre core area and aligning the axes.

Numerical aperture is a measure of light acceptance or light gathering capability of a fibre. It is a dimensionless quantity with values ranging from 0.14 to 0.50, and is related to the maximum acceptance angle of the fibre. We now derive an expression for the numerical aperture of a step index fibre using ray optics analysis. Consider the light ray entering the fibre core from a medium of refractive index n_0 at an angle α with respect to the fibre axis and falling on the core-cladding interface at a normal angle θ, as shown in Fig. 7.17. The ray will propagate along the fibre by means of multiple internal reflections provided that the angle is greater than the critical angle θ_c. This implies that angle ϕ is less than some maximum angle ϕ_m, where

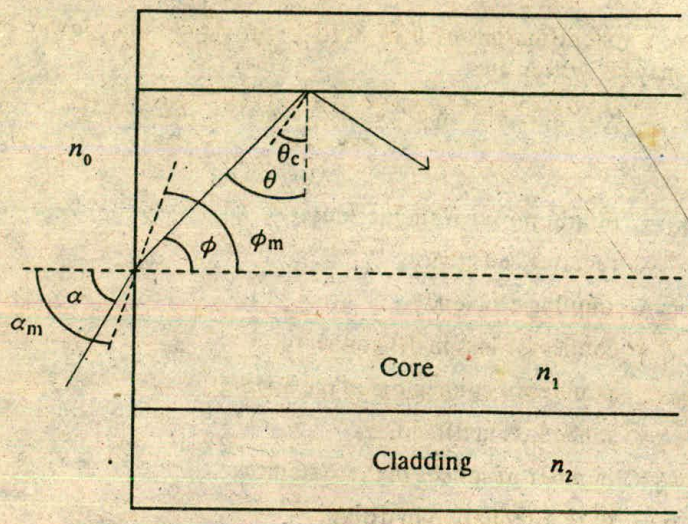


Fig. 7.17 Light propagation in a step index fibre.

$\phi_m = (\pi/2) - \theta_c$, and α is less than some maximum angle α_m. By applying Snell's law, we have

$$n_0 \sin \alpha = n_1 \sin \phi = n_1 \cos \theta \tag{7.51}$$

If we assume that the ray enters the medium from air, then $n_0 = 1$. For the critical rays,

$$\sin \alpha_m = n_1 \sin \phi_m = n_1 \cos \theta_c \tag{7.52}$$

From Eq. (7.8), we have

$$n_1 \sin \theta_c = n_2$$

Therefore,

$$\cos \theta_c = \frac{(n_1^2 - n_2^2)^{\frac{1}{2}}}{n_1}, \text{ or}$$

$$\sin \alpha_m = (n_1^2 - n_2^2)^{\frac{1}{2}} \tag{7.53}$$

Now, $n_1 - n_2 = \Delta n$. If we set $(n_1 + n_2)/2 = n$, then we have

$$\sin \alpha_m = (2n\Delta n)^{\frac{1}{2}} \tag{7.54}$$

The greater the value of α_m, the larger is the proportion of the incident light collected by the fibre. Hence, the term $n_0 \sin \alpha_m$ is known as the **numerical aperture (NA)** of the fibre. A mismatch between the numerical aperture of a fibre and the source beam results in coupling losses between the source and fibre. This type of mismatch is illustrated in Fig. 7.18 which shows a smaller acceptance cone and a wider source beam.

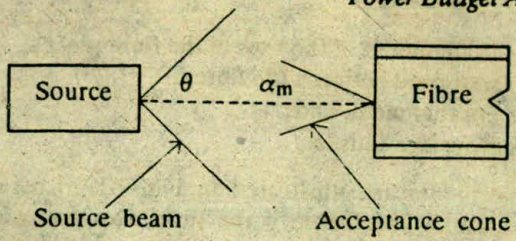


Fig. 7.18 Mismatch between source beam and fibre acceptance cone.

In order to calculate the coupling loss due to numerical aperture mismatch, the radiation characteristics of the source must be known. The fibre optic light sources may be placed under two categories with regard to their radiation characteristics: **lambertian** and **nonlambertian**. A lambertian source is an isotropic radiator whose radiation characteristics are governed by the equation

$$I(\theta) = I_0 \cos\theta \tag{7.55}$$

where

$I(\theta)$ = power radiated per unit solid angle in the direction of angle to the normal to the light emitting surface (see Fig. 7.18)

I_0 = power radiated per unit solid angle along the normal

Nonlambertian sources exhibit narrower complex radiation patterns which may be characterised by the equation

$$I(\theta) = I_0(\cos\theta)^m \tag{7.56}$$

where m is an integer greater than one. Surface emitting LEDs are modelled as lambertian sources, whereas edge emitting LEDs and ILDs are considered as nonlambertian sources.

Assuming that there is no loss due to unintercepted illumination, it can be shown that for a step index fibre and a lambertian light source,

$$\eta_c = (NA)^2 \tag{7.57}$$

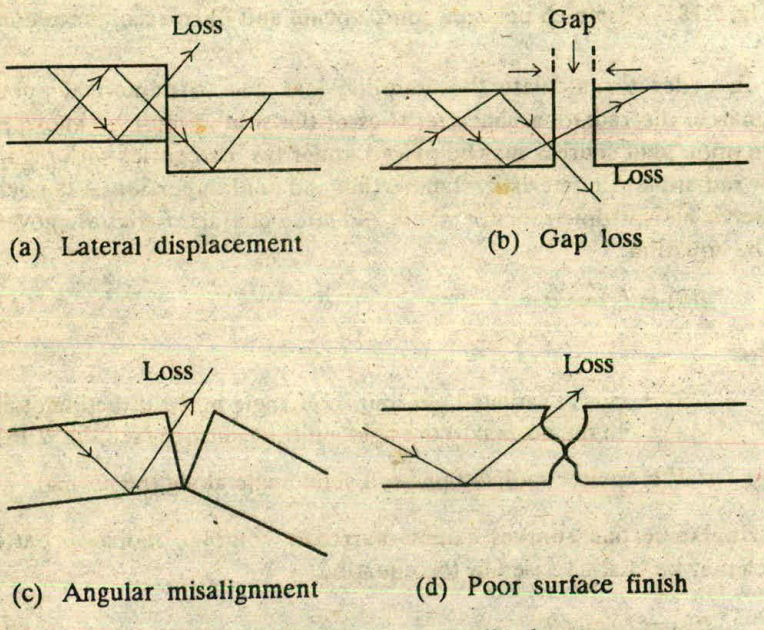
if the power is launched from air into the fibre. Otherwise,

$$\eta_c = (NA)^2/n_0^2 \tag{7.58}$$

The practical implementation of a fibre optic communication system calls for the use of connectors and splices. A connector is used in places where it is necessary to easily disconnect and reconnect fibres. A splice, on the other hand, is employed to permanently join lengths of fibres together. Losses at the connectors and splices are caused due to

1. lateral misalignment of the axes of the fibre pieces,
2. gap or separation between the fibres,
3. angular misalignment, and
4. imperfect surface finishes.

These impairments are illustrated in Fig. 7.19. Loss due to lateral displacement is usually negligible if the fibre axes are aligned to within



(a) Lateral displacement (b) Gap loss

(c) Angular misalignment (d) Poor surface finish

Fig. 7.19 Connector and splice losses.

five percent of the diameter of the smaller of the two fibres. Otherwise, the loss may run into several decibels. Splices are usually made by fusion of the two fibres. Hence, there are no gaps in splice joints. But, in the case of connector joints, the fibre ends are not made to touch each other to prevent damage to them due to rubbing while making and breaking a connection. The loss due to angular displacement is negligible if the misalignment is less than 2°. The ends of the two joining fibres should be highly polished and should fit together squarely. The imperfection tolerance in this case should be usually less than 3° off normal to ensure negligible losses.

Fibre attenuation is caused predominantly by the following factors:

1. Absorption loss
2. Scattering loss
3. Radiation loss.

Absorption loss is related to the material composition and the fabrication process of the fibre. The absorption loss may be caused by the

interaction between one or more major components of the glass or by impurities within the glass. Some of the transmitted optical power is absorbed by the atoms of the glass molecules, OH^- ions of trapped water molecules, atoms of the impurity material such as ion, copper and chromium or by valence electrons being excited to conduction band. The absorption results in random vibratory motion of the particles, leading to dissipation of heat in the fibre.

Scattering mechanisms cause the transfer of some or all of the optical power contained within one propagating mode to another mode. This tends to result in the attenuation of the transmitted light. Scattering results from nonideal physical properties of the manufactured fibre. During manufacturing, some microscopic irregularities are permanently formed in the fibre. When light rays propagating through the fibre strike some of these irregularities, they are diffracted. Diffraction causes light to disperse and spread out in many directions and some of the light escapes through the cladding. This is called **Rayleigh scattering loss** which is dependent on the wavelength.

Radiation losses are caused by small bends and kinks in the fibre. A part of the light energy which is on the outside of the bend is required to travel faster than that on the inside. As this is not possible, there occurs an energy loss through radiation.

EXAMPLE 7.8 Determine the required output power of an LED source in fibre optic communication link given that: $\eta_c = 0.6$, $P_r = -25$ dBm, $l_c = l_s = 0.5$ dB, $\alpha_f = 1$ dB/km, $L = 15$ km, $SM = 6$ dB and the number of connectors and splices put together is eight.

Solution From Eq. (7.49), we have

$$P_s - (1 - \eta_c)P_s = m_1 l_c + m_2 l_s + \alpha_f L + SM + P_r$$

$$0.6\,P_s = 0 \text{ dBm}$$

Therefore,

$$P_s = 1.678 \text{ mW}.$$

It may be noted that dB and dBm can be added to or subtracted from each other as one unit and the result is in dBm.

7.6 Telecommunication Applications

Optical fibre transmission systems are currently being favoured for a variety of applications in telephone and data networks throughout the world. Major telecommunication administrations are opting for fibre optic links as an alternative to coaxial and high frequency cable pair systems. Millimeteric waveguide transmissions are being abandoned in favour of optical fibre communications. Some of the current applications are as follows:

1. Undersea cables
2. Intercity long-haul trunks

3. Interoffice trunking in metropolitan areas
4. Local subscriber loop
5. CATV loop
6. Low and high speed data transmission
7. Metropolitan area voice/data networks
8. Local area networks.

In this section, we briefly discuss the applications enumerated in points 1–6 above. Data network applications mentioned in points 7 and 8 are discussed in Chapter 10.

Barring application mentioned in point 5, all other applications mentioned above use digital fibre optic communication systems. The most common modulation scheme used in digital systems is the pulse code modulation (PCM). In the analog domain, direct modulation of the source power is not efficient for optical systems and almost universally pulse intensity modulated wave is used as a subcarrier in the analog system. This subcarrier may, in turn, be modulated in amplitude, width, position or frequency rendering pulse amplitude modulation (PAM), pulse width modulation (PWM), pulse position modulation (PPM), and pulse frequency modulation (PFM) respectively. These schemes retain the analog modulation formats and at the same time offer advantages of digital detection.

Digital optical telecommunication systems may be considered to have gone through three generations of evolution, present day systems belonging to the third generation. Fourth generation systems are in the offing and fifth generation systems may appear by the end of this decade. A change from one generation to another is characterised by significant improvements in data rates or significant reduction in fibre losses or both. These changes are achieved by technological advances that make possible precision fabrication of different types of fibres using high purity materials, which operate at minimum attenuation and minimum dispersion wavelengths. Typical characteristics of different generations of optical fibres are summarised in Table 7.5. The data presented in the table are generally accepted as characterising the different generations. However, there are writers who tend to use somewhat different characterisation for describing the generations. The important thing to note is the development that has taken place with regard to fibre optic communication systems within a short span of about 20 years.

Digital fibre optic systems of different generations find wide spread use in trunk routes. Typical features of these systems in different trunk applications are presented in Table 7.6.

Undersea cable systems are an integral part of the international telecommunication networks. They are used as the primary means of communication over short distances (e.g. across the English Channel) and on longer routes (e.g. transAtlantic); they provide route diversity in conjunction with satellite links. Third generation fibre optic transmission systems are used as undersea cables.

Table 7.5 Characteristics of Different Generations of Optical Fibres

Generation	Typical maximum speed distance product (Mbps-km)	Attenuation (dB/km)	Fibre type	Nominal wavelength of operation (μm)
I	45	3	Multimode	0.85
	150	1	graded index	
II	500	< 1	Multimode	1.3
	10,000	< 0.3	graded index	
III	> 10,000	< 0.01	Single mode	1.3
IV			Single mode	1.55
V			Infrared fibre	2.0

Intercity trunk systems vary in distance from 50 km to over 1000 km. Both first and second generations are suitable for intercity trunk applications with typical speeds varying from 34 Mbps to 140 Mbps.

The interoffice networks span a distance of 5–20 km. The speed requirements of interoffice trunks are typically 8 Mbps and 34 Mbps which are met by the first generation transmission systems.

Table 7.6 Features of Different Digital Fibre Optic Systems

	Intercity trunk	Interoffice trunk	Undersea cable	High-speed data link
Source	Laser	Laser	Laser	Laser
Detector	APD/*p-i-n*	APD	*p-i-n*	*p-i-n*
Bit rate (Mbps)	90/140	45	274	100
Fibre	Multimode/ Single mode	Multimode	Single mode	Multimode
Attenuation (dB/km)	< 6/ < 1	< 6	< 1	< 10
Bandwidth (MHz-km)	> 300/ > 8000	> 300	> 8000	> 100
Repeater distance (km)	> 8 / > 20	> 5	> 30	< 0.5
Wavelength (μm)	0.82/1.3	0.82	1.3	1.3

Compared to their use in trunk networks, the rate of introduction of optical fibre systems in local subscriber loops is somewhat slow. However, the newly emerging demands for more wideband services like switched cable TV, videophone, teleconferencing and high speed data transmission are causing fibres to be used in subscriber loops.

Community antenna television (CATV) system is a typical application where the information is both available and required in analog form. Hence, analog fibre optic systems find widespread use in this application. By virtue of their high bandwidth, the fibre optical CATV systems are capable of supporting interactive video services. The system permits program selection through remote operation of a video switch using a key pad operated by the customer and is capable of giving access to a large number of channels. Fibre optic CATV systems are likely to bring in an era of large scale interactive video.

FURTHER READING

1. Cherin, A.H., *An Introduction to Optical Fibers,* McGraw-Hill, Tokyo, 1983.
2. Ghatak A.K., Fiber optics and optical communications: A perspective. *IETE Technical Review,* India, Vol.7, No.4, July–August 1990.
3. Ghatak, A.K. and Thyagarajan, K., *Optical Electronics,* Cambridge University Press, Cambridge, 1989.
4. Gowar, J., *Optical Communication Systems*, Prentice-Hall International (UK), London, 1984.
5. Hoss, R.J., *Fiber Optic Communications Design Handbook*. Prentice-Hall, Englewood Cliffs, New Jersey, 1990.
6. Keiser, G., *Optical Fiber Communications*, McGraw-Hill, Singapore, 1983.
7. Senior, J.M., *Optical Fiber Communication: Principles and Practices,* Prentice Hall International (UK), London, 1985.

EXERCISES

1. Describe any three advantages of optical fibre as a transmission medium when compared to copper cables.
2. Define the range of frequencies and wavelengths that are suitable for fibre optic transmission. Why is the wavelength region 1.1 to 1.6 μm preferred?
3. Cladding is not necessary for light to propagate along the fibre. Why?

4. A step index fibre has a refractive index of 1.4. If the index difference is 1%, what is the refractive index of the cladding?

5. A 50-μm graded index fibre has a triangular index profile. If the refractive index at the core axis is 1.8 and the index difference is 0.8%, estimate the slope of variation.

6. The incident angle of a light ray from air to a glass plate is 60°. On incidence, part of the ray is reflected and the other part is refracted. Reflected and refracted beams make an angle of 90° with each other. Determine the refractive index of the glass plate.

7. Calculate the numerical aperture of a step-index fibre having $n_1 = 1.45$ and $n_2 = 1.42$. What is the maximum entrance angle for this fibre?

8. In a multimode fibre, a beam of light energy experiences a total phase shift of 3080 radians. Determine the number of guided modes in this fibre.

9. Calculate the transmission frequency and the photon energy of an optical source that operates at 0.6 μm wavelength.

10. If the nonradiative recombination rate is 10% of the radiative recombination rate in an LED, determine the internal quantum efficiency. What is the bulk recombination lifetime if the radiative recombination lifetime is 5 ns?

11. A GaAs injection laser has an optical cavity of length 250 μm. Given that the reflection coefficient of the cavity is 0.32 and loss coefficient is 10 per cm, determine the gain threshold of the laser.

12. The gain threshold in an ILD is directly proportional to the current density. If the cavity width of the injection laser in Problem 11 is 100 μ m and the slope of the curve showing gain threshold as a function of current density is 0.023 A^{-1} cm, determine the threshold current for the device.

13. In Problem 11, if the refractive index of the lasing material is 2.4, determine the photon lifetime in the cavity.

14. Determine the threshold wavelength above which a photodetector will cease to operate if the bandgap energy of the photodiode material is 1.1 eV. What is the quantum efficiency if the responsivity of the device is 0.65 A/W?

15. A photodiode has a quantum efficiency of 65% when photons of energy 1.5×10^{-19} J are incident upon it. Determine the operating wavelength of the device. Calculate the incident optical power required to obtain a photocurrent of 2.5 μA.

16. An APD has a carrier multiplication factor of 20 and operates at 1.5 μm wavelength. Calculate the quantum efficiency and the output

photocurrent given that the responsivity of the device is $0.6AW^{-1}$ and 10^{10} photons are incident upon it per second.

17. A photodiode has a capacitance of 5 pF. Calculate the maximum load resistance which permits 8 MHz bandwidth.

18. The r.m.s. pulse spreading constant of an optical fibre is 0.6 ns/km. Determine the maximum bit rate for a 8-km fibre if NRZ coding is used.

19. A long haul single mode optical fibre system operating at a wavelength of 0.8 μm has the following parameters:

Mean power launched into the fibre	–3 dBm
Fibre attenuation	2 dB/km
Splice loss	0.3 dB/splice
Connector loss	1 dB/connector
Receiver sensitivity	–55 dBm
Required safety margin	9 dB

Determine the maximum possible link length without repeaters, if the link is formed by splicing 1-km fibre lengths. Assume that the connectors are used only at the ends.

8

Traffic Engineering

Traffic engineering provides the basis for the analysis and design of telecommunication networks. In Chapters 2–4 and 6 we calculated blocking probabilities due to the nonavailability of switching paths. The calculations were based on a quantity that specified the fraction of the time for which a subscriber line may be busy. In practice, the situation is much more complex. It is not only the switching elements but also many other common shared subsystems in a telecommunication network that contribute to the blocking of a subscriber call. In a telephone network, these include digit receivers, interstage switching links, call processors and trunks between exchanges. The load or the traffic pattern on the network varies during the day with heavy traffic at certain times and low traffic at other times. The task of designing cost effective networks that provide the required quality of service under varied traffic conditions demands a formal scientific basis. Such a basis is provided by traffic engineering or teletraffic theory. Traffic engineering analysis enables one to determine the ability of a telecommunication network to carry a given traffic at a particular loss probability. It provides a means to determine the quantum of common equipments required to provide a particular level of service for a given traffic pattern and volume.

8.1 Network Traffic Load and Parameters

In a telephone network, the traffic load on a typical working day during the 24 hours is shown in Fig.8.1. The originating call amplitudes are relative and the actual values depend on the area where the statistics is collected. The traffic pattern, however, is the same irrespective of the area considered. Obviously, there is little use of the network during 0 and 6 hours when most of the population is asleep. There is a large peak around mid-forenoon and mid-afternoon signifying busy office activities. The afternoon peak is, however, slightly smaller. The load is low during the lunch-hour period, i.e. 12.00–14.00 hours. The period 17.00–18.00 hours is characterised by low traffic signifying that the people are on the move from offices to their residences. The peak of domestic calls occurs after 18.00 hours when persons reach home and reduced tariff applies. In many countries including India,

the period during which the reduced tariff applies has been changed to begin later than 18.00 hours and one may expect the domestic call patterns also to change accordingly. During holidays and festival days the traffic pattern is different from that shown in Fig. 8.1. Generally, there is a peak of calls

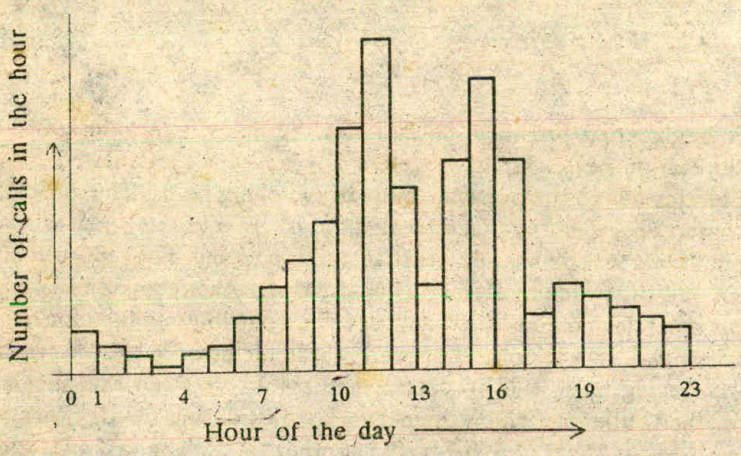


Fig. 8.1 Typical telephone traffic pattern on a working day.

around 10.00 hours just before people leave their homes on outings and another peak occurs again in the evening.

In a day, the 60-minute interval in which the traffic is the highest is called the **busy hour (BH)**. In Fig. 8.1, the 1-hour period between 11.00 and 12.00 hours is the busy hour. The busy hour may vary from exchange to exchange depending on the location and the community interest of the subscribers. The busy hour may also show seasonal, weekly and in some places even daily variations. In addition to these variations, there are also unpredictable peaks caused by stock market or money market activity, weather, natural disaster, international events, sporting events etc. To take into account such fluctuations while designing switching networks, three types of busy hours are defined by CCITT in its recommendations E.600:

1. **Busy Hour**: Continuous 1-hour period lying wholly in the time interval concerned, for which the traffic volume or the number of call attempts is greatest.
2. **Peak Busy Hour**: The busy hour each day; it usually varies from day to day, or over a number of days.
3. **Time Consistent Busy Hour**: The 1-hour period starting at the same time each day for which the average traffic volume or the number of call attempts is greatest over the days under consideration.

For ease of records, the busy hour is taken to commence on the hour or half-hour only.

Not all call attempts materialise into actual conversations for a variety of reasons such as called line busy, no answer from the called line, and blocking in the trunk groups or the switching centres. A call attempt is said to be successful or completed if the called party answers. **Call completion rate (CCR)** is defined as the ratio of the number of successful calls to the number of call attempts. The number of call attempts in the busy hour is called **busy hour call attempts (BHCA)**, which is an important parameter in deciding the processing capacity of a common control or a stored program control system of an exchange. The CCR parameter is used in dimensioning the network capacity. Networks are usually designed to provide an overall CCR of over 0.70. A CCR value of 0.75 is considered excellent and attempts to further improve the value is generally not cost effective. A related parameter that is often used in traffic engineering calculations is the **busy hour calling rate** which is defined as the average number of calls originated by a subscriber during the busy hour.

EXAMPLE 8.1 An exchange serves 2000 subscribers. If the average BHCA is 10,000 and the CCR is 60%, calculate the busy hour calling rate.

Solution

$$\text{Average busy hour calls} = \text{BHCA} \times \text{CCR} = 6000 \text{ calls}$$

$$\text{Busy hour calling rate} = \frac{\text{average busy hour calls}}{\text{total number of subscribers}} = 3$$

The busy hour calling rate is useful in sizing the exchange to handle the peak traffic. In a rural exchange, the busy hour calling rate may be as low as 0.2, whereas in a business city it may be as high as three or more. Another useful information is to know how much of the day's total traffic is carried during the busy hour. This is measured in terms of **day-to-busy hour traffic ratio** which is the ratio of busy hour calling rate to the average calling rate for the day. Typically, this ratio may be over 20 for a city business area and around six or seven for a rural area.

The traffic load on a given network may be on the local switching unit, interoffice trunk lines or other common subsystems. For analytical treatment in this text, all the common subsystems of a telecommunication network are collectively termed as **servers**. In other publications, the term **link** or **trunk** is used. The traffic on the network may then be measured in terms of the occupancy of the servers in the network. Such a measure is called the **traffic intensity** which is defined as

$$A_0 = \frac{\text{period for which a server is occupied}}{\text{total period of observation}} \tag{8.1}$$

Generally, the period of observation is taken as one hour. A_0 is obviously dimensionless. It is called **erlang (E)** to honour the Danish telephone engineer A.K. Erlang, who did pioneering work in traffic engineering. His paper on traffic theory published in 1909 is now regarded as a classic. A server is said to have 1 erlang of traffic if it is occupied for the entire period of observation. Traffic intensity may also be specified over a number of servers.

EXAMPLE 8.2 In a group of 10 servers, each is occupied for 30 minutes in an observation interval of two hours. Calculate the traffic carried by the group.

Solution

$$\text{Traffic carried per server} = \frac{\text{occupied duration}}{\text{total duration}}$$

$$= \frac{30}{120} = 0.25 \text{ E}$$

Total traffic carried by the group $10 \times 0.25 = 2.5$ E

Erlang measure indicates the average number of servers occupied and is useful in deriving the average number of calls put through during the period of observation.

EXAMPLE 8.3 A group of 20 servers carry a traffic of 10 erlangs. If the average duration of a call is three minutes, calculate the number of calls put through by a single server and the group as a whole in a one-hour period.

Solution

$$\text{Traffic per server} = \frac{10}{20} = 0.5 \text{ E}$$

i.e. a server is busy for 30 minutes in one hour.

$$\text{Number of calls put through by one server} = \frac{30}{3} = 10 \text{ calls}$$

$$\text{Total number of calls put through by the group} = 10 \times 20 = 200 \text{ calls}$$

Traffic intensity is also measured in another way. This measure is known as **centum call second (CCS)** which represents a call-time product. One CCS may mean one call for 100 seconds duration or 100 calls for one second duration each or any other combination. CCS as a measure of traffic intensity is valid only in telephone circuits. For the present day networks which support voice, data and many other services, erlang is a better measure to use for representing the traffic intensity. Sometimes, **call seconds (CS)** and **call minutes (CM)** are also used as a measure of traffic intensity. We may note that

$$1\,E = 36\,CCS = 3600\,CS = 60\,CM$$

EXAMPLE 8.4 A subscriber makes three phone calls of three minutes, four minutes and two minutes duration in a one-hour period. Calculate the subscriber traffic in erlangs, CCS and CM.

Solution

$$\text{Subscriber traffic in erlangs} = \frac{\text{busy period}}{\text{total period}} = \frac{3+4+2}{60} = 0.15\,E$$

$$\text{Traffic in CCS} = \frac{(3+4+2)\times 60}{100} = \frac{540}{100} = 5.4\,CCS$$

$$\text{Traffic in CM} = 3 + 4 + 2 = 9\,CM$$

Whenever we use the terminology "subscriber traffic" or "trunk traffic", we mean the traffic intensity contributed by a subscriber or the traffic intensity on a trunk. As mentioned above, traffic intensity is a call-time product. Hence two important parameters that are required to estimate the traffic intensity or the network load are

- Average call arrival rate, C
- Average holding time per call, t_h

We can then express the load offered to the network in terms of these parameters as

$$A = Ct_h \tag{8.2}$$

C and t_h must be expressed in like-time units. For example, if C is in number of calls per minute, t_h must be in minutes per call.

EXAMPLE 8.5 Over a 20-minute observation interval, 40 subscribers initiate calls. Total duration of the calls is 4800 seconds. Calculate the load offered to the network by the subscribers and the average subscriber traffic.

Solution

$$\text{Mean arrival rate } C = 40/20 = 2 \text{ calls/minute}$$

$$\text{Mean holding time } t_h = \frac{4800}{40\times 60} = 2 \text{ minutes/call}$$

Therefore,

$$\text{Offered load} = 2 \times 2 = 4\,E$$

$$\text{Average subscriber traffic} = 4/40 = 0.1\,E$$

In the above discussions, we have calculated the traffic in two ways: one based on the traffic generated by the subscribers and the other based on the observation of busy servers in the network. It is possible that the load

generated by the subscribers sometimes exceeds the network capacity. There are two ways in which this overload traffic may be handled: The overload traffic may be rejected without being serviced or held in a queue until the network facilities become available. In the first case, the calls are lost and in the second case the calls are delayed. Correspondingly, two types of systems, called **loss systems** and **delay systems** are encountered. Conventional automatic telephone exchanges behave like loss systems. Under overload traffic conditions a user call is blocked and is not serviced unless the user makes a retry. On the other hand, operator-oriented manual exchanges can be considered as delay systems. A good operator registers the user request and establishes connection as soon as network facilities become available without the user having to make another request. In data networks, circuit-switched networks behave as loss systems whereas store-and-forward (S&F) message or packet networks behave as delay systems. In the limit, delay systems behave as loss systems. For example, in a S&F network if the queue buffers become full, then further requests have to be rejected.

The basic performance parameters for a loss system are the grade of service and the blocking probability, and for a delay system, the service delays. Average delays, or probability of delay exceeding a certain limit, or variance of delays may be important under different circumstances. The traffic models used for studying loss systems are known as **blocking or congestion models** and the ones used for studying delay systems are called **queuing models**.

8.2 Grade of Service and Blocking Probability

In loss systems, the traffic carried by the network is generally lower than the actual traffic offered to the network by the subscribers. The overload traffic is rejected and hence is not carried by the network. The amount of traffic rejected by the network is an index of the quality of the service offered by the network. This is termed **grade of service** (GOS) and is defined as the ratio of **lost traffic** to **offered traffic**. Offered traffic is the product of the average number of calls generated by the users and the average holding time per call. This is given by Eq. (8.2). On the other hand, the actual traffic carried by the network is called the **carried traffic** and is the average occupancy of the servers in the network as given by Eq. (8.1). Accordingly, GOS is given by

$$\text{GOS} = \frac{A - A_0}{A} \tag{8.3}$$

where

A = offered traffic
A_0 = carried traffic
$A - A_0$ = lost traffic

The smaller the value of grade of service, the better is the service. The recommended value for GOS in India is 0.002 which means that two calls in every 1000 calls or one call in every 500 calls may be lost. Usually, every common subsystem in a network has an associated GOS value. The GOS of the full network is determined by the highest GOS value of the subsystems in a simplistic sense. A better estimate takes into account the connectivity of the subsystems, such as parallel units. Since the volume of traffic grows as the time passes by, the GOS value of a network deteriorates with time. In order to maintain the value within reasonable limits, initially the network is sized to have a much smaller GOS value than the recommended one so that the GOS value continues to be within limits as the network traffic grows.

The blocking probability P_B is defined as the probability that all the servers in a system are busy. When all the servers are busy, no further traffic can be carried by the system and the arriving subscriber traffic is blocked. At the first instance, it may appear that the blocking probability is the same measure as the GOS. The probability that all the servers are busy may well represent the fraction of the calls lost, which is what the GOS is all about. However, this is generally not true. For example, in a system with equal number of servers and subscribers, the GOS is zero as there is always a server available to a subscriber. On the other hand, there is a definite probability that all the servers are busy at a given instant and hence the blocking probability is nonzero. The fundamental difference is that the GOS is a measure from the subscriber point of view whereas tne blocking probability is a measure from the network or switching system point of view. GOS is arrived at by observing the number of rejected subscriber calls , whereas the blocking probability is arrived at by observing the busy servers in the switching system. We shall show later in this chapter, through analysis carried out on loss systems, that GOS and P_B may have different values depending upon the traffic characterisation model used. In order to distinguish between these two terms clearly, GOS is called **call congestion** or **loss probability** and the blocking probability is called **time congestion.**

In the case of delay systems, the traffic carried by the network is the same as the load offered to the network by the subscribers. Since the overload traffic is queued, all calls are put through the network as and when the network facilities become available. GOS as defined above is not meaningful in the case of delay systems since it has a value of zero always. The probability that a call experiences delay, termed **delay probability**, is a useful measure, as far as the delay systems are concerned. If the offered load or the input rate of traffic far exceeds the network capacity, then the queue lengths become very large and the calls experience undesirably long delays. Under such circumstances, the delay systems are said to be unstable as they would never be able to clear the offered load. In practice, it is possible that there are some spurts of traffic that tend to take the delay systems to the unstable region of operation. An easy way of bringing the system back to stable region of

operation is to make it behave like a loss system until the queued up traffic is cleared to an acceptable limit. This technique of maintaining the stable operation is called **flow control.**

Some authors use the term GOS in a broader sense, covering both loss systems and delay systems and representing both loss and delay probabilities in the respective systems. In this text, we use the term GOS in the conventional sense to represent the loss probability only as defined in Eq. (8.3). In more recent times, a new term called **quality of service** (QOS) is being used and it is considered to be more general than the GOS and includes other factors like quality of speech, error-free transmission capability etc. It is important that the reader be warned of the fact that these terms are used loosely by many writers and the reader will do well to verify what exactly an author means by different terms. It is essential to recognise that there are two different important performance measures, one obtained by observing the subscriber traffic and the other by observing the network behaviour. The terms that are used synonymously in this text from the subscriber and network viewpoints are

Subscriber viewpoint:

GOS = call congestion = loss probability

Network viewpoint:

Blocking probability = time congestion

8.3 Modelling Switching Systems

A telecommunication network carries traffic generated by a large number of individual subscribers connected to the network. Subscribers generate calls in a random manner. The call generation by the subscribers and therefore the behaviour of the network or the switching systems in it can be described as a random process. A **random process** or a **stochastic process** is one in which one or more quantities vary with time in such a way that the instantaneous values of the quantities are not determinable precisely but are predictable with certain probability. The quantities are called **random variables.** Thus, a stochastic process is a time indexed function of one or more random variables. It is generally possible to characterise the behaviour of a random process by some statistical properties and thus predict the future performance with a certain probability. The telephone traffic qualifies as a stochastic process where the number of simultaneously active subscribers and the number of simultaneously busy servers are random variables. It is not possible to precisely estimate the number of simultaneously active subscribers at a given instant of time but a prediction can be made with a certain probability. Figure 8.2 shows typical fluctuations in the number of simultaneous calls in a half-hour period. The pattern signifies a typical random process.

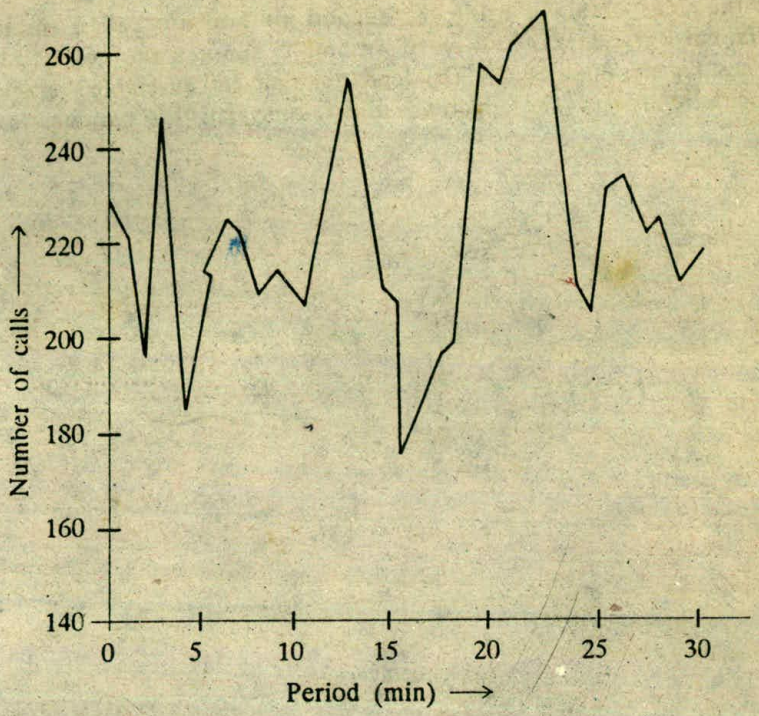


Fig. 8.2 Typical fluctuations in the number of telephone calls.

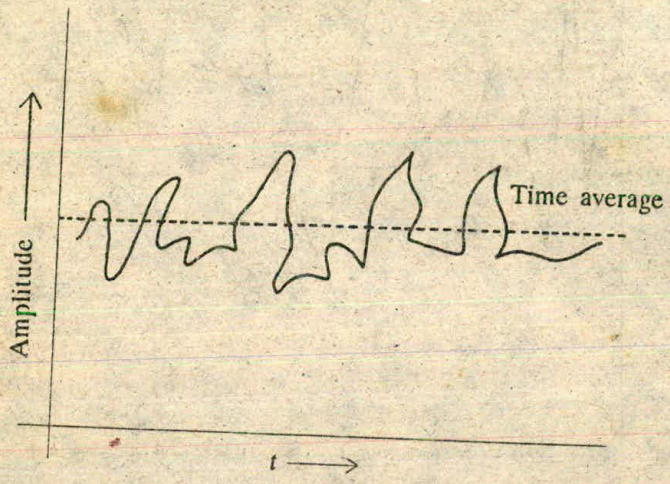
The values taken on by the random variables of a random process may be discrete or continuous. In the case of telephone traffic, the random variable representing the number of simultaneous calls can take on only discrete values whereas a random variable representing temperature variations in an experiment can take on continuous values. Similarly, the time index of the random variables can be discrete or continuous. Accordingly, we have four different types of stochastic processes:

1. Continuous time continuous state
2. Continuous time discrete state
3. Discrete time continuous state
4. Discrete time discrete state.

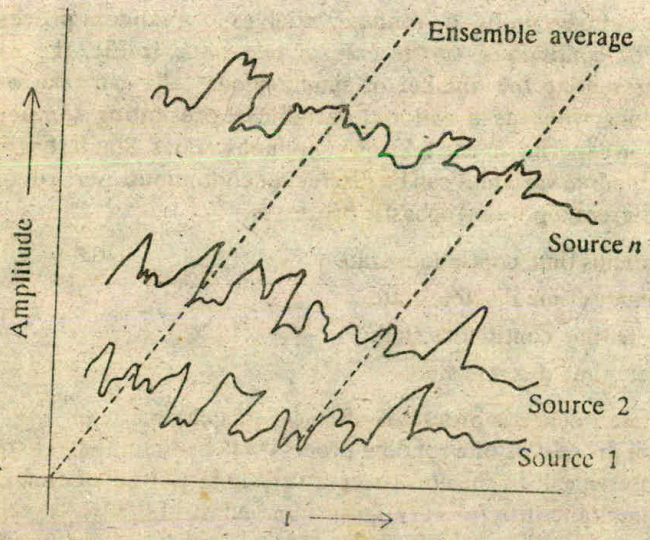
A discrete state stochastic process is often called a chain.

Statistical properties of a random process may be obtained in two ways: either by observing its behaviour over a very long period of time, or by observing simultaneously, a very large number of statistically identical random sources at any given instant of time. The statistical properties obtained using the first method are known as time statistical parameters,

and the ones obtained using the second method are called **ensemble statistical parameters**. The word ensemble denotes the collection of sources. Figures 8.3(a) and 8.3(b) depict the time and ensemble averages of a stochastic process. These methods of obtaining statistical parameters of a



(a) Time average



(b) Ensemble average

Fig. 8.3 Time and ensemble averages of stochastic averages.

andom process are analogous to determining the statistical behaviour of a lie by throwing one die a very large number of times or by throwing a very arge number of identical dice all together and observing the outcomes. In the first case, a particular outcome would occur one-sixth of the times the experiment is performed and in the second case one-sixth of the outcomes could correspond to a particular face value of the die. The statistical parameters obtained in either case are the same.

In the case of random processes, the ensemble and time statistics need not be the same for two reasons. First, the statistical properties of a random process may themselves vary with time in which case the ensemble statistics will be different at different instants of time. In order that the ensemble and time statistics are the same, it is necessary but not sufficient that the statistical parameters of the random processes under consideration remain unchanged over time. Random processes whose statistical parameters do not change with time are known as **stationary processes**. Second, the statistical properties of the ensemble of the sources, though stationary, need not be identical, in which case the time statistics will depend on the sources selected for observation. Thus, the second condition that must be satisfied is that all random processes in the ensemble must have identical statistical properties. When both the conditions are met, the time and the ensemble statistical parameters are identical. The random processes which have identical time and ensemble averages are known as **ergodic processes.** It may be noted that an ergodic process is stationary, whereas a stationary process need not necessarily be ergodic. In some random processes, the mean and the variance alone are stationary and other higher order moments may vary with time. Such processes are known as **wide-sense stationary processes.**

Telephone traffic is nonstationary as is seen from Fig. 8.1. The mean and variance of the traffic are different during different times of the day. However, if we consider segments of telephone traffic during a day, say, 6–9 hours, 10–11 hours, 12–13 hours etc., the traffic may be considered to be stationary. Nonstationary processes are difficult to model and analyse. Accordingly, we model and analyse telephone traffic in segments when they can be considered to be stationary.

The behaviour of a switching system can also be modelled as a stochastic process. For example, the number of servers busy simultaneously is a discrete random variable. The time at which a server becomes busy or free also exhibits a random behaviour, and hence the entire system can be modelled as a random process. It is also obvious that our interest lies in the discrete state random processes and, therefore, in our modelling we use discrete state stochastic processes.

8.3.1 Markov Processes

Markov Processes form an important class of random processes that have some special properties. These properties were first defined and investigated

by A.A Markov in 1907. He proposed a simple and highly useful form of dependency among the random variables forming a stochastic process. This class of processes is of great interest to our modelling of switching systems. A discrete time Markov chain, i.e., discrete time discrete state Markov process is defined as one which has the following property:

$$P[\{X(t_{n+1}) = x_{n+1}\}/\{X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \cdots, X(t_1) = x_1\}] = P[\{X(t_{n+1}) = x_{n+1}\}/\{X(t_n) = x_n\}] \quad (8.4)$$

where $t_1 < t_2 \ldots < t_n < t_{n+1}$ and x_i is the i-th discrete state space value.

Equation (8.4) states that the probability that the random variable X takes on the value x_{n+1} at time step $n+1$ is entirely determined by its state value in the previous time step n and is independent of its state values in earlier time steps; $n-1, n-2, n-3$ etc. We may also state that the entire past history of the process is summarised in its current state and hence the next state is determined only by the current state. If this is the case, the time period for which the process has stayed in the current state should play no role in determining the next state. We, however, require some specification of the time that elapses between state transitions. Since the process is random, the time specification also should be in terms of a random variable. This means that we are now seeking a probability distribution function for time which would in no way influence the state transitions of a process. In other words, the duration for which a process has stayed in a particular state does not influence the next state transition. There are only two distribution functions that satisfy this criterion. One is the exponential distribution, which is continuous, and the other is the geometric distribution which is discrete. Thus, the interstate transition time in a discrete time Markov process is geometrically distributed and in a continuous time Markov process it is exponentially distributed. In view of the above described property of these distributions, they are said to be **memoryless**.

8.3.2 Birth-death Processes

If we apply the restriction that the state transitions of a Markov chain can occur only to the adjacent states, then we obtain what are known as **birth-death (B-D) processes**. It is customary to talk of population in a B-D process. The number in the population is a random variable and represents the state value of the process. The B-D process moves from its state k to state $k-1$ if a death occurs or moves to state $k+1$, if a birth occurs; it stays in the same state if there is no birth or death during the time period under consideration. This behaviour of the birth-death process is depicted in Fig. 8.4.

Birth-death processes are very useful in our analysis of telecommunication networks. A telecommunication network can be modelled as a B-D

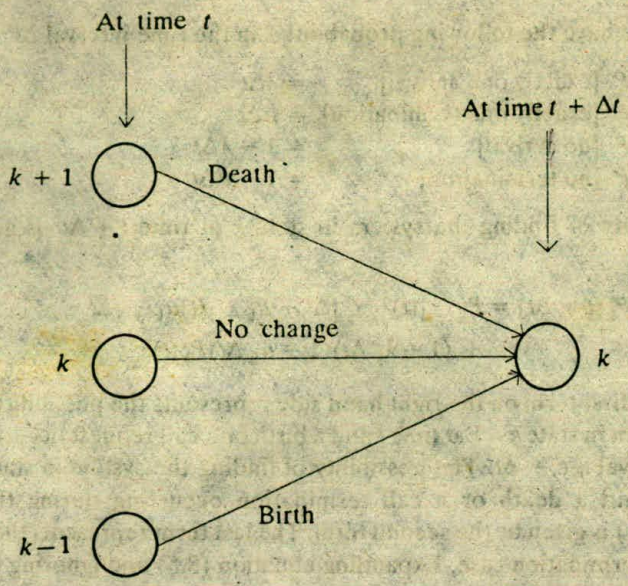


Fig. 8.4 State transitions at a Birth-Death process.

process where the number of busy servers represents the population, a call request means a birth and a call termination implies a death.

In order to analyse a B-D process, we shall choose a time interval Δt small enough such that

1. there can almost be only one state transition in that interval,
2. there is only one arrival or one termination but not both, and
3. there may be no arrival or termination leaving the state unchanged in the time interval Δt.

We further assume that

- the probability of an arrival or termination in a particular interval is independent of what had happened in the earlier time intervals and
- the probability of an arrival is directly proportional to the time interval Δt.

With these assumptions we now proceed to determine the dynamics of a switching system modelled as a birth-death process. Let

$P_k(t)$ = the probability that the system is in state k at time t, i.e. k servers are busy at time t

λ_k = call arrival rate in state k

μ_k = call termination rate in state k

Then, we have the following probabilities in the time interval Δt :

$$
\begin{aligned}
&P \text{ [exactly one arrival]} &&= \lambda \Delta t \\
&P \text{ [exactly one termination]} &&= \mu \Delta t \\
&P \text{ [no arrival]} &&= 1 - \lambda \Delta t \\
&P \text{ [no termination]} &&= 1 - \mu \Delta t
\end{aligned}
$$

Probability of finding the system in state k at time $t + \Delta t$ is given by the equation

$$
\begin{aligned}
P_k(t + \Delta t) = {} & P_{k-1}(t)\lambda_{k-1}\Delta t + P_{k+1}(t)\mu_{k+1}\Delta t \\
& + (1 - \lambda_k \Delta t)(1 - \mu_k \Delta t)P_k(t) \qquad (8.5)
\end{aligned}
$$

The first term on the right hand side represents the possibility of finding the system in state $k - 1$ at time t and a birth or a call request occurring during the interval $t, t + \Delta t$. The possibility of finding the system in state $k + 1$ at time t and a death or a call termination occurring during the interval $(t, t + \Delta t)$ is given by the second term. The last term represents the no arrival and no termination case. Expanding equation (8.5) and ignoring the second order Δt term, we get

$$
\begin{aligned}
P_k(t + \Delta t) = {} & P_{k-1}(t)\lambda_{k-1}\Delta t + P_{k+1}(t)\mu_{k+1}\Delta t \\
& - (\lambda_k + \mu_k)P_k(t)\Delta t + P_k(t) \qquad (8.6)
\end{aligned}
$$

Our interest lies in determining the dynamics of the system which is given by the rate of change of the probability P_k with time. Rearranging the terms in Eq. (8.6), we get

$$
\begin{aligned}
\frac{P_k(t + \Delta t) - P_k(t)}{\Delta t} = {} & P_{k-1}(t)\lambda_{k-1} + P_{k+1}(t)\mu_{k+1} \\
& - (\lambda_k + \mu_k)P_k(t) \qquad (8.7)
\end{aligned}
$$

In the limit $\Delta t \to 0$, we get

$$
\frac{dP_k(t)}{dt} = P_{k-1}(t)\lambda_{k-1} + P_{k+1}(t)\mu_{k+1} - (\lambda_k + \mu_k)P_k(t) \qquad (8.8)
$$

This is the differential equation governing the dynamics of a B-D process. Equation (8.8) applies for all values of $k \geq 1$. For $k = 0$, i.e. no calls in progress, there can be no termination of a call. In other words, $\mu_0 = 0$. Further, there can be no state with -1 as the state value. Therefore, for $k = 0$, Eq. (8.8) modifies as

$$
\frac{dP_0(t)}{dt} = P_1(t)\mu_1 - \lambda_0 P_0(t) \qquad (8.9)
$$

While the above equations actually give the rate of change of state probabilities, we are often concerned with the steady state operation of the networks. Under steady state conditions, the state probabilities reach an

equilibrium value and do not change with time, i.e. $P_k(t_1) = P_k(t_2) = P_k(t_i) = P_k$. For example, when commercial offices start functioning around 10.00 hours, switching networks pick up traffic and quickly reach a steady state condition when the state probabilities stabilize. Under these conditions, we have

$$\frac{dP_k(t)}{dt} = 0$$

and the B-D process becomes stationary. Therefore, the steady state equations of a B-D process are

$$P_{k-1}\lambda_{k-1} + P_{k+1}\mu_{k+1} - (\lambda_k + \mu_k)P_k = 0 \quad \text{for } k \geq 1 \tag{8.10}$$

$$P_1\mu_1 - \lambda_0 P_0 = 0 \quad \text{for } k = 0 \tag{8.11}$$

It may be noted that the steady state behaviour of a telecommunication switching system is governed by Eqs. (8.10) and (8.11), when the system is modelled as a B-D process.

8.4 Incoming Traffic and Service Time Characterisation

We have seen in Section 8.3 that a B-D process is useful in modelling telecommunication networks. Traffic in a telecommunication network is the aggregate of the traffic generated by a large number of individual subscribers connected to the network. Subscribers generate calls in a random manner and hence the telecommunication traffic is characterised as a random process. Whenever a subscriber originates a call, he adds one to the number of calls arriving at the network and has no way by which he can reduce the number of calls that have already arrived. We are thus in need of a model that describes an originating process. Interestingly, this process can be treated as a special case of the B-D process in which the death rate is equal to zero. In other words, there is no death occurring in the process. Such a process is known as a **renewal process**. It is a **pure birth process** in the sense that it can only add to the population as the time goes by and cannot deplete the population by itself. The equations governing the dynamics of a renewal process can be easily arrived at from the B-D process equations by setting $\mu_k = 0$. We thus have

$$\frac{dP_k(t)}{dt} = P_{k-1}(t)\lambda_{k-1} - \lambda_k P_k(t) \quad \text{for } k \geq 1 \tag{8.12}$$

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) \quad \text{for } k = 0 \tag{8.13}$$

Steady state equations are not relevant in a renewal process where we are counting the number of call originations or births. Let us say that we start

the observation at the time $t = 0$. As soon as a birth occurs, say at time $t = t_1$, it is impossible to find the system in state 0. Similarly, after i births, it is not possible for the system to be in state $i - 1$, $i - 2$ etc. Therefore, our concern is only in the dynamics of the system to answer questions like: What is the probability that there are k births in a given time interval t?

In Eqs. (8.10)–(8.13), the birth rate is dependent upon the state of the system. If we assume a constant birth rate λ which is independent of the state of the system, then we get a **Poisson process**. The governing equations of a Poisson process are

$$\frac{dP_k(t)}{dt} = \lambda P_{k-1}(t) - \lambda P_k(t) \quad \text{for } k \geq 1 \tag{8.14a}$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \quad \text{for } k = 0 \tag{8.14b}$$

In order to solve these equations, we have to assume certain boundary conditions, e.g. at time $t = 0$, the system is in state zero, i.e. no births have taken place. We then have

$$P_k(0) = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

With these conditions we get the solution for Eq. (8.14b) as

$$P_0(t) = e^{-\lambda t} \tag{8.15}$$

From Eq. (8.14a) and (8.15) we obtain, for $k = 1$,

$$\frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda e^{-\lambda t}$$

Solving this equation we get

$$P_1(t) = \lambda t e^{-\lambda t}$$

For $k = 2$, the solution is

$$P_2(t) = \frac{(\lambda t)^2 e^{-\lambda t}}{2!}$$

By induction we write the general solution as

$$P_k(t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{8.16}$$

Equation (8.16) is the most celebrated Poisson arrival process equation. The equation expresses the probability of finding the system with k members in the population at time t. In other words, it represents the probability of k arrivals in the time interval t. Equation (8.15) represents the probability of

zero arrival in a given time interval which is nothing but the probability distribution of interarrival times, i.e. the time that elapses between two arrivals. Thus, in a Poisson process, the interarrival time is exponentially distributed. This should be so, as the Poisson process is a Markov process which demands, as described earlier, that the interstate transition times be exponentially distributed.

EXAMPLE 8.6 A rural telephone exchange normally experiences four call originations per minute. What is the probability that exactly eight calls occur in an arbitrarily chosen interval of 30 seconds?

Solution

$\lambda = 1/15$ calls per second

When $t = 30$ s, $\lambda t = 2$. Therefore, the probability of exactly eight arrivals is given by

$$P_8(30) = \frac{2^8 e^{-2}}{8!} = 0.00086$$

In Example 8.6, eight arrivals in 30 seconds represent a four-fold increase in the normal traffic. The solution shows that the probability of such event occurring is very low. The reader would appreciate that such calculations are useful in sizing a switching system.

It is important to recognise that the assumption of an arrival rate independent of the state of the system in a Poisson process implies that we are dealing with infinite number of sources or essentially constant number of sources. If a number of arrivals occur immediately before any subinterval in question, some of the sources become busy and cannot generate further requests. The effect of busy sources is to reduce the average arrival rate unless the source population is infinite or large enough so as not to be affected significantly by the busy sources.

EXAMPLE 8.7 A switching system serves 10,000 subscribers with a traffic intensity of 0.1 E per subscriber. If there is a sudden spurt in the traffic, increasing the average traffic by 50%, what is the effect on the arrival rate?

Solution Number of active subscribers during

(a) normal traffic = 1000, (b) increased traffic = 1500

Number of available subscribers for generating new traffic during

(a) normal traffic = 9000, (b) increased traffic = 8500

$$\text{Change in the arrival rate} = \frac{500}{9000} \times 100 = 5.6\%$$

Figure 8.5 shows the relationship between the different Markov processes we have discussed so far, in the form of a Venn diagram. From the

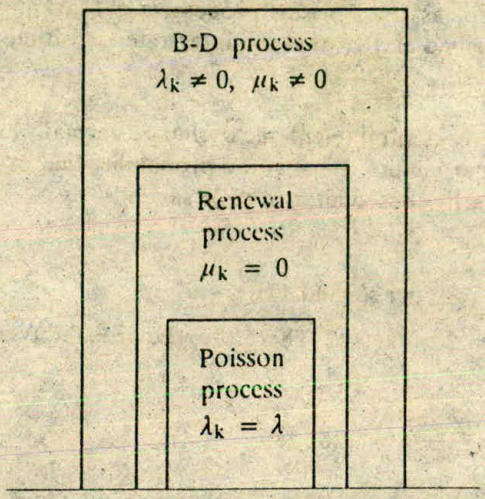


Fig. 8.5 Relationship among different Markov processes.

Venn diagram, we may describe the Poisson process as follows:

1. A pure birth process with constant birth rate
2. A birth-death process with zero death rate and a constant birth rate
3. A Markov process with state transitions limited to the next higher state or to the same state, and having a constant transition rate.

In real life, Poisson process occurs very often. Examples include:

1. Number of telephone calls arriving at an exchange
2. Number of coughs generated in a medical ward by the patients
3. Number of rainy days in a year
4. Number of typing errors in a manuscript
5. Number of bit errors occurring in a data communication system.

It is seen from the first example above that the telephone calls arriving at an exchange follow a Poisson process. But the process of call generation by the subscribers is not a Poisson process and is a renewal process. The question then arises: How is it that a renewal process at the subscriber end becomes a Poisson process at the exchange end? It can be shown that a superimposition of a large number of renewal processes results in a Poisson process. This is why we observe Poisson processes in nature, whenever we study the aggregate behaviour of a large population. All examples of Poisson processes stated above represent this phenomenon.

In a system modelled as a B-D process, the termination phenomenon can be characterised as a **pure death process**. We obtain a pure death process from a B-D process by setting the birth rate equal to zero. We thus obtain the equations governing the dynamics of a pure death process from Eqs. (8.8) and (8.9) as

$$\frac{dP_k(t)}{dt} = \mu_{k+1} P_{k+1}(t) - \mu_k P_k(t) \quad \text{for } k \geq 1 \tag{8.17}$$

$$\frac{dP_0(t)}{dt} = \mu_1 P_1(t) \tag{8.18}$$

We can further simplify the behaviour of the pure death process assuming a constant death rate similar to the assumption we made with regard to the birth rate in the case of Poisson process. We then obtain the governing equations as

$$\frac{dP_k(t)}{dt} = \mu P_{k+1}(t) - \mu P_k(t) \quad \text{for } k \geq 1 \tag{8.19}$$

$$\frac{dP_0(t)}{dt} = \mu P_1(t) \quad \text{for } k = 0 \tag{8.20}$$

The above equations can be solved by assuming suitable boundary conditions. We assume that at time $t = 0$ there are N members in the population. We then obtain the equations as

$$\frac{dP_k(t)}{dt} = \mu P_{k+1}(t) - \mu P_k(t) \quad \text{for } 0 < k < N \tag{8.21}$$

$$\frac{dP_N(t)}{dt} = -\mu P_N(t) \quad \text{for } k = N \tag{8.22}$$

$$\frac{dP_0(t)}{dt} = \mu P_1(t) \quad \text{for } k = 0 \tag{8.23}$$

Solving these equations, we get

$$P_N(t) = e^{-\mu t} \quad \text{for } k = N \tag{8.24}$$

$$P_k(t) = \frac{(\mu t)^{N-k}}{(N-k)!} e^{-\mu t} \quad \text{for } 0 < k < N \tag{8.25}$$

$$P_0(t) = \frac{(\mu t)^{N-1}}{(N-1)!} e^{-\mu t} \quad \text{for } k = 0 \tag{8.26}$$

Equation (8.25) expresses the probability of no termination or death in a given interval as the population remains at the initial level. This is nothing but the probability distribution of how long the system remains in a state without a death or termination occurring. In other words, this is the probability

distribution of the service times or the holding times in the case of calls in a switching system. Thus a pure death process has a service time distribution which is exponentially distributed, as required in the case of a Markov process.

The holding times of conventional voice conversations are well approximated by an exponential distribution. But for activities such as call processing and interoffice signalling, a constant holding time is a better model. In this case, it is straightforward to use Eq. (8.16), i.e. the Poisson equation, to determine the probability distribution of active channels. Let t_m be the constant holding time. Then, the probability of finding k busy channels at any instant of time is merely the probability that there are k arrivals during the time interval t_m immediately preceding the instant of observation, which is given by

$$P_k(t_m) = \frac{(\lambda t_m)^k \exp(-\lambda t_m)}{k!} \tag{8.27}$$

To sum up, in Section 8.3, we proposed the use of a birth-death process for modelling a telecommunication system. In this section, we have shown that the call arrivals can be modelled as a Poisson process and we describe how the holding times of the calls have an exponential distribution when the system is modelled as a B-D process. A constant holding time distribution is useful for modelling exchange functions.

8.5 Blocking Models and Loss Estimates

In Section 8.1, we pointed out that depending on the way in which overflow traffic is handled, telecommunication systems may be classified as loss systems or delay systems. The behaviour of loss systems is studied by using blocking models and that of the delay systems by using queuing models. In this section, we analyse loss systems; and in Section 8.6 we deal with delay systems.

From the discussion in Sections 8.3 and 8.4, it is apparent that we are concerned with three aspects while dealing with the analysis of the telecommunication systems:

1. Modelling the system
2. Traffic arrival model
3. Service time distribution

We model the system as a B-D process, the arrival as a Poisson process, and the holding time as an exponential or constant time distribution.

In loss systems, the overflow traffic is rejected. In other words, the overflow traffic experiences blocking from the network. What happens to the overflow traffic that is rejected? There are three ways in which overflow traffic may be handled:

1. The traffic rejected by one set of resources may be cleared by another set of resources in the network.
2. The traffic may return to the same resource after sometime.
3. The traffic may be held by the resource as if being serviced but actually serviced only after the resources become available.

It is important to recognise that case 3 above is not the one where a call is queued and then serviced. Here, a call is accepted and the customer is allowed to proceed with his information exchange process. No resources are allocated immediately although the call has been accepted. Some part of the initial information from the subscriber may be lost. But the resources are allocated soon enough so that the loss is unnoticeable by the subscriber or within acceptable limits. This aspect will be explained in more detail later. Corresponding to the above three cases, we consider three models of loss systems:

1. Lost calls cleared (LCC)
2. Lost calls returned (LCR)
3. Lost calls held (LCH).

8.5.1 Lost Calls Cleared System with Infinite Sources

We first analyse an LCC system assuming infinite number of subscribers. The assumption permits us to use Poisson arrival model for the traffic. The arrival rate is independent of the number of subscribers already busy and remains constant irrespective of the state of the system. Lost calls cleared model is well suited to the study of the behaviour of trunk transmission systems. Usually, there are many trunk groups emanating from a switching office and terminating on adjacent switching offices. Whenever a direct trunk group between two switching offices is busy, it is possible to divert the traffic via other switching offices using different trunk groups. In this way, the blocked calls in one trunk group are cleared via other trunk groups. In the context of subscriber calls, the LCC model assumes that a subscriber on hearing the engaged tone, hangs up and waits for some length of time before reattempting. He does not reattempt immediately or within a short time. Such calls are considered to have been cleared from the system and the reattempts are treated as new calls. The LCC model is used as a standard for the design and analysis of telecommunication networks in Europe, India and other countries that adopt European practices.

The LCC model was first studied comprehensively and accurately by A.K. Erlang in 1917. The main purpose of the analysis is to estimate the blocking probability, and the grade of service. Using Eq. (8.2), we can express the offered traffic A for Poisson arrival process as

$$A = \lambda t_h \tag{8.28}$$

where λ is the average Poisson call arrival rate. It be may be recalled that the quantity C in Eq. (8.2) represents the average call arrival rate irrespective of the arrival distribution, whereas λ represents the Poisson arrival rate here.

When all the servers in the system are busy, any traffic generated by the Poisson process is rejected by the system. Since the overflow traffic is lost, as far as the network is concerned, there is a different arrival rate which we call effective arrival rate. We denote the mean effective arrival rate as C_0, and the effective arrival rate in state i as C_i. The system is said to be in state j when j servers are busy. As long as all the servers in the system are not busy, the entire incoming traffic is carried by the network and when all the servers are busy no traffic is accepted by the network. Such a traffic on the network is known as **Erlang traffic** or **pure chance traffic of type 1**. In this case, we have

$$C_i = \lambda \quad \text{for } 0 \le i < R, \qquad C_R = 0$$

where R is the number of servers in the system. The mean effective traffic rate C_0 is calculated as

$$C_0 = \sum_{i=0}^{R-1} \lambda P_i$$

where P_i is the probability that the system is in state i. The system can be in any one of 0, 1, 2 ... R states. Therefore, we have

$$P_0 + P_1 + P_2 + \ldots + P_R = 1 \tag{8.29}$$

We may now write the expression of C_0 as

$$C_0 = \lambda(P_0 + P_1 + \ldots + P_{R-1}) = \lambda(1 - P_R) \tag{8.30}$$

The mean traffic carried by the network is given by

$$A_0 = C_0 t_h \tag{8.31}$$

$$= \lambda(1 - P_R) t_h \tag{8.31a}$$

By Eq.(8.28), this reduces to

$$A_0 = A(1 - P_R)$$

or

$$P_R = \frac{A - A_0}{A} \tag{8.32}$$

The blocking probability P_B is the same as the probability that all the servers are busy, i.e. P_R. Therefore, from Eqs. (8.3) and (8.32), we conclude that

$$\text{GOS} = P_B$$

for LCC model where the traffic arrival is characterised by Poisson process.

We now proceed to calculate the blocking probability. For this purpose, we perform the steady state analysis of the B-D process characterising the

LCC model. We have already discussed the arrival process and we now see the call termination process. A constant death rate may not be appropriate here. The call termination rate is likely to be dependent on the number of busy servers in the system, i.e. state of the system. If a large number of servers are busy, more calls are likely to terminate in a given time and the call termination rate will be higher. Hence, the termination rate may be considered to be directly proportional to the number of busy servers as given by

$$u_k = k\mu \quad \text{for } 0 \le k \le R \tag{8.33}$$

where

μ = mean call termination rate = $1/t_h$

μ_k = call termination rate in state k

Substituting the values of birth and death rates in Eqs. (8.10) and (8.11), we have

$$P_{k-1}\lambda + P_{k+1}\mu(k+1) - (\lambda + k\mu)P_k = 0$$

Using Eq. (8.28), we get

$$P_{k+1} = \frac{AP_k + kP_k - AP_{k-1}}{k+1} \quad \text{for } k > 0 \tag{8.34}$$

$$P_1 = AP_0 \quad \text{for } k = 0 \tag{8.35}$$

For $k = 1$, we have

$$P_2 = \frac{AP_1 + P_1 - AP_0}{2}$$

Substituting for P_1 from Eq. (8.35),

$$P_2 = \frac{A^2P_0}{2}$$

Similarly for $k = 2$, we have

$$P_3 = \frac{AP_2 + 2P_2 - AP_1}{3} = \frac{A^3P_0}{3 \times 2} = \frac{A^3P_0}{3!}$$

Generalising, we get

$$P_j = \frac{A^jP_0}{j!} \tag{8.36}$$

From Eqs. (8.29) and (8.36),

$$P_0 + AP_0 + \ldots + \frac{A^RP_0}{R!} = 1$$

Therefore,

$$P_0 = \frac{1}{1 + A + \frac{A^2}{2!} + \cdots + \frac{A^R}{R!}} \tag{8.37}$$

For $k = R$, substituting for P_0 from Eq. (8.37) into Eq. (8.36), we obtain

$$P_R = \frac{A^R/R!}{1 + A + \frac{A^2}{2!} + \cdots + \frac{A^R}{R!}} \tag{8.38}$$

This is the famous Erlang B formula or loss formula. The quantity P_R is the probability that all servers are busy in the system and hence is the blocking probability P_B of the system. We have already shown that for the Erlang traffic which occurs in the LCC model, the GOS and the blocking probability values are equal. Therefore, the value of GOS is given by Erlang B formula. CCITT has adopted Erlang B formula as the standard for estimating the grade of service of a switching system.

When R is large and A small, A/R is very small and the denominator in Eq. (8.37) reduces to e^A. Equation (8.36) then becomes

$$P_j = \frac{A^j}{j!} e^{-A} \tag{8.39}$$

which is the same as Poisson equation. In this case, the traffic is Poisson. In the limit R tending to infinity, we have

$$\lim_{R \to \infty} \frac{A^R e^{-A}}{R!} \to 0 \quad \text{or} \quad P_B \to 0 \tag{8.39a}$$

Thus, the blocking probability tends to zero with Poisson traffic.

8.5.2 Lost Calls Cleared System with Finite Subscribers

Erlang loss formula was derived in Section 8.5.1 under a fundamental assumption that call arrivals are independent of the number of active callers. Such an assumption is justified only when the number of sources is much larger than the number of servers. This may not always be the case in practice. For example, in a three-stage space division network, the number of subscribers connected to each input matrix is comparable to the number of servers (alternative paths) available in the network. In such cases, the arrival rate to the system is dependent on the number of subscribers who are not occupied as the busy subscribers do not generate new calls. The traffic in this case is known as **Engest traffic** or **pure chance traffic of type 2**. We now derive expressions governing blocking probabilities in such cases. The

blocking probabilities in finite source systems are always less than those for infinite source systems since the arrival rate decreases as the number of busy sources increases. Let

λ_s = arrival rate per subscriber
k = number of busy subscribers
N = total number of subscribers
R = number of servers

The offered traffic (arrival rate) when the system is in state k is given by

$$C_k = (N - k)\lambda_s \quad \text{for } k \leq 0 \leq R$$

The mean offered traffic rate is given by

$$C = \sum_{k=0}^{R} (N - k)\lambda_s P_k = N\lambda_s \sum_{k=0}^{R} P_k - \lambda_s \sum_{k=0}^{R} kP_k$$

$$= \lambda_s \left(N - \sum_{k=0}^{R} kP_k\right) \tag{8.40}$$

In Eq. (8.40), the quantity $\Sigma\, kP_k$ represents the average number of busy servers. The carried traffic in a network is the average number of calls accepted during the mean service time period. This is the same as the average number of busy servers at any given time. Hence, we have

$$C = \lambda_s (N - A_0) \tag{8.41}$$

The offered traffic is

$$A = Ct_h = \lambda_s t_h (N - A_0) \tag{8.42}$$

When the system is in state R, the offered traffic rate is $(N - R)\lambda_s$, but all the arrivals are rejected. Therefore, the lost traffic is obtained as

$$A - A_0 = (N - R)\lambda_s P_R t_h \tag{8.43}$$

The grade of service now works out to be

$$\text{GOS} = \frac{N - R}{N - A_0} P_R \tag{8.44}$$

Thus we see for Engest traffic, the blocking probability and the GOS are not the same, i.e. the time congestion and the call congestion values differ.

We now proceed to calculate the blocking probability. For this purpose, we analyse the steady state of the B-D process characterising this model. The arrival process has been discussed above. The termination process is the same as in the case of LCC model with infinite sources, and the termination rates is given by Eq. (8.33). Substituting the values of birth and death rates in Eqs. (8.10) and (8.11), we have

$$P_{k-1}\lambda_s(N-k+1) + P_{k+1}\mu(k+1) - (\lambda_s(N-k) + k\mu)P_k = 0$$

Rearranging the terms, we obtain

$$P_{k+1} = \frac{[(\rho(N-k)+k)]P_k - \rho(N-k+1)P_{k-1}}{k+1} \tag{8.45}$$

where $\rho = (\lambda_s/\mu)$. For $k = 0$, we have $P_1 = \rho N P_0$. For $k = 1$,

$$P_2 = \frac{[(\rho(N-1)+1)P_1 - \rho(N-1+1)P_0]}{2}$$

$$P_2 = \frac{\rho^2 N(N-1)P_0}{2}$$

For $k = 2$, we have

$$P_3 = \frac{\rho^3 N(N-1)(N-2)P_0}{3 \times 2}$$

Generalising, we get

$$P_j = \rho^j \binom{N}{j} P_0 \tag{8.46}$$

where $\binom{N}{j}$ is the binomial coefficient, i.e.

$$\binom{N}{j} = \frac{N!}{j!\,(N-j)!}$$

From Eqs. (8.29) and (8.46), we have

$$P_0 + \rho\binom{N}{1}P_0 + \rho^2\binom{N}{2}P_0 + \ldots + \rho^R\binom{N}{R}P_0 = 1$$

Therefore,

$$P_0 = \frac{1}{1 + \rho\binom{N}{1} + \rho^2\binom{N}{2} + \cdots + \rho^R\binom{N}{R}}$$

or

$$P_0 = \frac{1}{\sum_{k=0}^{R} \rho^k \binom{N}{k}} \tag{8.47}$$

The blocking probability P_B is given by

$$P_B = P_R = \frac{\rho^R\binom{N}{R}}{\sum_{k=0}^{R} \rho^k \binom{N}{k}} \tag{8.48}$$

From Eqs. (8.44) and (8.48), we have

$$\text{GOS} = \frac{N-R}{N-A_0}\,\rho^R \frac{N(N-1)!}{R!(N-R)(N-R-1)!} \frac{1}{\sum_{k=0}^{R} \rho^k \binom{N}{k}}$$

$$= \frac{N}{N-A_0}\,\rho^R \binom{N-1}{R} \frac{1}{\sum_{k=0}^{R} \rho^k \binom{N}{k}} \tag{8.49}$$

From Eqs. (8.40), (8.41) and (8.48),

$$N - A_0 = \frac{N \sum_{k=0}^{R} \rho^k \binom{N}{k} - \sum_{j=0}^{R} j\rho^j \binom{N}{j}}{\sum_{k=0}^{R} \rho^k \binom{N}{k}}$$

Limits being the same for summation, the terms in the numerator can be combined to give

$$N - A_0 = \frac{\sum_{k=0}^{R} \rho^k \binom{N}{k} (N-k)}{\sum_{k=0}^{R} \rho^k \binom{N}{k}}$$

Applying a simplification process similar to the one used in arriving at Eq. (8.49), we obtain

$$N - A_0 = \frac{N \sum_{k=0}^{R} \rho^k \binom{N-1}{k}}{\sum_{k=0}^{R} \rho^k \binom{N}{k}} \tag{8.50}$$

Substituting for $N - A_0$ in Eq. (8.49) from Eq. (8.50), we have

$$\text{GOS} = \frac{\rho^R \binom{N-1}{R}}{\sum_{k=0}^{R} \rho^k \binom{N-1}{k}} \tag{8.51}$$

An important case of the LCC model with finite sources occurs when $R \geq N$. In this case, there is no blocking. The state probability is given by

$$P_j = \binom{N}{j} \frac{\rho^j}{\sum_{k=0}^{N} \rho^k \binom{N}{k}} = \binom{N}{j} \frac{\rho^j}{(1+\rho)^N} \tag{8.52}$$

The traffic in this case is the Bernoulli traffic.

If we now define $a = \frac{\rho}{1+\rho}$, we have

$$\rho = a(1+\rho), \qquad 1 - a = \frac{1}{1+\rho} \tag{8.53}$$

Substituting from Eq.(8.53) in Eq. (8.52), we get

$$P_j = \binom{N}{j} \frac{a^j (1+\rho)^j}{(1+\rho)^N}$$

or

$$P_j = \binom{N}{j} a^j \frac{1}{(1+\rho)^{N-j}}$$

or

$$P_j = \binom{N}{j} a^j (1-a)^{N-j} \tag{8.54}$$

This is the well known binomial formula, which implies that servers are independent of one another.

We have defined ρ as $\lambda_s t_h$. The parameter $\lambda_s t_h$ does not by itself specify the average activity of a source and cannot therefore be measured directly. What can be observed is the average transmission rate per subscriber based on the total traffic. We define this as

$$a' = A/N \tag{8.55}$$

From Eq. (8.41),

$$a' = \rho(1 - A_0/N)$$

Substituting for A_0 from Eq. (8.3), we get

$$a' = \rho\left(1 - \frac{A}{N}(1 - \text{GOS})\right)$$

Simplifying, we obtain

$$a' = \frac{\rho}{1 + \rho(1 - \text{GOS})} \tag{8.56}$$

For the use of Bernoulli traffic, GOS is zero. Therefore, Eq. (8.56) reduces to

$$a' = \frac{\rho}{1+\rho} = a \text{ as defined in Eq. (8.53).}$$

EXAMPLE 8.8 In a telephone system, there are 20 servers and 100 subscribers. On an average, there are 10 busy servers at any time. The probability of all the servers being busy is 0.2. Calculate the grade of service assuming (a) Erlang traffic, and (b) Engest traffic.

Solution For Erlang traffic, GOS is the same as P_R. Therefore, GOS = 0.2.

For Engest traffic, GOS is given by Eq. (8.44).

Carried traffic = number of average busy servers = 10

Therefore,

$$\text{GOS} = \frac{100-20}{100-10} \times 0.2 = 0.1875$$

As expected, the GOS is lower in the case of LCC model with finite sources.

8.5.3 Lost Calls Returned System

In the LCC model discussed in Sections 8.5.1 and 8.5.2, it is assumed that unserviceable requests leave the system and never return. In other words, the arrival rate into system is in no way affected by the calls that are rejected. This is usually not the case, particularly in subscriber concentrator systems, call to busy lines etc. The rejected calls in such cases do return to the system in the form of retries with the result that the offered traffic now comprises two components:

Offered traffic = new traffic + retry traffic

In this section, we derive the blocking probability relationships taking into account the returning calls. The model used for analysis is known as **lost calls returned (LCR) model.** We proceed with the analysis by making the following assumptions with regard to the nature of the returning calls:

1. No new call is generated when a blocked call is being retried.
2. A number of retry attempts may be involved before a call eventually gets serviced.
3. Retries are attempted after a random time and each retry time is statistically independent of the others.
4. Typical waiting time before a retry is longer than the average holding time.

Assumption 3 says that the retries are not correlated. If they were, there would be traffic peaks at some intervals which would complicate the analysis.

Assumption 4 permits the system to maintain its statistical equilibrium even in the presence of retry. By assuming that the waiting time is longer than the average service time, the retry arrivals are made to have the same arrival characteristics as the new traffic. In effect, these assumptions make the retry traffic statistically indistinguishable from the new arrivals. Hence blocked calls merely add to the first attempt calls. Let λ be the arrival rate for new calls and let us denote the GOS as P_c, call congestion. Then P_c calls are rejected, which return to the system as retries. The retries will further experience blocking by a factor of P_c, i.e. $P_c \times P_c$ calls will be rejected. Reasoning thus, the effective arrival rate can be expressed as

$$\lambda' = \lambda + P_c\lambda + P_c^2\lambda + P_c^3\lambda \ldots = \frac{\lambda}{1 - P_c} = \frac{\lambda}{1 - \text{GOS}} \qquad (8.57)$$

Equation (8.57) relates the effective arrival rate to the call congestion. In this case, it is not possible to directly determine either λ' or P_c as one is expressed in terms of the other. They are estimated iteratively using Erlang B formula, starting with λ as the initial value for the arrival rate. A similar procedure can be adopted for the LCC model with finite sources as well.

The effect of returning traffic is insignificant when operating at low grade of service values. At high GOS values, the effect is noticeable. For a given traffic load and the capacity of the system, the estimated blocking probability has a lower value in the case of LCC model than in the case of LCR model. Stated differently, for a given value of GOS, the LCC model permits a larger offered load to the system than the LCR model.

8.5.4 Lost Calls Held System

As pointed out in the beginning of Section 8.5, lost calls held (LCH) systems are distinctly different from delay systems where the messages are queued and taken up for service as and when resources are available. The time a call spends in a delay system is the sum of the waiting time and the service time. In the LCH model, the total time spent in the system is independent of the waiting time and is only determined by the average service time required. As soon as a call arrives in the system, it requires service continuously for a period of time and terminates after that time irrespective of whether it is being actually serviced or not. If a call is blocked, a portion of it is lost until a server becomes free to service the call. This is depicted in Fig. 8.6.

An example of LCH system is the time assigned speech interpolation (TASI) system. In a TASI system, the number of conversations supported is larger than the number of transmission channels or servers in the system. The system exploits the fact that a speech conversation is interspersed with silence or inactivity periods when there is no need for transmitting the signal. The channel is deassigned during the silence period and used for supporting another conversation which is in the activity phase. Thus, more number

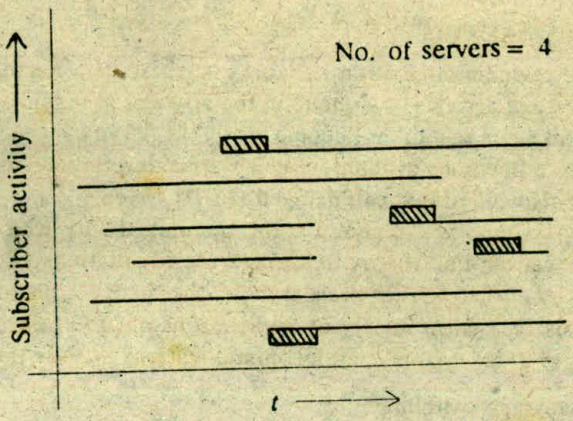


Fig. 8.6 Lost traffic in LCH model.

of conversations than the transmission channels can be supported on the system.

If a speech circuit becomes active when all the channels are busy, it is blocked and speech clipping occurs. A speech segment becomes active or inactive irrespective of whether a channel is available or not. In other words, the duration for which a source is active is independent of whether it is being serviced or not. As soon as a call arrives, it is accepted by the system whether a server is available or not. Hence, the number of active sources in the system is nothing but the number of call arrivals in a time duration t_h. For an infinite population, this is characterised by Poisson arrival process.

EXAMPLE 8.9 A TASI system has 10 channels and 20 sources connected to it. What is the probability of clipping if the activity factor for each source is 0.4?

Solution Average number of busy servers $= 0.4 \times 20 = 8$

Clipping occurs if 10 or more sources are in the activity phase simultaneously. Therefore, the probability that 10 or more sources are active gives the clipping probability. This is determined from the Poisson equation as:

$$\text{Prob (clipping)} = \sum_{j=10}^{\infty} P_j(t=8) = 1 - \sum_{j=0}^{9} P_j(t=8)$$

$$= 1 - e^{-8}\left(1 + 8 + \frac{8^2}{2!} + \ldots \frac{8^9}{9!}\right) = 0.284$$

8.6 Delay Systems

A class of telecommunication networks, such as data networks, places the call or message arrivals in a queue in the absence of resources, and services them as and when resources become available. Servicing is not taken up until the resource becomes available. Such systems are known as **delay systems** which are also called **lost call delayed (LCD)** systems.

Delay systems are analysed using queuing theory which is sometimes known as waiting line theory. Although the foundations of queuing theory were laid by early teletraffic researchers, the theory is now-a-days used for the analysis of a wide variety of applications outside telecommunications. Examples of delay systems in telecommunications include the following:

- Message switching
- Packet switching
- Digit receiver access
- Automatic call distribution
- Call processing.

The elements of a queuing system are shown in Fig. 8.7. There is a large population of sources that generate traffic or service requests to the network.

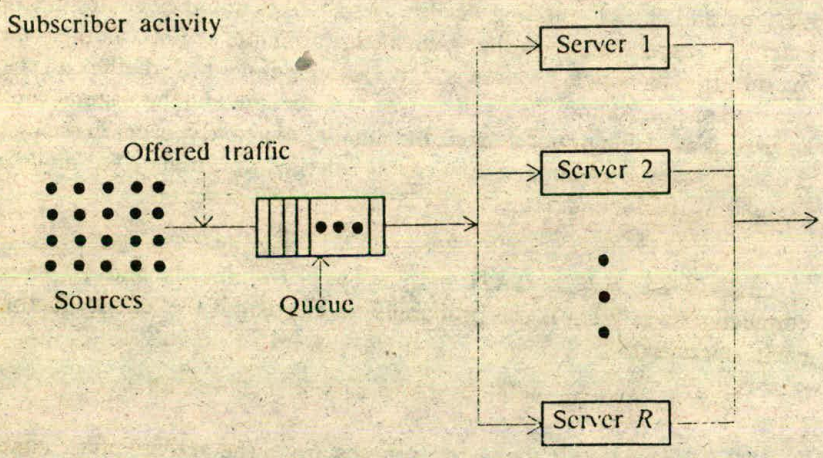


Fig. 8.7 Elements of a queuing system.

There is a service facility that contains a number of identical servers, each of which is capable of providing the desired service to a request. When all the servers are busy, a request arriving at the network is placed in a queue until a server becomes available. While analysing queuing systems, we have to deal with a number of random variables such as the number of waiting requests, interarrival times between requests, and time spent by a request in the system. Some of the important random variables associated with a queuing

system are listed along with the notation used to represent them in Table 8.1. The number of requests present in the system or the state of the system is given by the sum of the requests in the queue and those being serviced. No request can be pending in the queue unless all the servers are busy. Hence, we have

$$k = k_q + R \tag{8.58}$$

Table 8.1 Some Important Random Variables in a Queuing System

Random variable	Notation
State of the system: number of calls present in the system	k
Queue length : number of calls or requests waiting to be serviced	k_q
Number of calls in service	k_s
Mean wait time	t_q
Mean service time	t_k
Mean arrival rate	λ
Mean service rate	μ
Mean interarrival time	τ
Probability that there are k calls in the queuing system	P_k
Traffic intensity : λ/μ	ρ
Server utilisation : ρ/R	u

The mean time a call or a request spends in the system is the sum of the mean wait time t_q and mean service or holding time t_h. Readers may like to contrast this to the mean time spent in the system by a call in the LCH model discussed in Section 8.5.4.

A queued operation enables better utilisation of servers than does a loss system. Queuing has the effect of smoothening out the traffic flow as far as servers are concerned. Peaks in the arrival process build up the queue lengths. Since there is no statistical limit on the number of arrivals occurring in a short period of time, there is a need for infinite queuing capacity if there were to be no loss of traffic. In a practical system, only finite queue capacities are possible and hence there is a probability, however small it may be, of blocking in delay systems. Assuming that a delay system has infinite queue capacity in an operational sense, a necessary condition for its stable operation is as follows:

$$\frac{\text{Mean arrival rate}}{\text{Mean service rate}} < 1$$

or

$$\frac{\text{Offered traffic}}{\text{Number of servers}} < 1$$

If this condition is not satisfied, the queue length would become infinite sooner or later, and the system would never be able to clear the traffic offered to it. A queuing system is characterised by a set of six parameters. A concise 6-parameter notation, due to D.G. Kendall, is used to represent different types of queuing systems. This notation uses letters to identify the parameters. The notation reads as *A/B/c/K/m/Z*. The parameter specifications are as follows:

A = arrival process specification
B = service time distribution
c = number of servers
K = queue capacity
m = number of sources (input population)
Z = service discipline

The value of the parameters K and m may be either a finite number or an infinite number. Queue discipline is the rule used for choosing the next customer to be serviced from the queue. Commonly used queue disciplines include first-come-first-served (FCFS), random selection, and priority based selection. The parameters K, m and Z may be omitted from the queue specifications, in which case they assume some default values. For K and m, the default values are infinity, i.e. infinite queue capacity and infinite sources respectively. The default queue discipline is FCFS. The parameter c is a nonzero positive finite number. The parameters A and B may assume any one of the values shown in Table 8.2. As an example, the queue specification *M/D/4* means a queue system with Poisson arrival, deterministic service time distribution, four servers, infinite queue capacity, infinite sources and FCFS queue discipline. It is interesting to note that Kendall's notation may also be used to represent a loss system where the parameter K has a value zero.

When analysing delay systems in this section, we assume that infinite sources exist, infinite queuing capabilities exist, and the queue is serviced on FCFS basis. We assume a Poisson arrival process and concern ourselves with service times that are exponentially distributed or are constant. These service time distributions represent the most random and the most deterministic service times. Thus, a system that operates with some other distribution of service times has a performance somewhere between those produced by these two distributions. In effect, we shall model our telecommunication systems as *M/M/R* and *M/D/R* queuing systems. An important purpose of the analysis is to determine the probability distribution of waiting times and the associated mean values. Often, the probability that waiting time exceeds a specified limit is of interest. In particular, the probability that the waiting time is greater than zero represents the call congestion and hence is of immediate interest.

Table 8.2 Values for Queue Parameters A and B

Values for A and B	Meaning	Remarks
GI	Arrival process with general independent distribution for interarrival time	For A only
G	General (no assumptions) service time distribution	For B only
E_R	Erlang-k interarrival or service time distribution	
M	Poisson arrival and exponential service time distribution	M stands for Markov
D	Deterministic interarrival or service time distribution	e.g. constant time distributions
H_k	Hyperexponential (with k stages) interarrival or service time distribution	

As in the case of loss system analysis, we start with a B-D process and proceed to analyse the delay systems. For arrival and service rates, we have

$$\lambda_k = \lambda \quad \text{for } k = 0,1,2,\cdots,$$
$$\mu_k = k\mu \quad \text{for } k = 1,2,3,\cdots,R$$
$$\mu_k = R\mu \quad \text{for } k > R$$

The stability condition for the system demands that

$$\frac{\lambda}{R\mu} < 1 \quad \text{or} \quad A/R < 1 \tag{8.59}$$

Many of the results obtained for the LCC model apply here as well. The fundamental difference between the LCC system and the $M/M/R$ delay system is that the state of the $M/M/R$ system varies from zero to infinity whereas the loss system from 0 to R. Any number of calls may enter a delay system and be serviced or be in the wait state, whereas no calls can enter a loss system once all the servers are busy. Accordingly, Eq. (8.29) modifies to

$$P_0 + P_1 + \cdots + P_R + P_{R+1} \cdots = 1 \tag{8.60}$$

Therefore, we have

$$P_0 = 1 - \sum_{k=1}^{\infty} P_k \tag{8.60a}$$

The loss system and the delay system behave identically as long as the system is within the state R. Equation (8.36) applies to state probabilities except that the value of P_0 is governed by Eq .(8.60a). The systems behave differently for states equal to or greater than state R. Substituting the birth and death rates of the $M/M/R$ system in Eq. (8.10) for $k = R$ and rearranging the terms, we have

$$R\mu P_{R+1} = (\lambda + R\mu)P_R - \lambda P_{R-1}$$

Using the result of Eq. (8.36), we get

$$R\mu P_{R+1} = (\lambda + R\mu)\frac{A^R}{R!}P_0 - \lambda\frac{A^{R-1}}{(R-1)!}P_0$$

where the value of P_0 is governed by Eq. (8.60a). Simplifying, we get

$$R\mu P_{R+1} = \frac{\lambda A^R}{R!}P_0$$

or

$$P_{R+1} = \frac{\lambda}{\mu R}\frac{A^R}{R!}P_0 = \frac{A}{R}\frac{A^R}{R!}P_0 = \frac{A}{R}P_R \tag{8.61}$$

Similarly, for $k = R + 1$, we get

$$P_{R+2} = \left(\frac{A}{R}\right)^2 P_R$$

Generalising for $k > R$, we have

$$P_k = \left(\frac{A}{R}\right)^{k-R} P_R \tag{8.62}$$

Considering the inequality given in Eq. (8.59), we conclude that

$$\frac{A}{R} + \left(\frac{A}{R}\right)^2 + \left(\frac{A}{R}\right)^3 + \cdots = \frac{A/R}{1-(A/R)} = \frac{A}{R-A} \tag{8.63}$$

From Eqs. (8.60) and (8.63), we have

$$\sum_{k=0}^{R}\frac{A^k}{k!}P_0 + \frac{A}{R-A}\frac{A^R}{R!}P_0 = 1$$

Therefore,

$$P_0 = \frac{1}{\sum_{k=0}^{R}\frac{A^k}{k!} + \frac{A^R}{R!}\frac{A}{R-A}} \tag{8.64}$$

$$P_R = \frac{A^R}{R!} \frac{1}{\sum_{k=0}^{R} \frac{A^k}{k!} + \frac{A^R}{R!} \frac{A}{R-A}} \tag{8.64a}$$

Now, the probability of a message being delayed is nothing but the probability of finding the system in state R or above and is given by

$$P(\text{delay} > 0) = \sum_{k=R}^{\infty} P_k \tag{8.65}$$

Therefore, we have

$$P(\text{delay} > 0) = \left(1 + \frac{A}{R-A}\right) P_R = \frac{R}{R-A} P_R \tag{8.66}$$

where P_R is obtained by substituting the value of P_0 from Eq. (8.64) in Eq. (8.38). Equation (8.66) is variously known as Erlang second formula, Erlang delay formula and Erlang C formula.

We may express Erlang C formula in terms of Erlang B formula or first formula. If we term the value of P_R in Eq. (8.38) as P_{RB}, and P_R in Eq. (8.66) as P_{RC}, then P_{RC} is related to P_{RB} as

$$P_{RC} = \frac{P_{RB}(R-A)}{R - A(1 - P_{RB})} \tag{8.67}$$

Substituting for P_R in Eq.(8.66), we have

$$P(\text{delay} > 0) = \frac{RP_{RB}}{R - A(1 - P_{RB})} \tag{8.68}$$

where P_{RB} is the blocking probability given by Erlang B formula.

For $M/M/R$ queue, it can be shown that the waiting time distribution is exponential and is given by

$$P(\text{delay} > t) = P(\text{delay} > 0) \exp[-(R-A)t/t_h] \tag{8.69}$$

By integrating Eq. (8.69) over time, the average waiting time may be determined as

$$t_q = \frac{P(\text{delay} > 0) t_h}{R-A} \tag{8.70}$$

Equation (8.70) applies to all arrivals to the system, some of which enter the queue and others which do not. Given that a message is already put in the queue, the average waiting time may be expressed as

$$t_q' = \frac{t_h}{R-A} \tag{8.71}$$

EXAMPLE 8.10 A public call office (PCO) is installed in a busy part of a town. 150 persons use the booth everyday. The average holding time for a call is 1.5 minutes. There is a suggestion from the public that another PCO is required in the same locality as the wait times are unduly long. Analyse the situation using *M/M/*1 queue and determine if the suggestion deserves serious consideration.

Solution

Average arrival rate $\lambda = 0.104$ per minute

Offered load $A = \lambda t_h = 0.156$

From Eq. (8.64a), we have, for a single server case,

$$P_R = A\,(1-A)$$

and from Eq. (8.66),

$$P\,(\text{delay} > 0) = A = 0.156$$

From Eq. (8.70), $t_q = 0.278$ minutes, $t'_q = 1.78$ minutes

There is really no reason for the public to complain. But the problem statement implies that the traffic is spread over throughout the day which may not be true. The busy hour traffic is important.

FURTHER READING

1. Bear, D., *Principles of Telecommunications Traffic Engineering,* Peter Peregrinus Ltd., On behalf of the Institution of Electrical Engineers, Herts (UK), 1976.
2. Benes, V. E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.
3. Hebuterne, G., *Traffic Flow in Switching Systems,* Artech House, London, 1987.
4. Kleinrock, L., *Queuing Systems, Vol.1 : Theory,* John Wiley & Sons, New York, 1975.
5. Kleinrock, L., *Queuing Systems, Vol.2 : Computer Applications,* John Wiley & Sons, New York, 1976.
6. F lm, C., *Intensity Variations in Telephone Traffic: Studies in Telecommunication, Vol.10*, North Holland, Amsterdam, 1988.
7. Syski, R., *Introduction to Congestion Theory in Telephone Systems: Studies in Telecommunication, Vol.4,* North-Holland, Amsterdam, 1986.

EXERCISES

1. During a 2-hour busy period, 2400 calls arrive at an exchange. Average holding time per call is two minutes. What is the traffic load in (a) erlangs and (b) in CCS?

2. 10,000 subscribers are connected to an exchange. If the exchange is designed to achieve a call completion rate of 0.8 when the busy hour calling rate is 4.8, what is the BHCA that can be supported by the exchange? What should be the call processing time for this exchange?

3. A call processor in an exchange requires 120 ms to service a complete call. What is the BHCA rating for the processor? If the exchange is capable of carrying 700 erlangs of traffic, what is the call completion rate? Assume an average call holding time of two minutes.

4. The traffic statistics of a company using a PABX indicates that 180 outgoing calls are initiated every hour during working hours. Equal number of calls comes in. Each call lasts for 200 seconds on the average. If the GOS required is 0.05, determine the number of lines required between the PABX and the main exchange.

5. In an exchange, the calls arrive at the rate of 1100 calls per hour, with each call holding for a duration of three minutes. If the demand is serviced by a trunk group of 50 lines, determine the GOS.

6. An exchange is designed to handle 2000 calls during the busy hour. One day, the number of calls during the busy hour is 2200. What is the resulting GOS?

7. Consider a B-D process with coefficients

$$\lambda_k(0) = \begin{cases} \lambda & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

$$\mu_k(0) = \begin{cases} \mu & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

Give the differential-difference equations for $P_0(t)$ and $P_1(t)$. Solve these equations and express the answers in terms of $P_0(0)$ and $P_1(0)$.

8. Consider a B-D process in which the birth rate decreases and the death rate increases with the population k in the systems:

$$\lambda_k(0) = \begin{cases} (1-k) & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

$$\mu_k(0) = \begin{cases} k & \text{for } k \leq N \\ 0 & \text{for } k > N \end{cases}$$

Write down the differential-difference equations for $P_k(t)$.

9. A traffic load of two erlangs is offered to a full availability group of five trunks. (Full availability implies that there is no restriction on the way in which calls are allocated to particular trunks). The average call holding time is three minutes.
   (a) What is the probability that no calls arrive during a five-minute period?
   (b) Determine the value of call congestion.
   (c) Consider the case when the trunks are numbered 1,2,3,4,5 and a call is allocated to the lowest numbered free trunk always. How much traffic is carried by the first trunk? How much traffic is carried by the last trunk?

   Assume LCC model.

10. Between two end offices there is an average traffic of 24 erlangs. If CCITT standard 32-channel PCM link is used between the end offices, what is the probability of blocking? How much of traffic is cleared by other resources if LCC model is assumed?

11. With Poisson arrival of two calls per minute, what is the probability that more than three calls will arrive in two minutes? What is the time during which at least four calls will arrive with a probability of more than 95 per cent?

12. A PABX with five trunks supports 200 extensions. Each extension generates three external calls per 8 hour working day. Average call duration is two minutes. If LCR model is assumed, what is the blocking probability? What is the offered load?

13. A PABX which is designed to be nonblocking internaly serves 100 extensions and has four trunk lines. Busy hour calling rate is two. Thirty per cent of the total traffic is external. Average call holding time is three minutes. What is the probability that an incoming call finds all the four trunk lines busy?

14. A PABX provides queuing and automatic call back facility for outgoing calls. If there are 20 outgoing call requests per hour and if the average call duration is three minutes, how many trunks are needed to ensure delays less than 30 minutes for 90 per cent of the requests?

15. A traffic arrival stream is formed by merging the input from K independent Poisson sources with source i having an arrival rate of λ_i for all $1 \leq i \leq K$. Show that the merged stream is also Poisson with an arrival rate of $\lambda = \sum_{i=1}^{K} \lambda_i$

16. An exchange uses two small call processors, each capable of serving requests that arrive at the rate of 15 requests per second. The two pro-

cessors are to be replaced by a single higher capability processor. Characterise the arrival process for the new system.

17. In a group of 100 Poisson sources, each one generates messages at the rate of one in every 20 seconds to a single line. If the line rate is 1200 bps and the average message length is 200 bits, determine

    (a) the probability of a message entering the queue,

    (b) the average queuing delay,

    (c) the probability of delay being more than one second, and

    (d) the utilisation of the transmission link.

18. Calls arrive at a tandem exchange with R trunks at the rate of λ. If the calls are distributed to each trunk with probabilities $P_1, P_2, \ldots, P_R$, what is the request arrival rate at each trunk?

19. In an office, calls arrive at a PABX at the rate of 24 per hour during office hours. How much time is expected to elapse before the first call arrives at the exchange after the office opens in the morning?